Review of proposed CUAHSI Hydrologic Information System Hydrologic Observations Data Model

May 5, 2005

Review coordinated by David Tarboton, Utah State University.

The paper "A Data Model for Hydrologic Observations", by David Maidment, March 2005 (http://www.crwr.utexas.edu/cuahsi/symposium05/HODatabase/Documents/HydroObsDataModel.doc) presents the design for the integrated hydrologic observations database that is proposed for the CUAHSI Hydrologic Information System (HIS). An independent review of this design was undertaken to evaluate whether the data model for hydrologic observations meets the needs of the CUAHSI community. Specifically the review sought to address whether the model is simple, usable, and can be implemented in a variety of application systems, including relational databases, Excel, GIS, statistical packages and simulation systems like MatLab. This document presents this review.

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 the design and dissemination of data. A total of 21 individual review comments were received. Appendix 1 gives the complete text of all the reviews received. Appendix 2 gives the review questionnaire that reviewers were asked to respond to.

Review comments ranged from being delighted and indicating that everything is fine, to quite serious concerns over limitations. On the whole the review comments received were extremely thoughtful and represent significant community input to this process. In balance, I think the assessment is that this data model is a good start, but that there are a number of issues that need to be addressed. The issues that need to be addressed are mostly beyond simple fixes that can be suggested within the scope of this review and will require careful evaluation of design tradeoffs. It is impossible to in summary form convey fully the thought that went in to each review. Readers are therefore encouraged to view the reviews in their entirety in Appendix 1. There were a number of common themes in the reviews that are highlighted here, with some suggestions for how they might be addressed.

1. In the AHTSDM there is inadequate information to identify the source, heritage or provenance and give an exact definition of the data. There needs to be a mechanism for tracing the origin of any record in the database and any instrument, calibration or transformation information pertaining to the data values. Suggestions to accommodate this include expanding the TSType table to include this data source and heritage information. The TSType table could include by reference (e.g. URL) descriptive documents where necessary. (See detailed comments by reviewers 1, 2, 3, 8, 10, 14, 18, 21).

2. The schema of the tables received comments from a few of the more technically inclined. Reviewer #1 suggested a new design with additional tables. Reviewer #2 indicated that the three tables were unnecessarily complex. Reviewer #8 suggested that it is inefficient to have a feature identifier in every time series row when one would almost never look at an individual time series value in isolation. He suggested a single TSTypeID in the time series table to reference a table

that then points to the monitoring point and other data heritage information. There are design tradeoffs to be considered in deciding how much detail to include in the critical TimeSeries table. Reviewer #1 was most concerned with the efficiency of querying for records of a certain variable or from a certain source and to facilitate this suggests adding a number of qualifier codes to TimeSeries to facilitate these queries. Reviewer #8 on the other hand sees multiple identifier attributes in TimeSeries as inefficient so suggests reducing the identifier attributes in TimeSeries to one with other site and heritage information linked off the TSType table.

3. The monitoring point table does not contain enough information to fully spatially locate a measurement. Questions were raised regarding the encoding of GIS information in formats readable only by a specific commercial GIS system (ESRI geodatabase format). There are a number of spatial reference frame (projection) issues that arise in the context of specifying measurement locations that need to be addressed. Questions were also raised that relate to the addressing of measurement locations. Does it make most sense to have a measurement point located using X and Y coordinates, or relative to the (possibly hierarchical) overall measurement framework, i.e. on such and such a reach within such and such a watershed. These contextual spatial linkages may be more useful than simple X-Y locations. (Reviewers 1, 3, 7, 14, 15, 18)

4. It is important that the scale of measurements, defined in terms of their support (averaging domain), spacing and extent (e.g. Blöschl and Sivapalan, 1995) be quantified and associated with measurements. The data model needs to accommodate and be specific about measurements that apply to a point, length or area. (Reviewers 2, 3, 4, 7, 10, 15, 17, 20)

5. The monitoring point table does not readily accommodate measurements associated with a domain other than a point, or measurements where the instrument or observation moves (e.g. sensor on an aircraft or boat, or moving with a flood front). Related to this is the question of data that is changing both in space and time. Reviewers were concerned that the data model be able to store spatially varied time series and link this to point time series. For example, could spatial data, for example of a changing snowpack, be linked to the associated point discharge. (Reviewers 3, 14, 17)

6. The location of a monitoring point does not readily incorporate depth information. Elevation can be included in the specification of monitoring point "shape", but often is not. Furthermore, absolute elevation (relative to say the Geoid) is not the most natural way to specify the location of measurements along a profile (say through the soil, or snow or atmosphere) and once measurement records are separated by different elevations, reconstruction of profile information is cumbersome, and may be imprecise or inefficient depending on the precision with which elevation is resolved. (Reviewers 1, 11, 19)

7. The question of depth information and profile measurements is one example associated with the issue of how measurements should be grouped. The disaggretation of measurements down to the level of single values exposes individual measurements and is presumably intended to facilitate querying and analysis across alternative directions. However it can make reconstruction of related measurements difficult, so consideration needs to be given to the degree of parceling of related information. For example remote sensing images can be represented by a single value for each point, but may be more logically represented as a vector or array retained

and stored together. Similarly certain profile or time series information may be more logically represented as vectors always bound and stored together. Care needs to be exercised in designing this system to select the most appropriate size for bound groupings, so that one does not for example have to extract an image of the entire US, or time series of the entire record, when only interested in a part of it. Another issue in aggregation is linking together all the measurements at a single monitoring point. (Reviewers 3, 11, 18, 19, 20)

8. Many measurements have associated supplementary information that can be quite diverse, may change and needs to be incorporated somehow. Examples of this include cross sections of a stream, lithology or human subjects data. Some sort of catch-all category for qualitative or otherwise awkward data seems necessary. (Reviewers 4, 9, 10, 11, 16, 19)

9. The data structure needs a way to accommodate censored data, i.e. data below or above a detection limit. (Reviewers 1, 8)

10. 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 or could reasonably be interpreted as continuous, so that operations such as aggregation or interpolation are meaningful, or whether the data is categorical. Also, certain types of measurements do not easily fit into this classification. For example the number of daylight hours. (Reviewer 3, 4, 8, 15, 17, 19)

11. There was a suggestion that TSDateTime be indexed to a table containing times so that concurrent measurements could be more directly identified, rather than relying on logical tests as to whether TSDateTime has the same value. There were also questions about the resolution of TSDateTime. Is 1 second sufficient, for example if measurements are made of turbulence? Should offsets to UTC be provided to be specific about time zones? (Reviewers, 3, 5)

12. The focus on a single proprietary and favored set of software raises concerns (Reviewers 1, 6, 11, 12)

13. There needs to be a way to indicate the quality of the data. This can apply down to the scale of individual measurements so consideration needs to be given to a TSValueQualifier field in the time series table, to allow for example flagging data as provisional, corrected (and what the original value was etc.). (Reviewers 1, 8, 10, 14)

Reference

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

Acknowledgements

Thank you to the following who provided reviews (listed in alphabetic order different from the order in which reviews are listed): Tom Ballestero, Roger Bales, Matthew Cohen, Ralph K. Davis, F E Harvey, John Helly, Jeff Horsburgh, Norm Jones, Venkat Lakshmi, Ken Lanfear, Dennis Lettenmaier, Ricardo Mantilla, Andrew J. Miller, Glenn Moglen, N Leroy Poff, Kelly Redmond, Yoram Rubin , John Selker, David Steward , Enrique R. Vivoni, Brian A Waldron, Ross Woods

Appendix 1. Compilation of Individual Review Responses

Reviewer #1

The following are considered in this write-up:

1. What are the limitations to the existing database structure?
2. What changes/additions need to be made to the existing tables in the database?
3. What additional tables need to be added to the database structure?
4. What other issues need to be addressed in the database structure?

Limitations to the Existing Database Structure

The following limitations were identified in the existing time series database structure. Some of these come from Maidment's paper:

1. Integer values in coded value domains are not easily interpretable when viewed within the database or in other data systems once exported.
2. There is inadequate information to identify the source of the data within the existing time series database structure. It is not adequate to assume that the site identifier code will uniquely identify the source of the data because multiple data collection organizations may use the same site identifier (HydroCode).
3. The limitation of the Origin of the data to two values "Generated" and "Recorded" is too restrictive. This is related to number 2. The origin of the data should consider who collected or generated the data and how they did so (by making measurements or by running a model simulation, etc.) and should likely be described by more than one field in the database.
4. The current MonitoringPoint table does not contain enough information to spatially locate a monitoring point outside of ArcGIS. In addition, more descriptive fields in this table (like state, county, elevation, etc.) would allow querying data using criteria other than X/Y coordinates.
5. The current TimeSeries table does not have a depth field to store the depth at which measurements were made. This is especially important for lakes, reservoirs, estuaries, etc. where depth actually makes a difference.
6. The current database structure does not contain adequate data qualifying information. This information is important in interpreting the data, especially where values are estimated or where holding times or other sampling requirements have not been met because these values may be questionable.
7. The current database structure is not clear on how to deal with censored data (data that is above or below a detection limit).
8. When the TimeSeries table grows large (~millions of records) queries to retrieve individual time series may become slow because of the sheer number of records that must be sifted through to get the specific information that has been requested. Creation of additional indexes for the data may improve the speed and performance of the database – especially when serving the data over the internet.

In an effort to address the limitations listed above, the following suggestions have been made regarding changes to the existing time series database structure.

Additional Information Needed in the Time Series Data Model

The following sections provide for each table in the Arc Hydro time series data model some suggestions for additional required fields, additional fields that would be helpful but are not essential, and suggestions for existing fields that should be considered for deletion or moving to another table.

MonitoringPoint Table

Additional Required Fields

Field Name	Description/Justification
Latitude (decimal degrees)	Provides general spatial reference for monitoring locations that is independent of which GIS you are using (i.e., I don't have to be able to read the Shape field to get the spatial information for the point).
Longitude (decimal degrees)	""
LatLong Datum	Identifier for datum of latitude/longitude values (i.e., NAD 1927 or NAD 1983)
Local X	Provides an HO or HIS specific projected spatial reference for monitoring locations that is independent of which GIS you are using (i.e., I don't have to be able to read the Shape field to get the spatial information for the point).
Local Y	""
Local Projection Info	Identification of local projection or datum (may need more information here)

Additional Suggested Fields

Field Name	Description/Justification
State	Provides additional information on which you can specify criteria when you query the database (i.e., give me all data for stations in Idaho, or all data for stations above a certain elevation, etc.)
County	""
Elevation	""
Drainage Area	""

TimeSeries Table

Additional Required Fields

Field Name	Description/Justification
Organization Code	Code that links each observation to the organization that collected the data. Each time series value should be tagged with an organization code because it is possible for multiple organizations to collect data at one location and use the same station identifier (HydroCode). It is not adequate to assume that the HydroCode is enough information to tie the data back to the original source or data system.
Depth	Depth at which the observation was collected – especially important for lakes/reservoirs/estuaries, etc.
Data Qualifier Code	Code that stores any data qualifying information that accompanies the data (i.e., E for estimated)
Analysis Procedure Code	Code that identifies the method that was used to make the measurement (i.e., Dissolved Oxygen by field measurement using a DO probe versus DO by Winkler Titration).
Source Database Code	Code that identifies where the data originated (i.e., USEPA STORET, USGS NWIS, etc.) and that ties the data to the original data file in the HIS Digital Library. It is not adequate to assume that the HydroCode is enough information to tie the data back to the original source or data system.
Sample Medium	Medium of the sample that was collected (i.e., water, sediment, fish tissue, etc.)
Value Type	Code identifying the value as observed, calculated, simulated, etc.

Additional Suggested Fields

Field Name	Description/Justification
QAQC Code	Code that could be used to categorize the quality of the data (i.e., 1 for USGS because they have high quality data, 2 for a local organization that does not have rigorous data standards, etc.). Provides additional information for which criteria can be specified in queries.

TSType Table

Additional Required Fields

Field Name	Description/Justification
Censored	Code or value that specifies whether censored values (below or above detection limit values) can occur. Could just be a yes/no.

Additional Suggested Fields

Field Name	Description/Justification
Time Series Category	Code that could be used to categorize the different time series types (i.e., Climate as a code for all climate related parameters, water quality for all water quality related parameters, etc.). Would provide additional information for which criteria can be specified in queries.

Fields to Move/Delete

Field Name	Description/Justification
Origin	Air temperature is air temperature regardless of who collected the data or where the data originated. This field should accompany each observation in the TimeSeries table (see the suggested use of Source Database Code field, which would specify where the data came from and Value Type field, which would specify whether the data are observations, simulation results, etc.) but should not be in the TSType table to avoid duplication or difficulty querying data out of the database.

Proposed Additional Tables

AnalysisProcedureCodes – contains descriptions of the codes used to represent the analysis procedures using in measuring the data.

SourceDatabaseCodes – contains descriptions of the codes used to represent the source database from which the data were obtained (i.e., USGS NWIS = United States Geological Survey National Water Information System). Also could contain a link to the original data file stored in the HIS Digital Library.

OrganizationCodes – contains descriptions of the codes used to represent the organizations that have collected the data (i.e., USGS = United States Geological Survey).

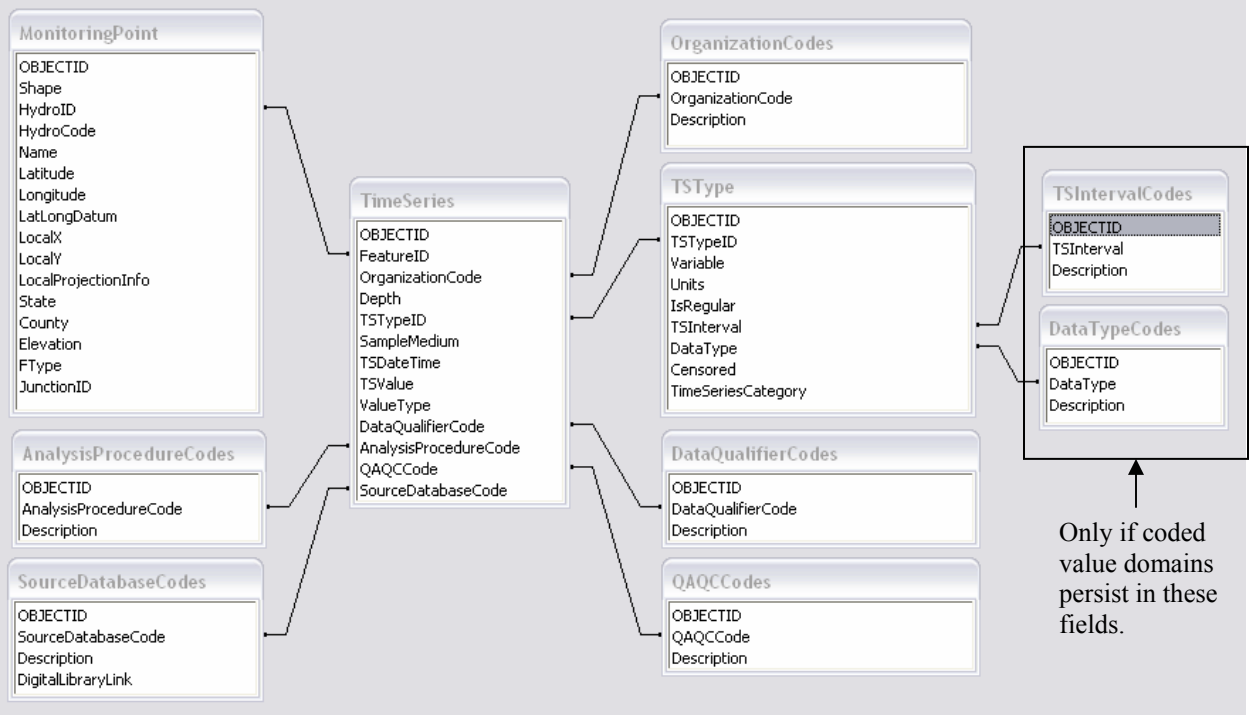
DataQualifierCodes – contains descriptions of the codes used to represent the data qualifying comments (i.e., E = Estimated value).

QAQCCodes – contains descriptions of the codes used to represent the quality of the data (i.e., 1 = high quality, 5 = low quality).

TSIntervalCodes – contains descriptions of the 17 codes used to represent the time interval over which the data were collected (only if the TSInterval code remains as a coded value domain).

DataTypeCodes – contains descriptions of the 6 different types of data (only if the DataType field remains a coded value domain)

Proposed New Database Structure



Additional Issues

1. Spatial Reference of Monitoring Locations

It is my opinion that the database structure should support a variety of 3rd party software programs – including those used to display the geographic information contained in the database (i.e., monitoring point locations). It is unfortunate that ESRI has chosen to make the "Shape" field that holds the actual geographic information associated with a shape into a proprietary binary object. Since I don't anticipate ESRI publishing this any time soon, it is imperative that the database hold enough spatial reference information for monitoring points to be located geographically. This should likely include a latitude and longitude and/or an X and Y coordinate from whatever local coordinate system the user has chosen to use for their data.

Map projections are an issue here – if a user is not using ArcGIS it is a problem if all of the spatial datasets they are using (including the locations of monitoring points) are not in a common, projected coordinate system.

2. How to Deal With Censored Data

Censored data are an issue with water quality data, for both surface water and groundwater. If censored data are left out, descriptive statistics calculated from the remaining data are skewed. There are many statistical methods that can be used to do censored data analysis (i.e., calculation of the mean of a distribution that includes censored data points), but all depend on how these values are represented in the database.

My suggestion is to use the "Censored" field in the TSType table to indicate whether censored data can be present in a particular time series, and then if Censored = True store below detection limit data points as negative numbers in the database where the number stored represents the censoring level (i.e., a reported value of < 0.05 mg/L for say total phosphorus would be stored in the database as -0.05). This way, queries can know whether to look for censored data, they can distinguish between above and below detection limit values (positive versus negative values), and they can know what the censoring level is. Above detection limit values (such as a fecal coliform bacteria concentration reported as "too numerous to count") are a bit more tricky and I don't have a great solution for these yet.

Another option is to store the value of the censoring level in the TSValue field (as a positive number) and then flag the values as above or below the detection limit in the proposed "Data Qualifier Code" field. The only problem with this approach is that there may already be a comment in the qualifying comment field. If this is a problem, an additional field could be added to the TimeSeries table indicating whether the value stored in the TSValue field is censored or not (true/false). Obviously this would have some implications on the size and possibly the performance of the TimeSeries table.

3. *All fields with coded value domains should have a linked table that defines the codes (i.e., TSInterval, DataType, etc. in the TSType Table).*

As a general practice for easing the interpretability of the data this rule should be followed. As an alternative, these fields could simple be changed from integer values to text values that are more easily interpreted (i.e., you don't need a linked table if the information in the field is self explanatory). This is also critical when exporting the data. If you provide people with a delimited text file that only contains numbers they will not know how to interpret it.

In all tables except for the TimeSeries table I would suggest not using coded value domains because it increases the number of tables in the database that must be managed. Expanding information in fields in all tables but the TimeSeries table will likely not have much effect on the size or performance of the database because of the relatively small number of records in these tables. However care should be taken with fields in the TimeSeries table. This table could become quite large (millions of records) and so coded valued domains become much more feasible and necessary to keep the size of the database down and to keep performance up.

4. *Flexibility in addition of tables*

Applications and tools build on top of this database structure may have some specific needs in terms of adding additional tables or information to the database. For example, the Time Series Viewer tool that we have built at USU adds an index table to the database that lists all of the time series types that have been collected at each monitoring point so that the tool doesn't have to query the whole TimeSeries table each time it is populating the list boxes where the user selects a monitoring point or time series type. This is done purely to increase the performance of the tool over the internet.

I don't see where this would be a problem if people are just adding tables to the database because the core data structure remains unchanged. However, there may be issues if someone wants to add an additional field to one of the core tables.

5. Spatial Referencing/Linking of Monitoring Locations

It makes a lot of sense to assign a JunctionID to a streamflow gage to link it to the stream network because streamflow gages are located on streams. The same is true for water quality stations. However, things like climate stations, groundwater wells, etc. are not located on the stream network, but you may still want to link them to things like which watershed they are in, etc.

Additional discussion comments by reviewer #1 following up on comments 1 and 2 in the summary.

Comment 1 is good, but I have a bit of an issue with the suggestions under this comment in the summary - I would not expand the TSType table to include information about data source and heritage. For example, water temperature is water temperature regardless of who measures it. Data source and heritage information should accompany the individual observations.

Imagine trying to query information out of the database for a particular parameter (lets keep with our water temperature example) - if you include data source and heritage in the TSType table you will have to figure out how many TSTypes are associated with water temperature (i.e., TSType 1 may be USGS temperature, TSType 2 may be Utah DWQ temperature, etc.). It makes much more sense to have a single TSType for water temperature, with source and heritage information associated with the individual observations (i.e., a query to the database for all data with TSType = 1 = Water Temperature returns all of the water temperature data - regardless of who made the measurements). Additional criteria can be added to the queries to limit the data to a single source, etc.

Reviewer #2

The contribution of this paper is mainly in that it lays out some of the characteristics of hydrologically-relevant data (e.g., NWIS, Storet) and presents a simple data model for time series data that has been used as the basis of the ArcHydro extension for ArcGIS.

The data model it presents is predicated on the conceptual model of the data cube that depicts a measurement as a point in continuous 3-space. Only one of these dimensions is actually continuous and ordered, time, while the space and parameter dimensions are encoded in discrete, nominally-valued parameters that have no natural ordering. Even though the feature-id is numbered it is not clear what information is contained in the number other than its uniqueness as a name within the ESRI application. It is, therefore, nominally-valued. The data parameter dimension uses values that are arbitrarily named and also have no natural order and it is not clear that there is actually a meaningful axis for this classification scheme other than alphabetical or subjective grouping. Consequently, the data cube gives the false impression of continuous space-time while it actually hides the measurement data in metadata that is not directly stored in the data model (cf. P 15).

While this data model is relatively simple and can be stored in any relational database, it is not clear why one would want to do that given its structure. It is unnecessarily complex in requiring at least three tables. The data model actually requires more information or tables since the coding scheme has to be stored somewhere; in other tables, for example.

As a data model for time-series data it is not one that I would readily use or recommend. A more general model would explicitly encode space and time (x, y, z, t) with explicit units, relevant datum's and projection information, site identification or naming, parametric values and units and data heritage or provenance. This could all be contained in one simple table for water quality and water quantity for example, although it is my opinion that no single data model actually exists that is sufficient for all purposes.

Other observations:

- This data model depends heavily on a highly idiosyncratic classification scheme (p. 16) that is dependent on classification rules that are not described. It appears to be designed for certain types of analysis and many of the assumptions that are implicit in that analysis are not captured in the data model.

- comments on the importance of preserving the differences between instantaneous and average data are useful but nothing is addressed about the statistical character of the average data in terms of errors or sample size. These differences are not addressed in the data model

- There is no treatment of the transformation of raw data to that of the data model. There is no mechanism for tracing the origin of any record in the relational database to its source data not to the processing needed to obtain the conversion.

Reviewer #3

1. Is the AHTSDM **capable of storing** all hydrologic observation data that you think this database should contain? If not please indicate the data that does not fit this data structure, explaining why and providing suggestions regarding a data structure that could accommodate this data.

- see my points 4, 7, 8, 11

2. Is the AHTSDM an **efficient way to store** all the hydrologic observation data that you think this database should contain? If not please indicate the data and situations for which this structure is inefficient and provide suggestions for making the data more efficient.

- Not clear for time series at single points, because the temporal indexing system is not described in any detail.
- This would be best tested with ~50 years of hourly flow or rainfall

3. Is the structure of AHTSDM **efficient for the querying and analysis** of hydrologic observation data? If not please explain the shortcomings and indicate how they might be overcome.

- Very hard to tell without using the system
- Depends on the queries the system is intended to address – is there a design specification?
- Some kind of time indexing will be needed for high-resolution temporal data.
- In the Tideda system in New Zealand, the database internally stores time measured in (8-byte) integer seconds since a time origin – this ensures that floating-point errors never contaminate time-based calculations

4. Spatial information is represented in the data model through the **monitoring point** (table 9 page 15) being a feature within a GIS database. **Is this sufficient** or is additional spatial information necessary as part of the Hydrologic Observations data model.

- see my point 4
- No evidence that time series data on spatial grids is well-catered for – this could be problematic for weather radar and for some types of simulation model output
- Perhaps the TSValue attribute could be generalised to an array or vector of values, in one-to-one correspondence with point locations within the spatial feature (a single value corresponding to a measurement point, a vector of values corresponding to points along a transect, an array of values corresponding to remote sensing pixel footprints.)

5. The **TsType table** in the AHTSDM (table 9 page 15) is effectively the location where metadata about the data values in the TimeSeries table are stored. Are the **contents of this table sufficient** for providing the ancillary information that needs to be kept in the database with hydrologic observations? If not what additional ancillary information should be part of the database and how should it be stored?

- No. see my points 7, 8, 9

6. The abstract indicates that the goal is to synthesize observations of **streamflow, precipitation, climate, water quality and groundwater** into a single database. Is this a **sufficient set** of "hydrologic observations" or should others be considered? Would other observations fit this model?

- Others should be included – even if the data model does not accommodate them initially, try to make the model sufficiently extendable that they can be included later.
- Most point observations of a single environmental parameter would fit – e.g. soil moisture, isotopic concentration of rain water
- In some cases there may be observations which only occur as two or more values recorded simultaneously – should the definition of TSValue be extended to a vector of elements at each timestamp?

7. Please provide any additional comments or suggestions.

- Interaction of data in this system with models is important, and yet you cannot reasonably expect to forecast all the future spatial structures that might be used by those models. It would help to describe which types of hydrology models might be compatible with the spatial data structures currently available in Arc Hydro.
- As hinted at in the paper, specialised analysis tools for temporal data will be needed. It is likely that some data users will want a basic set of timeseries tools available within the same environment that stores the data (i.e. ArcGIS). Other users will want to link their own tools to the data (preferably by accessing the data directly from the database by an industry standard access method which does not depend on operating system. Alternatively by exporting the data in ascii or other standard format (netcdf?))

Review comments

1. P1 ""“get me the all **the** hydrologic data for my region in a consistent format” rather than having to search all over the internet for data sources, spending long periods of time learning how to operate these various web sites, and synthesizing the data in many different formats that the web sites provide. If such a service could be created, professional hydrologists in government agencies and consulting firms would also find it useful." This seems to address a different question than that posed by the review

2. P2 "Hydrologic observation data have the following characteristics:

- They are indexed in *space* by the latitude and longitude of a point representing where the observations were made;
- They are indexed in *time* by the date and time at which each observation was made;
- They are indexed as to the type of *variable* being observed, such as streamflow discharge, water surface elevation, water quality concentration, etc."

This description is a good start, but omits important qualifications, and thus underestimates the information needed to define spatial or temporal measurements. A measurement has support, extent and spacing. These comments apply to both spatial and temporal data. Without knowledge of the support, extent and spacing, the data cannot be interpreted correctly in space and time. The standard reference is Bloschl and Grayson. I recommend that an additional bullet point be added, noting the importance of support, extent and spacing.

3. P3 "This distinction between *instantaneous* and *average* data …" This makes the point that I note as missing on P2, but I think it needs to be made as just one example of a more general principle.

4. P13: "any point location in space where hydrologic observations have been made": There seems to be an implicit assumption that it is a single time series which needs to be associated with a single spatial entity. This is true for data from a stream gauge, water quality monitoring site, raingauge or climate station. It is not true for a remote sensing image, a radar reflectivity image or any other data set which has both spatial and

temporal structure. Nor is it true for observations made from a moving instrument, such as an airborne sensor.

5. P13: Although it may be obvious to the author, it is important to mention that by associating the time series data model with the Arc Hydro data model, this enables a link between temporal and spatial data (such as GIS maps).

6. P15: "A critical point is to specify when the data value occurred and to what time interval it applies" AND also to specify what types of temporal operations (e.g. interpolation) are legitimate

7. P16 "has a time stamp of 2004-01-21 00:00:00" This timestamp implies that time resolution is one second in the data model. Is this fine enough for all likely data sources? For example, will sub-second information from turbulence observations be included?

8. P16: it is not clear how the time origin (or time zone & daylight time) is stored in this data model. Perhaps the TimeSeries table needs another attribute to define this? Enough information needs to be provided to allow time information to be transformed from any timezone (and time of year) to any other. Providing the offset from UTC would be one way to achieve this.

9. P16 "Arc Hydro classifies time series data into six types: … *Instantaneous* … *Cumulative* … *Incremental* … *Average* … *Maximum* … *Minimum*". I think that the type "Instantaneous" type needs to be subdivided, because the observations at a water level recording station will have the same type as spot gaugings or water quality samples. I believe that they should be considered differently, because it is meaningful to carry out many temporal operations on the data from the water level recording station, but there are very few meaningful temporal operations which can be carried out on the spot samples. The Tideda system uses Instantaneous for the recording data, and Gauging for the spot data. The attribute *IsRegular* is not sufficient to resolve this, because regular monthly flow measurements are definitely not adequate to calculate daily average flows

10. P18 "stores its Nexrad data using the Arc Hydro format with a value for each Nexrad cell in which rainfall occurs on a grid" This gives the impression that time series could be extracted at any individual point. Is it also possible to extract a grid of Nexrad data at a single time? And to take this to a larger scale, is it possible to extract data records not for a single spatial feature, but for many spatial features simultaneously (e.g. all flow recording sites within a specified space-time extent)?

11. There is no discussion of a facility for storing the calibration or instrument history. As well as standard measurements made by organizations such as USGS, there will be experimental measurements, for which the measurement method and/or sensor deployment needs to be recorded. There may also be instruments which do not sense a hydrologic quantity directly (e.g. a TDR probe), and for which a site-specific calibration curve is needed. Will the raw data (e.g. frequency) and calibration curve be stored in the

system, or will the calibrated series be stored? How will improvements to the calibration be managed? What if the calibration curve varies over time? These comments also apply to river flow data obtained from river sites where stream gaugings are used to develop a stage-discharge rating curve. For new sites typical of experimental studies an initial calibration is often developed early in the study, is refined over time, and may always be provisional. Since calibration curves come in many different forms, and may be applied in many ways, it may be more convenient to store only data with the calibration applied. In this case the details of the calibration need to be available as metadata.

12. Does ArcGIS provide a database connection (if appropriate) between a *MonitoringPoint* and the boundary of the catchment that drains to the *MonitoringPoint*?

Followup information from Reviewer #3
In regard to CUAHSI, see also the comment below from a database architect.

I've only had a brief look at this but it seems that they're trying to cobble together a relational time series data structure, which has been tried before with minimal success. Perhaps you could point them to Informix IDS, which offers purpose-built spatial and time series blades? These provide integration (spatial, time series, RDBMS), performance and scalability (among other things).

The web site explaining Informix IDS and "spatial and time series blades" is
http://www.ibm.com/software/data/informix/blades/
The spatial data blade provides relational-style efficient access to GIS data. We are using it to serve GIS data on the net. A time series datablade would provide relational-style efficient access to time series data.

This is much more of the classic IT professional approach, rather than a scientist posing as database expert (which I sometimes do). It's up to CUAHSI to decide what approach they want to take.

Reviewer #4

1. It is not clear how ArcHydro let's the user know the actual time base of the data (daily average, instantaneous, etc.). If data was averaged, obviously the raw data would be extremely helpful to researchers, for example if event mean concentrations (EMCs) are reported for stormwater data, the flow and concentration instantaneous data were synthesized into the EMC..

2..How does the database inform the user that discretized data (say daily rainfall or streamflow) was or was not continuous from time step to time step (just because 2 days in a row have reported precip or streamflow, does not mean that the process was continuous)?

3. Not much was said about the characteristics of the atmospheric processes, such as daylight hours, cloud cover, etc. Has such data been identified and will it be included in the database? Along these same lines, I would expect that, much like groundwater, snowpack, snow water equivalent, depth of frost, and soil moisture, is sparse data....but with many similar issues.

4. It is not uncommon that a raingage was moved "a little" and therefore the reporting organization considers the data as being continuous. How will the database handle this?

5. I might be way out on a limb on this one but.....Little was said about attendant geographic hydrologic information that may also be time variable. For example, sediment transport or velocity measured at a cross section might be reported, but if the cross section geometry is changing in time (for example, as in a meander bend), how might the time-variable characteristics of geospatial referencing reflect this?

6. Is all of the historic data available truly in the same datum's?

Reviewer #5

In general, I like Maidment's time series design. I think it is well laid out, clear, and will suffice for a broad range of data types. My two concerns are related to timestamping and metadata.

1. It would be really nice if the TSDateTime was input using a relationship class to a table containing times, as opposed to using an entered time. Let's say the TimeSeries table includes two records, one for the gage height of a stream and one for the discharge of the same stream at the same time. Currently, the only way to know that the gage height and discharge are related at exactly the same time would be to logically test whether the TSDateTime had the same value. Using a timestamp from a table with a unique ID could enable clearer relationship structures.

2. It would be nice to include a metadata identifier table for each time series measurement. This metadata table would include an ID, URL, etc. necessary to identify how the original data was collected and how the generated data was obtained from a model.

The structure is robust enough to be utilized for most groundwater data. An example of where we could run into computational limits would be, for example, if we put a set of data loggers into wells to measure head at say 10 minute intervals (which was done recently by the KGS at a number of wells in NW KS), giving about 50K measurements per year per well.
If we start dealing with cases where there are many measurements related to a single object, it may be necessary to devise a structure whereby we partition TimeSeries and TSType into sets of tables related to individual MonitorPoint features.

Thank you for your work chairing the review committee and for allowing me to make comments.

Reviewer #6

Cover remarks
I am not a GIS expert and thus feel less competent to comment on all that ArcWhatever is being asked to do for CUAHSI. But what experience I have had, and vicariously through others who

have tried to work with both space and time, and on relatively simple and more constrained sets of issues than what will be asked of HIS, leads me to be pretty cautious about simply embracing this approach, crossing our fingers, and hoping for the best.

Also, what CUAHSI adopts will serve as a kind of forerunner to other NSF efforts such as NEON, and the sense I get from the several of ~those~ meetings I have attended is that there is a fair degree of sentiment for local autonomy and involvement, and a reaction against perceived excessive planning at the national scale, although it needs to work that way under their (NSF) rules. I think these feelings on the part of potential participants need to be carefully considered, even though I do like the idea of a national framework, resulting in a kind of federalist system, if that analogy makes sense.

I like a significant element of the tried-and-true, as a starting point, because we cannot afford to have something that doesn't work at all.

Review comments

This article begins with a trip to some of the other communities that deal with environmental data. This is about as good as any way to start, but should soon ask what qualities and behavior would the potential user community like to see from a Hydrologic Information System.

A good point of departure for a user-driven and user-responsive system is the set of surveys conducted and presented at the HIS meeting in Austin. There was a lot of very good and useful material contained in these. The gist of these surveys was, as expected, that there are two main desires for information: 1) data and observations, and 2) derived summaries and manipulated quantities, generally referred to as "products," which are nothing more than functions of the data and observations. And the simplest such function consists simply of the data themselves, so in effect 2) can encompass 1).

Since much of hydrology deals with basins and areas, and their surface and subsurface properties, information on this constitutes a third need. Hydrology is not alone in this regard, but has had a "head start" in terms of needing to meet a demand for geographic information. The sentiments and thoughts expressed in these user surveys should be a significant driver for the development of an information delivery system. This means close and frequent contact with these users, and iteration, and early involvement from the start. This seems especially important if CUAHSI is to serve as infrastructure to facilitate innovative research.

Of the various examples discussed, the USGS streamflow system is the most intuitive and easiest to use for the uninitiated, and for return users. But it has relatively limited scope, and is not hooked to a set of analysis tools or products. The NCDC approach is described quite well, including the numerous frustrations encountered in getting through the asteroid belt that seems to protect it from users. This system was designed with almost no user involvement and it shows. It also knows only about NOAA data, which is just a small (though important) component of the available weather and climate data so needed by hydrology. Also, this system provides almost no real products, distillations, summaries, visualizations, probabilities, frequency distributions, wind roses, and so forth, which is the main reason many users are after the data to begin with.

These examples are illustrative of where things are now, but clearly we want to get beyond the current state.

A well-designed system should help a user in need of data or information to first sharpen the question about what it is they want or need. There is an interplay between what is desired and what is available, and arriving at an accommodation of these two sometimes conflicting concerns, helping a user learn what is available, is a major need in a user friendly system.

There are multiple ways in which people search for data and information, some spatial, some temporal, some topical, some by element, some using other mental approaches. A robust system needs to be able to cater to all of these.

With every data set there are numerous subtleties and arcane qualities, not the least of which is an understanding of how the collection and archive system works, because these can be immensely complicated and often involve a large cast of disparate characters, a fact little understood by users. This strongly argues that to the extent possible the human expertise of this "data priesthood" (as it has been called) be folded into the data archive and distribution system. Geographic Information Systems have a significant role to play as part of a Hydrologic Information System, and a generic ability along these lines seems essential. The focus on a single proprietary and favored set of software, however, raises lots of warning flags. The playing field should be as level as possible. The examples given all relate to a particular commercial vendor. Unlike the system for access to weather and climate information at NCDC (built around Oracle) the suggested hydro approach seems to require that the user have this software on their own computer. But those who use other systems should not be at a disadvantage when trying to extract information for their hydrologic need.

There are certainly a large number of users that like the idea of a geographic interface to point data, and this needs to be developed for all of the allied data sets that hydrologists will want to access. he ability to obtain time series from multiple sites at once ("all sites with groundwater values during the drought of 1988 for a five-state area") is frequently requested. User requests will range from individual small point data sets to large "give me everything" requests. Methods to keep such requests manageable and practical for the provider, and to prevent abuse or degrade system performance, are needed.

Along this line, a hybrid between a distributed and centralized system is probably best. Each has their advantages. As an analog, the monolithic CDO system at NCDC sometimes goes offline or suffers from poor connectivity, and acts as a single point of failure. It also does not have an army of innovative enthusiasts writing code to offer improved products or access. The distributed Applied Climate Information System (ACIS) of the Regional Climate Centers stands in counterpoint, as it ingests data from a variety of federal and state platforms reporting at a variety of intervals, and maintains multiple synchronized copies of the historical (obtained from NCDC) and recent provisional data bases.

So far we have not been impressed with the ability of Geographic Information Systems to handle both spatial and temporal duties equally well. Environmental data systems have tended to be good at one or the other but not both. But GIS could serve as a spatial interface to data.

Also, whether all the data need to be available locally, and thus collected beforehand, or else obtained from the various disciplinary data centers on demand, needs to be discussed. Overall, the latter approach seems preferable because it builds on existing systems and takes advantage of existing expertise and infrastructure. But all of these centers need to know and be involved so that they can be prepared for increases in load, or other demands related to reliability, speed, search capability, and so forth.

So far I have not mentioned much about the ArcHydro time series data model, introduced at Austin during an impromptu lunchtime meeting. This is relatively new, still under development, has no track record, does not have an active and engaged external community building applications centered on this software, and it seems unwise to put all the eggs in this one basket. It does seem worth exploring to what degree it can assist with the issues of concern to the user community. But I would have a lot of reservations about building everything around it without a considerable track record.

As a special case, many of the Hydrologic Observatories, especially in the East, might want to utilize Nexrad radar data, or other remotely sensed spatial or spatio-temporal data. This involves large volumes of data. We do not know if AHTSDM is up to the task of handling this efficiently.

Several questions were posed as to the suitability of AHTSDM for HIS usage. This is so new that it seems we cannot answer them very well at the moment, and maybe not for some time to come. This reinforces the thought above to be on guard against a single approach that is not robust enough to handle a variety of situations and data.

Reviewer #7

1) Biological (and habitat) data can fall into 3 spatial scales: point, reach, and watershed. The AHTSDM structure seem fully appropriate for point-based biological and habitat data. Water quality measurements are a good example. Many biological databases, however, are not strictly point based, but come from some kind of averaging over a larger spatial scale, e.g., a "reach." The USGS NAWQA and USEPA EMAP data fall into this category. Habitat data supporting these biological measurements are also typically at the reach scale. Therefore, the AHTSDM format needs to allow for these kind of data (perhaps using a river reach code as a locator). Many biological datasets also include watershed scale habitat measurements (e.g., %land use cover) that may influence rainfall-runoff processes and these should also be included as possible fields. Some of these watershed-scale measures are also biological (e.g,. %riparian cover in drainage network) and need to be captured, especially as LIDAR and other technologies facilitate whole watershed assessments.

2) It seems very well thought out and flexible, if the scale issues above (#1) can be resolved.

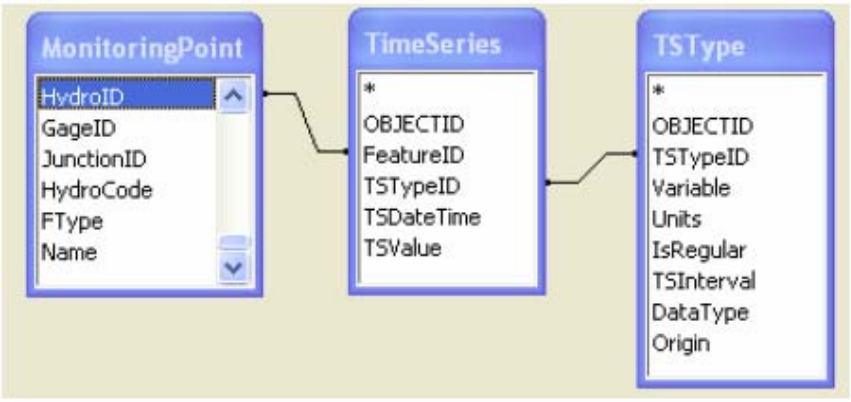
3) See point below (in #4)

4) I would like to see additional fields for reach (USGS/EPA format?) and possibly (sub)watershed (8th code or finer HUC or HUB?).

5) I don't see much issue for biological data here.
6) Obviously, I think biological data should also be included! 'Water quality' (basically water chemistry) data alone are not adequate. Variable types under 'biology' include: nutrients, algae/periphyton, macroinvertebrates, fish, and riparian indicators.

7) Ideally, it would be really good to have a delineation of the upstream watershed for every sampled point/reach on the map. This would allow searches for information within the watershed (e.g., stream gauges, precip gauges, other biological samples) for modeling and hypothesis-testing purposes. Also, some kind of simulated hydrograph for point/reach locations would be ideal. Obviously, these data generally don't exist for all points/reaches, so some modeling would be required. It may not be the role of HIS to provide this model support, but I think it would be extremely helpful if there were some supporting documentation on 'HOW' to go about accessing models and associated GIS coverages, databases, etc. that can delineate watershed boundaries and generate synthetic hydrographs for ungauged locations.

Reviewer #8



TimeSeries Table

(Critical!) TimeSeries is missing a critical element to deal with censored values, e.g. values below a detection limit. Although rarely needed for streamflow, this capability is essential for water quality. TSValueModifier should be added, with allowed values, “gt” or “lt”. (The biologists may add other values.) An alternative would be to allow the modifier within TSValue, but this would turn the element into a string to be parsed.

It also would be helpful to include some indicator of data quality to distinguish, for example, provisional data from final or approved data. I recommend adding a TSValueQualifier element to hold provider-supplied qualifiers, such as “provisional”. An alternative could be to include this element within the TSType table, referring to the whole TimeSeries, but that would not allow marking individual values.

TSType Table

I recommend adding TSSourceURL element for those series that are directly imported from places like NWISWeb.

The purpose of the IsRegular element is not clear. Can the IsRegular value conflict with the TimeSeries/TSDateTime element which may have irregular intervals or missing values? My recommend solution is to delete the IsRegular element and let whatever software is analyzing the date use the TimeSeries/TSDateTime element to figure out if the series is regular for its purposes.

If you are dealing with a regular reporting interval, then how do you handle missing values? Do you add an element, TSMissingValues, or do you compute missing values as you populate the dataset? Having a TimeSeries/TSValueQualifier element would help in this regard.

We already know that the DataType element is too confining. Aren't both TSInterval and DataType implied in the Variable element?

I think the real problem with the TSType table is that it's trying to be metadata but isn't going far enough. It works if you're importing data for immediate use, where the source and limitations are fresh in your mind, but will not stand the test of time for true metadata. The Origin element, for example, is a little bit helpful by distinguishing model data from sample data, but tells nothing about the model or the sample. Adding a TSSourceURL element will help somewhat, but you probably want to think of a pointer to fully-documented metadata, TSSMetaURL. I suggest looking at the Advisory Committee For Water Information data elements, http://wi.water.usgs.gov/methods/tools/wqde/index.htm .

Basic Relationships

The data schema seems designed for a limited number of TSType rows. However, there needs to be a TSType row for at least every station-parameter combination. Including a TSSourceURL element would add even more rows. It seems inefficient to put FeatureID in every TimeSeries row when you'd almost never look at an individual value in isolation.

Why not remove TimeSeries/FeatureID and place it into TSType/FeatureID? MonitoringPoint/HydroID would point to many TSType rows, which would point to many TimeSeries rows. The new structure would be:

MonitoringPoint	**TSType**	**TimeSeries**
	OBJECTID	OBJECTID
HydroID ←>>	FeatureID	
	TSTypeID ←>>	TSTypeID
GageID	Variable	TSDateTime
JunctionID	Units	TSValue
HydroCode	…	TSValueQualifier
FType	TSSourceURL	
Name	TSMetaURL	

Reviewer #9

As usual, a delight to read Maidment's prose. It seems clear that point-wise and area data are well considered. I wonder about human subjects survey data (e.g., a meeting of people from a region who are asked a set of questions with the possibility of essay answers; stories of residents about the region in years past (floods etc.); educational materials that are specific to an area). I think we need to talk to folks on the social sciences side to get their feedback. At a minimum it seems that we need a catch-all category of "qualitative regional data" which is flexible in data format and metadata.

Reviewer #10

Here are a few comments I hope you will find useful. Not being a frequent user of large databases, I have but a limited vantage point. Also, these are just personal notes, and should not be interpreted in any way as an endorsement of some sort from NCHS. Overall, I am very impressed by this effort, and my comments are intended to help, not criticize. Thank you for coordinating the review effort.

1. I would think that the desired database ought to provide an exact definition of the variables contained in the database(s), as well as the measurement procedure(s). For example, concentration. In subsurface applications, concentration can be given as mass per volume, or as mass per (solid) mass. They can be volume averaged or flux averaged (different concepts altogether). Or velocity measured in a river. One would need to know how it was measured, at what locations, depth etc, what instruments were used, etc. In conclusion, one should also be careful not to throw data of all sorts into the mix.
2. A few of the data mentioned in the discussion are obtained by averaging of some sort, or integration, etc. I imagine that one would need to know what's the algorithm used (not so much for averaging, but numerical integration can be quite tricky, as you well know).
3. Many studies provide images, cross sections etc. These elements should be easily retrievable. They are particularly important in subsurface applications, where data are often

given in the form of well-logs, or geophysical images, not just numbers (see for example, Lunt et al., Sedimentology, 2004, 51, 377-414. I assume also that geomorphologists have similar interests. Providing links to where the data acquisition procedures are discussed may help. I suppose the generalization of that is that research communities other than surface hydrologists may have different needs, and they should be consulted.

4. There may be a need to QA the generated time series by the user (for example, by comparing the generated time series with the source, just to make sure that the query indeed performed what it was supposed to do). I, for one, would like to QA the data I am working with before embarking on an analysis. So there should be an easy way to do that by viewing the data at the source and compare it with the data provided by ARC Hydro.

5. I would think that one would want to adhere to some commonly used practices. Are there any such standards?

Reviewer #11

1 Is the AHTSDM capable of storing all hydrologic observation data that you think this database should contain? If not please indicate the data that does not fit this data structure, explaining why and providing suggestions regarding a data structure that could accommodate this data.

- Concern about time series data for subsurface profiles of moisture, temperature etc. Typically we ill have observations of the type (depth, value) for each location. This can be very well represented by AHTSDM. In some cases, especially hydrologic modeling outputs, data corresponding to subsurface layers is reported. In such a case we need a representation like depth1, depth2, value) corresponding to the depth extent for each location. Spatial attributes of he data are represented using ESRI shape files in AHTSDM, but shape files do not yet have a very good data model for representing "volumes" in case of multi-layered data. The (space, time, variable-type, variable value) representation is excellent due to its broad scope over all types of hydrologic data. Please note that variables such as soil type obtained from STASGO or other SGC data bases can be reported as SOILTYPE, UPPER_DEPTH-LOWER_DEPTH. Observations of soil moisture and temperature from USDA-NRCS SCAN networks are also referenced as DEPTH, SOIL_MOISTURE, SOIL_TEMPERATURE

2 Is the AHTSDM an efficient way to store all the hydrologic observation data that you think this database should contain? If not please indicate the data and situations for which this structure is inefficient and provide suggestions for making the data more efficient.

- AHTSDM seems to be an efficient way to store point observations. As mentioned above an efficient way to store subsurface layered data may be needed. How can AHTSDM handle spatial map-like data such as that from satellite r/s which may or may not be gridded?

3 Is the structure of AHTSDM efficient for the querying and analysis of hydrologic observation data? If not please explain the shortcomings and indicate how they might be overcome.

- Querying of database will see a significant improvement through the AHTSDM framework due to the underlying geodatabase. However, what software platforms will be required for hydrological modeling? What will be typical steps in integrating my hydrological model written in say FORTRAN with data input in the AHTSDM framework. How will existing hydrological models ingest hydrological data formatted according to AHTSDM specifications? For example,

if I need to retrieve the spatial attributes, say latitude and longitude, will I need to use an ESRI application for the shape file format? HDF – EOS data for example provides a solution in the form of providing the users with the HDF library that allows subsetting, retrieval and such operations on the data set using simple programming statements. The users do not have to worry about the internal format of HDF-EOS data. They can just be familiar with the general structure of the data set as in is it swath data, grid data or point data. Since AHTSDM is built on an underlying ArcHydro data model, it will be required to provide users with such a library (Open Source preferably), which will allow them to write simple scripts for very basic operations. Basically, apart from the data model, a library for accessing and manipulating data will also be required.

4 Spatial information is represented in the data model through the monitoring point (table 9 page 15) being a feature within a GIS database. Is this sufficient or is additional spatial information necessary as part of the Hydrologic Observations data model.
When you have remotely sensed data, aircraft or satellite, how can you represent the interface to do the same as Figure 8 on page 14? Can the query directly access the stored images (mostly in binary format due to size)? I would very much like to see this come thru as it represents a sufficient breakthrough for this visualization tool!

5 The TsType table in the AHTSDM (table 9 page 15) is effectively the location where metadata about the data values in the TimeSeries table are stored. Are the contents of this table sufficient for providing the ancillary information that needs to be kept in the database with hydrologic observations? If not what additional ancillary information should be part of the database and how should it be stored?
Data model for time series is complete and encompasses all observation types that I can think of.

6 The abstract indicates that the goal is to synthesize observations of stream flow, precipitation, climate, water quality and groundwater into a single database. Is this a sufficient set of "hydrologic observations" or should others be considered? Would other observations fit this model?
- Other station data such as air temperature. Remotely sensed data? Vegetation, Air and surface Temperature, Soil Moisture, Snowmelt, atmospheric water vapor and temperature profile and cloud fraction and height, each would fit well into the (TSDateTime, TSTypeID, FeatureID, TSValue) data cube representation. Only concern would be the representation of spatial resolution in terms of FeatureID framework. Basically how would a single grid cell (with a location and cell size) be represented by the FeatureID specification? Does AHTSDM need a raster data model also?

7 Please provide any additional comments or suggestions.
Each data set should have a field for Keywords just as in a journal publication; keywords describe the broad scope of the paper. This will be very useful from data mining, web search results point of view.

Reviewer #12

let me give you a very brief response, i may have more comments later.
my thoughts are based on what i've been able to learn about the his from the review meeting last fall, discussions with david m., discussions with other members of the standing committee, and a meeting i had with john helly at sdsc last week. there are many aspects of the his project, and this short note should not be construed as reflecting on the merits of the work that has been done to date. instead, it deals with a major issue that i think needs to be addressed, sooner rather than later -- specifically, the role of proprietary software. the david m. document you attached seems to do a reasonable job of laying out the problem and a pathway to solving it. the problem i have (and i believe i share the concern with many others in the hydrologic community) is that the pathway that is laid out is based on a commercial software family. i don't think that this necessarily has to be cast in an "either-or" framework, but i see real resistance developing if this is the only pathway. i think that it is important that a parallel open source pathway be provided by the his team -- even if it is not as well developed as the arc hydro one. how to conceptualize that , and make it feasible, is a critical issue that i think needs to be addressed -- certainly in the time frame of a "phase 2" his cyberinfrastructure proposal.

Reviewer #13

I just read through the paper and everything looks fine to me. I have already seen much of this since I have been working with David over the years off and on. I can't think of anything I would change.

Reviewer #14

Response: I think David Maidment has done an excellent job of proposing a structure that can accommodate many different forms of hydrologic data. As I think of it, trying to develop an all-encompassing format for such data is a daunting task. This is a great start.

Question 1: Capable of storing…

The system seems well set up for point data, but my concern is for data that are changing both in time and space such as:

- Radar rainfall (minutes to hours)
- Snow pack (weeks)
- Soil moisture (days to weeks)
- Land use (years to decades)

It is not clear to me if these spatial time series are part of the “digital watershed” (shown on Page 1 of David Tarboton’s review document) or not. Digital watershed means static things to me like topography and soils information but perhaps it also includes information on spatially and temporally varied quantities.

In any case, I would be curious about how spatially/temporally varying quantities such as those itemized above should be stored. Also, it seems to me that spatial time series give context to point information. Two examples:

- Daily streamflow at a gage. How might/would the spatial data (of a changing snowpack) and the point discharge data be linked?
- An annual or decadal time series of land use. How might this time series explain or be linked to non-stationarities in the gage annual maximum flood series.

Linkage here is an important point. Is there a way for this spatially varied information to be linked to otherwise related point information? (see my expansion on this idea under Question 3)

Question 2: Efficient storage…

none

Question 3: Efficient queries…

Is there a planned linkage between multiple layers of point data that might be related? For example, if my point of interest is a streamflow gage, the watershed draining to that gage might contain precipitation gages, dams, water/wastewater treatment plants, etc. that I might want to know about. Could I query that gage and immediately know the relevant other data sources within that watershed? This is a bit above and beyond the scope here, but would be very helpful.

Question 4: Spatial information representation…

<<Please see my discussion on spatial data in item 1>>

Question 5: Time Series Table…

Should a Boolean variable be added such as: TValueUpdateable? I know firsthand that the USGS posts preliminary values for streamflow values which are subject to revision at later times (possibly even years later if an error is detected). This leads to the possibility that I could do an analysis one day and get one result (because of one streamflow estimate, for example) and get a different result the next day (because the streamflow was updated).

I'm imagining that the solution here will be to develop computer models that work with these hydrologic data series and will produce both the standard output we're familiar with, but also some form of output that provides a "metadata" style summary of the input data that would distinguish otherwise identical analyses that differ in the result because the hydrologic data itself has been updated.

Question 6: Other types of observations?

none

Question 7: Additional Comments…

Two comments:

- Perhaps this is "out-of scope" as well, but it seems the system we're evolving towards is one where we should be able to "wire" a computer program where the inputs are high-level references to web addresses where files with predictable names and content could be available. This would allow us to run out computer models without having to download the input data to our local machines. The effort here to develop a consistent structure would certainly facilitate this. Is this proposal to actually make this possible, or is that just a dream for somewhere down the road?
- I'm wondering about the state and local folks that are gathering hydrologic data of their own. Will this proposed framework be fairly easy to work with? If we really want these folks to contribute data in a meaningful way, it would be helpful to actually provide them with simple tools that will organize and format their own data into appropriate formats. I wonder if such tools should be designed/developed by us (the CUASI community) or whether this represents a niche-market for somebody like Haestad or Boss to step in and provide software/expertise in the organization of data for general consumption compatible with this data model.

Reviewer #15

The data model for time series data allows spatial features (polygons, lines, points) to link with a very general schema for reporting variation through time. The resulting system is a preliminary hydrologic space-time model.

1) The time series model can deal well with point and reach data. The model for dealing with time series raster data is a bit unusual, but should work as a general approach for dealing with continuous spatially distributed phenomenon. (The application, at SFWMD, assigned NEXRAD predicted rainfall to each cell polygon as a time series attributes is a generalizable approach that can be used for water budget components in general (e.g. soil moisture, ecosystem ET, runoff) – This is quite different than storing multiple raster maps for each time interval).
2) Without some actual testing the data model, its efficiency is not obvious. However, relational databases store non-spatial data in a similar way and are clearly efficient for storage and retrieval.
3) Not sure
4) Point data alone are probably insufficient. It would be best if "monitoring point" can be generalized to "monitoring shape". This would allow a generalization of the SFWMD NEXRAD time series model to pre-assigned polygons (e.g., soil series) in thematic maps.
5) This appears to be sufficient. Perhaps a type field that would define categorical from continuous data would be useful for qualitative vs. quantitative data sets. Another small point: many studies suggest that variance in flows through time is an important ecological descriptor, especially for estuarine ecosystems where salinity fluctuation is a driver (like the Suwannee River Estuary). How feasible is it to include among the time series data types some measure of variance in previous time steps? These would mostly be derived data, but should still be useful inputs for coupled hydrologic-ecological models.

6) Distributed surface runoff seems absent (i.e., non-channelized surface flows and interflow – the latter being particularly important in the Suwannee Basin). These are not often measured, but may be modeled.
7) What are the capabilities for displaying time series data sequentially for each of the spatial components for which they are stored? In other words, can ArcHydro, and this data model in particular, be used to simplify the visualization of changes in the system in space and time? Automated procedures to develop animations have many applications.

Some specific issues in the Suwannee basin that may or may not be general to the other proposed HO sites are:

1) The current mode greatly simplifies groundwater systems, and in particular doesn't handle subsurface conduit flow well. This is not a limitation of the data model per se, since groundwater flow data may be coded within the schema as well as groundwater level, but may necessitate greater flexibility in the time intervals or TSTypeID.
2) Landscapes with internal closed depressions (sinkhole lakes, isolated wetlands) cause major problems for the digital terrain modeling algorithms. This is not really relevant the time series data model, but is a significant problem for applying ArcHydro modeling tools in the Suwannee River Basin.

Reviewer #16

The statement "get me all the hydrologic data for my region in a consistent format is absolutely correct and certainly what is needed.

1. Yes, the AHTSDM is capable of storing the hydrologic observations. One data set that is always problematic is lithology. Here you need a 3-d spatial rather than 2-d structure. I'm not sure the current AHTSDM readily handles 3-d spatial data (x,y,z,t). Another example of 3-d data would be groundwater data from multiple depths, either one point in time or time series data. Incorporating another field to make the spatial component 3-d would really enhance the data structure.

2. AHTSDM seems to be an efficient way to store the data.

3. AHTSDM is efficient for data query, data download, and data analysis.

4. The monitoring point is sufficient as long as the monitoring point is stationary. It is not clear how the system would function if the monitoring point is not static. An example would be fish migration in response to some hydrologic perturbation.

5. AHTSDM - the TSType table is adequate for the meta data.

6. Yes, others should be considered. It will be necessary to incorporate lithologic data and biologic data. The HO's are to provide data sets related to hydrology within the HO, which includes bioassessment at streams, springs and groundwater. As discussed under item 1, both

will require 3-d spatial data and the biologic component will likely require variable spatial locations.

7. This is not stated in the paper and it probably needs to be articulated. Are we proposing to change the existing data structure (i.e. USGS, NOAA, etc) to this format, or are we proposing that the HO teams routinely compile data from these sources into the AHTSDM format for use HO participants? The first will likely result in significant resistance, while the second creates likely data conversion errors.

Thanks for the chance to review.

Reviewer #17

(0) this is an "ad-hoc" data model: while it facilitates the handling and analysis of some specific kinds of data within a GIS, it is not sufficiently generic or powerful to handle unusual queries about hydro events.

(1) the most important limitation of the model is its apparent reliance on discrete static spatial objects, thus it is not possible to represent a flooding front moving into a floodplain or a changing extent of droughty area through time.

(2) minor point: on page 16 in the description of the types of time series handled by Arc Hydro there is no mention of the (necessary?) granularity of the time series. while integrals are used to describe the cumulatives, they assume a continuity that may not exist.

Reviewer #18

I think the ArcHydro Time Series Data Model is an adequate model for most hydrologic time series data. There are other kinds of field data that are not time series data that might require other approaches but those will be generated locally within the HO rather than taken from agency files. I would like to see spatial information keyed to a watershed hierarchical structure that includes DEM data, but that is not part of the time series data model and tools for doing it are already incorporated within other programs being developed by the HIS team for use by the community as demonstrated for us at the workshop in Austin. One issue that occurs to me is that the conversion of geographic coordinates to a variety of projections for local use requires additional manipulation, and although ArcGIS can do this it's not as intuitive a process to do the conversion as it could be. Individual HO teams may want to have a tool that is user-friendly to allow anyone using their data to make such conversions easily But I guess this point is not pertinent to the time-series data per se.

The TSType table on p. 15 of the report looks ok to me, but to be honest I can't give it a fair evaluation without actually trying to use it with all the kinds of data we might ultimately want to collect. To evaluate it at this point almost seems premature. It might make more sense to accept

the data structures provisionally and then have the HO "testbeds" do the road test and recommend changes at the end of a year of actual use.

Hydrologic observations not listed in question 6 include soil texture, hydraulic conductivity and moisture content as well as data pertinent to the calculation of evapotranspiration. One problem with such data is that they may be gathered for a point but applied to an area. If others besides those who generate these data want to be able to use them, there needs to be some consideration of how best to represent point data that are going to be applied over some spatial domain and to include information about how that can be done, and the degree of reliability, in some form of metadata. The same may be said about aquifer data; perhaps in addition to point data, individual aquifers ought to have their own metadata indicating not only where the actual measurements are, but what the best professional judgment is about how those can be interpolated for modeling purposes and what degree of uncertainty should be attached to any such interpolations. Obviously different researchers will make different choices. But the data model might incorporate a way of allowing researchers to communicate to other potential users any caveats or cautions about how the data should be interpreted for modeling purposes.

The same goes for rainfall fields. If NEXRAD data are going to be published, there should be metadata indicating what processing or bias correction has been done over what spatial domain for particular time periods and the level of confidence attached either to the original data or to the bias-corrected data. I can think of individual storms we have worked on, some of which show a very good spatial correlation with ground measurements after bias correction and some of which still have problems that are not easily corrected.

I think the data model as presented is a good approach, but I'll reiterate my statement above that we should accept it provisionally and then allow the different groups to work with it for a while in real time, outside of the workshop environment, in order to identify any shortcomings or additions that might be desirable.

Reviewer #19

In general, the document reads well and is convincing of its approach for handling hydrologic time series data within a geodatabase, a current limitation of most GIS systems. The following points address deficiencies I have observed if this hydrologic observations data model is to be inclusive of all hydrologic data necessary for a model application or part of an observatory.

1) Partial or sporadic data: An explicit account is not made for time series data that may be of partial duration (several months or years) or sporadic (point measurement taken once). For many hydrologic observations, sparseness is a fundamental trait. For example, soil samples may be taken for analysis, pore water sampled for isotopic composition, vegetation rooting depth measured. These can possibly occur only once, but are spatially distributed and can form part of an HO, especially when investigating geomorphology, paleoclimate or paleohydrology.

2) Raster Datasets: Is the proposed data structure the best for raster data? An example is provided that the NEXRAD data (time series of spatial maps) can be stored in ArcHydro, as time series at

each individual cell. Is this the most efficient means of storage? Many distributed models would better use the raster data itself (grids for all times) in the application. In addition, what is done with projections and coordinate transformations when raster grids of rainfall are represented as thousands of individual points?

Other raster data sets are equally important and not discussed in the document. For example, topography, land cover, vegetation, soils, geology can all be represented as raster data, typically as single snapshots in time. How are these going to form part of ArcHydro? Where are the metadata going to be stored? How are the digital values related to actual classifications (say for land cover)? This also applies to other time varying remote sensing data, for example, NDVI, albedo, brightness temperature, that can be obtained from various sensors. How are these going to be stored, accessed, documented, queried?

3) Vertical variations: The current formulation does not consider vertical variations in sampling locations as the descriptions were only indicated in planform. What happens when 10 soil moisture probes are installed at identical X,Y but different Zs? How is this stored, queried, displayed in the data model. The same applies for atmospheric observations, groundwater well observations, within stream or lake measurements, etc. Will this data be viewed in vertical section at some point?

4) Hydrogeological Data: While the current data model seemed to apply to all hydrology, a definite bias is noted toward surface hydrology. I am aware of a separate Hydrogeology Data Model effort. How are these going to tie together? The hydrologic observations data model should include observations of hydrostratigraphy, well logs, groundwater level and chemistry. I think that our data models should not reinforce perceived and real differences between surface and groundwater hydrologists.

Reviewer #20

The AHTSDM revolves around point data. Raster formats are not discussed, but mentioned only once with regard to South Florida's NEXRAD data. I'm not giddy about stripping a data's natural format such as with GIS rasters, but across the board we'll see this with other remotely sensed data. I would recommend that HIS produce at least two pre-processing software programs that will massage point series data that represents a continuous surface into a format readable by GIS and MATLAB.

There needs to be a means of indicating what observed parameters are present with a point. I may have 100 rain gages, but only 30 of them include the additional parameters of wind speed, humidity, solar radiation, etc. Do you expect the SQL query to be used for this purpose?

What about the storage of spatial data that has an associated length or area? Will these best hosted by the individual HO, by HIS, or both through something like an enterprise geodatabase?

What about single point data that has a range of values (not parameters) – TSValue? Hydraulic conductivity of a soil is a good example where at a single observation point 5 different people

will get 5 different K's. All are important, especially for future science endeavors including parameter estimation or model sensitivity analyses.

I did not read this or maybe misunderstood something, but is there a limit to the number of parameters associated with a single point or is there one record for each parameter for a single point? How is the proposed data structure prepared for flexibility with regard to discontinued or new parameter observation (e.g., I'm observing rainfall amount, then later add wind speed and direction)?

Is there a means of correction?

With regard to question 6, one needs to include geology and seismology.

Is there a "no-data" identifier?

Reviewer #21

Overall, we feel that what HIS is doing is fine. However one key issue that you don't address is data accuracy. Following your suggestions, we are responding to the questions that you have raised.

Q1. Is the AHTSDM capable of storing all hydrologic observation data that you think this database should contain? If not please indicate the data that does not fit this data structure, explaining why and providing suggestions regarding a data structure that could accommodate this data.

R1. The system of relational database is appropriate to handle time series. One thing that is not addressed in the paper is meta information about data accuracy. From my perspective in addition to the field TSValue a TSAccuracy field must be added. This is because different instruments have different precision and this information is necessary. This is specially true when data is used in the context of model calibration. In most cases users will go to extremes to try to match their model output to data, without taking into account that data can contain uncertainty, and that this errors could even be systematic for extremes (low or high) values of data.

Q2. Is the AHTSDM an efficient way to store all the hydrologic observation data that you think this database should contain? If not please indicate the data and situations for which this structure is inefficient and provide suggestions for making the data more efficient.

R2. The relational database model is ideal in terms of efficiency. However a clear treatment to missing data must be established. Missing data have different meanings that depend on the type of time series being considered. Tables must be designed in a way that missing data doesn't have to be stored. If this is not carefully crafted the database may get filled out with -9999 in the TSValue field, decreasing its efficiency in terms of storage.

Q3. Is the structure of AHTSDM efficient for the querying and analysis of hydrologic observation data? If not please explain the shortcomings and indicate how they might be overcome.

R3. YES. At first glance looks like tables would be efficiently accessed with SQL queries.

Q4. Spatial information is represented in the data model through the monitoring point (table 9 page 15) being a feature within a GIS database. Is this sufficient or is additional spatial information necessary as part of the Hydrologic Observations data model.

R4. It is fine.

Q5. The TsType table in the AHTSDM (table 9 page 15) is effectively the location where metadata about the data values in the TimeSeries table are stored. Are the contents of this table sufficient for providing the ancillary information that needs to be kept in the database with hydrologic observations? If not what additional ancillary information should be part of the database and how should it be stored?

R5. I think information about the recording instrument should be added. You can never trust data 100%.

Q6. The abstract indicates that the goal is to synthesize observations of streamflow, precipitation, climate, water quality and groundwater into a single database. Is this a sufficient set of "hydrologic observations" or should others be considered? Would other observations fit this model?

R6. Evapotranspiration, both potential and actual, are two key fluxes and should be included if they are not.

Q7. Please provide any additional comments or suggestions.

R7. The metadata should contain information about the typical spatial and temporal variability of the variable that is being measured.

Reviewer #22

Cover remarks
I have asked many folks who are contributing to the HO planning but found few who are interested in ArcHydro or were willing to consider a data model that incorporates it. Nevertheless, we have compiled a few comments based on conversations and on our analysis of David's paper & references therein. Basically, we do recognize the power of the proposed data model, but are still quite concerned about the investment needed to adapt it to the diversity of data types we expect, and to its efficiency for the many users who just want data and are not interested in doing analysis with ArcHydro.

Comments

1. We are concerned about both the flexibility of the AHTSDM given the somewhat restrictive nature of the observational data used in the analysis and of the need for it for the variety of hydrologic data expected. For example, if measurements are clustered horizontally and vertically within an area of interest, e.g. sensor web, but reported as both individual nodes and spatial measurements how would this show up in he current model? How about spatial but multi-scale and multi-layer measurements? Until we can invest some more significant time into testing the system we cannot build the confidence needed to say that it has the flexibility to incorporate data from more advanced measurement systems.
2. Our impression is that AHTSDM may require several steps to retrieve data, which makes it inefficient for the applications not using ArcGIS. That is, for the ArcGIS application it may be fine, but most uses of hydrologic data do not involve ArcGIS. Again, some more significant investment of time into testing the system is needed before we can gauge if there are efficient ways to use AHTSDM, but our first thought is to just skip it & go directly into the digital library.
3. Same answer as #2.
4. Quite a bit of additional spatial information will be necessary given the coming of sensor networks and instrument clusters. See answer to #1.
5. Need a flexible metadata structure. Perhaps this category could be sufficient if the categories under TSType were expanded into subcategories, providing more detailed information when necessary.
6. The set of "hydrologic observations" considered within this manuscript is somewhat restrictive. Would flux tower data or aircraft data fit this model? How about interpolated products? Dense instrument clusters? Redundant measurements?
7. Our obvious concern is that ArcHydro is only set up to ingest a certain subset of potential HO data. Thus it has a niche, but should it be a foundation? A case was not presented on the advantage of this product, thereby, posing a question: is a more flexible, robust application available that provides flexibility for diverse data types, and user groups, allowing for flexible data management system and access to data that is downloadable into multiple data analysis platforms including ArcHydro, thereby eliminating the need for specialized commercal software?

Appendix 1. Review Questionnaire.

CUAHSI Hydrologic Information System Hydrologic Observations Data Model Review

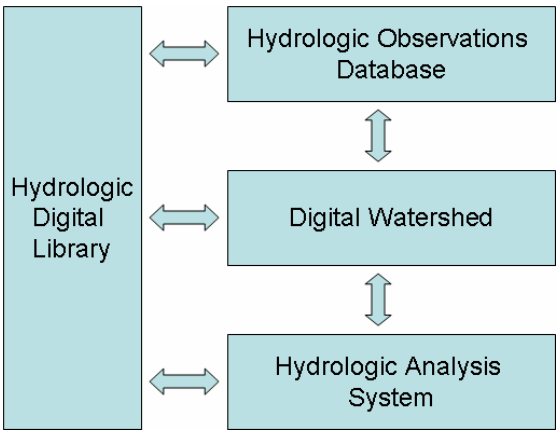
The attached paper "A Data Model for Hydrologic Observations", by David Maidment, March 2005 presents the design for the integrated hydrologic observations database that is proposed for the CUAHSI Hydrologic Information System (HIS). David Maidment has asked me to undertake an independent review of this design to ensure that we have a data model for hydrologic observations that is simple, usable, and can be implemented in a variety of application systems, including relational databases, Excel, GIS, statistical packages and simulation systems like MatLab. After giving a brief background on CUAHSI and the HIS project this document gives the questions I would like this review to address and the process and schedule I hope to follow in conducting this review. We seek your advice and participation in this review so that the CUAHSI HIS Hydrologic Observations data model can be designed to fulfill the needs of the CUAHSI community. We appreciate you taking the time to participate in this review. If you have any other comments or guidance to offer, please contact us.

David G Tarboton, dtarb@cc.usu.edu, Data Model Review Coordinator.
David Maidment, maidment@mail.utexas.edu, HIS project PI.
Rick Hooper, rhooper@cuahsi.org, CUAHSI president.
Wendy Graham, wgraham@ufl.edu, Chair, CUAHSI board of directors.

Background

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 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. As presently conceived, the CUAHSI Hydrologic Information System has four components:

- 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;
- 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.



The HIS development team consists of a team of academic hydrologists collaborating with the San Diego Supercomputer Center. More information is available from http://www.cuahsi.org, http://cuahsi.sdsc.edu/html/, and http://www.crwr.utexas.edu/cuahsi/symposium05/index.htm.

Hydrologic Observations Data Model review

What are we reviewing? It is not the purpose of this review to evaluate the paper that describes this model, such as might be done if considering this paper for publication. Rather I would like to focus on the Arc Hydro time series data model (AHTSDM) defined on pages 12-18 of this paper with extensions proposed on pages 18-19 and address the question as to whether this model is sufficient for the representation of hydrologic observations of interest to CUAHSI, and specifically for the representation of observations from the proposed CUAHSI Hydrologic Observatories. The paper concludes "It appears that this data model is appropriate for constructing a hydrologic observations database for the CUAHSI Hydrologic Information System". It is this specific conclusion that we need to evaluate. If this data model is not sufficient then we need to provide suggestions as to how it can be fixed.

1. Is the AHTSDM capable of storing all hydrologic observation data that you think this database should contain? If not please indicate the data that does not fit this data structure, explaining why and providing suggestions regarding a data structure that could accommodate this data.
2. Is the AHTSDM an efficient way to store all the hydrologic observation data that you think this database should contain? If not please indicate the data and situations for which this structure is inefficient and provide suggestions for making the data more efficient.
3. Is the structure of AHTSDM efficient for the querying and analysis of hydrologic observation data? If not please explain the shortcomings and indicate how they might be overcome.
4. Spatial information is represented in the data model through the monitoring point (table 9 page 15) being a feature within a GIS database. Is this sufficient or is additional spatial information necessary as part of the Hydrologic Observations data model.
5. The TsType table in the AHTSDM (table 9 page 15) is effectively the location where metadata about the data values in the Time Series table are stored. Are the contents of this table sufficient for providing the ancillary information that needs to be kept in the database with hydrologic observations? If not what additional ancillary information should be part of the database and how should it be stored?
6. The abstract indicates that the goal is to synthesize observations of streamflow, precipitation, climate, water quality and groundwater into a single database. Is this a sufficient set of "hydrologic observations" or should others be considered? Would other observations fit this model?
7. Please provide any additional comments or suggestions.

Hydrologic Observations Data Model Review Procedure and Schedule

Following are the steps I propose to use for this review.

April 2005.	Request comments on Data Model. Comments summarized and circulated/posted on a web site. Please provide your written review comments in electronic form by April 18. Please also indicate the times that you would be available for a conference call the Week of April 25.
Week of April 25.	Conference call to discuss comments and formulate initial recommendation. Maidment not participating.
May 6, 2005	Draft review circulated for comment. Copy provided to Maidment.
Week of May 9.	Conference call to discuss draft review with Maidment.
May 23, 2005	Review completed and submitted to Maidment and CUAHSI.