

Progress on water data integration and distribution: a summary of select US Geological Survey data systems

David Blodgett, Jessica Lucido and James Kreft

ABSTRACT

Critical water-resources issues ranging from flood response to water scarcity make access to integrated water information, services, tools, and models essential. Since 1995 when the first water data web pages went online, the US Geological Survey has been at the forefront of water data distribution and integration. Today, real-time and historical streamflow observations are available via web pages and a variety of web service interfaces. The Survey has built partnerships with Federal and State agencies to integrate hydrologic data providing continuous observations of surface and groundwater, temporally discrete water-quality data, groundwater well logs, aquatic biology data, water availability and use information, and tools to help characterize the landscape for modeling. In this paper, we summarize the status and design patterns implemented for selected data systems. We describe how these systems contribute to a US Federal Open Water Data Initiative and present some gaps and lessons learned that apply to global hydroinformatics data infrastructure.

Key words | data integration, groundwater, surface water, water budget, water quality, web services

David Blodgett (corresponding author)

Jessica Lucido

James Kreft

USGS Center for Integrated Data Analytics,

8505 Research Way,

Middleton,

Wisconsin 53562,

USA

E-mail: dblodgett@usgs.gov

INTRODUCTION

In recent years, the idea of an open water data infrastructure has surfaced as an authoritative set of water information assembled from best available sources and made freely available through the internet. The idea builds on the US National Spatial Data Infrastructure (NSDI), which has been evolving for more than two decades (Federal Geographic Data Committee 1994, 2013). According to the US Office of Management and Budget, 'The NSDI facilitates efficient collection, sharing, and dissemination of spatial data among all levels of government institutions, as well as the public and private sectors, to address issues affecting the Nation's physical, economic, and social well-being' (Office of Management & Budget 2002). The hydrologic community has extended and focused the NSDI concept to include observational and spatial data to address water-resources issues. In early 2014, a charge was put forward proposing 'a new Open Water Data Initiative that will integrate currently fragmented water information into a connected, national water data framework and leverage

existing systems, infrastructure and tools to underpin innovation, modeling, data sharing, and solution development' (Castle *et al.* 2014). This charge provided an opportunity for broad coordination of ongoing activities and an opportunity to identify and implement missing components of water-data infrastructure.

The US Geological Survey (USGS) maintains the largest network of surface-water (Norris 2010) and groundwater (Dennehy 2005) monitoring stations in the United States. The USGS also contributes to national data network projects focused on surface water, water quality, and groundwater; and has played a leading role in implementation and operations of their computing infrastructure. This paper reviews technical challenges and lessons learned through implementation and involvement with several national hydroinformatics projects. It is not meant to be an exhaustive overview of activities in this field. Systems and projects summarized were chosen because of their national scope, broad coverage of water-resources

disciplines, and familiarity to the authors. Hydrographic data such as stream networks are not discussed in this paper. We recognize that many of the data sources presented here can be integrated using streams and watersheds to put observations and model estimates into their hydrographic context. Discussion of these concerns is beyond the scope of this paper, but they are a critical component of an open water-data infrastructure. The concepts presented here outline progress made to date on the selected USGS projects and provide ideas for how others can be informed collaborators in a global hydroinformatics data infrastructure.

WATER DATA SYSTEM SUMMARIES

While the emphasis of this paper is on water-resources disciplines, such as surface water and groundwater, issues relating to the basic structure and metadata of continuous time-series data, temporally discrete sample data, and geospatial landscape coverage data are also discussed. Two technical strategies for aggregation of water data are presented in the water quality and groundwater sections. Efforts to integrate landscape data for the purpose of hydrologic and water-quality modeling in order to assess national water availability and water quality are also shared.

Each of these systems contributes to the Open Water Data Initiative (OWDI). Following is a summary of each system, the ways in which it contributes to the OWDI, and enhancements that would help to improve it. The principles of the OWDI have been summarized as: (1) the information owner is responsible for and maintains control of data; (2) data are available in common and/or standard formats requiring no license for access; (3) machine interfaces, typically web services, are generalized according to a standard where possible; and (4) data use machine-interpretable documentation for things like controlled vocabularies and methods. The OWDI discussion of each system is framed using these four principles. Greater technical detail is presented where it is needed, e.g. web service and data-format standards. In other cases, like system architecture and common vocabularies, less technical detail is needed.

NATIONAL WATER INFORMATION SYSTEM

The USGS National Water Information System (NWIS) (<http://waterdata.usgs.gov/nwis>) is an enterprise water data management system that is the USGS Water Mission Area's primary offering to the OWDI. Preceded by the USGS Water Data Storage and Retrieval System (STORET) (Hutchison 1975), the use of NWIS became a core operational capability in the 1980s. Since the data stored in the system were first made available on the Internet in 1995, the NWIS web interface has grown to provide tens of millions of web-page views and web-service requests per month. The diversity of human- and machine-oriented service offerings from NWIS has grown over that time to include numerous variables for multiple time resolutions at thousands of monitoring sites (Bales 2014; Hirsch & Fisher 2014). NWIS web services have been integrated into the National Groundwater Monitoring Network (NGWMN) (<http://cida.usgs.gov/ngwmn>) and Water Quality Portal (WQP) (<http://waterqualitydata.us>), presented below, and other water-data networks like the Consortium of Universities for the Advancement of Hydrologic Sciences, Inc. (CUAHSI) Hydrologic Information System Catalog (<http://hiscentral.cuahsi.org/>) and the international Global Earth Observations System of Systems (GEOSS) (www.geoportal.org/).

Until recently, NWIS has been a collection of regional databases such that national aggregation has been costly. Coordination across these databases has been accomplished with shared software and limited content-sharing between databases. This arrangement is well suited to field data collection around the country. However, national USGS systems have had to maintain copies of the list of sites and particular data in all the regional databases to accomplish their goals. The NWIS web system aggregates the publicly available real-time and archived data from all the regional databases. Another example requiring aggregation of the regional databases is a data system to store and disseminate evaluations of the nation's water quality. In this case, a national subset of NWIS data that passed quality checks and matched characteristics needed for particular evaluations was assembled. See Figure 1 for an architecture diagram summarizing the current NWIS aggregation. While necessary because the databases are decentralized

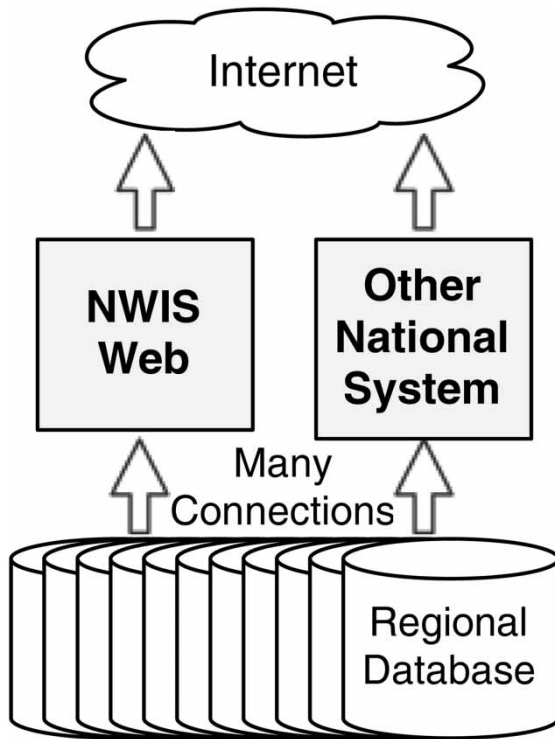


Figure 1 | Architecture diagram of NWIS distributed database aggregation.

and contain differing content, this requirement has had significant operational costs for these and other projects.

In recent years, a project to aggregate the regional databases to a single national database has been undertaken. Since the regional databases share database schema and software to update them, aggregating them is possible. However, because hydrologic boundaries do not follow the State boundaries that separate the USGS Water Science Centers, many databases have sites that are duplicated from one state to the next with conflicting information. It has been difficult to identify the canonical site where duplicates are found and to mediate changes to the source databases. The migration to a single, national database is a necessary step toward providing public service access or national synthesis capabilities into the future.

An important aspect of NWIS's public offerings to the OWDI is the adoption of standardized definitions of data elements and the format used to distribute data. In 2007, the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) presented a discussion paper to the Open Geospatial Consortium (OGC) Hydrology Domain Working Group entitled CUAHSI WaterML

(Water Markup Language) (Zaslavsky *et al.* 2007). It set the stage for development of an international standard for content and encoding of hydrologic time-series data. A project to harmonize various implementations of time-series data transport formats was published in 2010 (Taylor *et al.* 2010). It was followed by interoperability experiments (Broderic & Booth 2011; Fitch 2013) to test the in-development hydrologic time-series standard. The USGS took part in both of these experiments; contributing content from the NWIS database and implementing services to test the draft data-exchange formats. The result of this work was published as OGC WaterML 2.0 Part 1 – Timeseries in 2012 (Taylor 2012).

The WaterML 2.0 Part 1 standard focuses on the fundamental structure and definition of time-series data associated with a monitoring location such as a stream gage. It defines the basic relationship between a site at which observations are being made, the organization responsible for the site and observations, the water body being monitored, the phenomenon or variable being observed, and the method being used to observe the phenomenon. At the heart of the standard is a comprehensive time-series data model, accounting for particulars of interpolation between time steps, per time step changes in metadata, and practical implementability of the standard. In 2014, members of the OGC Meteorology and Oceanography Domain Working Group proposed a rebranding of the WaterML 2.0 Part 1 – Timeseries standard to TimeseriesML. The change will be primarily semantic, preserving the structure and breadth of WaterML 2.0 (Tomkins 2014). This will greatly enhance the potential for cross-discipline use of time-series data.

NWIS web services are provided free of charge or restriction on data use. Data formats provided via machine-readable web services conform to common information standards in the cases where they exist. NWIS services providing time-series data support a basic delimited text format, the WaterML 1.1 format (Valentine & Zaslavsky 2009), and the WaterML 2.0 Part 1 – time-series format. The OGC Hydrology Domain Working Group has vetted and published a best practice for implementing the Sensor Observation Service standard for the hydrology domain (Andres *et al.* 2014). NWIS took part in development of the best practice and is implementing the capability at the

time of writing. The USGS has also taken part in development of WaterML2.0 Part 2 Ratings and Gagings (Taylor 2013), implementing experimental versions of the standard and providing feedback to the standardization process.

While NWIS service offerings provide great access to the data in NWIS, the data use proprietary code lists and vocabularies. Machine-interpretable metadata are available for some of these such as site details, parameter codes, and time-series summary statistics, but a comprehensive service to allow machine interpretation of all metadata elements for data in NWIS has yet to be designed and developed. Publication of these metadata as services that relate NWIS-specific code lists to common hydrology domain vocabularies would be of great value.

NATIONAL GROUNDWATER MONITORING NETWORK

In addition to the management of surface-water sites and observation data, NWIS contains information about groundwater wells and springs. The NWIS Ground-Water Site Inventory system contains information about well and spring box construction, well logs, well discharge, hydrogeology, aquifers, aquifer tests and groundwater levels (US Geological Survey 2012). Most of these data are stored in the NWIS central database(s) and served to the public through the NWIS web interface (Hirsch & Fisher 2014). Although this data management and dissemination model is adequate for the present needs of USGS scientists and its cooperators, groundwater monitoring networks are operated by numerous federal, state, tribal, and local agencies (Advisory Committee on Water Information| (ACWI) – Subcommittee on Groundwater (SoGW) 2013). A National Ground-Water Monitoring Network (NGWMN) was initiated in 2007 to bring together disparate groundwater-level observations, water-quality samples, lithology and construction information from contributing agencies, including the USGS, to attain a nationally consistent picture of the nation's groundwater.

A publicly accessible data portal was built to serve groundwater data from distributed national, state and local databases through a map-based graphical user interface. Additionally, the Network's data are provided via web

services as well as the NGWMN Data Portal. An Open Geospatial Consortium (OGC) – Sensor Observation Service (OGC-SOS) (Bröring 2014) that serves groundwater levels in OGC – WaterML2 (Taylor 2012) and an OGC – Web Feature Service (Vretanos 2005) that serves well-characterized data in GroundWater Markup Language (Boisvert & Brodaric 2006) were implemented to provide well-documented data and achieve interoperability with the international water-resources community.

Given the large number of agencies that conduct groundwater-monitoring efforts and manage these data locally, that the data is small enough to be rapidly transferable over the internet (usually less than 1 MB per site), and the requirement for contributing agencies to maintain ownership and control over their data sets, a distributed web service-oriented architecture was implemented. The architecture is composed of four discrete components: the 'hub' consists of the web application, the data mediator, and a registry of basic site information. Web services from data providers form the 'spokes'. Users request data for sites in the registry displayed in the web application; the mediator aggregates and transforms the data sets from cooperating agencies' web services into a standard format and returns the aggregated results to the user. This design, shown graphically in Figure 2, allows data providers to retain ownership and control over their data and removes the need for data providers to standardize their output, providing flexibility and a lower barrier for participation. This differs from NWIS, where data storage has also been distributed across a number of databases, however these databases share a common structure and can be aggregated directly. The adherence to a common database model is possible because all of the data belong to a single organization with shared management, something not possible with the NGWMN.

Distributed service-oriented architecture fits the OWDI principle of maintaining data-providers' ownership of their data, but heavy reliance on the Internet when data must be retrieved from providers is a weakness. A centralized system makes many fewer external requests and is much less prone to errors caused by Internet issues. Outages of collaborators' web services can also cause issues with reliability and stability of the system. Complications such as changes in the format of data returned from network

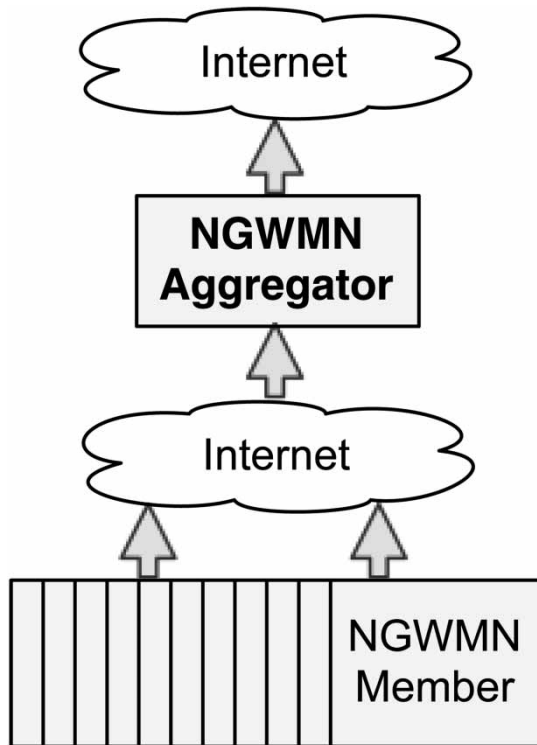


Figure 2 | Diagram of NGWMN web service based architecture.

member services can cause a breakdown in the dynamic mediation and aggregation. To address such challenges, a web service cache was implemented. It stores web service responses from collaborator services to elevate the reliability and stability of the system. The cache is updated frequently in order to keep data current. General changes to data provider web services and network or data center outages remain as significant challenges, although their impact is diminished since previously fetched data can be served in the absence of normal web service function. This pattern has proven effective for integration of NGWMN data and would be suited to other similar systems.

The NGWMN design is consistent with principles of the OWDI but there are a number of enhancements that would improve its offerings. It provides data via services and has adopted community standards for requesting and transmitting data. It also integrates data while keeping data providers in control of their data. One necessary improvement to the NGWMN service holdings is the adoption of the OGC-SOS hydrology profile, mentioned above. The existing NGWMN OGC-SOS does not implement all of

the data discovery and metadata providing functions that are part of the hydrology profile. Another expected enhancement to the NGWMN is implementation of a new version of GroundWater Markup Language that has been harmonized and enhanced to include concepts from the European INSPIRE directive. The new standard will include more complete descriptions of well lithology and construction. Groundwater data systems in the USA, including the NGWMN, have generally not incorporated common vocabularies, which are of particular value for lithology information. A significant enhancement to the NGWMN would be incorporation of geologic and lithologic vocabularies. A further improvement to the Network would be generation of metadata for each dataset in the Network documenting the provenance, data processing, collection methods, and aspects of data quality in a metadata format conducive to cataloging, discovery, and understanding a dataset's contents.

WATER QUALITY PORTAL

In addition to surface and groundwater time-series data, NWIS is also the repository for nearly 90 million temporally discrete water-quality measurements collected by USGS scientists starting in the early 1900s. While the data have been available through NWIS web interface web pages since the late 1990s, there was no web-service access for discrete water-quality measurements. Similar to groundwater data, the USGS is not the only collector of water-quality data. States, tribes, municipalities, consultants, and others have been collecting and submitting data to the US Environmental Protection Agency's (EPA) STORage and RETrieval (STORET) (www.epa.gov/storet/) database for decades. In 2004, leaders from the EPA and USGS signed a memorandum of understanding to create a tool to integrate and serve water-quality data from both agencies under the aegis of the National Water Quality Monitoring Council.

This shared tool required a common vocabulary to describe water-quality samples. Coincident with planning of the integrated water quality access tool, the EPA supported development of a Water Quality Exchange (WQX) through the Environmental Information Exchange Network

(www.exchangenetwork.net/data-exchange/wqx/), an EPA-affiliated organization that facilitates data transfer and sharing to provide better access to high-quality environmental data of all types. Part of the development of the Water Quality Exchange was the collaborative development of a standard data format for submission of water-quality sampling data known as WQX. While the WQX standard was initially designed to facilitate data submission to STORET by organizations collecting water-quality data through EPA-provided grants, it was recognized that the data format was well suited to form the foundation of a standard to integrate data from EPA STORET and USGS NWIS systems.

With a data format standard in place, a collaborative effort between the USGS and EPA was initiated to establish shared query parameters, or domain values. For example, this shared domain value vocabulary allows a search for 'stream' sites to also include 'river', 'creek', and 'river/stream' sites in the other systems. This serves to enhance a user's ability to find potentially similar data and should be a capability across systems under the OWDI. The USGS-EPA collaboration also led to an expansion of WQX into what is now called WQX-Outbound to support additional values important to successfully mapping NWIS data to WQX.

As a first step toward web-service integration of USGS and EPA data, two 'mini-portals' were developed that provided access to STORET and NWIS data through separate services with common query interfaces and output formats. In April 2012, the Water Quality Portal (WQP) was launched; delivering a single interface that provides both a common user interface and a single web service interface that combined the data available from the two 'mini-portals' in one service endpoint. Although the WQP is now able to serve data from multiple databases in a unique common format, the WQP system is not the primary system of record for these datasets, instead relying on the underlying systems to both maintain data quality and integrity over time.

When introduced, the WQP could provide a unified response with content from the STORET and NWIS water-quality databases. The WQP's architecture is similar to the NGWMN in that multiple providers send data to an aggregator. It is also different from NGWMN in that data

providers present a common protocol and data-transfer format to the aggregator, whose job is only to combine structurally consistent data while mediating some data values to provide consistent semantics in the output. This aggregator implementation was able to respond relatively quickly and stream the combined query results from database to saved file with good scalability. In 2014, the US Department of Agriculture's Sustaining the Earth's Watersheds Agricultural Research Data System data were added to the WQP.

The introduction of a third data system with the accompanying unique data semantics combined with a massive increase in data use exposed multiple weaknesses in the aggregation-focused architecture of the WQP. While the idea of maintaining multiple independent database and web-service instances was appealing from a theoretical perspective, where additional standards-compliant services could be simply 'bolted on' to the existing aggregator, operational challenges have been significant. The Water Quality Portal's use has grown to satisfy tens of thousands of web-service requests serving hundreds of millions of rows of water-quality sample data per week. The holdings have also grown from 150 million to over 250 million records in three years. With this growth, the distributed architecture and aggregation concept has not been able to keep up with demand as well as could be hoped. A major re-design of the portal, focused on a single, high-performance data warehouse, is currently underway. See [Figure 3](#) for a graphical description. Although the data are now in a single data store, the data are explicitly linked to the data system of origin. The WQP system is effectively a cache, refreshed on a regular basis. This work will allow the WQP to serve massive aggregated datasets from many data providers based on custom queries in a way that satisfies user demand into the future.

Similar to the NGWMN, the WQP is consistent with principles of the OWDI in that it provides free machine-interpretable data and metadata services. As described above, the development of the WQX water-quality data exchange format included definition of a vocabulary for common water-quality query parameters. Mediation of query and response values of these parameters is a significant value-added service that is a model for other systems to follow as the OWDI progresses. While this functionality exists and is useful for querying and using returned data

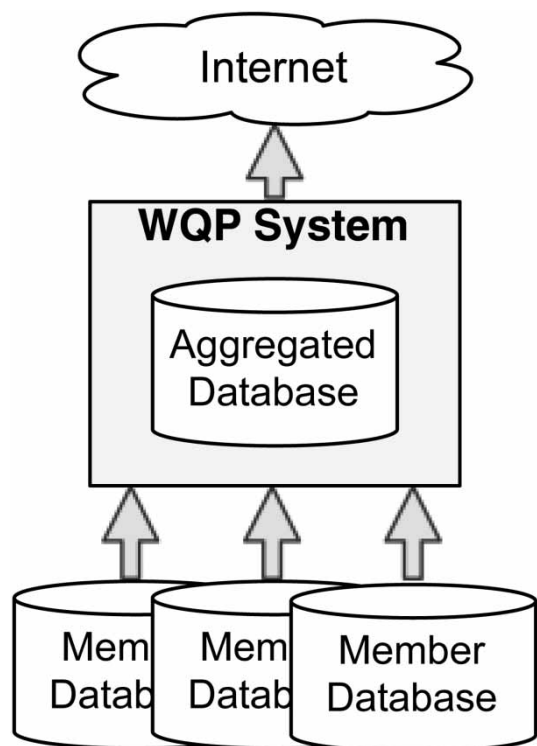


Figure 3 | Diagram of the WQP aggregated database architecture.

from multiple sources, the ontology used behind the service is not exposed in a way that external systems can build upon. Establishing data quality or comparability is also a significant issue when using data collected by multiple agencies for use in a wide variety of applications. While systems such as the National Environmental Methods Index (www.nemi.gov/home/), which is used by the WQP when possible, provide information to inform assertions about data quality and comparability, significant improvements are needed to ensure that all data in the portal are described fully in this regard.

WATER USE AND AVAILABILITY DATA SYSTEMS

Water use and availability data are handled differently from surface-water or groundwater data in that they are not typically direct observations. Rather, they are derived from demographic or economic information or provided by an organization such as a city, county, irrigation district, or industrial association. The USGS has, for many decades,

prepared a compilation of water use estimates by county every five years; the latest report was released in 2010 (Maupin *et al.* 2014). These reports contain trend information relating to previous compilations and up-to-date summaries for the compilation year. They represent the only source for nationally consistent water use information for setting policy and analyzing national water-use trends (National Research Council 2002). More of these summaries are available in report form from the USGS Publications Warehouse (<https://pubs.usgs.gov/>). The data in the reports are held in an NWIS dataset known as the Aggregate Water-Use Data System (AWUDS) and are made available to the OWDI through the NWIS web interface (<http://waterdata.usgs.gov/nwis/wu>) for particular counties, the National Water Use Program's web page (<http://water.usgs.gov/watuse>) for entire compilations, and the National Water Census (NWC) data portal (<http://cida.usgs.gov/nwc/>) in graphical form and via web services.

The data published with these compilation reports are also distinct from data collected at river monitoring sites or groundwater wells in that they pertain to spatial polygon-reporting units. Since 1995, when watershed-based estimates were phased out, the reporting units have been counties. The compilation reports have used this information to report water-use status and trends and it has been published on the web in tabular format on the water-use program's web page. For a once-every-five-year compilation associated with a few thousand reporting units, this approach has been sufficient; however, with site-specific water-use estimates becoming a more common way of managing the information, new methods for data archiving and dissemination will be needed.

Site-specific water-use estimation is being introduced as the preferred method of archiving water-use information that is collected by the USGS (Alley *et al.* 2013). As part of an effort to build a national assessment of water availability and use, the USGS is implementing a shift from aggregated county and watershed water use estimates to more site-specific estimates. Withdrawal locations, conveyance connections to systems that use water, and any return flows will be catalogued to support future analysis on any given reporting area. This work has begun through an analysis to estimate water use at thermoelectric power plants (Diehl & Harris 2014), and will continue with an initial emphasis

on large public-supply water systems. The USGS has an ongoing partnership with the Western States Water Council to integrate with their Water Data Exchange project (*Western States Water Council 2014*). The project is at the forefront of standardizing and making the data more broadly available across numerous Western States stakeholders. This water-use data integration and sharing project is an example of an existing collaboration that is coming under the OWDI.

The USGS has begun creation of a system to automatically attribute best estimates of water availability and use to any unit as part of the National Water Census. Early work toward this goal has focused on observed and modeled precipitation and evapotranspiration with the intention of using site-specific water-use information as it becomes available from USGS and partner-developed systems. The work is being pursued using the water budget as a unifying theme for water availability estimation. Ongoing USGS research seeks to identify the best available sources of water budget terms and to develop infrastructure that will allow automatic attribution of water budget terms to any reporting unit (*Alley *et al.* 2013*). See the National Water Census web page (<http://water.usgs.gov/watercensus/>) for the latest research from the program.

Work to automatically summarize water-budget information has initially focused on precipitation and evapotranspiration data because it is already available as nationally consistent datasets. Precipitation summaries are derived from the DayMet dataset (*Thornton *et al.* 2012*) that interpolates and extrapolates rainfall observations to a regular grid in space and time. Similarly, evapotranspiration estimates are provided by the operational simplified surface-energy-balance model, which relies on remote sensing and ground-based observations to produce a national gridded data product (*Senay *et al.* 2013*). Data services for these two products in gridded format and attributed to watersheds are available through the NWC data portal and services. Tools described in the following section can be used to subset the data and summarize them to any polygon area of interest.

USGS water-use data offerings available for the OWDI are varied and could be improved under the OWDI. AWUDS data for counties and watersheds are made available free of charge via simple tabular data downloads

(<http://water.usgs.gov/watuse/data/index.html>). Water-use data summaries are available via NWIS web pages (<http://waterdata.usgs.gov/nwis/wu>) for states and counties. However, given that the data are not too large or varied, web services providing query capabilities or multiple formats have not been a priority as they have for time-series and other NWIS data. The NWC data portal recently implemented web services for county water use that it uses for plotting and displaying data on the Internet, but the data behind the service are not large or dynamic enough to require services for other uses. The polygons and non-water-use attributes of the counties and watersheds included in this dataset are not available and are an important addition to fully describe the water-use estimates. Water-use data attributed to sites or areas representative of the use has typically not been archived consistently.

An integrated facility for sharing water-use data would have many things in common with the NGWMN or WQP. Mappable point and polygon locations of water uses and water allocations would be aggregated with metadata to help discover and understand the available time-series or other summary data. As outlined above, the Western States Water Council Water Data Exchange, the USGS, and others are beginning to focus on such water-use data services. It is expected that the OWDI will result in significant progress on this front.

WATERSHED MODELING DATA SYSTEMS

To fully characterize and understand the past, present, and future of the nation's water resources, landscape and climate data such as land use, land cover, soil information, precipitation, temperature, evapotranspiration, and other spatially continuous data must be considered. These data are important for both water-quantity and water-quality characterization and modeling. Accessing and summarizing these data to hydrologic and other landscape models is often a considerable task that takes time away from model calibration and results interpretation. Two examples of tools developed at the USGS to bring automated data manipulation for landscape water-resources modeling to the Internet are presented in the following paragraphs. The Geo Data Portal (GDP) (<http://cida.usgs.gov/gdp/>) helps

modelers access data to build and run models. The SPATIally Referenced Regressions On Watersheds (SPARROW) Decision Support System (<http://cida.usgs.gov/sparrow/>) runs pre-existing models for scenario testing and decision support.

The GDP is a tool that provides per-watershed (or other spatial unit) summaries of gridded data using standard web services (Blodgett *et al.* 2011). Typically it is used to generate time series of precipitation and temperature in order to drive watershed models, but can be used to summarize a wide range of data. A catalog of datasets that follow particular data structure and metadata standards has been assembled for the tool's interface with references to data hosted by the Federal agencies, the US National Labs, and academia. Interoperability of the datasets in the catalog is made possible by the Open Source Project for a Network Data Access Protocol (OPeNDAP) web service interface and the Network Common Data Format Climate and Forecasting Conventions. While scalability and performance of this system of web service and data content standards has been found to be highly dependent on several implementation details, interoperability can be accomplished quite easily.

The GDP can access and process data from any publically available host via OPeNDAP. The catalog mentioned above is a convenient utility for users so they do not need to have knowledge of the data-access methods of the system. The GDP has been successful at brokering data access for users from multiple water-resources disciplines who need access to large gridded time-series data sources (Francy *et al.* 2013; LaFontaine *et al.* 2013; Daraio & Bales 2014; Duvenick *et al.* 2014; Read *et al.* 2014; Williamson *et al.* 2014; Creutzberg *et al.* 2015; Newman *et al.* 2015), but good performance has been difficult to achieve for all data-access patterns. For data that are large both spatially and temporally, users that want to access a very small area for all time steps experience poor performance with datasets that are designed to work well for accessing large spatial areas. This problem has proven very difficult to solve without storing two separate copies of a dataset, one for time-series access and another for map access. This approach is not always feasible given financial and resource limitations and the hundreds of gigabytes to terabyte size of many gridded time-series datasets.

The GDP processing service is a web application that has minimal external software dependencies. The primary

deployment of the software is at a USGS data center alongside a large USGS-maintained archive of compatible data. While this USGS-supported deployment is the primary access point for the GDP project, some projects have deployed parts of the system themselves and written their own user interfaces to execute the service interfaces. This model, deploying data subsetting and summarization software near the very large data that water-resources studies require is novel, but this system shows that it should be considered for OWDI and other similar data-integration activities.

SPARROW models link landscape characteristics to in-stream water-quality characteristics (Preston *et al.* 2009). When building a SPARROW model, numerous spatial coverages of landscape characteristics are compiled and regression equations that relate those characteristics to observed stream water quality are developed. Once the model has been developed, the calibrated regression equations can be used to perform landscape management scenario testing. Because the model is spatially explicit, landscape change scenarios can be applied to targeted areas or to entire watersheds. Effects of changes can be observed as they propagate downstream and mix with downstream tributaries.

USGS scientists and GIS specialists for particular projects develop SPARROW models, often as part of nationwide water-quality assessment. See the decision support system, linked above, for a list of available models.

The decision support system allows users to visualize and run spatially explicit scenarios by modifying landscape inputs to the archived models. To build the decision support system, the finished models' river network, catchments, and regression coefficients are first loaded into a database system. This database and associated services are used by a web user interface that provides a catalog of models, map views of the models' spatial units, and a variety of ways to summarize model results for current conditions or test-scenario conditions. See Booth *et al.* (2011) for technical details of the system. Without this model archiving and execution facility, accessing these model results and executing test scenarios would be a task limited to only the most capable and motivated users.

In order to provide maps of model results dynamically, regression equations are executed with changed parameter

values, results must be accumulated downstream, and calculated results need to be rendered into a map. The former two steps can be performed very rapidly within a database. The latter rendering step has been difficult to implement such that it performs well for many users at once. This type of processing may be common in OWDI systems and the lessons learned here can be used elsewhere. Dynamically rendering content that is unique to a particular user, while possible, requires dedicated computing capacity for every user, or a way to have users' requests queued and completed some time after a request is made.

The initial approach implemented for the system used a technology that performed map-rendering computations on the server for every request from any user. To perform well for multiple users, this required significant computing power. Subsequent work has focused on rendering and caching content that displays for default settings of all models in the system such that most users who use the system experience better performance and rendering computing resources are used only for users who request custom scenarios. This smart caching to provide uniform performance for default settings of an application with load-dependent performance of custom-computed content is increasingly being viewed as a best practice for applications that include large, computationally intensive data and processing, but also responsive visualizations.

Landscape characteristic and climate data should be available according to principles of the OWDI because of its importance to watershed modeling. Watershed modeling, focused on characterizing quantity and quality of all terms of a water budget, must rely on numerous physical characteristics of the landscape. These data come from a network of data providers, each with their particular specialty. The USGS National Geospatial Program provides elevation, a consortium of agencies compiles a national land-cover database, the US Department of Agriculture Natural Resources Conservation Service compiles national soils databases, and the National Oceanographic and Atmospheric Administration (NOAA) collects and distributes various precipitation data products. Remote sensing and land-surface reanalysis models, from the National Aeronautics and Space Administration and NOAA, are another very important source of information to build and drive watershed models. The GDP project has begun bringing together many of these data

sources and the SPARROW decision support system exposes model results that build off these and other data products of the OWDI.

CONCLUSIONS

In this paper, we have summarized selected recent experience with publishing national-scale hydrologic data and data services via the Internet at the USGS. The discussion spanned a wide range of water-resources disciplines, highlighting general hydroinformatics lessons learned, and detailed how each fulfills the principles of the US Open Water Data Initiative. The systems that integrate data from multiple partners for groundwater, water quality, and hydrologic model input data demonstrate how web services can be used to allow data providers to maintain control and be responsible for availability of their data. NWIS and the other systems discussed shared progress on infrastructure and system designs that provide web service access to their data holdings. Good progress has been made on common standardized formats for data with specific formats being used for all the systems discussed. Provision of machine-interpretable documentation such as controlled vocabularies for metadata, standardized methods for water quality, and data provenance are areas where we have noted gaps. Community consensus and facilities to host reference lists and documentation are needed to lay a foundation for systems like those described here to build upon.

REFERENCES

- Advisory Committee on Water Information (ACWI) – Subcommittee on Groundwater (SoGW) 2013 A National Framework for Ground-Water Monitoring in the United States. Available from: http://acwi.gov/sogw/ngwmn_framework_report_july2013.pdf.
- Alley, W. M., Evenson, E. J., Barber, N. L., Bruce, B. W., Dennehy, K. F., Freeman, M. C., Freeman, W. O., Fischer, J. M., Hughes, W. B., Kennen, J. G., Kiang, J. E., Maloney, K. O., Musgrove, M., Ralston, B., Tessler, S. & Verdin, J. P. 2013 Progress toward Establishing a National Assessment of Water Availability and Use. U.S. Geological Survey Circular 1384. Available from: <http://pubs.er.usgs.gov/publication/cir1384>.

- Andres, V., Jirka, S. & Utech, M. 2014 OGC Sensor Observation Service 2.0 Hydrology Profile. Available from: <http://docs.opengeospatial.org/bp/14-004r1/14-004r1.html>.
- Bales, J. D. 2014 Progress in data collection and dissemination in water resources, 1974–2014. *Water Resour. IMPACT* **14** (3), 18–23.
- Blodgett, D. L., Booth, N. L., Kunicki, T. C., Walker, J. I. & Viger, R. J. 2011 *Description and Testing of the Geo Data Portal: Data Integration Framework and Web Processing Services for Environmental Science Collaboration*. U.S. Geological Survey Open-File Report 2011–1157. US Geological Survey, Reston, VA, USA. Available from: <http://pubs.er.usgs.gov/publication/ofr20111157>.
- Boisvert, É. & Brodaric, B. 2006 *GroundWater Markup Language Specification v. 1.0*. Report. Natural Resources Canada, Ottawa, ON, Canada. Available from: http://ngwd-bdnes.cits.rncan.gc.ca/service/api_ngwds/en/gwml.html.
- Booth, N. L., Everman, E. J., Kuo, I., Sprague, L. & Murphy, L. 2011 A web-based decision support system for assessing regional water-quality conditions and management actions. *JAWRA J. Am. Water Resour. Assoc.* **47** (5), 1136–1150.
- Brodaric, B. & Booth, N. 2011 *OGC Groundwater Interoperability Experiment – Final Report*. Open Geospatial Consortium, Wayland, MA, USA. Available from: http://portal.opengeospatial.org/files/?artifact_id=43545&version=1.
- Bröring, A. 2014 Open Geospatial Consortium Sensor Observation Service Interface Standard OGC 12-006. Available from: https://portal.opengeospatial.org/files/?artifact_id=47599.
- Castle, A., Caramanian, L., Johnston, J., Werkheiser, B., Bales, J., Dalton, J., Maidment, D., Gallagher, K. & Deloatch, I. 2014 Open Water Data Initiative Proposal and Charge for Discussion. Available from: http://acwi.gov/acwi-minutes/acwi2014/Open_water_data_proposal_to_acwi-7-21-14.pdf.
- Creutzburg, M. K., Scheller, R. M., Lucash, M. S., Evers, L. B., LeDuc, S. D. & Johnson, M. G. 2015 Bioenergy harvest, climate change, and forest carbon in the Oregon coast range. *Glob. Change Biol. Bioenergy* DOI: 10.1111/gcbb.12255.
- Daraio, J. A. & Bales, J. D. 2014 Effects of land-use and climate change on stream temperature i: daily flow and stream temperature projections. *JAWRA J. Am. Water Resour. Assoc.* **50** (5), 1155–1176.
- Dennehy, K. F. 2005 Ground-Water Resources Program U.S. Geological Survey Fact Sheet 2005-3097. Available from: <http://pubs.er.usgs.gov/publication/fs20053097>.
- Diehl, T. H. & Harris, M. A. 2014 *Withdrawal and Consumption of Water by Thermoelectric Power Plants in the United States 2010*. U.S. Geological Survey Scientific Investigations Report 2014-5184. US Geological Survey, Reston, VA, USA. Available from: <http://pubs.er.usgs.gov/publication/sir20145184>.
- Duveneck, M. J., Scheller, R. M. & White, A. 2014 Effects of alternative forest management strategies in the face of climate change in the Northern Great Lake Region. *Can. J. Forest Res.* **44**, 700–710.
- Federal Geographic Data Committee 1994 *The 1994 Plan for the National Spatial Data Infrastructure: Building the Foundation of an Information Based Society*. Federal Geographic Data Committee, Reston, VA. Available from: www.fgdc.gov/policyandplanning/NSDI%20Strategy%201994.pdf.
- Federal Geographic Data Committee 2013 *National Spatial Data Infrastructure Strategic Plan 2014–2016*. Federal Geographic Data Committee, Reston, VA. Available from: www.fgdc.gov/nsdi-plan/draft-nsdi-strategic-plana8093public-comment-version-7-31-2013.pdf.
- Fitch, P. 2013 *OGC Surface Water Interoperability Experiment Final Report*. Open Geospatial Consortium, Wayland, MA, USA. Available from: www.opengis.net/doc/ER/surface-water-ie.
- Francy, D. S., Brady, A. M. G., Carvin, R. B., Corsi, S. R., Fuller, L. M., Harrison, J. H., Hayhurst, B. A., Lant, J., Nevers, M. B., Terrio, P. J. & Zimmerman, T. M. 2013 *Developing and implementing predictive models for estimating recreational water quality at Great Lakes beaches*. U.S. Geological Survey Scientific Investigations Report 2013-5166. US Geological Survey, Reston, VA, USA. Available from: <http://dx.doi.org/10.3133/sir20135166/>.
- Hirsch, R. M. & Fisher, G. T. 2014 Past, present, and future of water data delivery from the U.S. Geological Survey. *J. Contemp. Water Res. Edu.* **153**, 12.
- Hutchison, N. E. 1975 *WATSTORE: National Water Data Storage and Retrieval System of the U.S. Geological Survey; User's Guide*. U.S. Geological Survey Open-File Report 75-426. US Geological Survey, Reston, VA, USA. Available from: <http://pubs.er.usgs.gov/publication/ofr75426>.
- LaFontaine, J. H., Hay, L. E., Viger, R. J., Markstrom, S. L., Regan, R. S., Elliott, C. M. & Jones, J. W. 2013 *Application of the Precipitation-Runoff Modeling System (PRMS) in the Apalachicola-Chattahoochee-Flint River Basin in the southeastern United States*. U.S. Geological Survey Scientific Investigations Report 2013-5162. US Geological Survey, Reston, VA, USA. Available from: <http://pubs.er.usgs.gov/publication/sir20135162>.
- Maupin, M. A., Kenny, J. F., Hutson, S. S., Lovelace, J. K., Barber, N. L. & Linsey, K. S. 2014 Estimated Use of Water in the United States in 2010. U.S. Geological Survey Circular 1405. Available from: <http://pubs.er.usgs.gov/publication/cir1405>.
- National Research Council 2002 *Estimating Water Use in the United States: A New Paradigm for the National Water-Use Information Program*. The National Academies Press, Washington, DC.
- Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., Viger, R. J., Blodgett, D., Brekke, L., Arnold, J. R., Hopson, T. & Duan, Q. 2015 Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance. *Hydrol. Earth Syst. Sci.* **19**, 209–223.

- Norris, J. M. 2010 U.S. Geological Survey Streamgage Operation and Maintenance Cost Evaluation from the National Streamflow Information Program. U.S. Geological Survey Fact Sheet 2010-3025. Available from: <http://pubs.er.usgs.gov/publication/fs20103025>.
- Office of Management, Budget 2002 Circular No. A-16. Available from: www.whitehouse.gov/omb/circulars_a016_rev/.
- Preston, S. D., Alexander, R. B., Woodside, M. D. & Hamilton, P. A. 2009 SPARROW MODELING – Enhancing Understanding of the Nation’s Water Quality. U.S. Geological Survey Fact Sheet 2009–3019. Available from: <http://pubs.er.usgs.gov/publication/fs20093019>.
- Read, J. S., Winslow, L. A., Hansen, G. A., Van Den Hoek, J., Hanson, P. C., Bruce, L. C. & Markfort, C. D. 2014 *Simulating 2,368 temperate lakes reveals weak coherence in stratification phenology*. *Ecol. Model.* **291C**, 142–150.
- Senay, G. B., Bohms, S., Singh, R. K., Gowda, P. H., Velpuri, N. M., Alemu, H. & Verdin, J. P. 2013 *Operational evapotranspiration mapping using remote sensing and weather datasets: a new parameterization for the SSEB approach*. *JAWRA J. Am. Water Resour. Assoc.* **49** (3), 577–591.
- Taylor, P. 2012 OGC WaterML 2.0: Part 1 – Timeseries. OGC 10-126r4. Available from: https://portal.opengeospatial.org/files/?artifact_id=57222.
- Taylor, P. 2013 WaterML2.0 – Part 2: Ratings, Gaugings and Sections Discussion Paper. Available from: https://portal.opengeospatial.org/files/?artifact_id=54423.
- Taylor, P., Valentine, D., Walker, G., Sheahan, P., Pratt, A., Dornblut, I. & Grellet, S. 2010 Harmonizing Standards for Water Observation Data-Discussion Paper. Available from: http://portal.opengeospatial.org/files/?artifact_id=39090.
- Thornton, P. E., Thornton, M. M., Mayer, B. W., Wilhelm, N., Wei, Y., Devarakonda, R. & Cook, R. 2012 *Daymet: Daily Surface Weather on a 1 km Grid for North America, 1980–2008*. Oak Ridge National Laboratory (ORNL) Distributed Active Archive Center for Biogeochemical Dynamics.
- Tomkins, J. 2014 TimeSeriesML 1.0 Standards Working Group Charter. Available from: https://portal.opengeospatial.org/files/?artifact_id=60856&version=1.
- US Geological Survey 2012 User’s Manual for the National Water Information System of the U.S. Geological Survey: Groundwater Site-Inventory System Version 5.2. Available from: http://nwis.usgs.gov/nwisdocs5_2/gw/GW.user.book.html.
- Valentine, D. & Zaslavsky, I. 2009 CUAHSI WaterML 1.1. San Diego Super Computing Center. Available from: http://his.cuahsi.org/documents/WaterML_1_1_part2_v1.pdf.
- Viger, R. J. & Bock, A. 2014 GIS Features of the Geospatial Fabric for National Hydrologic Modeling. US Geological Survey. Available from: <http://dx.doi.org/10.5066/F7542KMD>.
- Vretanos, P. A. 2005 OGC Web Feature Service Implementation Specification. Open Geospatial Consortium. Available from: http://portal.opengeospatial.org/files/?artifact_id=8339.
- Western States Water Council 2014 Western States Water Council 2013: WaDE Program Highlights. Available from: www.westernstateswater.org/wp-content/uploads/2012/10/WaDE-2014-Update-Factsheet-for-Spring-Council-Meeting-in-DC3.pdf.
- Williamson, T. N., Lee, B. D., Schoeneberger, P. J., McCauley, W. M., Indorante, S. J. & Owens, P. R. 2014 *Simulating soil-water movement through loess-veneered landscapes using nonconsilient saturated hydraulic conductivity measurements*. *Soil Sci. Soc. Am. J.* **78** (4), 1320–1331.
- Zaslavsky, I., Valentine, D. & Whiteaker, T. 2007 CUAHSI WaterML OGC 07-041r1. Available from: http://portal.opengeospatial.org/files/?artifact_id=21743.

First received 8 April 2015; accepted in revised form 30 June 2015. Available online 31 July 2015