Environmental Modelling & Software 37 (2012) 146–156

Contents lists available at SciVerse ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

HydroDesktop: Web services-based software for hydrologic data discovery, download, visualization, and analysis

Daniel P. Ames [a],*, Jeffery S. Horsburgh [b], Yang Cao [a], Jiří Kadlec [a], Timothy Whiteaker [c], David Valentine [d]

[a] *Geospatial Software Lab – Ctr. for Adv. Energy Studies, Idaho State Univ., Idaho Falls, ID 83402, USA*
[b] *Utah Water Research Laboratory, Utah State Univ., Logan, UT, USA*
[c] *Center for Research in Water Resources, Univ. of Texas at Austin, Austin, TX, USA*
[d] *San Diego Supercomputer Center, University of California, San Diego, USA*

ARTICLE INFO

Article history:
Received 24 November 2011
Received in revised form
21 March 2012
Accepted 24 March 2012
Available online 27 April 2012

Keywords:
Data management
Geographic information system
Hydrologic information systems
Hydrologic modeling
Observation data
Web services

ABSTRACT

Discovering and accessing hydrologic and climate data for use in research or water management can be a difficult task that consumes valuable time and personnel resources. Until recently, this task required discovering and navigating many different data repositories, each having its own website, query interface, data formats, and descriptive language. New advances in cyberinfrastructure and in semantic mediation technologies have provided the means for creating better tools supporting data discovery and access. In this paper we describe a freely available and open source software tool, called HydroDesktop, that can be used for discovering, downloading, managing, visualizing, and analyzing hydrologic data. HydroDesktop was created as a means for searching across and accessing hydrologic data services that have been published using the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) Hydrologic Information System (HIS). We describe the design and architecture of HydroDesktop, its novel contributions in web services-based hydrologic data search and discovery, and its unique extensibility interface that enables developers to create custom data analysis and visualization plug-ins. The functionality of HydroDesktop and some of its existing plug-ins are introduced in the context of a case study for discovering, downloading, and visualizing data within the Bear River Watershed in Idaho, USA.

Software availability

All CUAHSI HydroDesktop software and documentation can be accessed at http://his.cuahsi.org. Source code and additional documentation for HydroDesktop can be accessed at the HydroDesktop code repository website http://hydrodesktop.codeplex.com. HydroDesktop and its source code are released under the New Berkeley Software Distribution (BSD) License which allows for liberal reuse of the software and code.

1. Introduction

As scientists begin to investigate complex hydrologic processes at expanding spatial and temporal scales, integration of data from multiple sources, projects, and research efforts becomes critical. The last several years have seen a significant improvement in hydrologic and environmental data availability (Beran and Piasecki, 2009); however, most observational data published on the Internet are not inherently discoverable using automated systems because they are usually encapsulated within files or databases, the contents of which cannot easily be discovered or cataloged by web crawler technologies employed by major web search engines.

Another challenge is that syntactic and semantic heterogeneity in data from different sources make data discovery, integration, and synthesis difficult. Observational data are rarely annotated with sufficient attribute information, or metadata, to make their interpretation unambiguous by investigators other than those who collected the data, and semantic differences in metadata from different sources can limit both data discovery and data integration. Syntactic heterogeneity, or differences in the way data are encoded or organized, makes it difficult for data consumers to mediate across the various data formats that they download to organize data and get it into their data analysis software of choice.

The Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) Hydrologic Information System (HIS) is focused on the development of cyberinfrastructure for hydrologic science. Its vision is to bring together hydrologic observations from multiple sources across the globe into a uniform, standards-based, service-oriented environment where heterogeneous data can be seamlessly integrated for advanced computer-intensive analysis and modeling. The HIS team has developed several software tools and

* Corresponding author. Tel.: +1 208 533 8141.
E-mail address: dan.ames@isu.edu (D.P. Ames).

1364-8152/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.envsoft.2012.03.013

standards that together enable publishing and accessing hydrologic data collected at point locations (e.g., time series of observations from stream gages, water quality monitoring sites, weather stations, etc.). The availability of the CUAHIS HIS tools has led to a network of hydrologic data servers that hosts data from a number of different sources, including the United States Geological Survey's (USGS) National Water Information System (NWIS), the U.S. Environmental Protection Agency's (USEPA) STORage and RETrieval (STORET) system, and academic research groups collecting data within experimental watersheds across the United States. HIS compliant data servers have also been set up in the Czech Republic and other countries (Kadlec and Ames, 2011).

Each hydrologic data server participating in the CUAHSI HIS hosts one or more water data web services that publish observational data on the Internet in a standard format, or syntax (Beran et al., 2009). This federated system of observational data web services comprises the largest repository of syntactically homogenous hydrologic observations of its kind within the U.S. and perhaps the world. As of June 30, 2011, there were over 70 water data web services registered with the CUAHSI HIS, having data for approximately 2 million sites and observations of over 13,000 variables. The CUAHSI HIS provides access to over 23 million observational time series comprised of more than 5 billion individual observations.

The CUAHSI HIS network has increased the availability and accessibility of hydrologic observations to support research and management. However, the volume of data now available via water data services and the distributed nature of the servers on which the data are hosted can present a challenge for individual scientists who seek to discover and use data from one or more of these services. Tools are needed to assist data consumers in discovering data of interest, for managing the download and organization of the data, and for facilitating the import of data into analytical and modeling software where they can be used.

In this paper, we describe the design, architecture, implementation, and a case study involving a newly developed open source HIS software tool called HydroDesktop intended to meet this need. HydroDesktop was designed to enable the discovery and retrieval of syntactically homogenous data hosted on any of the distributed hydrologic data servers registered with the CUAHSI HIS system using user-specified spatial, temporal, and keyword based constraints to narrow search results. Through an extensible graphical user interface (GUI), HydroDesktop provides many capabilities needed by hydrologic data consumers, including: discovery of hydrologic time series data; map-based visualization of monitoring locations and other geographic information systems (GIS) data; download, organization, visualization, editing, and maintenance of hydrologic time series; linkage with integrated modeling systems such as OpenMI (e.g. see Castronova and Goodall, 2010); and linkage with common data analysis and modeling software such as the R statistical computing environment.

Section 2 of this paper provides background and describes the need for software tools like HydroDesktop. Section 3 describes the CUAHIS HIS, including its common information model and services oriented architecture (SOA). Section 4 describes the design and key capabilities of HydroDesktop and its extensibility architecture and GUI. Section 5 describes a case study for discovering, downloading, organizing, and visualizing hydrologic data using HydroDesktop. Section 6 provides discussion and conclusions related to HydroDesktop. Section 7 describes the availability of the HydroDesktop software, documentation, and source code.

2. Background

The challenges of discovering and integrating disparate data and schemas from physically distributed sources are not unique. Finding solutions to interoperability problems is a common component of large cyberinfrastructure projects being conducted within many scientific domains, including geology (Nambiar et al., 2006), oceanography (Chave et al., 2009), and atmospheric sciences (Droegemeier et al., 2004). Within the hydrology domain, there have been, to date, no public standards for data organization, formats, or publication mechanism that would increase the interoperability of water observations data expressed as time series. Consequently, there has been no means of unified discovery or access to water observations information.

Water observations data are series or sets of time-indexed values collected at point locations such as stream gages, water quality sampling sites, and weather stations. They describe the quantity and quality of water or the characteristics of weather and climate that influence water conditions. Such data are collected by many agencies and also by water and atmospheric scientists, and are often made publicly accessible through individual websites from each data source. However, most observational data published on the Internet are not discoverable using web search engines because the data are usually encapsulated within files or databases, the contents of which cannot be easily discovered or cataloged by web crawler technologies employed by search engines.

A user of water data, such as a water scientist or student, needs access to water observations from one or more data servers seamlessly. That is, the user needs to simply search for water data of the required type on the Internet, obtain a description of datasets that may meet his or her needs, and then download data that are most relevant to his or her search criteria. There are several challenges to overcome in creating such capability. For example, querying the data holdings of large, nationwide agencies like USEPA and the USGS, as well as those of regional or local water management agencies or research groups currently requires that a data consumer first learn about a data source, then find the query functionality provided by each data source, and finally learn how to operate each data discovery and retrieval system. Each data source may have different functionality, software interfaces, and query criteria, making discovery of data difficult.

Several existing hydrologic software tools and products (both commercial and open source) have capabilities for downloading spatial and temporal data from predetermined data sources. For example, the USEPA Better Assessment Science Incorporating Point & Nonpoint Sources (BASINS) system includes a data download tool with support for USGS streamflow, water quality, and watershed boundary data, as well as many other datasets (http://www.epa.gov/ost/basins). Another open source project, Water Resources Database (WRDB – http://wrdb.codeplex.com) provides support for downloading data from selected U.S. government sponsored data repositories and has extensive capabilities for managing a local database of hydrologic time series data. These and similar tools typically require custom code for each website from which they retrieve data. Such an approach is maintainable only for a small number of well-defined websites because of the unique format and structure of each data provider website.

A well-established solution to the problem of providing consistent web-based data access is to implement a system of "web services" on each data server. A web service is a standardized method for communication between two devices or software applications over the internet (W3C, 2004). One of the leading organizations for establishing standardized web service definitions is the Open Geospatial Consortium (www.opengeospatial.org), which has defined web service standards for sharing geospatial data such as Web Coverage Services (WPS), Web Map Services (WMS), Web Feature Services (WFS), and other standard service interfaces.

Web services have been applied in a number of environmental modeling and data analysis studies. For example, Granell et al. (2009)

implemented a web services-based modeling system wherein key-geospatial data processing tasks were carried out on distributed servers accessed by an alpine runoff model. Feng et al. (2011) created a web services based system using OGC standards for sharing computational models such that different models that implement the WPS web service standard can be accessed and integrated through a common user interface. Castronova et al. (2012) take significant steps toward integrating the CUAHSI HIS web services with component based modeling using the OpenMI standard. The present work builds heavily on web services-based data sharing advancements made by Goodall et al. (2008) and Horsburgh et al. (2010).

Another challenge in seamlessly integrating multiple data sources is resolving heterogeneity issues (Beran and Piasecki, 2009; Piasecki and Beran, 2009; Horsburgh et al., 2009). In the example above, each data source may use different vocabularies to describe data collection locations or measured variables, making it difficult to search multiple systems for similar data. This can be particularly difficult when a study requires data from multiple scientific domains, and the scientist is not familiar with the vocabulary used by one or more of the domains whose data is being accessed. Performance of queries and search mechanisms for data discovery can be significantly improved when semantic heterogeneity in data among datasets is overcome (Madin et al., 2007).

Even when scientists can find needed data, there remains the sometimes difficult challenge of mediating across the various formats, or syntaxes, of the data for integration purposes. Syntactic heterogeneity includes both the way data and metadata are organized (e.g., rows versus columns) and how they are encoded (e.g., text files versus Excel spreadsheets). Syntactic heterogeneities tend to arise from differing methods of either collecting data or publishing them on the Internet. Given the large variety in computer hardware, software, and file formats used by hydrologists, there are relatively few general-purpose data storage formats and data management tools available.

Beran and Piasecki (2009) described several innovations within this problem space, and indeed, much of the work related to discovery of hydrologic data described in this paper is a direct outgrowth of their work. They described an ontology-aided search engine website called HydroSeek, which was developed to enable users to query multiple hydrologic data repositories simultaneously using keywords. They created this functionality by developing what they called a “knowledge base” that covered the water quality, meteorology, and hydrology domains, and that provided a linkage between scientific or everyday language (e.g., the keywords or terms that scientists would use to search for data) and the language and variable codes used by repositories to store data.

HydroSeek used its knowledge base to resolve semantic heterogeneity issues between data repositories as well as for clustering search results. Metadata catalogs were created for each data source, which were similar to web indexes compiled by conventional web search engines and contained a description of the data holdings of each source to facilitate the search. Development of the knowledge base, the metadata catalogs, and the semantic mappings between variables from each data source and the keywords or concepts in the knowledge base was done by the developers of HydroSeek. HydroSeek did not have data analysis or extensive visualization capabilities, but data discovered using HydroSeek could be downloaded to a user’s computer for later analysis.

The data discovery and download functionality that was pioneered by HydroSeek has now been generalized and expanded within the CUAHSI HIS system through the creation of two major components. The first is a central web service registry, metadata catalog, and data discovery service called HIS Central (described in more detail in the next section). The second is HydroDesktop, which is end-user client software that has data discovery and download capabilities similar to HydroSeek as well as a number of data organization, management, analysis, visualization, and management tools.

3. The CUAHSI HIS: a service-oriented architecture for hydrologic observations

The design of the CUAHSI HIS (Fig. 1) follows an open, service-oriented architecture (SOA) model. SOA relies on a collection of loosely-coupled, self-contained services that communicate with each other through the Internet and can be called from multiple clients in a standard fashion (Goodall et al., 2008). Common benefits associated with a SOA include: scalability, security, easier monitoring and auditing, standards-reliance, interoperability across a range of resources, and plug-and-play interfaces (Josuttis, 2007; Goodall et al., 2010). Internal service complexity is hidden from service clients, and backend processing is decoupled from client

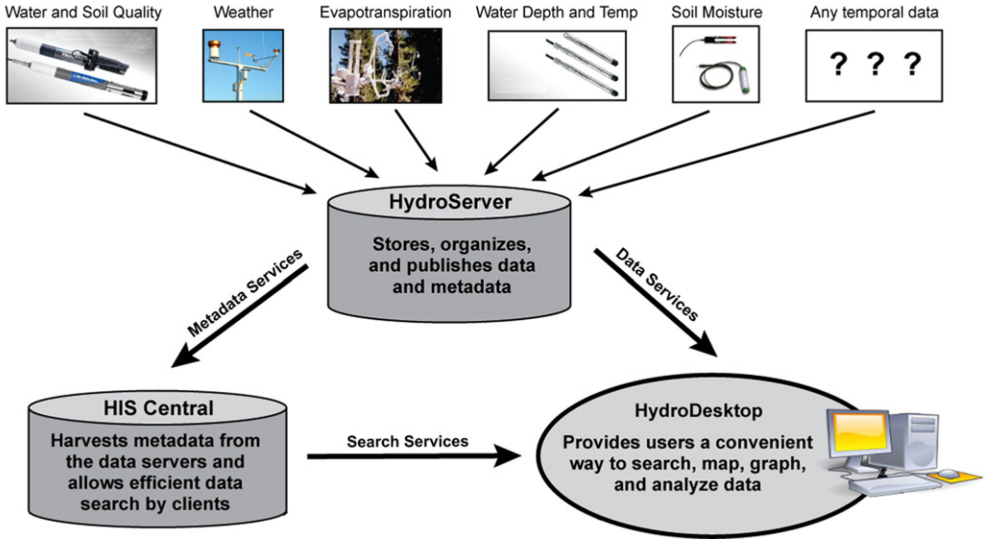

Water and Soil Quality
Weather
Evapotranspiration
Water Depth and Temp
Soil Moisture
Any temporal data
? ? ?
HydroServer
Stores, organizes, and publishes data and metadata
Metadata Services
Data Services
HIS Central
Harvests metadata from the data servers and allows efficient data search by clients
Search Services
HydroDesktop
Provides users a convenient way to search, map, graph, and analyze data

Fig. 1. Key components of the CUAHSI HIS include a distributed network of HydroServers sharing data that have been cataloged at HIS Central and which are retrievable through HydroDesktop (Figure courtesy of Stephen Brown, Univ. of New Mexico.).

applications, making the core of the system independent of a specific platform or implementation (Huhns and Singh, 2005; Granell et al., 2009). As a result, different desktop and online client applications are able to access the same service functionality, leading to a more modular, transparent, and easier to manage system. HydroDesktop is presented here as a desktop client application that fully exploits the modular design of the CUAHSI HIS.

At the physical level, the CUAHSI HIS infrastructure represents a collection of computer servers, referred to as HydroServers, which support publishing hydrologic observations data (Horsburgh et al., 2010). Over the past several years, HydroServers have been installed at various universities and public agencies, and there is now a large and growing number of hydrologic observations data available via web services published on HydroServers (Horsburgh et al., 2009).

The core of the HIS SOA is a collection of hydrologic web services, called WaterOneFlow, that provide uniform access to multiple repositories of water observations data. Each HydroServer hosts one or more WaterOneFlow web services, each of which contains two types of web service methods: 1) a data delivery method called GetValues, which publishes the values of water observations; and 2) metadata delivery methods, including GetSites, GetSiteInfo, and GetVariableInfo, which identify and describe collections or series of data values associated with particular spatial locations. WaterOneFlow web services are used to publish hydrologic observations on a HydroServer using a markup language called Water Markup Language (WaterML) (Zaslavsky et al., 2007). These are key components of CUAHSI HIS, ultimately enabling data search and discovery through the HydroDesktop application.

Data published on HydroServers using WaterOneFlow web services are indexed within a central web service registry and metadata catalog called HIS Central. HIS Central regularly harvests the metadata describing published time series of hydrologic observations from registered WaterOneFlow web services by calling their metadata delivery methods. The metadata are compiled within a central metadata catalog database, which is then exposed to search queries via a data discovery web service. HIS Central also contains a variable name ontology that is used to semantically tag variable metadata harvested from WaterOneFlow web services and enable mediation across the vocabularies used by different data sources. This data discovery web service is publicly accessible and can be called by any client application that wishes to incorporate data discovery capabilities. It supports spatial, temporal, and variable keyword constraints to narrow search results.

The CUAHSI HIS SOA is completed by client applications that use the data discovery service available at HIS Central to enable discovery of relevant time series of hydrologic observations and then use the metadata returned by data discovery queries to access the data via the WaterOneFlow web services on the HydroServer that hosts the data. Client applications are then free to download, store, and use the data. The primary client application that has been developed by the CUAHSI HIS project is HydroDesktop.

3.1. A common information model

Unifying data discovery and access across data from many different sources requires identification of the informational elements that are common across all data sources and an information model that can represent the semantics of the data from each source. The CUAHSI HIS relies on a common information model for organizing, storing, and publishing observational time series collected at point locations, a simplified version of which is depicted in Fig. 2. In this high-level view of the information model (Panel a), an organization operates a network of monitoring sites. At each monitoring site a number of variables are measured resulting in a time series of data values for each variable at each site. Each data series (Panel b) is made up of individual, time-indexed values. This information model is generic and can be used regardless of the source of the data.

On HydroServers, the information model has been implemented within the design of the Observations Data Model (ODM) (Horsburgh et al., 2008), ensuring that hydrologic observations are stored with their associated metadata. It also serves as the basis for the WaterOneFlow web service methods and the WaterML schema, which ensures that observations are published and transmitted with all appropriate metadata. These tools make it possible for disparate investigators and organizations to publish their hydrologic observations using a common mechanism, in a common language, with common syntax and semantics, thus alleviating the heterogeneity in hydrologic datasets from different sources (Horsburgh et al., 2009).

The consistency enforced by use of WaterOneFlow web services has also been key to enabling data discovery and access across the many different sources of hydrologic data within the CUAHSI HIS. Indeed, availability of the data via common web services (WaterOneFlow) that use a common syntax (WaterML) for publishing and transmitting the data means that data can be indexed consistently, in much the same way that the standardization of the Internet on HTML enabled modern Internet search providers (e.g., Google, Yahoo, Bing) to catalog and index the contents of the Internet and enable sophisticated searches. It also opens the door for client software applications to discover and access data using these services.

4. Architecture, design, and key capabilities of HydroDesktop

The general architecture of HydroDesktop and its relationship to other HIS project components are shown in Fig. 3. HydroDesktop serves as a common window into observational data published using

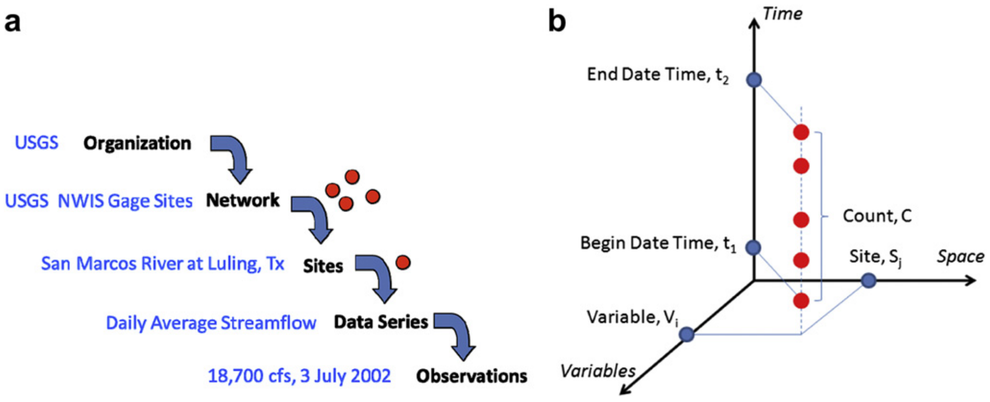


Fig. 2. Simplified information model for point hydrologic observations.

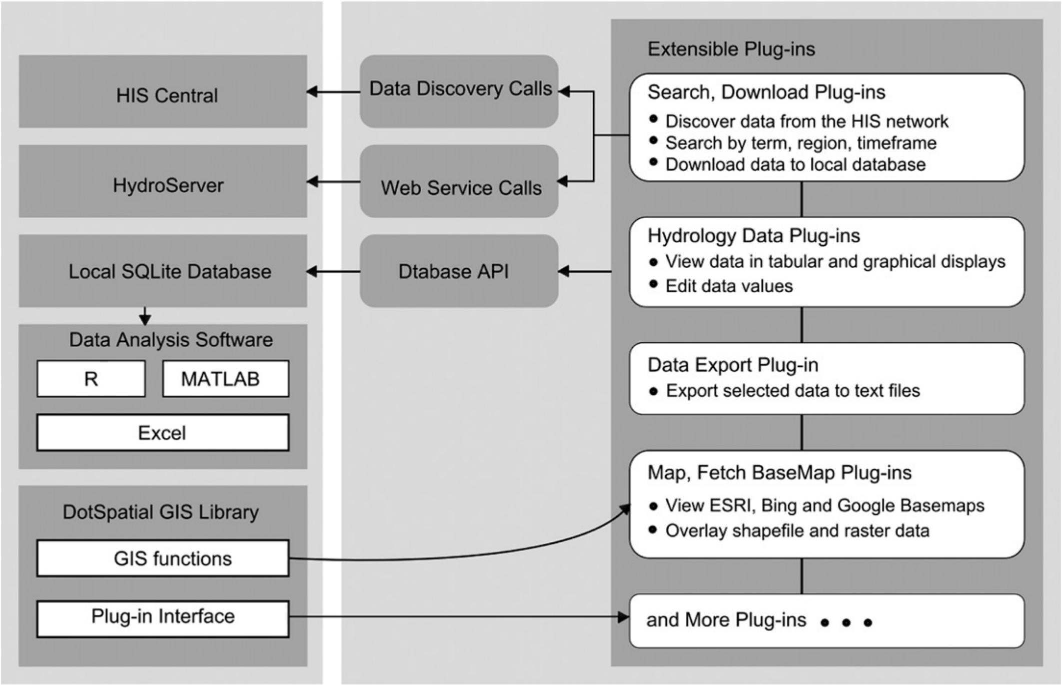

HIS Central
HydroServer
Local SQLite Database
Data Analysis Software
R
MATLAB
Excel
DotSpatial GIS Library
GIS functions
Plug-in Interface
Data Discovery Calls
Web Service Calls
Dtabase API
Extensible Plug-ins
Search, Download Plug-ins
• Discover data from the HIS network
• Search by term, region, timeframe
• Download data to local database
Hydrology Data Plug-ins
• View data in tabular and graphical displays
• Edit data values
Data Export Plug-in
• Export selected data to text files
Map, Fetch BaseMap Plug-ins
• View ESRI, Bing and Google Basemaps
• Overlay shapefile and raster data
and More Plug-ins • • •

Fig. 3. HydroDesktop design and relationship to other CUAHSI HIS components.

WaterOneFlow web services (Ames et al., 2009). Data discovery is accomplished through searches across a comprehensive metadata catalog maintained at HIS Central and/or individual HydroServers hosting WaterOneFlow web services. These searches are facilitated by additional web services that expose the metadata catalog and the Hydrologic Ontology maintained at HIS Central. Search results can be further refined to specify datasets that a user would like to download. Data downloads are performed by making GetValues calls (part of the WaterOneFlow web services definition) to the appropriate WaterOneFlow web services. Downloaded data are stored in a desktop data repository database following a relational database schema. This database is accessible to additional tools and software either through an application programmer interface (API) or directly.

Visualization and analysis tools that are part of HydroDesktop are developed using the API data access method to maintain a level of data access consistency and integrity as well as abstraction from the HydroDesktop database. Additionally, users can access data through third party data analysis applications that have the ability to read from a relational database (e.g. R and MATLAB). HydroDesktop includes a number of plug-ins developed by the HIS team, and also supports third party plug-ins that follow a standard, well-defined plug-in interface described at the project website.

HydroDesktop has a familiar primary interface similar to most desktop GIS programs with the addition of tools and forms specifically related to time series data visualization and analysis. Included are a ribbon-style main menu, legend, and a main map display in a tabbed interface with movable/dockable panels (Fig. 4). The map display is the main visualization element, while the other portions of the interface provide tools for searching, obtaining, and managing data. With a wide base of users in mind, HydroDesktop was developed with a simple interface that should be easily usable regardless of the operator's technical background. The GIS capabilities are powered by the open source DotSpatial GIS components (http://www.dotspatial.org).

The primary purpose of HydroDesktop is to facilitate discovery and access of hydrologic data. A secondary purpose is to provide support for data manipulation and synthesis. The user primarily interacts with HydroDesktop via a GUI with the functionality described below.

4.1. Data discovery

HydroDesktop supports two different methods of data discovery: 1) ontology-based discovery across all WaterOneFlow web services that have been registered at HIS Central and for which metadata has been harvested and stored in the HIS Central metadata catalog; and 2) discovery of data within a single WaterOneFlow web service that has not been registered at HIS Central. The first type of data discovery is supported by HIS Central metadata web services that expose the contents of the HIS Central metadata catalog. The second type of data discovery involves making data discovery calls directly to the web service that has not been registered with HIS Central. This approach facilitates both the use of datasets cataloged and documented at HIS Central, as well as use of datasets stored on individual or regional HydroServers but not registered with HIS Central.

HIS Central includes a metadata catalog describing the time series datasets served by registered WaterOneFlow web services. This catalog includes the mappings between variables and HIS Ontology concepts. This catalog is automatically updated weekly and represents a comprehensive listing of data published using WaterOneFlow services and registered at HIS Central. The contents of the HIS Central metadata catalog are exposed by a web service API that provides methods for retrieving the following information: (i) the full metadata description (including the URL to the WaterOneFlow service) for all WaterOneFlow web services registered at HIS Central, (ii) a listing of all searchable keywords/concepts from the HIS Ontology, and (iii) the full metadata description for all data that meet certain spatial, temporal, and variable search criteria.

HydroDesktop uses the methods from the HIS Central metadata catalog API to search for data series that meet criteria input by a specific user. HydroDesktop presents users with a search tool that supports the following search criteria: (i) a latitude/longitude-bounding box to serve as the spatial constraint on the query.

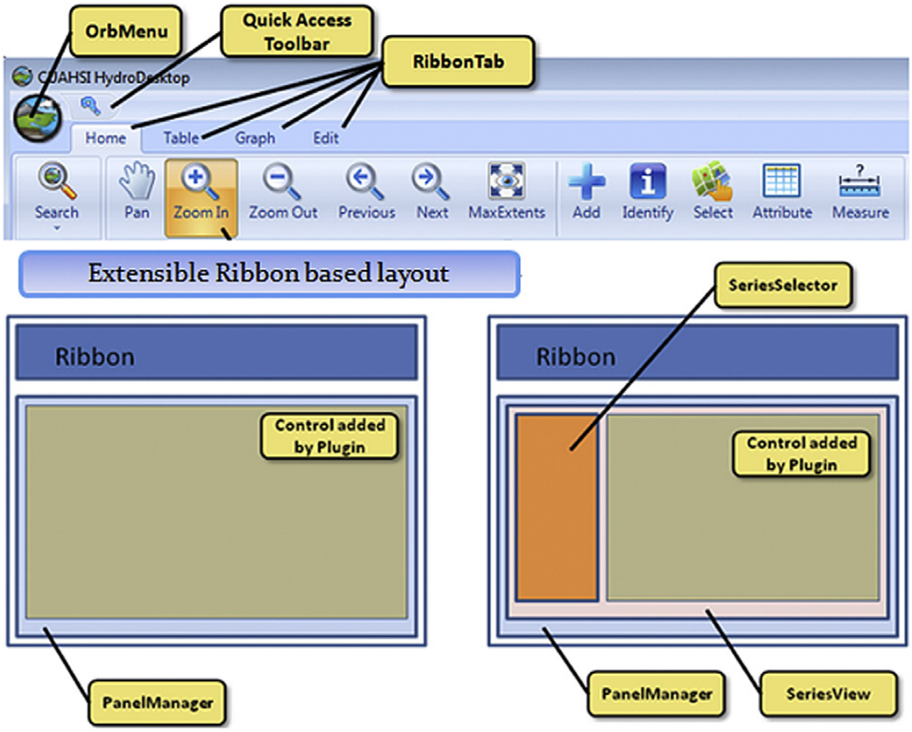


Fig. 4. Extensible ribbon based layout design of HydroDesktop.

The box can be input by typing in coordinates, by drawing a rectangle on the HydroDesktop map, or by selecting a polygon feature from one of the layers in the HydroDesktop map (e.g., a watershed boundary – the extent of the feature would be converted to a latitude/longitude box), (ii) a searchable concept from the HIS Ontology (to be input by the user or selected from a list), (iii) a begin date and end date to serve as the temporal constraint on the query, (iv) a minimum number of observations (only data series that have more than this minimum number for the entire data record will be selected, regardless of time window specified), and (v) a list of WaterOneFlow web services to include in the search. This will be a user-specified subset of the web services registered at HIS Central that constrains search results to only a selected set of web services.

The result of a data discovery query using the HIS Central metadata catalog is the full metadata description for a listing of all of the data series cataloged at HIS Central that meet the search criteria. For example, a user may choose to search the entire state of Utah for streamflow data. The results of the search will be a list of sites and data series that meet the criteria. The user can then subset the results to the data series of particular interest, i.e. after seeing a map of the locations of several hundred streamflow gauge sites in Utah, the user may choose to only retrieve data for sites that meet some additional condition. The user then organizes data into a thematic dataset on the local machine for layered GIS viewing and interaction.

4.2. Data download

The goal of the HydroDesktop data download functionality is to retrieve observational data series that have been identified for download using the data discovery tools described above and to create a local cache copy of the data in the local database. The metadata resulting from the discovery process consist of a descriptive list of data series identified by a user for download. Using this list, HydroDesktop issues GetValues calls to retrieve each data series in WaterML format. As a user selected option, HydroDesktop saves a copy of the result of each GetValues call as a WaterML formatted XML file on the user's hard drive. Next, HydroDesktop parses each of the WaterML results into the HydroDesktop data repository database. The purpose of saving the WaterML files is to preserve the data as they were retrieved from the web service when the GetValues call was made as part of data provenance. The purpose of loading the data into the data repository database is to facilitate and enable analysis and manipulation of the data.

The data repository database has a relational structure and is implemented within a relational database management system (RDBMS), serving as a local cache copy of the data that have been retrieved. The relational schema of the data repository database is semantically similar to the CUAHSI ODM database design (Horsburgh et al., 2008), with similar naming conventions and data types, but has been modified and extended to facilitate management of the data series that have been downloaded and storage of provenance information. The relational database schema of the HydroDesktop data repository database is available for viewing at http://hydrodesktop.codeplex.com/ documentation.

The data repository database is capable of storing all of the information encoded within WaterML files resulting from GetValues calls and also supports the storage of provenance information, including: (i) where was the data obtained, i.e., which web service? (ii) the query that resulted in the data that was loaded (the GetValues call used to get the data); (iii) a pointer to the WaterML file from which the data originated (the file is cached locally); (iv) the date on which the data were loaded; (v) the last date on which the data were checked for updates; (vi) the last date on which the data were updated with new data; and (vii) what has been done to the data since it was added to the database.

4.3. Data visualization, editing, and export

HydroDesktop supports visualization of both geospatial and time series data. Geospatial data visualization is enabled through an

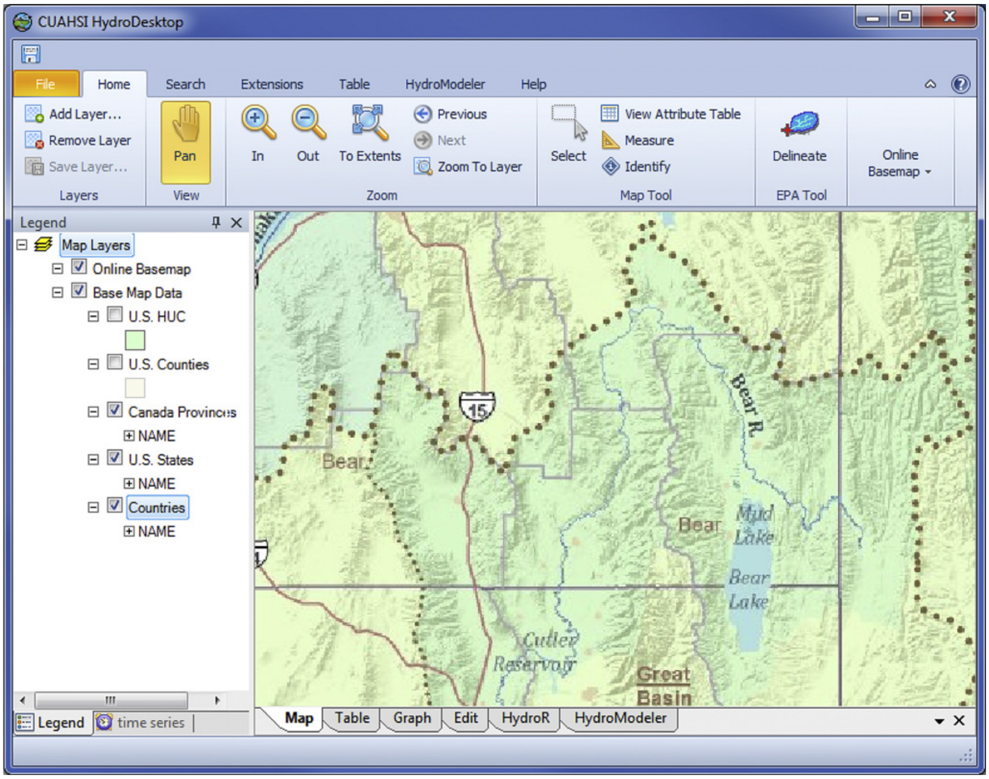
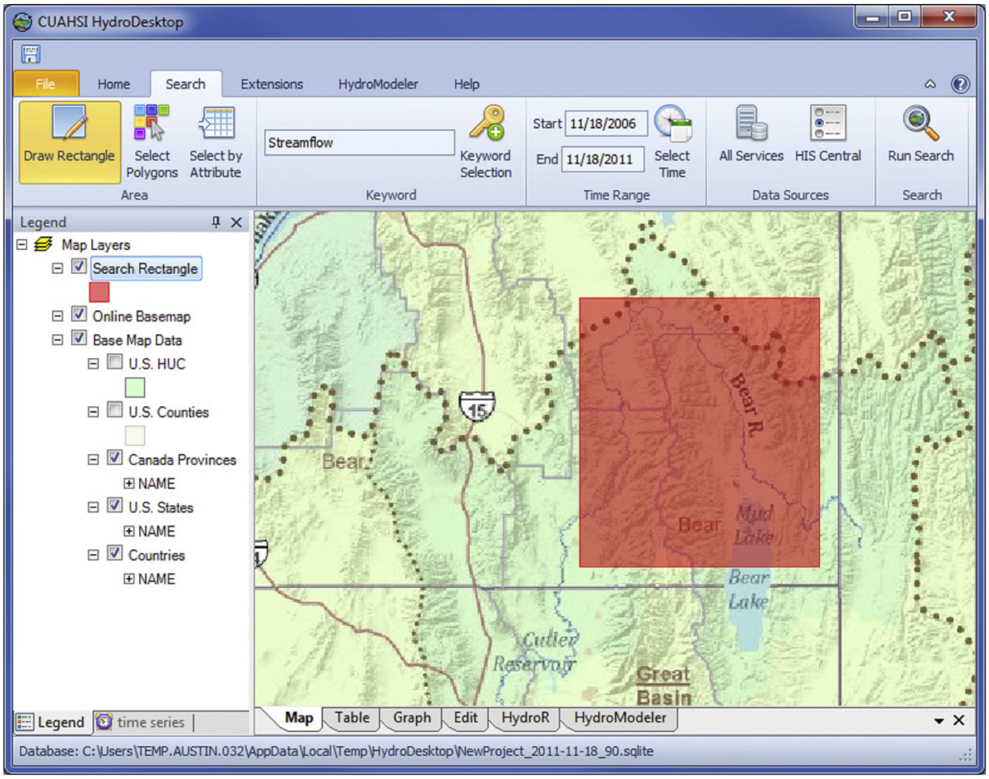


Fig. 5. HydroDesktop map interface showing the Bear River near Mud Lake.



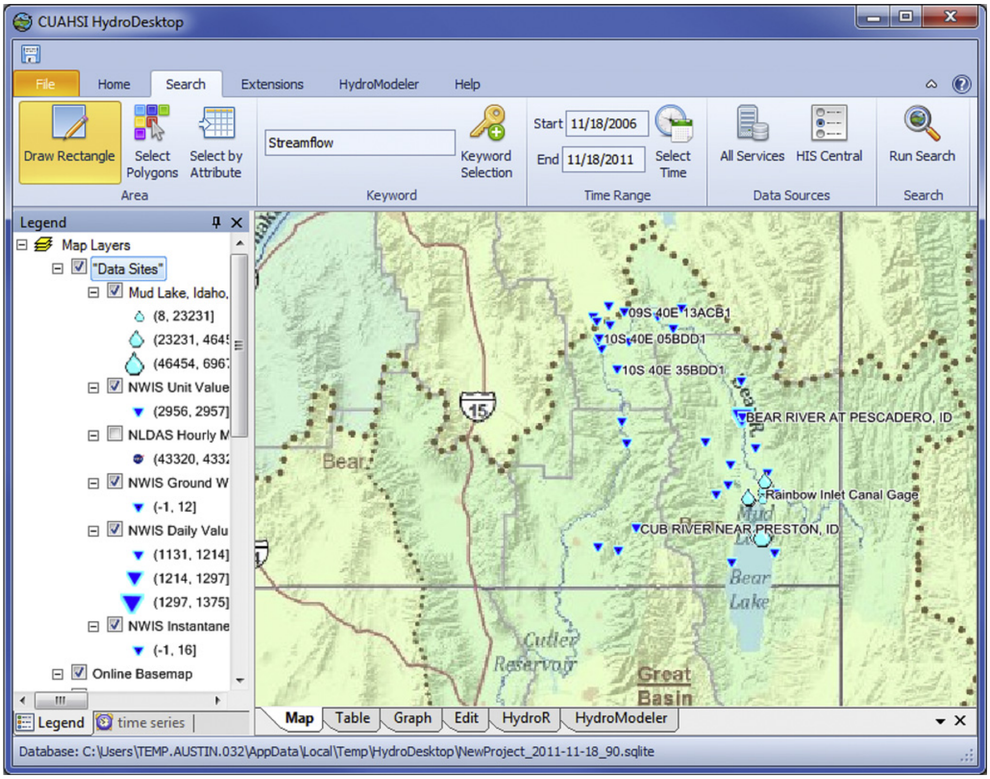
Streamflow

Start 11/18/2006

End 11/18/2011

Database: C:\Users\TEMP.AUSTIN.032\AppData\Local\Temp\HydroDesktop\NewProject_2011-11-18_90.sqlite

Fig. 6. HydroDesktop search tab includes options for specifying search location, keyword, date range, and web services (optional).



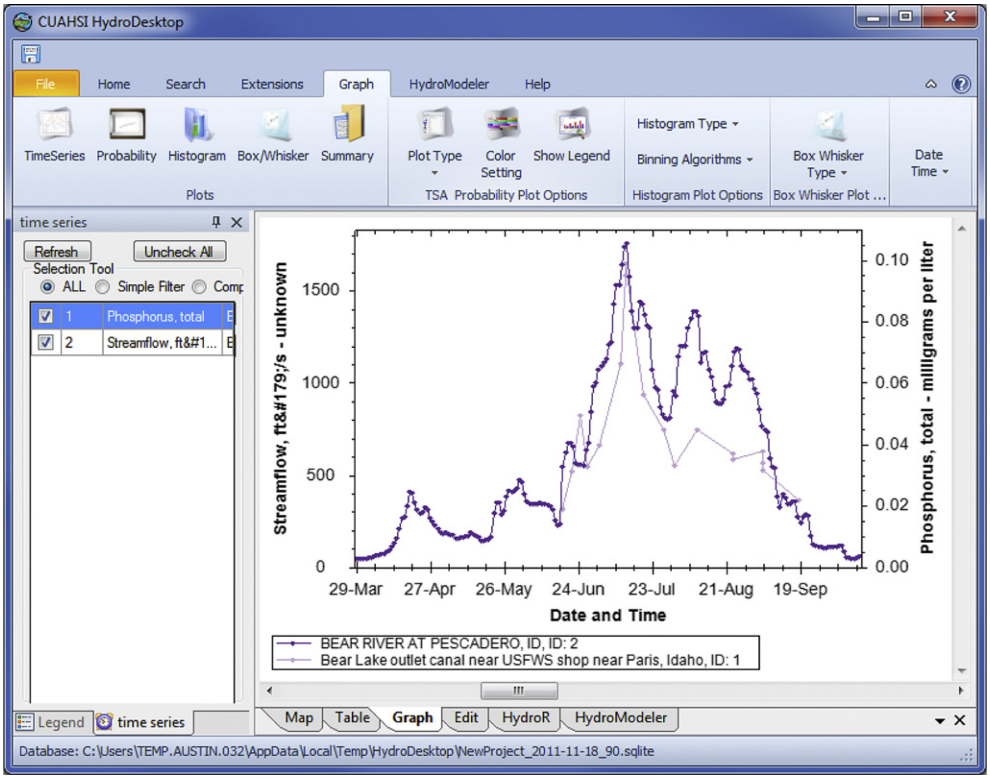
Streamflow

Start 11/18/2006

End 11/18/2011

Database: C:\Users\TEMP.AUSTIN.032\AppData\Local\Temp\HydroDesktop\NewProject_2011-11-18_90.sqlite

Fig. 7. Locations of streamflow time series matching the search parameters.



● ALL ○ Simple Filter ○ Comp

☑	1	Phosphorus, total	E
☑	2	Streamflow, ft&#1...	E

Database: C:\Users\TEMP.AUSTIN.032\AppData\Local\Temp\HydroDesktop\NewProject_2011-11-18_90.sqlite

Fig. 8. Graph of streamflow and phosphorus concentration within the HydroDesktop user interface.

interactive GIS map using the open source DotSpatial GIS components which are based on the MapWindow GIS system (Ames et al., 2008) and 3rd party DotSpatial plug-ins. DotSpatial supports a variety of vector, raster, and image GIS data types, and includes functionality for navigating the map as well as many other GIS tools and features. The HydroDesktop interactive map is used for displaying and manipulating spatial datasets as well as for setting the context for data discovery. As described in the sections above, an area of interest is often used as a spatial filter for narrowing a search for data. The HydroDesktop interactive map enables the user to set the geographic context for data discovery and access by enabling users to draw a bounding box or select a polygon feature from one of the GIS layers in the map (e.g., state boundaries, watershed boundaries, etc.) within which they would like to conduct their search.

Once time series of observational data have been retrieved and stored in the desktop data repository database, HydroDesktop provides users with tools for visualizing and analyzing the data. HydroDesktop maintains a GIS data layer showing the locations of the sites for which data have been downloaded to the desktop data repository database. This layer is dynamically built from the data repository database each time data are downloaded. Visualization of observational data is provided through a variety of plots using the open source Zed Graph plotting package and is focused on exploratory data analysis for data series that are downloaded and stored in the HydroDesktop data repository. Plot types available for visualizing time series data at a selected site include time series, histogram, box-and-whisker, and probability plots. The HydroDesktop time series visualization and analysis tool also enables users to view a selected time series in a simple tabular view and calculates simple descriptive statistics (minimum, maximum, mean, percentile values, etc.) for the selected time series.

Additionally, HydroDesktop includes an R statistics plug-in that supports manipulation and transformation of data, statistical analysis, and modeling using data from the HydroDesktop database. A data export plug-in allows users to export selected observational data from the local database to a delimited text file.

5. A case study for using HydroDesktop

In this section, a brief case study describing use of the HydroDesktop software for accessing and visualizing water quantity and quality data from the Bear River Watershed in Idaho, USA is presented. This case study presumes that the user has installed version 1.3 or later of HydroDesktop. In this case study, the goal is to identify patterns in phosphorus loading into the Bear River from Mud Lake. Data pertaining to phosphorus concentrations and streamflow will need to be obtained. The acquisition of these data using HydroDesktop is given emphasis in this case study.

5.1. Searching for hydrologic data

When HydroDesktop opens, the user is presented with a map showing political boundaries included as a GIS dataset in HydroDesktop. The user can augment this map by displaying a variety of online basemaps from ESRI, Google, Bing, OpenStreetMap, and more. By manipulating and navigating the map, the user sets the spatial context for the study (Fig. 5).

Once zoomed in to the portion of the Bear River near Mud Lake in Idaho, USA, the user activates the Search tab to specify filters for the data search. In this case, the user draws a box (shown in red in Fig. 6) around the area of interest, specifies a keyword such as “streamflow,” and enters a date range in order to constrain the search. The user could further filter the search to only include specific data sources of interest, but in this case will leave the default of search across all WaterOneFlow web services registered at HIS Central.

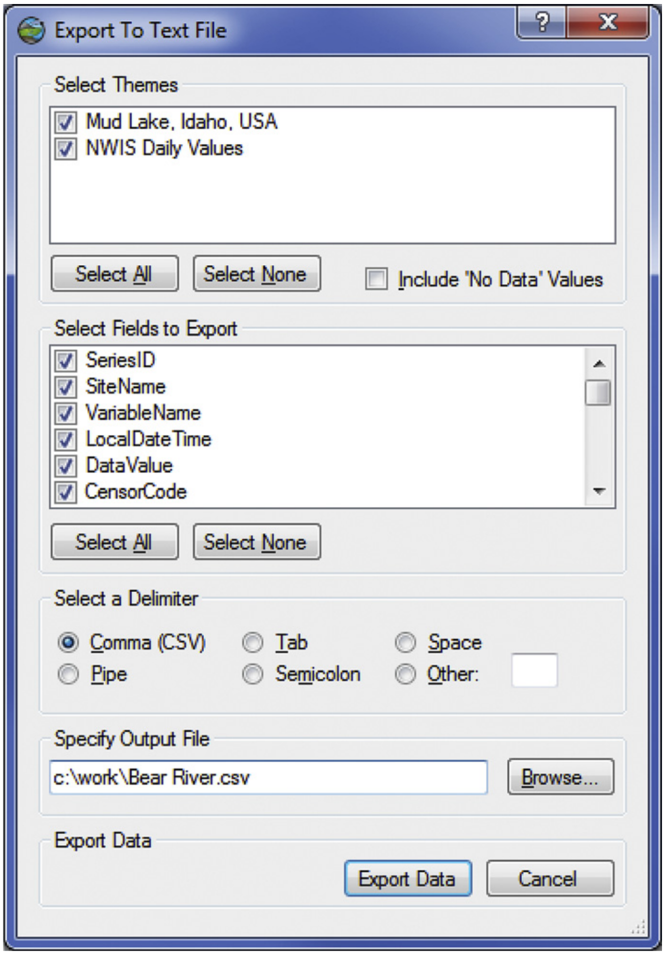
With the search parameters set, the user clicks Run Search, which triggers a query to HIS Central for metadata about time series that match the search criteria. HydroDesktop presents the search results to the user as points on the map where each point is symbolized by data source. Larger points indicate a greater number of time series values available at that location compared to locations with smaller points (Fig. 7). The user can hover the mouse over a point to view metadata about the time series available at the point or open a table to view metadata about the time series at all locations.

5.2. Downloading time series

In this case study, the user selects a site with data from the USGS National Water Information System (NWIS) and clicks to download the time series. HydroDesktop identifies the URL to data source from the metadata and makes a connection to that data source’s WaterOneFlow web service to download the data, saving the result to the local database. A second search is carried out for total phosphorus concentrations, and a time series at the outlet of Mud Lake is selected and downloaded.

5.3. Visualizing time series data

With the data retrieved, the user clicks the Graph tab and plots the streamflow and phosphorus concentration time series on the same graph. After zooming in on the graph, the user notices that



☑ Mud Lake, Idaho, USA
☑ NWIS Daily Values

□ Include 'No Data' Values

☑ SeriesID
☑ SiteName
☑ VariableName
☑ LocalDateTime
☑ DataValue
☑ CensorCode

● Comma (CSV) ○ Tab ○ Space
○ Pipe ○ Semicolon ○ Other:

c:\work\Bear River.csv

Fig. 9. Options for exporting data to a text file.

phosphorus concentrations increase with increased streamflow during the summer (Fig. 8), indicating a correlation between phosphorus concentration and streamflow. At this point, the user could use the HydroR plug-in, which provides integration with the R statistical computing software (http://www.r-project.org/), to perform more in-depth analyses with the data. A growing number of open source hydrologic modeling and time series analysis tools have been developed using R (e.g. Andrews et al., 2011), and integration with R greatly extends the capabilities and utility of HydroDesktop.

5.4. Exporting time series data

To export the data, the user clicks the Table tab and then clicks the Export button on the ribbon. In the “Export To Text File” dialog, the user specifies which data series to export, a delimiter, and an output file location (Fig. 9). The user then clicks Export Data, and HydroDesktop saves the data to a delimited text file. The user can then import the data into other programs that can read text files.

6. Discussion and conclusions

The main contributions of this work are: (i) HydroDesktop provides free access to data from distributed data services that are part of the CUAHSI HIS Internet-based, service-oriented architecture (SOA) and its 23 million data series; (ii) the HydroDesktop software interface enables end users that include university faculty, graduate and undergraduate students, K-12 students, engineering and scientific consultants, and others to operate within a relatively uncomplicated software environment; (iii) as an open source, free software application, HydroDesktop does not require use of commercial, third party software beyond the operating system and hence is expected to facilitate growth of a community of users and developers who can maintain and enhance the software. An on-going usability study focused on improving HydroDesktop and demonstrating/quantifying its efficiencies and performance over legacy methods is also underway, and results will be published.

While the core HydroDesktop software is complete and available for use (over 33,000 downloads as of March 2012), new plug-ins and extended capabilities are under active development at http://hydrodesktop.codeplex.com/. Here project participants, both from the CUAHSI HIS team and volunteers from the hydrologic sciences community share a discussion forum, bug tracking system, documentation WIKI, and an open Mercurial code-sharing repository. User support and documentation for HydroDesktop is provided informally by the open source and volunteer development community at the project website (including step-by-step tutorials) as well as formally through a series of workshops, webinars, and outreach activities sponsored by CUAHSI (see http://his.cuahsi.org) and through the detailed help system included with the software.

Any interested parties are invited to visit the project website, download the source code and join in the development and testing activities related to this project. It is expected that the simple plug-in architecture will encourage and facilitate third party development of plug-ins that significantly extend the base HydroDesktop application, making full use of all of the data retrieval and storage mechanisms in the initial version of HydroDesktop.

Specific future development plans for HydroDesktop include: support for new data sources and formats (including the OGC WaterML 2.0 standard); entry and upload of data into a HydroServer via HydroDesktop (e.g. for data collection purposes); ability to find and view metadata for datasets with limited access rights; a number of geospatial data analysis tools provided through the DotSpatial toolbox (e.g. geostatistical interpolation, clipping); and new time series management tools (e.g. unit conversion).

Acknowledgments

The software described in this paper was developed as part of the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) Hydrologic Information System (HIS) project whose goal is to advance information system technology for hydrologic science. This work was supported by the National Science Foundation under grant EAR 0622374 and under the Idaho NSF EPSCoR grant EPS 0814387. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We also acknowledge the extensive programming, testing, and documentation work of the entire CUAHSI HIS team, without whose help this project could not have been completed.

References

Ames, D.P., Michaelis, C.D., Anselmo, A., Chen, L., Dunsford, H., 2008. MapWindow GIS. In: Shekhar, S., Xiong, H. (Eds.). Encyclopedia of GIS. Springer, New York, pp. 633–634.

Ames, D.P., Horsburgh, J.S., Goodall, J., Tarboton, D.G., Whiteaker, T., Maidment, D.R., 2009. . Introducing the open source CUAHSI hydrologic information system desktop application (HIS Desktop). In: AnderssenR.S., R.D.Braddock, L.T.H. Newham (Eds.), 18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand and International Association for Mathematics and Computers in Simulation, July 2009, ISBN 978-0-9758400-7-8, pp. 4353–4359. http://www.mssanz.org.au/modsim09/J4/ames.pdf.

Andrews, F.T., Croke, B.F.W., Jakeman, A.J., 2011. An open software environment for hydrological model assessment and development. Environmental Modelling & Software 26 (10), 1171–1185. doi:10.1016/j.envsoft.2011.04.006.

Beran, B., Piasecki, M., 2009. Engineering new paths to water data. Computers and Geosciences 35, 753–760. doi:10.1016/j.cageo.2008.02.017.

Beran, B., Valentine, D., Zaslavsky, I., Cox, S.J.D., McGee, J., 2009. Web services solutions for hydrologic data access and cross-domain interoperability. International Journal on Advances in Intelligent Systems 2 (2 & 3), 317–324.

Castronova, A.M., Goodall, J.L., July 2010. A generic approach for developing process-level hydrologic modeling components. Environmental Modelling & Software. ISSN: 1364-8152 vol. 25 (7). ISSN: 1364-8152, 819–825. doi:10.1016/j.envsoft.2010.01.003.

Castronova, A.M., Goodall, J.L., Ercan, M.B., 2012. Integrated modeling within a hydrologic information system: an OpenMI based approach. Environmental Modelling & software. Available Online 9 March 2012, doi:10.1016/j.envsoft.2012.02.011.

Chave, A.D., Arrott, M., Farcas, C., Farcas, E., Krueger, I., Meisinger, M., Orcutt, J.A., Vernon, F.L., Peach, C., Schofield, O., Kleinert, J.E., 2009. . Cyberinfrastructure for the US Ocean observatories Initiative: enabling interactive observatories in the ocean. In: Proceedings of the OCEANS '09 IEEE Conference, Bremen, Germany. IEEE Ocean Engineering Society, pp. 1–10. doi:10.1109/OCEANSE.2009.5278134.

Droegemeier, K.K., Chandrasekar, V., Clark, R., Gannon, D., Graves, S., Joseph, E., Ramamurthy, M., Wilhelmson, R., Brewster, K., Domenico, B., Leyton, T., Morris, V., Murray, D., Plale, B., Ramachandran, R., Reed, D., Rushing, J., Weber, D., Wilson, A., Xue, M., Yalda, S., 2004. Linked Environments for Atmospheric Discovery (LEAD): a cyberInfrastructure for mesoscale meteorology research and education. In: Proceedings of the 20th Conference on Interactive Information Processing Systems for Meteorology, Oceanography, and Hydrology, American Meteorological Society, Seattle, Washington. Amer. Meteor. Soc., CD-ROM, S6.1.

Feng, M., Liu, S., Euliss Jr., N.H., Young, C., Mushet, D.M., 2011. Prototyping an online wetland ecosystem services model using open model sharing standards. Environmental Modelling & Software 26 (4), 458–468. doi:10.1016/j.envsoft.2010.10.008.

Goodall, J.L., Horsburgh, J.S., Whiteaker, T.L., Maidment, D.R., Zaslavsky, I., 2008. A first approach to web services for the National Water Information System. Environmental Modelling & Software 23 (4), 404–411. doi:10.1016/j.envsoft.2007.01.005.

Goodall, J.L., Robinson, B.F., Castronova, A.M., 2010. Modeling water resource systems using a service-oriented computing paradigm. Environmental Modelling & Software 26 (5), 573–582. doi:10.1016/j.envsoft.2010.11.013.

Granell, C., Diaz, L., Gould, M., 2009. Service-oriented applications for environmental models: reusable geospatial services. Environmental Modelling & Software 25 (2), 182–198. doi:10.1016/j.envsoft.2009.08.005.

Horsburgh, J.S., Tarboton, D.G., Maidment, D.R., Zaslavsky, I., 2008. A relational model for environmental and water resources data. Water Resources Research 44, W05406. doi:10.1029/2007WR006392.

Horsburgh, J.S., Tarboton, D.G., Piasecki, M., Maidment, D.R., Zaslavsky, I., Valentine, D., Whitenack, T., 2009. An integrated system for publishing environmental observations data. Environmental Modelling & Software 24 (8), 879–888. doi:10.1016/j.envsoft.2009.01.002.

Horsburgh, J.S., Tarboton, D.G., Schreuders, K.A.T., Maidment, D.R., Zaslavsky, I., Valentine, D., 2010. Hydroserver: A platform for publishing space-time hydrologic datasets. In: Proceedings 2010 American Water Resources Association Spring Specialty Conference Geographic Information Systems (GIS) and Water Resources VI, Orlando, Florida. American Water Resources Association, Middleburg, Virginia, TPS-10-1. http://www.awra.org/orlando2010/doc/abs/JefferyHorsburgh_7cb420e3_6602.pdf (last accessed 16.11.10.).

Huhns, M., Singh, M., 2005. Service-Oriented computing: key concepts and Principles. IEEE Internet Computing 9 (1), 75–81.

Josuttis, N.M., 2007. SOA in Practice: The Art of Distributed System Design. O'Reilly Press, Sebastapol, CA. 324 p.

Kadlec, J., Ames, D.P., 2011. Design and development of web services for accessing free hydrological data from the Czech Republic. In: Hrebicek, Jiri, Schimak, Gerald, Denzer, Ralf (Eds.), Environmental Software Systems. Frameworks of Environment. Proceedings of the 9th IFIP WG 5.11 International Symposium, ISESS 2011, Brno, Czech Republic, June 27–29, 2011. Springer, Boston, pp. 581–588. doi:10.1007/978-3-642-22285-6_63.

Madin, J., Bowers, S., Schildhauer, M., Krivov, S., Pennington, D., Villa, F., 2007. An ontology for describing and synthesizing ecological observation data. Ecological Informatics 2 (3), 279–296. doi:10.1016/j.ecoinf.2007.05.004.

Nambiar, U., Ludaescher, B., Lin, K., Baru, C., 2006. The GEON portal: accelerating knowledge discovery in the geosciences. In: WIDM '06 Proceedings of the 8th Annual ACM Workshop on Web Information and Data Management, Arlington, Virginia, USA. Association for Computing Machinery Press, pp. 83–90. doi:10.1145/1183550.1183567.

Piasecki, M., Beran, B., 2009. A semantic annotation tool for hydrologic sciences. Earth Science Informatics 2 (3), 157–168. doi:10.1007/s12145-009-0031-x.

W3C, 2004. Web Services Glossary, W3C Working Group Note. http://www.w3.org/TR/2004/NOTE-ws-gloss-20040211/ (last accessed 16.3.12.).

Zaslavsky, I., Valentine, D., Whiteaker, T. (Eds.), 2007. CUAHSI WaterML. OGC Discussion Paper OGC 07-041r1. Version 0.3.0. http://portal.opengeospatial.org/files/?artifact_id=21743 (last accessed 16.11.10.).