

A Data Model for Hydrologic Observations

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

Paper prepared for presentation at the
CUAHSI Hydrologic Information Systems Symposium
University of Texas at Austin, March 7, 2005

Abstract

The goal of constructing a hydrologic observations database for a region is to acquire streamflow, precipitation, climate, water quality and groundwater data measured at observation sites and synthesize them into a single database. This paper examines how the USGS, EPA and NCDC web sites deliver tabular hydrologic information to users, shows that this information can be synthesized using the Arc Hydro time series model, gives an example of an ArcIMS web site for serving an integrated hydrologic observations database for the Neuse basin in North Carolina, and describes a possible approach to automated data harvesting across the tabular data holdings of the USGS, EPA, NCDC and related organizations serving hydrologic observation data.

Introduction

The Consortium of Universities for the Advancement of Hydrologic Science, Inc (CUAHSI) is an organization supported by the National Science Foundation to provide infrastructure and services to advance the development of hydrologic science and education in the United States. The National Science Foundation has also sponsored a CUAHSI Hydrologic Information System (HIS) project to examine how hydrologic data can be better assembled and analyzed to support hydrologic science and education. Hydrologic scientists and students want a hydrologic information system that will just “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.

Many organizations and individuals measure hydrologic variables such as streamflow, water quality, groundwater levels and precipitation. For the continental United States, the USGS has available in its National Water Information System, data from surface and groundwater observations at about 1.4 million distinct geographic locations. The National Climate Data Center has more than 8000 locations where it is recording data on precipitation and air temperature. The EPA Storet system contains a very large repository of water quality data. Additional data are measured by state and local agencies and by individual hydrologists in universities and other research institutions.

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.

There may also be other information associated with the data, for example, it may be useful to record the agency which made the measurements and kind of instrument used, and also to record quality control measures that describe how accurate the measurements are. For the moment, let us set aside these additional considerations, and consider simply the triplet {space, time, variables}. These three indices can be visualized as the axes of a *data cube*, as shown in Figure 1, where a particular observed data value D is located as a function of where it was observed, L , its time of observation, T , and what kind of variable it is, V , thus forming $D(L, T, V)$.

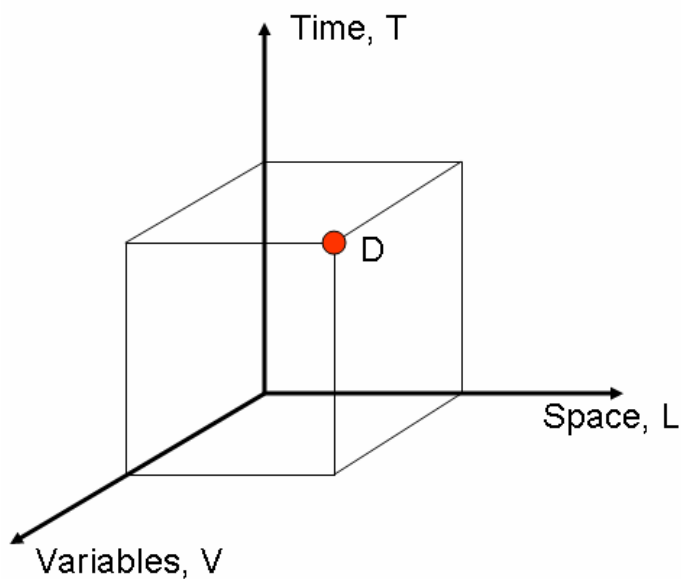


Figure 1. The data cube. A measured value D is indexed by its spatial location, L , its time of measurement, T , and what kind of variable it is, V .

For example, a download of daily streamflow data from the USGS streamgage site on the Neuse River near Clayton, NC, yields the information shown in Table 1. The USGS site number is 02087500, the date of each observation is given as a value under the column header `dv_dt`, along with the streamflow value under the column header `dv_va`. The fact that this represents daily mean discharge can be inferred from the metadata description given at the head of the discharge data. A similar download of real-time streamflow data from the same stream gage yields an output shown in Table 2. There are now two variables specified, discharge and gage height, each having its own variable code. The

time of measurement is now specified down to the minute at which the measurement was made. Comparison of Tables 1 and 2 shows that even for a single measurement device, such as a USGS stream gage, there are subtle differences in the way that the data are presented depending on the database from which the data were extracted (historical daily flow or real-time flow), the type of data being extracted and the number of variables involved.

And, in truth there are important differences between these two kinds of data – the daily mean discharge values are averages over the day of the 15-minute or “unit-value” measurements. Hence the daily mean discharge measurements do not, strictly speaking, apply at any instantaneous time point during a day, but rather they are a single value representing the flow during the day as a whole. By contrast, the unit value data each 15 minutes represent the flow at that particular instant in time, not an average over a 15 minute time interval. This distinction between *instantaneous* and *average* data is an important characteristic that needs to be preserved in a hydrologic observations data model.

```
# This information includes the following fields:
#
# agency_cd    Agency Code
# site_no     USGS station number
# dv_dt       date of daily mean streamflow
# dv_va       daily mean streamflow value, in cubic-feet per-second
# dv_cd       daily mean streamflow value qualification code
#
# Sites in this file include:
# USGS 02087500 NEUSE RIVER NEAR CLAYTON, NC
#
agency_cd  site_no    dv_dt  dv_va  dv_cd
USGS 02087500 2003-09-01 1190
USGS 02087500 2003-09-02 649
USGS 02087500 2003-09-03 525
USGS 02087500 2003-09-04 486
USGS 02087500 2003-09-05 733
USGS 02087500 2003-09-06 585
USGS 02087500 2003-09-07 485
USGS 02087500 2003-09-08 463
USGS 02087500 2003-09-09 673
USGS 02087500 2003-09-10 517
USGS 02087500 2003-09-11 454
```

Table 1. Daily streamflow data from Neuse River near Clayton, NC.

Source: <http://nwis.waterdata.usgs.gov/nc/nwis/discharge>

```

# The column headers include the following fields
#
# column          column definition
# -----
# agency_cd       Agency collection or maintaining the site
# site_no         USGS site identification number
# datetime        date and time in ISO format (YYYY-mm-dd HH:MM:SS)
#
# Data for the following stations is contained in this file
# -----
# USGS 02087500 NEUSE RIVER NEAR CLAYTON, NC
#
# DD parameter - Description
# -- -----
# *05  00060    - Discharge, cubic feet per second
# *06  00065    - Gage height, feet
#
agency_cd  site_no    datetime    06_00065    05_00060
USGS 02087500  2005-02-23 16:15  1.75  372
USGS 02087500  2005-02-23 16:30  1.75  372
USGS 02087500  2005-02-23 16:45  1.76  376
USGS 02087500  2005-02-23 17:00  1.77  380
USGS 02087500  2005-02-23 17:15  1.78  385
USGS 02087500  2005-02-23 17:30  1.79  389
USGS 02087500  2005-02-23 17:45  1.79  389
USGS 02087500  2005-02-23 18:00  1.80  393

```

Table 2. Real-time streamflow data from Neuse River near Clayton, NC.
Source: <http://waterdata.usgs.gov/nc/nwis/uv?02087500>

Site Information

Suppose a hydrologist wishes to find out more information about the stream gage on the Neuse River near Clayton, NC. The USGS maintains site information about its gages, which is summarized for this gage in Figure 2. Gage site information specifies the Latitude, Longitude and elevation of the site, tells what county it is in, and what Hydrologic Unit Code watershed this gage resides within. That information is helpful in locating related rainfall and water quality information from NCDC and EPA.

LOCATION

Latitude 35°38'50", Longitude 78°24'19" NAD83,
Johnston County, North Carolina , Hydrologic Unit 03020201

DRAINAGE AREA

1,150.00 square miles

GAGE

Datum of gage is 128.41 feet above sea level NGVD29.

SITE TYPE:

Stream / River

AVAILABLE DATA:

| Data Type | Begin Date | End Date | Count |
|---------------------------------------|--------------------------|------------|-------|
| Real-time | This is a real-time site | | |
| Peak streamflow | 1919-07-23 | 2003-03-21 | 77 |
| Daily streamflow | 1927-08-01 | 2003-09-30 | 27820 |
| Water Quality Samples | 1955-02-01 | 1999-09-28 | 237 |

OPERATION:

Record for this site is maintained by the USGS office in North Carolina

Figure 2. Site information for the stream gage on the Neuse River near Clayton, NC.

Source: http://nwis.waterdata.usgs.gov/nwis/inventory/?site_no=02087500

Rainfall

A hydrologist may also decide to get some rainfall data for gages in the same area. A search on the National Climate Data Center website for gages in Johnston County, NC, reveals two stations, Clayton, Wtp; and Smithfield. The site description for Clayton Wtp is given in Figure 3. If the hydrologist clicks on the DATA label presented in this Figure, he or she will go on another path involving many more screens, and if this is done from a computer operating through a commercial internet service provider, there will come a request for payment for the data while if the request is made from a University computer on campus, the data are available without charge – the NCDC website is sophisticated enough that it can detect whether the request is being made from a computer within a .edu domain (in which case the information is free) or if it comes from a .com domain for a commercial internet service provider (in which case the information must be paid for). It is possible to get data from the NCDC website but the labyrinth

through which the web navigation has to proceed is much more complex than is the case with the USGS NWIS website.

Clayton Wtp
 Clayton, NC, United States

[Other Stations in Clayton](#)
[List Stations in Division NC-07](#)
[List Stations in Johnston County](#)
 List Stations within [5](#), [10](#), [25](#), or [50](#) miles

Type*: COOP
 Call Sign/ICS*:
 WBAN*:
 COOP ID*: **311820**
 Climate Division*: NC-07 - Central Coastal Plain
 WMO ID*:
 In Service*: **01 Nov 1955 to Present**
 Elevation*: **91.4m (300') above s/l**
 Lat/Lon*: **35°38'N / 78°28'W**
 County*: **Johnston**

- [Data Inventories](#)
- [DATA](#)
- [Station History](#)
- [State Climatologist](#)
- [Regional Climate Center](#)

Information at Other Sites

- [National Weather Service](#)
- [Eastern Region Headquarters](#)
- [Regional Forecast Office](#)
- [Local Weather Conditions](#)
- [Local Weather Forecast](#)
- [Current Radar](#)
- [Current Satellite](#)
- [Sunrise/Sunset](#)

Figure 3. Site information for a rain gage at Clayton Wtp, NC

| COOPID | WBANID | Prelim | year | month | day | Tmax | Tmin | Tobs | Tmean | Cdd | Hdd | Prcp |
|--------|--------|--------|------|-------|-----|------|------|------|-------|-----|-----|------|
| 311820 | 99999 | | 2004 | 3 | 5 | 80 | 55 | 60 | 68 | 3 | 0 | 0 |
| 311820 | 99999 | | 2004 | 3 | 6 | 78 | 60 | 68 | 69 | 4 | 0 | 0 |
| 311820 | 99999 | | 2004 | 3 | 7 | 71 | 44 | 45 | 58 | 0 | 7 | 0.12 |
| 311820 | 99999 | | 2004 | 3 | 8 | 71 | 36 | 40 | 54 | 0 | 11 | 0.22 |
| 311820 | 99999 | | 2004 | 3 | 9 | 55 | 28 | 28 | 42 | 0 | 23 | 0 |
| 311820 | 99999 | | 2004 | 3 | 10 | 54 | 28 | 35 | 41 | 0 | 24 | 0.13 |
| 311820 | 99999 | | 2004 | 3 | 11 | 56 | 26 | 27 | 41 | 0 | 24 | 0 |
| 311820 | 99999 | | 2004 | 3 | 12 | 61 | 27 | 34 | 44 | 0 | 21 | 0 |
| 311820 | 99999 | | 2004 | 3 | 13 | 66 | 34 | 42 | 50 | 0 | 15 | 0 |
| 311820 | 99999 | | 2004 | 3 | 14 | 57 | 31 | 41 | 44 | 0 | 21 | 0 |
| 311820 | 99999 | | 2004 | 3 | 15 | 64 | 41 | 54 | 53 | 0 | 12 | 0 |
| 311820 | 99999 | | 2004 | 3 | 16 | 64 | 54 | 58 | 59 | 0 | 6 | 0.94 |
| 311820 | 99999 | | 2004 | 3 | 17 | 66 | 41 | 41 | 54 | 0 | 11 | 0.35 |

Table 3. Daily weather observations for the Clayton weather station.

Source: <http://cdo.ncdc.noaa.gov/CDO/cdo>

Individual daily values for climatic variables are obtained from the NCDC Climate Data Online (CDO) system. The file format that is obtained is somewhat different than for streamflow information, as shown in Table 3, but the same general elements appear, namely a CoopID, and WbanID, which index this station within the NWS weather networks, the year, month and day of the measurements and then a string of values representing the recorded values of the observed variables on that day at that station.

These daily values include Tmax, Tmin, and Tmean. Tmax and Tmin are the daily *maximum* and daily *minimum* temperatures, which occurred at two different moments in the day but when they occurred is unknown; Tmean is an *average* value of temperature over the day, like daily mean discharge. The two days of rain, March 7-8, have 0.12 and 0.22 inches respectively – these values represent the *incremental* amount of rainfall that fell in each day and were parts of a storm whose *cumulative* rainfall over two days was 0.34 inches. There is a distinction between an average value of temperature or streamflow, which is the average of a large number of values recorded during the day and the incremental value of the rainfall for the day which represents the amount of rain that fell in 24 hours – it is not an average of any quantity which varied through time during the day (there probably were long periods with no rain at all). Another characteristic of daily rainfall data from Cooperative Observers is that the gages are read at 9AM each morning so the rainfall value shown is actually a 9AM to 9AM value not strictly a 24 hour 12 midnight to 12 midnight value!

Water Quality

Besides climate and streamflow, hydrologists also wish to access water quality data. The EPA Storet water quality website <http://www.epa.gov/STORET/> has two variants – Legacy Storet and Modernized Storet (not available for all US states). A search made in Modernized Storet on dissolved oxygen data in water in Johnston County, NC, from 2000 to 2005 yields 272 data values which download fairly easily into a delimited text format which is readily imported into Excel. The resulting information for a few samples at one station is summarized in Table 4. In this instance the site information is downloaded with the data rather than separately as with the NCDC and NWIS sites, there is a StationID (J417000), the data are *irregularly* recorded in time (9:40AM on Jan 11, 2000; 10:04AM on Feb 29, 2000), and the time zone in which the measurements were made is noted – either EST for standard time or EDT for daylight time. These are *instantaneous* data – they represent the dissolved oxygen value at that point in time (actually at the time of beginning the sampling procedure). What is also striking about these data is the amount of descriptive information that is carried along with each measurement about the nature of the analytical procedure, the depth at which the measurement was taken, and what kind of a sample was analyzed.

Water quality data can also be downloaded from the USGS NWIS web site. Table 5 shows a selection of these data recorded in samples taken at the Neuse River near Clayton, NC, and Table 6 shows some parameter codes for the water quality constituents in the dataset. Parameter 300 is Dissolved Oxygen, water, unfiltered, milligrams per liter, so the dissolved oxygen value at this sampling station at 12:50PM on 13 September 1996 was 5 mg/l. The USGS water quality data is described by coded values rather than by descriptive means as the Storet data are. USGS data are organized by measurement site and the information about the measurement site is summarized in the header file as it is with USGS streamflow data. Storet water quality data carry the descriptive information about the measurement site along with each measurement, rather than in a header file. It is notable that the time format in the USGS water quality information is not the same as for the real-time streamflow information – for the USGS water quality data, the day and

| Org Name | Station ID | State | County | HUC |
|-----------------------|--------------------|--------------------------|----------------|----------------------|
| NCDENR-DWQ (2nd) | J4170000 | NORTH CAROLINA | JOHNSTON | 3020201 |
| NCDENR-DWQ (2nd) | J4170000 | NORTH CAROLINA | JOHNSTON | 3020201 |
| NCDENR-DWQ (2nd) | J4170000 | NORTH CAROLINA | JOHNSTON | 3020201 |
| NCDENR-DWQ (2nd) | J4170000 | NORTH CAROLINA | JOHNSTON | 3020201 |
| NCDENR-DWQ (2nd) | J4170000 | NORTH CAROLINA | JOHNSTON | 3020201 |
| NCDENR-DWQ (2nd) | J4170000 | NORTH CAROLINA | JOHNSTON | 3020201 |
| Station Latitude | Station Longitude | Station Horizontal Datum | | |
| 35.64732 | -78.40567 | NAD83 | | |
| 35.64732 | -78.40567 | NAD83 | | |
| 35.64732 | -78.40567 | NAD83 | | |
| 35.64732 | -78.40567 | NAD83 | | |
| 35.64732 | -78.40567 | NAD83 | | |
| 35.64732 | -78.40567 | NAD83 | | |
| Visit Num | Activity ID | Activity Start | Zone | Activity Medium |
| 1 | 00-FLDW-0.1 | 1/11/2000 9:40 | EST | Water |
| 2 | 00-FLDW-0.1 | 2/29/2000 10:04 | EST | Water |
| 3 | 00-FLDW-0.1 | 4/5/2000 10:00 | EDT | Water |
| 4 | 00-FLDW-0.1 | 4/26/2000 10:24 | EDT | Water |
| 5 | 00-FLDW-0.1 | 5/11/2000 10:30 | EDT | Water |
| 7 | 00-FLDW-0.1 | 6/16/2000 11:30 | EDT | Water |
| Activity Type | Activity Depth | Activity Depth Unit | | |
| Field Msr/Obs | 0.1 | m | | |
| Field Msr/Obs | 0.1 | m | | |
| Field Msr/Obs | 0.1 | m | | |
| Field Msr/Obs | 0.1 | m | | |
| Field Msr/Obs | 0.1 | m | | |
| Field Msr/Obs | 0.1 | m | | |
| Characteristic Name | Sample Fraction | Value Type | Statistic Type | Result Value as Text |
| Dissolved oxygen (DO) | Dissolved | Actual | | 10 |
| Dissolved oxygen (DO) | Dissolved | Actual | | 10.1 |
| Dissolved oxygen (DO) | Dissolved | Actual | | 7 |
| Dissolved oxygen (DO) | Dissolved | Actual | | 8.8 |
| Dissolved oxygen (DO) | Dissolved | Actual | | 7 |
| Dissolved oxygen (DO) | Dissolved | Actual | | 4.7 |
| Units | Analytical Proc ID | | | |
| mg/l | 21NC02WQ -WQS SOP | | | |
| mg/l | 21NC02WQ -WQS SOP | | | |
| mg/l | 21NC02WQ -WQS SOP | | | |
| mg/l | 21NC02WQ -WQS SOP | | | |
| mg/l | 21NC02WQ -WQS SOP | | | |
| mg/l | 21NC02WQ -WQS SOP | | | |

Table 4. An edited Excel file of dissolved oxygen data for station J4170000 in Johnston County, NC downloaded from the EPA Storet system.

Source: http://oaspub.epa.gov/storpubl/DW_resultcriteria_geo

| agency_cd | site_no | sample_dt | sample_tm | parameter_cd | result_va | remark_cd |
|-----------|-------------|------------|-------------|--------------|-----------|-----------|
| USGS | 2087500 | 9/13/1996 | 12:50 | 10 | 24 | |
| USGS | 2087500 | 9/13/1996 | 12:50 | 25 | 740 | |
| USGS | 2087500 | 9/13/1996 | 12:50 | 27 | 1028 | |
| USGS | 2087500 | 9/13/1996 | 12:50 | 28 | 81213 | |
| USGS | 2087500 | 9/13/1996 | 12:50 | 61 | 8180 | |
| USGS | 2087500 | 9/13/1996 | 12:50 | 300 | 5 | |
| USGS | 2087500 | 9/13/1996 | 12:50 | 310 | 1.9 | |
| USGS | 2087500 | 9/13/1996 | 12:50 | 410 | 3 | |
| USGS | 2087500 | 9/13/1996 | 12:50 | 610 | 0.11 | |
| USGS | 2087500 | 9/13/1996 | 12:50 | 615 | 0.02 | |
| | | | | | | |
| qw_method | anl_stat_cd | anl_src_cd | hyd_cond_cd | samp_type_cd | hyd_event | medium_cd |
| | H | 9 | 9 | 9 | 7 | 9 |
| | H | 9 | 9 | 9 | 7 | 9 |
| | H | 9 | 9 | 9 | 7 | 9 |
| A | H | 9 | 9 | 9 | 7 | 9 |
| | H | 9 | 9 | 9 | 7 | 9 |
| | H | 9 | 9 | 9 | 7 | 9 |
| A | H | 9 | 9 | 9 | 7 | 9 |
| | H | 9 | 9 | 9 | 7 | 9 |
| F | H | 9 | 9 | 9 | 7 | 9 |
| F | H | 9 | 9 | 9 | 7 | 9 |

Table 5. Water quality data downloaded from NWIS for the Neuse River near Clayton, NC

The following parameters are included:
00010 - Temperature, water, degrees Celsius
00025 - Barometric pressure, millimeters of mercury
00027 - Agency collecting sample, code
00028 - Agency analyzing sample, code
00061 - Discharge, instantaneous, cubic feet per second
00300 - Dissolved oxygen, water, unfiltered, milligrams per liter

Table 6. USGS Water Quality Parameter Codes

time of the day are recorded as separate fields which then have to be combined to get a complete date-time timestamp like the real-time streamflow data have. The time zone is not noted directly in the USGS data as it is in the Storet data. In Storet the units for dissolved oxygen are mg/l not milligrams per liter. A human reads and understands that these are the same thing, but a computer usually recognizes only the literal text string that it is looking for and may miss the fact that mg/l and milligrams per liter are really the same thing.

Groundwater

Additionally, a hydrologist wants data on groundwater levels. A search on the USGS NWIS site for groundwater data in Johnston County, NC, yields a total of 71

measurement sites, each of which has a single measured value of water level at a particular point in time, as shown in Figure 4. USGS groundwater sites are described by a 15 character text string, as compared to 8 characters for surface water sites. It appears that the USGS maintains a much larger number of groundwater sites than surface water sites but the average number of data values at each site is small. In some cases, the day that the groundwater level measurement was made is known, in other cases only the month. USGS site information for one of the groundwater wells is shown in Figure 5.

Site Selection Results -- 71 sites found

County = Johnston

[Save file of selected sites](#) to local disk for future upload

Data for individual sites can be obtained by selecting the site number below

| Agency | Site Number | Site Name | From | To | Count |
|--------|---------------------------------|----------------------|------------|------------|-------|
| USGS | 351630078285501 | JH-140 A JOHNSON | 1958-10 | 1958-10 | 1 |
| USGS | 351947078285001 | JH-137 H A TART | 1955-06 | 1955-06 | 1 |
| USGS | 352236078322001 | JH-132 BENSON MUN | 1969-04-03 | 1969-04-03 | 1 |
| USGS | 352606078343201 | JH-124 H A PENNY | 1958-11-06 | 1958-11-06 | 1 |
| USGS | 352630078252201 | JH-096 FOUR OAKS MUN | 1957-09 | 1957-09 | 1 |
| USGS | 352654078254301 | JH-098 FOUR OAKS MUN | 1957-08-09 | 1957-08-09 | 1 |

Figure 4. Data recorded at USGS groundwater sites in Johnston County, NC

USGS 351630078285501 JH-140 A JOHNSON

Johnston County, North Carolina
 Latitude 35°16'30", Longitude 78°28'55" NAD27
 The depth of the well is 200 feet below land surface.
 This well is completed in TUSCALOOSA FORMATION
 (211TSCL)

Output for

| | | |
|------------------------------------|--|--|
| Table of data | | |
| Tab-separated data | | |
| Graph of data | | |
| Reselect period | | |

| Date | Time | Water level, feet below land surface |
|---------|------|--------------------------------------|
| 1958-10 | | 41 |

Figure 5. Groundwater site information for a USGS groundwater site

Characteristics of a Hydrologic Observations Database

An exercise for gathering data like the one just undertaken, could be carried on indefinitely, with data ingested from other federal, state and local sources and gradually compiled and integrated. What the exercise just described makes obvious is that there is no standard form for the data files that are obtained – they differ from different agencies and variables; so to obtain a complete inventory of information and to enable a traceback to original data sources, it will be necessary to store both the directly ingested file that was obtained from the data site (or directly from the instrument if this is recorded information not derived from a web site), and also to transform the data and synthesize them into a common format.

Some things that emerge in common from these different data collection programs and web sites for the USGS NWIS, EPA Storet, and NCDC CDO data systems:

- All of the observation sites are located by a specific point in space described by latitude and longitude;
- Files are outputted with data in delimited ascii format;
- Observation sites have identifiers such as the USGS site number, the NCDC CoopID, and the EPA StationID, that serve to link the description of the site with the data measured at that site but the identifier is specific to the agency creating it, and its format may be different even for a given agency between surface and groundwater stations;
- There is a significant variety in the manner in which time stamps for data values are given – some systems have year month and day in separate fields, others have them in a single text field, the USGS real-time system uses a datetime field in which date and time are specified in an ISO format (YYYY-mm-dd HH:MM:SS);
- The EPA Storet data are presented with one record per measured value, NCDC files have many values in each record, the USGS water quality system has one value per record, but the streamflow system can have both flow and stage on the same record;
- Each agency has a standard set of codes or headers that identify the variables and sometimes the units of those variables, but these codes and headers change from one agency to another even for the same variable (such as dissolved oxygen values stored by USGS and by EPA);
- The EPA Storet and USGS NWIS web sites are relatively easy to navigate and to download data from. The NCDC web site is more complicated to use and many products require payment unless the site is accessed from a computer inside the .edu domain;
- For rainfall, temperature, and streamflow data there are not many variables involved but the number of measured values can be very large at a given site, and they are recorded regularly;
- For water quality, there are many variables but not many measurements per variable when compared to streamflow and rainfall, and the measurements are made irregularly in time;
- For groundwater there are a lot of sites but few variables and measurements at each site;

- Some water quality data values have the qualifier < for less than a detection limit;
- The USGS NWIS system serves data for each station individually and puts out meta data at the head of the file for each station describing the measurements from that station that this file contains. The data are served immediately;
- The NCDC server serves data from stations collectively but puts the results in several files, one describing the stations, another describing the number of data available for variables at these stations, and a third file giving the data itself. These files are generated asynchronously and the requester gets a message via email when they are available;
- The EPA Storet system serves data for stations collectively and all the information is in a single record for each measured value, including details about the station it was measured at and the character of the measurement. The data are served immediately.

Designing and building an integrated hydrologic observations database is a fairly formidable task. Fortunately, some operational experience has been gained using the Arc Hydro time series data format, and the strengths and weaknesses of that format have been evaluated.

Arc Hydro Time Series Data

In 1999, the Environmental Systems Research Institute (ESRI) of Redlands, California, decided to reengineer their geographic information software products ArcInfo and ArcView and to produce a new software system called ArcGIS, of which ArcView and ArcInfo are now the low-end and high-end variants, respectively. During that process, the author was asked to design a customized version of ArcGIS for water resources, now called Arc Hydro, and he convened a GIS in Water Resources Consortium of industry, government and academics to help accomplish this task. Several national meetings were held to discuss variants on this design, and in 2002 it was published (Maidment, 2002). One of the key aspects of Arc Hydro is that it provides a means of integrate geospatial features with time series observations by uniquely labeling all the geospatial features with an identifier, their HydroID, and the labeling each time series observations associated with those features with this identifier, in much the same way as the USGS data are labeled with site number, and the EPA data with Station ID.

In Arc Hydro, the data value is called a TSValue, and the three axes indexing that value are named FeatureID for space, TSTypeID for the variable type and TSDateTime for the time index, as shown in Figure 6.

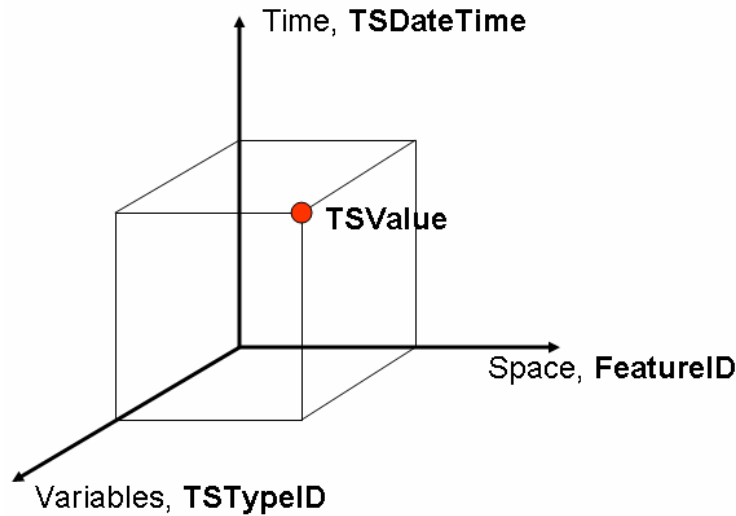


Figure 6. Representation of the data cube in Arc Hydro.

Time series information are then presented in a tabular format with one column per descriptive field, as shown in Figure 7. The hydrologic observations are contained in the fields {TSDateTime, TSValue} and the other two columns, FeatureID and TSTypeID are integer index fields used to relate the time series table with the geographic features described by those time series, and with a table describing the type of data. The ObjectID field shown in Figure 7 is a standard ArcGIS field that is added to all tables to index the rows of the table.

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

Figure 7. An Arc Hydro time series of daily streamflow for the Neuse River near Clayton, NC

Most water resources agencies use *relational databases* to store and provide internet access to their hydrologic data. A relational database comprises a set of tables like the one shown in Figure 7 connected by *relationships* which are links between key fields or columns whose values are shared by two or more tables. For example, the USGS site number is a key field within NWIS and the value 02087500 for the Neuse River gage near Clayton NC links the site information shown in Figure 2 with the observational data given in Tables 1 and 2. Relational databases include commercial systems such as Microsoft Access (which comes as part of Microsoft Office), Microsoft SQL/Server, or Oracle (which is probably the most widely used system for large databases), and include

open source databases such as MySQL or PostgreSQL. The terminology SQL refers to Structured Query Language, a standardized protocol for querying a set of related tables and extracting data records that satisfy a particular search criterion. For example, the query “find all the daily streamflow records at site 02087500 between 1 January 2003 and 31 December 2003” would extract a year of streamflow data for this stream gaging site from a relational database containing many stations, years of data, and perhaps types of streamflow data, such as stage and flow.

ArcGIS is implemented as a *geodatabase*, which is a relational database that stores geographic features as well as tabular values. In the terminology of this system, an *object class* is a table that does not contain geographic information, such as the time series table shown in Figure 7, while a *feature class* is an object class that also has a field called *shape* containing the geographic coordinates of the point, line or area, as shown in Figure 8 for a feature class of monitoring points in the Neuse basin. The term *MonitoringPoint* is used in Arc Hydro to mean any point location in space where hydrologic observations have been made, such as a stream gage, water quality measurement site, groundwater well, or a precipitation gage. When ArcGIS plots the points on a map, it looks up the geographic location in the shape field. In addition, the latitude and longitude, or easting and northing, coordinates of that point can be stored as additional fields or *attributes* of each monitoring point feature so that non-GIS data systems also know where the station is located without having to interpret the shape field.

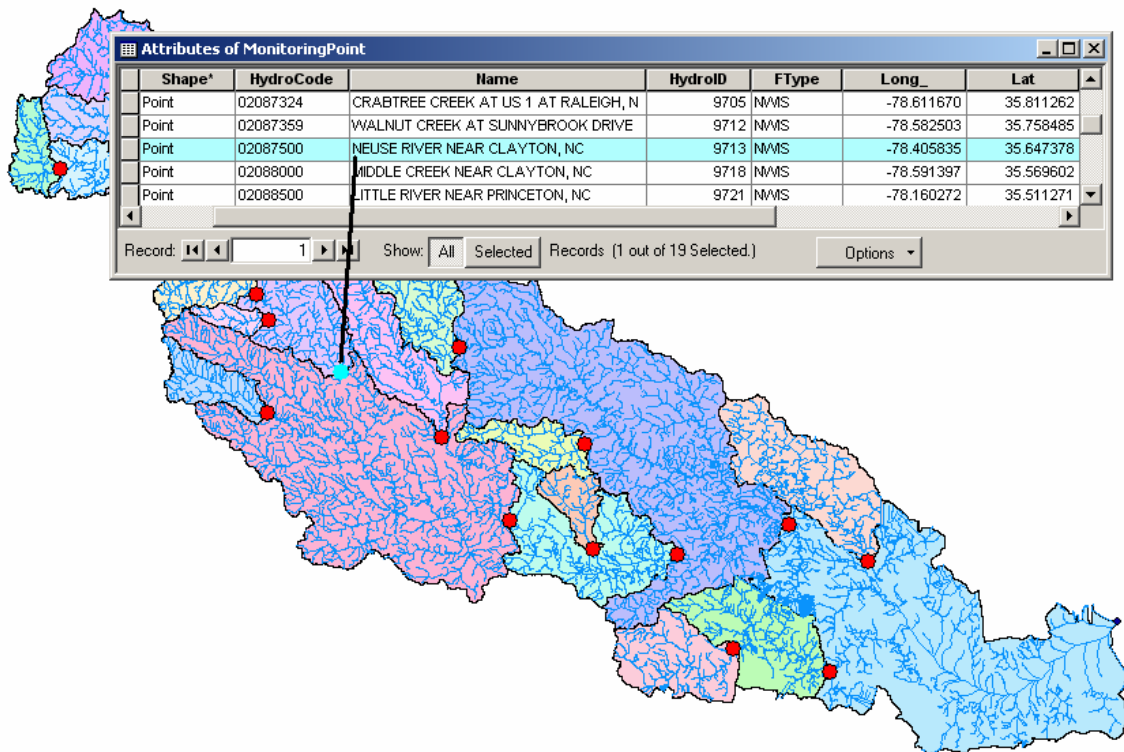


Figure 8. A feature class of MonitoringPoints in the Neuse Basin

All geographic features in Arc Hydro are *HydroFeatures*, which means that they have a HydroID and a HydroCode. The HydroID is a long integer identifier that serves as the relational connection in Arc Hydro between that feature and related tables and features. Arc Hydro has a special toolset that assigns HydroIDs uniquely within a given geodatabase so that a number once assigned is never assigned again. The HydroCode is a text field that stores the permanent public identifier of that feature in an external database system, such as the USGS site number, NCDC CoopID or the EPA Station ID. Using the combination of HydroID for internal data linkages, and HydroCode for external data linkages, Arc Hydro can index measurement sites from many different public data systems, and connect hydrologic observations extracted from a particular source back to the original data system.

The complete description of a set of hydrologic observations in Arc Hydro requires three tables and two relationships: the TimeSeries table containing the actual data, the MonitoringPoint table containing the location and description of the measurement sites, and a TSType table that describes the character of the time series data. In Arc Hydro TimeSeries tables the field FeatureID is the HydroID of the feature described by the time series values, and serves to relationally connect the MonitoringPoint feature class and the TimeSeries object class in a *one to many* relationship (one monitoring point has many time series records associated with it). Likewise the field TSTypeID serves to relationally connect the TSType table in a one to many relationship with the TimeSeries table (each type of time series data describes many data values). Figure 9 shows the three tables and two relationships in the Arc Hydro time series model as they appear in Microsoft Access – an ArcGIS *personal geodatabase* is actually an MS Access file that can be opened independently of ArcGIS.

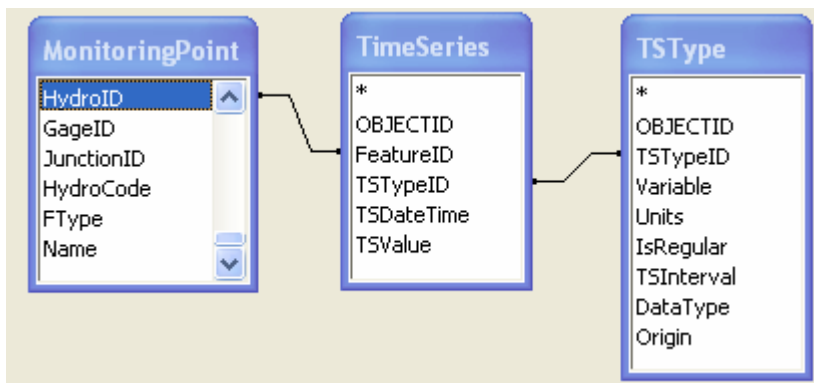


Figure 9. Arc Hydro time series data tables and relationships viewed using Microsoft Access

Specification of Time Series Data

A great deal of care must be exercised in describing time series data to ensure that the data are interpreted correctly and that functions operate appropriately on data series. A critical point is to specify when the data value occurred and to what time interval it applies. As the previous discussion of NWIS, Storet, and NCDC systems has shown,

there are many ways that time is indicated in public hydrologic data delivery systems. In Arc Hydro, the TSDatetime field is a *time stamp* which is an instantaneous time which records the instant the data value was measured or the *beginning* of the time interval to which a data value applies. For example, a daily streamflow value for 1 January 2004 has a time stamp of 2004-01-21 00:00:00, or at 12AM on 1 January 2004. The same Arc Hydro time stamp is used for the instantaneous flow value recorded at that instant, and for monthly data for January 2004, and for annual data for 2004. The custom of time stamping data with the instant at the beginning of a time interval is a common procedure that is also used in weather forecast models, where the time stamp on a data value refers to the beginning of the forecast interval it describes. It is thus necessary to know more than the time stamp to understand the temporal character of the data.

Arc Hydro classifies time series data into six types:

1. *Instantaneous* data – the phenomenon, such as streamflow, $Q(t)$ is specified at a particular instant in time;
2. *Cumulative* data – the data represents the cumulative value of a variable measured or calculated up to a given instant of time, such as cumulative volume of flow or cumulative precipitation: $V(t) = \int_0^t Q(\tau) d\tau$, where τ represents time in the integration over the interval $[0,t]$;
3. *Incremental* data – the value represents the incremental value of a variable over a time interval Δt such as the incremental volume of flow, or incremental precipitation: $\Delta V(t) = \int_t^{t+\Delta t} Q(\tau) d\tau$;
4. *Average* data – the value represents the average over a time interval, such as daily mean discharge or daily mean temperature: $\bar{Q}(t) = \frac{\Delta V(t)}{\Delta t}$;
5. *Maximum* data – the value is the maximum value occurring at some time during a time interval, such as annual maximum discharge or a daily maximum air temperature;
6. *Minimum* data – the value is the minimum value occurring at some time during a time interval, such as 7-day low flow for a year, or the daily minimum temperature.

These time series types are illustrated in Figure 10. It follows from their mathematical definition that the dimensions of instantaneous and average data are the same and that cumulative and incremental data are integrated through time, so their dimensions are different (e.g. cubic feet instead of cubic feet per second). It also follows that as $\Delta t \rightarrow 0$, $\bar{Q}(t) \rightarrow Q(t)$.

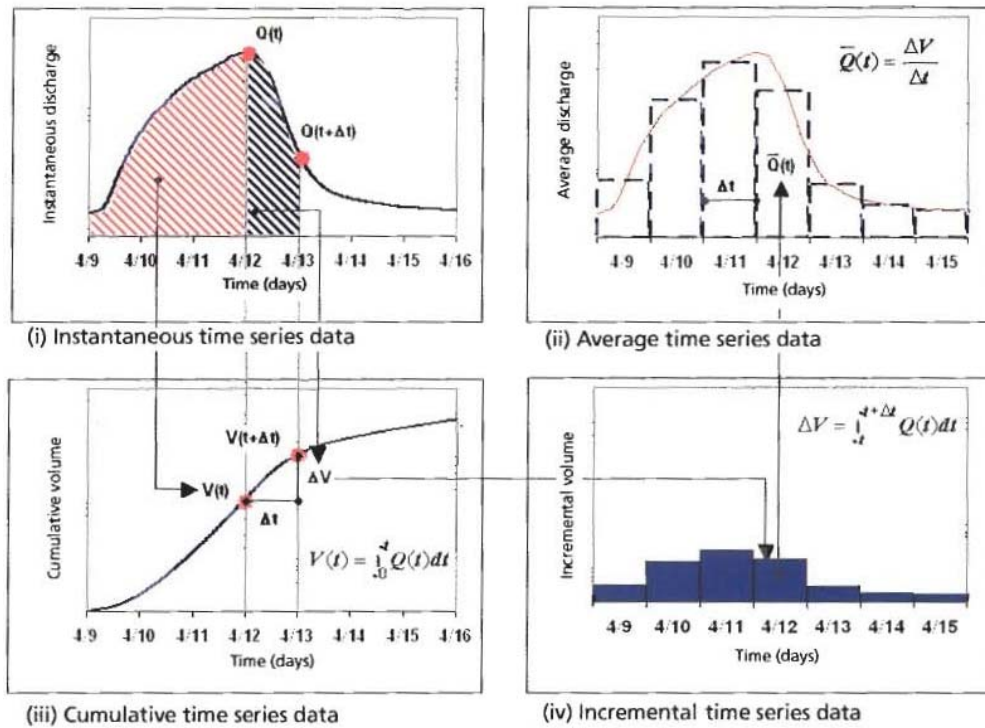


Figure 10: Time series data types. Source: Maidment (2002).

The Arc Hydro TSType table contains the information describing the time series data attributes, Figure 11. The attributes of this table are:

- *TSTypeID* – a long integer key field which indexes the type of time series;
- *Variable* – a text field describing the variable being presented;
- *Units* – a text field giving the units of the data;
- *IsRegular* – a Boolean (True/False) field that specifies whether the data are regularly recorded or not;
- *TSInterval* – a coded value domain that specifies the time interval of the data as one of a set of 17 specified time intervals;
- *DataType* – a coded value domain specifying the time series data as one of the 6 types just described (instantaneous = 1, cumulative = 2, incremental = 3, average = 4, maximum = 5, minimum = 6).
- *Origin* – a coded value domain that specifies whether the data were recorded or generated using a model;

Coded Value Domains are a device used in ArcGIS to associate integer values with a text field as a single field attribute in a data table. These are the only admissible values for this field.

| OBJECTID* | TSTypeID* | Variable | Units | IsRegular | TSInterval | DataType | Origin |
|-----------|-----------|------------------|--------|-----------|------------|----------|--------|
| 1 | 2 | Daily Streamflow | cfs | 1 | 17 | 4 | 1 |
| 2 | 3 | Hourly Rainfall | inches | 1 | 10 | 3 | 1 |
| 3 | 4 | Daily Rainfall | inches | 1 | 17 | 3 | 1 |

Figure 11. An Arc Hydro TSType table

Advantages and Limitations of the Arc Hydro time series data model

The main advantage of the Arc Hydro time series model is that provides a way of forming “geospatial time series”, that is, time series linked to the spatial features they describe. Although the discussion here has been limited to MonitoringPoint features, in fact any feature class of points, lines, areas or volumes may be used as the spatial indexing field, and a given time series table may contain data that relate to HydroFeatures from many feature classes, all uniquely indexed by their HydroID. This creates a simple but powerful model for describing time variation on discrete spatial objects.

The careful specification of Arc Hydro time series types allows for proper construction of functions operating over the database to ensure that they adapt themselves to the character of the time series data they are operating on (e.g. the accumulation of instantaneous streamflow discharge values to get a cumulative volume of flow has to be handled differently than the accumulation of incremental precipitation values to get cumulative precipitation).

The structure or *schema* of the time series data model shown in Figure 9 is very simple and it can be implemented in any relational database system including open source databases such as MySQL and PostgreSQL. The tables that it involves can be opened and interpreted as Excel spreadsheets.

The storage of all time series information in one very large table allows for SQL queries to operate on that table to extract many different combinations of data. An Arc Hydro geodatabase with more than 5 million records of hydrologic observations for the Rio Grande basin has been built and is operating successfully. The South Florida Water Management District stores its Nexrad data using the Arc Hydro format with a value for each Nexrad cell in which rainfall occurs on a grid of 30,000 cells at a time interval of 15 minutes for several years duration, and serves the data out in delimited text format and Excel format using the ArcIMS Internet Map Server.

During the several years that have passed since the Arc Hydro time series format has been established, operational experience has pointed to some limitations:

- It is difficult to index individual time series within a large data table, and it is also difficult to view and interpret the data in this form without the assistance of a software application such as a querying and graphing routine;

- The use of Coded Value Domains leaves integer values for domain indices stored in the time series attribute fields that are not particularly interpretable when viewed in other data systems, somewhat like the USGS water quality codes shown in Table 5.
- The limitation of time series intervals to 17 specified values is too restrictive and a more flexible scheme is needed whereby a time interval type is specified (second, minute, hour, day, month, year), and a separate field stores the number of the specified intervals this series is using (e.g. 1 or 3.4);
- The limitation of the Origin of the data to two values Generated and Recorded is too restrictive, and a more general text field should be used for this attribute.

The Arc Hydro time series format provides a solid foundation for storing Hydrologic Observational data in the CUAHSI Hydrologic Information System but detailed review and examination by the CUAHSI community of the format and attributes used in this structure is needed.

ArcIMS Service for Publishing Arc Hydro Time Series

Observational data on rainfall, streamflow, water quality and groundwater levels have been compiled for the Neuse basin into the Arc Hydro time series format, and an ArcIMS internet web service has been constructed to display the measurement points and to enable downloading of the time series information in a comma delimited form as a .csv file. It is normal to think of a map service as serving and delivering maps on the internet but in this case, the map serves as a backdrop for the display and downloading of time series information on sets of observational data. Figure 12 shows the map interface for the ArcIMS web service for the Neuse basin hydrologic observational data. The panel of tools on the lower right of the map display is specially constructed to enable selection and downloading of sets of hydrologic data that can be specified for any combination of stations, types of data, time period and limits on the range of the values of the variables selected. The user may select from the types of data shown in Figure 13. A more elaborate interface would allow multiple variables to be selected in a single query but the present interface just allows for one variable at a time to be selected.

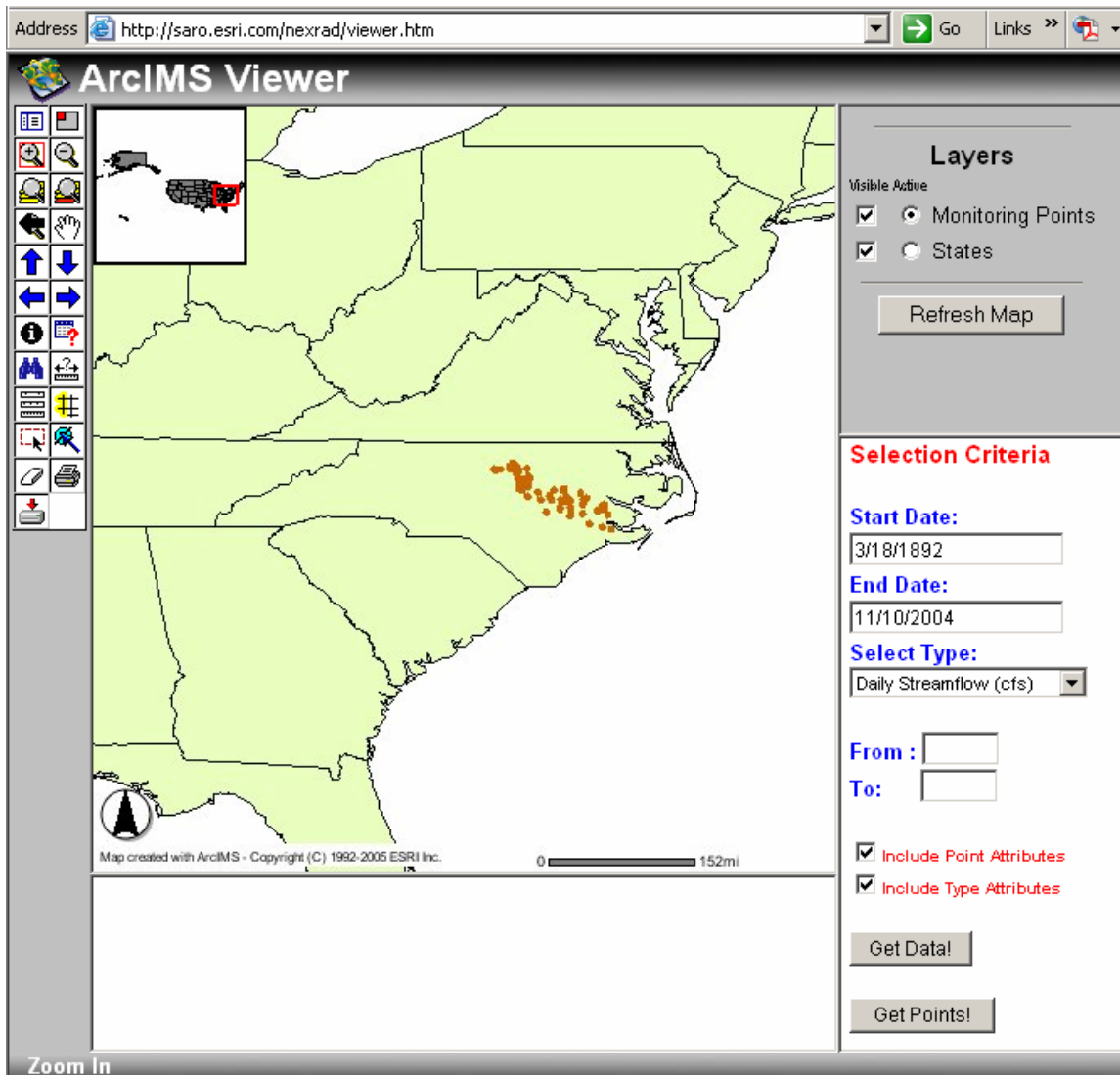


Figure 12. An ArcIMS web service for downloading hydrologic observational data for the Neuse basin.

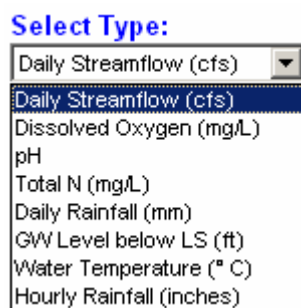


Figure 13. Hydrologic observational data for the Neuse basin served using ArcIMS

Data downloaded from the ArcIMS web server for daily streamflow and water quality from the USGS gage on the Neuse river near Clayton NC are shown in Tables 7 and 8, respectively, in Excel format. The ArcIMS web site allows for saving of the file in .csv format which opens directly in Excel without requiring any importing or interpretation. In Table 6, the TSDatetime field for the daily streamflow data has been formatted in Excel to show only the day, while in Table 7 the corresponding field for the dissolved oxygen data has been formatted to show both date and time of day since the dissolved oxygen varies throughout the day and the measurements are time stamped with the time at which the sampling began.

| HydroCode | FeatureID | TSTypeID | TSValue | TSDatetime | Variable | Units |
|-----------|-----------|----------|---------|------------|------------------|-------|
| 02087500 | 4608 | 1 | 2930 | 1/1/03 | Daily Streamflow | cfs |
| 02087500 | 4608 | 1 | 2680 | 1/2/03 | Daily Streamflow | cfs |
| 02087500 | 4608 | 1 | 2180 | 1/3/03 | Daily Streamflow | cfs |
| 02087500 | 4608 | 1 | 2090 | 1/4/03 | Daily Streamflow | cfs |
| 02087500 | 4608 | 1 | 1930 | 1/5/03 | Daily Streamflow | cfs |
| 02087500 | 4608 | 1 | 1850 | 1/6/03 | Daily Streamflow | cfs |
| 02087500 | 4608 | 1 | 1760 | 1/7/03 | Daily Streamflow | cfs |
| 02087500 | 4608 | 1 | 826 | 1/8/03 | Daily Streamflow | cfs |
| 02087500 | 4608 | 1 | 579 | 1/9/03 | Daily Streamflow | cfs |
| 02087500 | 4608 | 1 | 549 | 1/10/03 | Daily Streamflow | cfs |

Table 7. Daily streamflow data for the Neuse River near Clayton NC

| HydroCode | FeatureID | TSTypeID | TSValue | TSDatetime | Variable | Units |
|-----------|-----------|----------|---------|------------------|------------------|-------|
| 02087500 | 4608 | 3 | 7 | 9/17/96 1:30 PM | Dissolved Oxygen | mg/l |
| 02087570 | 4606 | 3 | 4.4 | 9/17/96 2:15 PM | Dissolved Oxygen | mg/l |
| 02087500 | 4608 | 3 | 5.7 | 9/17/99 11:30 AM | Dissolved Oxygen | mg/l |
| 02087570 | 4606 | 3 | 5.3 | 9/17/99 2:00 PM | Dissolved Oxygen | mg/l |
| 02087500 | 4608 | 3 | 7.2 | 9/19/99 10:30 AM | Dissolved Oxygen | mg/l |
| 02087570 | 4606 | 3 | 4.5 | 9/19/99 2:00 PM | Dissolved Oxygen | mg/l |
| 02087500 | 4608 | 3 | 7.5 | 9/20/99 11:00 AM | Dissolved Oxygen | mg/l |
| 02087570 | 4606 | 3 | 4.9 | 9/20/99 1:30 PM | Dissolved Oxygen | mg/l |
| 02087500 | 4608 | 3 | 7.7 | 9/28/99 9:00 AM | Dissolved Oxygen | mg/l |
| 02087570 | 4606 | 3 | 7.4 | 9/28/99 9:30 AM | Dissolved Oxygen | mg/l |

Table 8. Water quality data for the Neuse River near Clayton NC

The example just shown of data downloading is rather elementary but it is sufficient to demonstrate that hydrologic observations can be treated in a consistent manner across the main kinds of data required, and data can be readily acquired from the ArcIMS server from one or a group of stations in a region. If the appearance of Tables 6 and 7 is contrasted with the variable appearance of data in Tables 1-5, the value is readily apparent of having a systematic structure to the downloading and presentation of hydrologic observations that does not vary according to the original web site or source of the data.

Automated Data Harvesting

Suppose now, that the intent is not to acquire hydrologic observations in an existing database like the one for the Neuse basin just illustrated but instead to use a common web interface to interrogate the NWIS, Storet and NCDC web sites directly. What is then required is that the points on the map represent locations where measurements are available at those web sites, and that there is also an inbuilt routine customized for each individual website that knows how to construct the appropriate query language to download the data automatically into the Arc Hydro time series format. This requires first of all a *observation site map*, which is a map of all the observational sites whose data are available from a give web site. The USGS has provided CUAHSI with interpretive code that has enable the CUAHSI HIS team to query and download the observational site characteristics of about 1.4 million sites in the continental United States whose data are available from NWIS. The CUAHSI HIS team also has a map of more than 8000 NCDC precipitation and temperature measurement sites. The USGS also publishes their site map as an ArcIMS web service from the USGS seamless server <http://seamless.usgs.gov> A map of streamgaging sites with real-time telemetry for North Carolina is shown in Figure 14

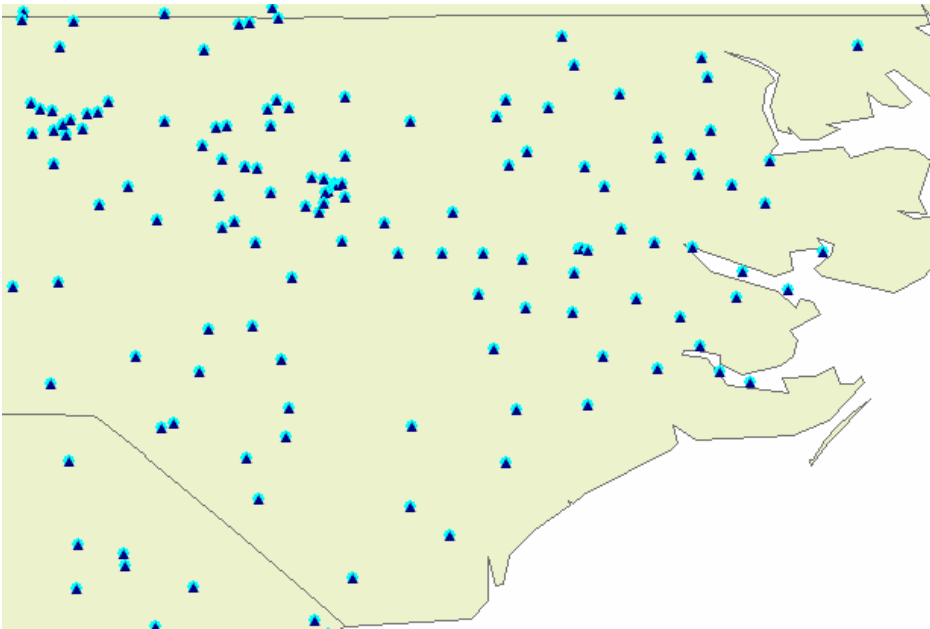


Figure 14. Streamgaging sites with real time telemetry in North Carolina
Source: GIS Servers\seamless.usgs.gov\NWIS_FS\Sites with real-time telemetry

Associated with each web site is a protocol for downloading data from that site. For example, to obtain the most recent 7 days of real time streamflow data for the Neuse River near Clayton NC, when accessed through the USGS Water Watch web site, requires the following html query:

http://waterdata.usgs.gov/nc/nwis/uv?format=rdb&period=7&site_no=02087500

This query can be reproduced programmatically, so it is possible to automatically execute this query, obtain the resulting ASCII file, extract the required data from it into Arc Hydro format and store the result in a geodatabase. This task has been carried out for daily streamflow data at more than 1300 stations across the United States as part of the construction of Arc Hydro USA, a geodatabase of the United States prepared at the time the original Arc Hydro book was published in 2002.

It is unknown whether the EPA and NCDC websites can also be programmatically accessed as the NWIS site can be. In the event that is possible, then an automated data harvesting tool for the main federal sources of hydrologic information could be prepared, and perhaps extended to include similar information from state and local web sites.

Conclusions

This paper has examined the way in which hydrologic observations for water quality, streamflow, rainfall, air temperature and groundwater are presented on the USGS National Water Information System (NWIS), the NCDC Climate Data Online (CDO) and the EPA Storet system. It is shown that the format in which information is presented from these web sites in ascii format is similar structurally from one web site to another although there are many variations even within a single agency in the details of how data are presented. The Arc Hydro time series data model is presented as an example of geospatial time series where a table of time series measurements is relationally connected to a table describing the monitoring points at which the measurements were made and to another table which describes the character of the time series data. It is demonstrated that this structure captures the main features of the output of the NWIS, CDO and Storet data systems. The advantages and limitations of the Arc Hydro time series model are evaluated and some extensions to it are suggested. It appears that this data model is appropriate for constructing a hydrologic observations database for the CUAHSI Hydrologic Information System but independent review by the CUAHSI community is needed to ensure that the resulting structure is sufficient and flexible.

A prototype ArcIMS web site is described that serves an integrated database of hydrologic observations for the measurement sites in the Neuse basin and produces delimited ascii files in .csv format that are readily read into Excel spreadsheets. It may be possible to build an automated data harvesting system by executing automated queries from a map of observation sites like the one presented in the ArcIMS web service, except that instead of obtain the data from its own geodatabase, the data harvester would go out to a series of federal web sites and download the information directly.

Reference

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