HARNESSING THE ENVIRONMENTAL DATA FLOOD

A Comparative Analysis of Hydrologic, Oceanographic, and Meteorological Informatics Platforms

BY ANDREW K. DOW, ELI M. DOW, THOMAS D. FITZSIMMONS, AND MAURICE M. MATERISE

The evolution of software for sharing and consuming hydrologic, oceanographic, and meteorological data will be defined by comprehensiveness of content; fluidity of interchange; and flexible, multifunctional visualization.

The challenge and opportunity presented by “big data,” as they apply to hydrology, oceanography, and their meteorological components (referred to under the umbrella of “water data” or “water-science data”), lies in providing access to the growing quantity and breadth of data while allowing data consumers to make use of such vast supplies of information more efficiently (Beran and Piasecki 2009; Hart and Martinez 2006; Horsburgh et al. 2009, 2011; Megler 2013). New tools for visualizing sensor observations data (Horsburgh et al. 2011) and descriptive (or

predictive) model-generated data (Goodall et al. 2011) facilitate the informed decision making needed to manage water quality and availability in a period of continued population growth, land development, and climate change (Gold et al. 2013; Hart and Martinez 2006; Vörösmarty et al. 2000). While the Internet provides widespread access to a large, heterogeneous pool of sensor observations and model-generated data (Maier et al. 2012), successful application will depend on the ability to sort through the inherent complexity of this data landscape so as to integrate, organize, and visualize the combinations of data needed to provide actionable insight (Horsburgh et al. 2009; Tarboton et al. 2009). In the water sciences, such abilities help researchers focus their efforts on analysis of existing datasets and provide them with opportunities to draw insights from much larger datasets than they would be otherwise able to generate independently.

As researchers and government organizations involved in water-data collection and publication continue to discuss best practices for data management and retrieval (Beran and Piasecki 2009; Megler and Maier 2011, 2013; Horsburgh et al. 2008, 2009; Ramachandran et al. 2005; Tarboton et al. 2009), sharing (Horsburgh et al. 2011; Kanwar et al. 2010), and visualization (Ames et al. 2012; Domenico et al.

AFFILIATIONS: DOW, DOW, FITZSIMMONS, AND MATERISE—IBM Observations Data Model-Based Visualization Portal, IBM Systems & Technology Group, Poughkeepsie, New York

CORRESPONDING AUTHOR: Eli M. Dow, Building: 801, Office: 30-217B, 1101 Kitchawan Road, Yorktown Heights, NY 10598
E-mail: emdow@us.ibm.com

The abstract for this article can be found in this issue, following the table of contents.

DOI:10.1175/BAMS-D-13-00178.1

A supplement to this article is available online (10.1175/BAMS-D-13-00178.2)

In final form 6 October 2014

©2015 American Meteorological Society

2002; Horsburgh et al. 2011; Jeong et al. 2006), this essay complements that dialogue with a holistic survey of these practices and a quantitative discussion of their prevalence. We do so with a focus on the data consumption experience, considering an array of features that give users access to more data while, most importantly, affecting their ability to access and analyze the data more effectively. This essay provides current observations on water-data content, sharing, and visualization among water-data informatics platforms, along with considerations for the successful evolution of these platforms in a future shaped by measurement technology advances, machine-driven data use, and cloud-based data storage and discovery. Detailed information about the features of specific platforms can be found in the supplemental material.

Our goals are twofold: to identify opportunities for improved data discovery and application and to further promote the multidisciplinary dialogue required to address these opportunities. In addition, we hope that this essay and the detailed information about the surveyed platforms (provided in the supplemental material) may help students, educators, and citizens interested in environmental monitoring. Our goal is to help them explore the relationships between weather conditions and the quality and safety of local water bodies, identify public software tools most relevant to their interests, and explore the full range of their functionality in a data landscape otherwise intimidating in size and scope.

The growing landscape of publicly available data for water is complex in content and construction. The datasets describe inland and coastal water bodies and the weather conditions that affect their characteristics, while the procurement, publication, and use of such data spans the fields of hydrology, oceanography, and meteorology as well as computer science, the physical sciences, engineering, and public policy. The multidisciplinary community shaping this data landscape includes government research organizations, industry data-analytics researchers, academic institutions, and consortia. Among the institutions surveyed for this essay, the U.S. Geological Survey (USGS), the National Oceanic and Atmospheric Administration (NOAA), and the National Aeronautics and Space Administration (NASA) offer some of the largest sources of continuous hydrologic, oceanographic, and meteorological data at national and global scales. These sources are supplemented by water-quality measurements provided by the U.S. Environmental Protection Agency (EPA), physical snow and water data from the U.S. Department of Agriculture (USDA) Natural Resources Conservation

Service (NRCS), and diverse data collected by regional research entities, many of which are affiliated with larger-scale research initiatives such as NOAA's Integrated Ocean Observing System (IOOS). The quantity of data available and the multitude of participants and measurements reflect the complexity inherent in big data opportunities as they apply to the water-science community.

Solutions for managing heterogeneous water data have evolved within disciplines and provide examples for achieving wider standardization and integration. In hydrology, for example, efforts by the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) have been instrumental in steering the hydrological community toward common standards for data formatting and interchange. Additionally, CUAHSI offers examples of cyber infrastructure and software platforms for comprehensive data aggregation, retrieval, and visualization. In particular, the CUAHSI Hydrologic Information System (HIS) (Maidment 2008; Tarboton et al. 2009) is a National Science Foundation (NSF)-supported initiative addressing hydrological data organization and access with the aim of "creating a common window on water information for the United States" (Maidment 2008). To improve data organization and access, the CUAHSI HIS has 1) integrated multiple data sources and streamlined data access using a service-oriented architecture, while establishing web services that allow machine-to-machine data communication through Simple Object Access Protocol (SOAP) (Goodall et al. 2008); 2) defined a relational database structure known as the Observations Data Model (ODM) to organize and store data from heterogeneous sources (Horsburgh et al. 2008); 3) defined a community-specific extensible markup language (XML) schema, known as WaterML, to homogenize the format and vocabulary of observations data from such heterogeneous sources (Zaslavsky et al. 2007); and 4) thereby facilitated data retrieval through web services into a variety of user-preferred environments (Tarboton et al. 2009).

Realizing the potential of big data in the water sciences for a wide user base requires the data community to build on these best practices, including integration of data sources and consistent data presentation within publicly accessible platforms for discovery and visualization. We define these platforms as collections of content and interfaces, connected over the web, that allow users to access, analyze, and export data for a variety of purposes.

Furthermore, we approach this discussion as participants in the water-data community, with firsthand experience developing a data visualization platform for the

Beacon Institute for Rivers and Estuaries (BIRE) (Dow et al. 2011). The authors' relevant experience, standards emerging from the water-science community, and paradigms applied externally in data management and software services are reflected in this platform, World Water Window. The World Water Window, like many of the data platforms surveyed, combines independent datasets into a common data storage model and shares the combined data through web services and through a single graphical user interface (GUI) for browser-based searching and visualization.

We structure the remainder our discussion in terms of three informatics functions: providing data of value, supporting exchange of data, and helping users analyze the data. We propose that the water informatics community can collectively provide better data discovery by identifying and implementing best practices in each area.

DATA. High-quality, useful data are the foundation for any water informatics platform. We assert that the data provided are more likely to be of value to a wide user base when the content is comprehensive and when it is present in sufficient quantities to derive statistically meaningful conclusions. While data quantity is relatively trivial to measure, water-data content can be comprehensive in several dimensions: breadth in the variables measured, geographic coverage, and temporal span and continuity. For many users, the value of a dataset is likely to be a function of strength along two or more of these dimensions.

Surprisingly, such combinations are difficult to find in individual data providers. The majority of platforms we surveyed provided data for several variables across multiple measurement categories. For instance, 77% of the studied platforms provide some combination of water and meteorological measurements and 77% provide some combination of physical and compositional water measurements. This diversity can be attributed to two main factors: sensor technologies that collect multiple types of information in

real time and the increasingly common practice of integrating data from multiple observation networks or primary data sources into centralized retrieval and visualization platforms. Nearly half of the surveyed data suppliers (47%) aggregate their data from multiple sources of published data and this practice was observed at global, national, and regional scales.

However, the nominal breadth in content displayed at a qualitative, holistic level may not reflect the quantitative distribution of information available in practice. For instance, the USGS, one of the nation's largest collectors and providers of hydrologic data, allows users to select from more than 50 parameters that describe meteorological conditions or the physical state and composition of water throughout the United States via the National Water Information System web interface. The measurements collected consistently across sites, however, are primarily physical measurements describing water quantity: namely, discharge and gage height. Water-quality measurements are available only at locations in which special sensors are in place or at which manual testing campaigns were conducted for specific periods of time. Furthermore, the specific compositional parameters measured in either scenario vary on a site-by-site basis. Metadata analysis of a random 137 million data-point sample of USGS data provided through the National Water Information System (NWIS), indicated that four variables, all representing physical characteristics, account for approximately 84% of the total data values, as shown in Fig. 1.

FIG. 1. Distribution of data values by parameter within a sample set of 137 million data values obtained through the USGS NWIS. Approximately 84% of the data describe purely physical characteristics (i.e., not related to chemical or biological constituents), reflecting the limited availability of continuous compositional data in large-scale environmental monitoring systems.

Quantitatively understanding the metadata of such datasets helps recognize their value and clarify their limitations. The USGS NWIS, for example, represents a hydrologic data source of tremendous geographic scope, time span, and total volume. Users can perform large-scale streamflow modeling and computations with such a dataset and it serves as a central source of such information for locations across the nation. However, users may have difficulty finding water-quality measurements of sufficient geographic density or time span to incorporate chemical or compositional factors into their models or to build a detailed understanding of water-quality changes and constituent transport in a given watershed.

That said, measuring more parameters consistently across sites can be financially and logistically challenging, particularly at large scales. Another possible approach for making data access more comprehensive is the practice of aggregating information from numerous sources and integrating it for users through web services or through data discovery interface software (Horsburgh et al. 2009). CUAHSI provides one of the largest implementations of this practice in the water-data community through their WaterOneFlow web services (Horsburgh et al. 2008) and the HydroDesktop data discovery client (Ames et al. 2012), respectively, which provide data through more than 100 individual web services. Thus, CUAHSI web service metadata comprises a rather representative picture of an aggregated hydrologic data landscape. However, it can be difficult to provide data that are comprehensive in all dimensions even in an aggregated collection, such as that provided through the CUAHSI HIS. For example, 95% of the total data provided through the CUAHSI web services are provided by the top 10 web services out of the 115 web services whose metadata we queried for this study. Of these top 10, 8 services provide data for five or fewer variables and 4 services provide data for one variable, precipitation. As reflected in these metadata characteristics, users may be challenged to find centralized datasets that are voluminous as well as comprehensive in all three dimensions of geography, measurement breadth, and time.

The aggregate data collection provided by CUAHSI web services reflects the challenges of the general data landscape in which big data that are large in volume, time continuity, and (or) spatial coverage are concentrated among relatively few sources with limited content diversity, while the tail of “little data” is rich in content diversity but marked by complexity, inconsistency, and context specificity that make integrating such data challenging. The log-scale volume-sorted

distribution of total data points among 115 CUAHSI web services shown in Fig. 2 is symbolic of the “big data, little data” landscape, with a cluster of voluminous providers and exponential decay in volume beyond these top sources.

The content limitations of big data in water science pose an important question. How can the water-science communities make their collective datasets more comprehensive? One possible approach is to maximize the integration of small, complementary datasets, essentially increasing the contribution of the data tail. However, the gains from this approach can be limited in practice by sheer lack of data volume and by complex metadata requirements to accurately integrate values among sources. Another approach is to make the big data piece more diverse, essentially by deploying more comprehensive sensor networks or even consistent manual testing regimes that provide long-term, continuous monitoring of multiple parameters. Such an approach may not be feasible on large geographic scales, which Hart and Martinez (2006) implicitly point out as they contrast the singular (or limited) measurement functionality of large-scale sensor networks with “localized multifunction sensor networks” that tend to measure a variety of environmental and meteorological properties.

With this in mind, a combination of the two approaches may be the most feasible mechanism for achieving more diverse, consistent, and continuous water data. Consistent, multiparameter measurement efforts may be more financially and logistically feasible to conduct on the scale of regional networks. It may also be easier to garner funding and strategic support for such monitoring programs at the regional level. The programs can be owned and overseen by local research universities or institutes, and the concept of understanding and effectively managing the local water bodies may appeal to regional public interests. For example, the environmental and ecological health of the Hudson River, notorious for a legacy of polychlorinated biphenyl (PCB) contamination from the mid-twentieth century (HRNRT 2013), is a topic of economic relevance, public health concern, and scenic pride for communities along the Hudson River basin. Today, the Hudson is one of the best-studied rivers in the world (Strayer 2012). Sensor networks such as the Hudson River Environmental Conditions Observing System, Clarkson University’s Rivers and Estuaries Observatory Network (REON), and the Stevens Institute for Technology’s New York Harbor Observing and Prediction System (NYHOPS) (Georgas et al. 2009) measure conditions throughout the Hudson River with sensor networks, providing near-real-

time information for many measurement sites. These regional data providers lack the geographic scope of USGS data but, in some cases, provide more consistency in compositional measurements across monitoring locations.

The information from these regional networks could ideally be integrated through collections of web services (such as the CUAHSI HIS services) as well as common software portals for browser-based access and visualization. Ideally, regional networks would coordinate to share technology, practices, and quality-control standards, perhaps overseen by members of a larger governing or coordinating body, such as the USGS, EPA, and NOAA. The Integrated Ocean Observatory, developed and overseen by NOAA, is a current example (NOAA 2009). In reference to estuarine monitoring efforts, Boicourt et al. (2012) emphasize that “cooperative enterprises” among regional observatory networks may be key to the long-term success of such environmental data collection programs in the face of future budget uncertainties.

As mentioned in our discussion of integrating complementary datasets, the confidence with which users can jointly apply data from multiple primary sources depends on whether the associated metadata contains enough detail for end users to assess reliability and compatibility. Even in the case of the regionally executed but broadly coordinated environmental monitoring programs described above, in which best practices for data collection and quality assurance may be shared and standardized, documentation of such standards remains useful for data users and curators. Applying the “role based” definition of metadata preferred by Frew and Dozier (2012), we would consider such metadata to include any information that aids in the interpretation and proper use of the data it describes. In the context of water-science data, information that would help end users determine suitability and comparability of data for various applications includes a provenance trail indicating how the data has been obtained and produced as well as applicable notation of measurement methodology, technology, and quality screening status (e.g., raw,

FIG. 2. Data volume for 116 CUAHSI web services sorted by total data values. The top 6 web services by volume provide 82% of the total data values available through the web service collection, and the top 10 web services represent 95% of the total data volume.

filtered, or validated data). While such information ideally is embedded in a machine-parsable metadata package associated with any retrieved data object, it is also desirable to make such information visible to human users exploring data holdings via a web interface. Given the prevalence of platforms providing data in near-real time, much of the data available through web-based discovery portals can be raw (i.e., minimally processed) or at least provisional (i.e., not validated). Furthermore, there is the risk that platforms displaying data aggregated from multiple primary sources may not carry over the same contextual detail provided by the original source. When examining current practice with respect to these issues, we found that the type of information provided among the studied platforms varies significantly. Observed examples range from a simple and general disclaimer about the provisional nature of the data provided to data pertaining to the operational status of the measurement instrument (e.g., sensor battery voltage over time) and to display of (or links to) detailed text-based metadata. Such metadata may include descriptions of the data source, measurement methods, quality assurance and quality-control methods applied to the data, and definitions of quality labels or codes applied to the data. In rare cases, users are given the option to retrieve or view data categorized by different levels of quality validation. Attention to quality-related metadata and contextual details when building tools for aggregated data discovery and display will help water-data providers achieve not only centralized but more integrated datasets for purposes of downstream ease of use.

A key component of this metadata challenge is maintaining clear chain of custody or provenance as data values are imported from various government agencies. For a definition of how provenance information is useful in contexts such as the BIRE World Water Window and the other data-aggregating portals that informed this discussion, we submit the following excerpt taken from the PROV documentation (Groth and Moreau 2013):

Provenance is information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness. The PROV Family of Documents defines a model, corresponding serializations and other supporting definitions to enable the inter-operable interchange of provenance information in heterogeneous environments . . .

These concepts are relevant to the water-science community when integrating data from disparate sources with varying definitions and standards, the provenance of which can be encoded using the PROV-O ontology recommended by the World Wide Web Consortium (Lebo et al. 2013). This area of research seems fruitful and timely as big data

becomes socialized and remixed into new platforms and derivative datasets.

INTERCHANGE. While the nature of the data provided to users through water informatics platforms is of primary importance, those data also need to be accessible and easily shared. Web services (Cerami 2002) are a very machine-accessible means of exporting data, and many data-intensive sectors have adopted the web services paradigm. Members of the water-science community have advocated for a web service model for delivering customizable derived-data outputs while minimizing cumbersome transfer of raw data (Kanwar et al. 2010), integrating data from heterogeneous sources (Horsburgh et al. 2009), and integrating environmental models or processing capabilities (Goodall et al. 2011; Granell et al. 2010). By supporting machine-to-machine communication and interoperability, web services facilitate data aggregation, allowing data providers to build more centralized, comprehensive tools and repositories for data discovery (Foster 2005). They also benefit the research community by making it easy for users to export the data into a preferred software environment for analysis and modeling work (Kanwar et al. 2010; Tarboton et al. 2009). Of the 30 platforms we surveyed, 57% currently use web services for exporting data, suggesting that the water community stands to benefit from increased use of web services.

The issue of using web services to share data goes hand in hand with the file formats used to share that data. XML is one of the formats best suited to web service exchange, because of its excellent machine readability, self-describing nature, and interoperability (Cerami 2002). The percentage of platforms providing XML formatting (63%) is understandably similar to the number that support export through web services. Many platforms provide multiple formats for exported data files, and the prevalence of various file formats among the platforms we surveyed

FIG. 3. Prevalence of various file formats for data provided by sample group of 30 surveyed water informatics platforms, as available directly from the respective visualization interfaces or by programmatic access via applicable web services. Many platforms provide data in multiple formats, with the relative overall availability by format reflected in this figure. The legacy and continued relevance of separated-value formats in user practices and preferences, such as for manual manipulation and analysis of data by individual researchers, is apparent, as is the opportunity to increase the availability of data in machine-friendly formats such as XML and JSON. Increasing the availability of such formats facilitates the software-driven data exchange and consumption integral to cutting-edge data integration and modeling capabilities.

is shown in Fig. 3. The flexibility offered by multiple format choices can be useful in a time of heterogeneity and transition to common standards. However, the diversity of formats can hinder data integration if data sources do not share formats in common. As presented in Fig. 3, comma- and tab-separated value formats remain the most common. This reflects the preference among many researchers to export data into tables and work with local copies for final analysis; largely for this reason, such formats remain useful. As data access and analysis become increasingly software-driven and the demand for feeding data into modeling software increases, data providers can support and increase use of their data by providing machine-readable formats such as XML and JavaScript Object Notation (JSON) alongside legacy separated-value formats. Although machine friendly rather than directly user friendly, such formats ultimately benefit end users by facilitating the integration of data across multiple sources for purposes of increasingly centralized and seamless retrieval.

Thus, by adopting the formatting and exchange paradigms that have become standard in data-intensive sectors and that facilitate powerful software-driven content integration and analysis, data providers can supply content that is not only useful but also highly accessible. Translation of existing datasets to machine-readable and community-standardized XML schema, such as recent efforts by the USGS to transition datasets to WaterML, will be an important component of a community evolution in data sharing, going hand in hand with machine-executed data interchange. The adoption of such formats by new data sources is equally important to the future of data fluidity. In addition to promoting fluid interchange of water data based on machine readability, formats such as XML (and particularly its discipline-specific variants) facilitate joint application of disparate datasets based on their potential to incorporate detailed metadata (Hart and Martinez 2006; Taylor 2012). However, such detail can come at the price of relative verbosity or a degree of specificity that precludes easy interpretation and readability by those outside of a particular discipline. In our own implementation, we adopted a community standard (WaterML) while also using and providing a more compact version for performance reasons. Hart and Martinez (2006) note that the definition of different XML schema by individual projects can perpetuate data integration challenges; however, ontology-based semantic frameworks and search strategies are emerging as a promising solution to seemingly inevitable heterogeneity by making it easier for software to relate dif-

ferent data descriptions to common conceptual categories (Hart and Martinez 2006; Piasecki and Beran 2009). The semantic communication incorporated in the Open Geospatial Consortium's WaterML 2.0 information model (Taylor 2012) and the approach presented by Zheng et al. (2011) reflect a growing interest in ontological solutions for environmental data interoperability. In capitalizing on the potential benefits of mature, community-tailored XML schema, the water-science community can strive to balance these benefits with a simplicity and versatility that facilitate data use across disciplines while also contributing to evolving, unifying ontologies.

VISUALIZATION. This essay has thus far addressed opportunities to improve the availability and accessibility of water data for a variety of user needs. Multisource data integration and the data-management practices that facilitate that integration save users time by centralizing data access and creating more comprehensive aggregate datasets, but this yields a new set of challenges to overcome. Consider that this also gives rise to datasets that are overwhelmingly large, more complex, and heterogeneous in content (Megler and Maier 2011). Informatics platforms make data available but also approachable and usable, by providing an agile data discovery experience. That is, they offer users the flexibility required to hone in on information of interest as well as view that information in ways that make it easier to understand.

The ability to customize search criteria when querying within a data discovery interface can make the search process more efficient and user friendly (Horsburgh et al. 2011). Criteria that may be of interest when filtering data include geographic location, variables (parameters) measured, time period during which measurements were obtained, and the status of datasets as continuously updated (providing current or near-real-time measurements) or historical. The prevalence of these and other filtering options among the data discovery and visualization portals for the surveyed water informatics platforms is summarized in Fig. 4.

All of the surveyed platforms offer the ability to filter by location. In fact, geography-based searching often dominates the data discovery experience, as it is presented to users in the form of a map-based interface in 97% of the surveyed tools. The popularity of map-based interfaces reflects the centrality of geography in the majority of user needs. For instance, researchers may wish to study a specific watershed and would use such an interface to view

the concentration of data in a region of interest or to hone in on an example location that offers a sufficient density of measurements.

However, geography alone may not determine the scope (or focus) of researchers' questions. Big data facilitates the study of big questions, and users may have an interest in visualizing data from a variety of locations pertaining to a particular time period or particular measurement parameters. Additionally, researchers may need to view data pertaining to a specific climate phenomena or weather event. The ability to work with a package of information appropriately tailored to user needs therefore depends largely on the combination of options provided for specifying and filtering data content of interest. Building better data discovery experiences will require that more platforms provide combinations of parameter- and time-based searching in the initial user interface in addition to geographic searching. Such flexibility can be particularly useful when integrated within a map-based interface, allowing users to filter by factors such as specific parameters (i.e., measurement variables) while viewing the availability of relevant data geographically.

While the above features are related to visualizing data availability for ease of discovery, platforms also vary with respect to the flexibility they provide for viewing graphical representations of the information.

As datasets become increasingly integrated and heterogeneous, customization becomes more important not only in data retrieval but also in data visualization. Effective data-driven decision making is facilitated by the ability to select chart formats and statistical analysis tools that quantify relationships. Embedding such visualization and analysis tools in data discovery interfaces can make datasets accessible to less experienced audiences (Horsburgh et al. 2011) and helps users evaluate the relevance of datasets and optimize their queries before exporting the data for local use (Jeong et al. 2006; Maier et al. 2012). Among the surveyed platforms, time series plots are, by far, the most commonly available data display format (Fig. 5). This chart format is simple for software developers to implement, reflects the dominant model of water-observations data collection, and is effective for displaying and studying trends in observations at discrete locations. A discrete-location-based data collection model underlies the majority of data provided through the surveyed tools.

The only other visualization feature present in at least half of the surveyed tools (56%) is a summary of current measurements, offering users a snapshot of current conditions. This measurement snapshot can allow public stakeholders with commercial or recreational concerns to stay aware of conditions at sites of interest or can help researchers identify the onset of

conditions that may be important to their research. The most practical means of achieving these goals is to allow users to set up automated alerts that are customized based on user-defined thresholds for particular parameters of interest or concern, a feature present in only 20% of surveyed platforms offering real-time data. Given the prevalence of visualization tools publishing near-real-time data, the lack of custom alerting features among those real-time data providers is a notable opportunity. Part of the power of real-time data streaming is the opportunity to remain aware of conditions and react to them quickly. More widespread

FIG. 4. Prevalence of various search filters for structuring data queries among 30 surveyed water informatics platforms. These filters apply only to the primary interfaces presented to users for structuring an initial data search. Percentages are calculated out of total platforms eligible for each type of filter based on the nature of their respective offerings. These statistics reflect the prominence of geography in the options provided for data discovery. The largest opportunities for increased flexibility and specificity in data discovery are in wider implementation of filters pertaining to two key aspects of potential user questions: the “what” (i.e., measured variable or parameter) and “when” (i.e., point in time or date range of interest).

implementation of customized alerting features can help researchers stay informed of phenomena of interest and optimize use of resources by deploying sensors or collecting environmental samples in circumstances most relevant to their research questions. It could also help commercial or recreational stakeholders, including water-treatment facilities, fisheries, and boaters, plan for favorable or problematic conditions as they propagate through a region or water body (Ruberg et al. 2008).

Charts facilitating spatial analysis, such as color-coded maps or charts indicating the magnitude of a particular variable over a two-dimensional region (referred to as a “heat map”), are less common because of the unique data collection requirements. Visuals of this type are present in only 40% of the sample portals. Only data produced by models or acquired by remote sensing technology [e.g., NASA’s ocean surface measurements via satellite-mounted Moderate Resolution Imaging Spectroradiometer (MODIS) technology] tend to possess the degree of spatial resolution and continuity appropriate for heat map generation. Some of the platforms analyzed for this study incorporate remote sensing observations into their visualization portals through the option to overlay heat maps on the backgrounds of their map-based interfaces. This practice allows users to view site-specific sensor or sample observations in the context of spatially continuous measurements such as surface temperature, chlorophyll concentration, or relative levels of dissolved solids. Such a combination of features can help researchers evaluate model skill (Maier et al. 2012) and provides an example of how to effectively integrate diverse types of measurements into cohesive visualization tools.

While time series plots are a common and effective way of studying point observations data, there are additional related features that can help users derive understanding and identify relationships between variables or phenomena. Such features help enhance visualization functionality beyond

FIG. 5. Prevalence of visual output formats available among 30 surveyed water informatics platforms. Heat maps and animations (which are often time-evolving heat maps or vector fields) are found in platforms providing gridded or spatially continuous remotely sensed or model-generated data and, in several platforms, are integrated with point observations data in a customizable map interface. Such integration of time-series observations within a context of more spatially continuous data fields is an example of the potentially useful combinations and visualization flexibility made possible by multisource data platforms. Built-in tools for visualizing statistical analysis of datasets are noticeably rare, suggesting an opportunity for wider implementation to aid in data interpretation.

simple display of information, aiding interpretation. For example, trends in a given parameter over time may be more important to researchers in the context of other parameters. The emerging model of integrated observations data offers an unprecedented opportunity to study related trends in multiple types of measurements that are temporally or spatially related, even if from diverse collection sources. Such relationships can be studied efficiently if the data visualization portals allow users to view such measurements concurrently on a customizable time series plot or to visualize parameter correlation directly using scatterplots. Surprisingly, only 42% of the surveyed tools that provide time series plots allowed users to visualize multiple parameters on the same graph, representing an important opportunity for future tool development (Fig. 6). Furthermore, only four of the applicable platforms (15%)—the CUAHSI HydroDesktop, the BIRE World Water Window, the USGS NWIS Web Interface, and the Stevens Institute for Technology NYHOPS—offer users the ability to visualize time series data from multiple locations (potentially across multiple primary sources) simultaneously. This also presents an important gap relative to potential user needs. Such flexibility better integrates complex, multisource datasets at the user end, allowing more seamless analysis of information across sources (Domenico et al. 2002).

Another visualization feature that can provide context and support data interpretation is the ability to view statistical analyses for selected datasets. Only 13% of the surveyed tools allow users to generate statistical charts for this purpose, in the form of either scatterplots for quantifying variable correlation or histograms and related tools for quantifying frequency or probability-based distributions of values.

In addition to visualizing historical trends and gaining a statistically rigorous understanding of current and historical conditions, users may be interested in estimates of future conditions projected from observed patterns. Predictive analytics support proactive rather than reactive application of data and the ability to abstract more broadly applicable patterns of interaction and behavior. Some data visualization tools apply built-in predictive algorithms to extend existing user-selected datasets into the future, while others integrate and publish values generated by external models alongside real-time and historical observations data. Such predictive visualization features are present in one-third of the sample portals in this study. One explanation for their limited presence would be the multidisciplinary knowledge required to apply appropriate algorithms to existing datasets while accurately assessing the validity of model-generated values. Another explanation is a potential functional separation within the data informatics community between data visualization for analysis and for modeling. By taking an increasingly holistic view and combining the expertise and datasets needed to integrate modeling alongside simpler analytics, the community can provide greater research versatility through a given set of centralized software portals.

SUMMARY. The water-data informatics community is poised to progress in serving a wider audience of users and truly harnessing the insights available through observational data by addressing opportunities to enrich data content, to optimize data access and fluidity, and to provide an agile data discovery and visualization experience. Our analysis shows

FIG. 6. Prevalence of various user interface and visualization features among 30 surveyed water informatics platforms. Potential low-hanging opportunities to facilitate data use include 1) providing more flexibility in combining data of interest in concise visual outputs, such as options to concurrently plot measurements of multiple variables and from multiple locations, and 2) increasing the availability of automated user alerting based on user-customized data interests.

that, in current open-access water-data repositories, volume does not imply comprehensiveness. Efforts to monitor a wider set of parameters more consistently at manageable scales, combined with integration of these and other complementary datasets, may provide a feasible strategy for making data repositories more comprehensive. In addition, data providers will best support content integration and serve the evolving needs of the user community by offering web services for data sharing and by aligning on formats for data exchange. We especially encourage the use of machine-readable, general-purpose XML and its community-specific versions, such as WaterML.

Furthermore, to truly realize the value of even the most comprehensive and accessible information, water-data informatics platforms will provide tools that aid in understanding and applying the data through flexible data discovery and integrated visualization capabilities. Such integrated visualization capabilities can include combinations of geospatial and point observations data, multiparameter and multilocation plotting, statistical analysis tools, and even predictive visualization and modeling.

Last, we emphasize that the nascent software-as-a-service (SaaS) paradigm, in which service providers make applications available to users over the Internet, is shaping opportunities in the water-data community for providing these tools through web clients. Providing data discovery and visualization in the cloud, as the vast majority of platforms do, centralizes the maintenance and computing functions, thereby

minimizing user requirements and ensuring access to dynamic, up-to-date data repositories. This model brings such tools to a wider audience by maximizing their accessibility and compatibility with a variety of operating systems and devices. With the proliferation of Internet-enabled mobile computing devices, the ability to access data informatics tools on these devices significantly broadens the scope of their use cases, including fieldwork guided by real-time conditions. In this spirit, delocalization and distribution of content, access, and analytical services may well define the future of water-science data as an evolving public resource.

ACKNOWLEDGMENTS. The authors wish to thank the IBM managers and colleagues who supported this effort. Thanks to Dr. Harry Kolar for his feedback and suggestions on early drafts of this work and to Mr. Nickalaus Painter for his assistance with data collection in early stages of this work. We would also like to thank Mr. Michael Desens, Dr. Wayne Howell, and Mr. Gary Anderson for their support of water-data informatics research and platform development in partnership with the Beacon Institute for Rivers and Estuaries. We thank the Beacon Institute staff, including Dr. Tim Sugrue, Mr. John Cronin, Dr. James Bonner, and Dr. Mohammad Shahidul Islam, for their efforts to facilitate development of the Beacon Institute World Water Window. Their vision and support provided the opportunity to conduct this comparative review of the state of the art in water-data discovery and visualization. Observations, opinions, and recommendations are the authors' alone and do not necessarily reflect those of IBM, the Beacon Institute, the American Meteorological Society, or any of the persons acknowledged above.

REFERENCES

- Ames, D. P., J. S. Horsburgh, Y. Cao, J. Kadlec, T. Whitaker, and D. Valentine, 2012: HydroDesktop: Web services-based software for hydrologic data discovery, download, visualization, and analysis. *Environ. Modell. Software*, **37**, 146–156, doi:10.1016/j.envsoft.2012.03.013.
- Beran, B., and M. Piasecki, 2009: Engineering new paths to water data. *Comput. Geosci.*, **35**, 753–760, doi:10.1016/j.cageo.2008.02.017.
- Boicourt, W. C., M. Li, N. Nidzieko, A. F. Blumberg, N. Georgas, E. J. Kelly, T. G. Updyke, and W. D. Wilson, 2012: Observing the urban estuary: Review and prospect. *Proc. Oceans*, Hampton Road, VA, IEEE, 1–9, doi:10.1109/OCEANS.2012.6405120.
- Cerami, E., 2002: *Web Services Essentials*. 1st ed. O'Reilly, 304 pp.
- Domenico, B., J. Caron, E. Davis, R. Kambic, and S. Nativi, 2002: Thematic Real-Time Environmental Distributed Data Services (THREDDS): Incorporating interactive analysis tools into NSDL. *J. Digit. Inf.*, **2**, [Available online at <https://journals.tdl.org/jodi/index.php/jodi/article/viewArticle/51/54>.]
- Dow, E. M., T. Fitzsimmons, S. Schmidt, and D. Milvaney, 2011: Data architecture for location aware smarter water sensor visualization. *Proc. 12th Irish National Hydrology Conf.*, Athlone, Ireland, Irish Joint Committees of the International Hydrological Programme and International Commission on Irrigation and Drainage, 6 pp. [Available online at <http://people.clarkson.edu/~dowem/downloads/DataArchitectureForLocationAwareSmarterWaterSensorVisualization.pdf>.]
- Foster, I., 2005: Service-oriented science. *Science*, **308**, 814–817, doi:10.1126/science.1110411.
- Frew, J. E., and J. Dozier, 2012: Environmental Informatics. *Annu. Rev. Environ. Resour.*, **37**, 449–472, doi:10.1146/annurev-environ-042711-121244.
- Georgas, N., A. F. Blumberg, M. S. Bruno, and D. S. Runnels, 2009: Marine forecasting for the New York urban waters and harbor approaches: The design and automation of NYHOPS. *Proc. Third Int. Conf. on Experiments/Process/System Modelling/Simulation & Optimization*, Athens, Greece, Learning Foundation in Mechatronics, 345–352. [Available online at www.stevens-tech.edu/ses/documents/fileadmin/documents/pdf/Georgas_et_al_3rd-IP-EpsMsO.pdf.]
- Gold, A. J., and Coauthors, 2013: Advancing water resource management in agricultural, rural, and urbanizing watersheds: Why land-grant universities matter. *J. Soil Water Conserv.*, **68**, 337–348, doi:10.2489/jswc.68.4.337.
- Goodall, J. L., J. Horsburgh, T. Whiteaker, D. Maidment, and I. Zaslavsky, 2008: A first approach to web services for the National Water Information System. *Environ. Modell. Software*, **23**, 404–411, doi:10.1016/j.envsoft.2007.01.005.
- , B. F. Robinson, and A. M. Castronova, 2011: Modeling water resource systems using a service-oriented computing paradigm. *Environ. Modell. Software*, **26**, 573–582, doi:10.1016/j.envsoft.2010.11.013.
- Granel, C., L. Diaz, and M. Gould, 2010: Service-oriented applications for environmental models: Reusable geospatial services. *Environ. Modell. Software*, **25**, 182–198, doi:10.1016/j.envsoft.2009.08.005.
- Groth, P., and L. Moreau, cited 2013: PROV-Overview: An overview of the PROV family of documents. [Available online at www.w3.org/TR/2013/NOTE-prov-overview-20130430/.]

- Hart, J. K., and K. Martinez, 2006: Environmental Sensor Networks: A revolution in the earth system science? *Earth-Sci. Rev.*, **78**, 177–191, doi:10.1016/j.earscirev.2006.05.001.
- Horsburgh, J. S., D. G. Tarboton, D. R. Maidment, and I. Zaslavsky, 2008: A relational model for environmental and water resources data. *Water Resour. Res.*, **44**, doi:10.1029/2007WR006392.
- , —, M. Piasecki, D. R. Maidment, I. Zaslavsky, D. Valentine, and T. Whitenack, 2009: An integrated system for publishing environmental observations data. *Environ. Modell. Software*, **24**, 879–888, doi:10.1016/j.envsoft.2009.01.002.
- , —, D. R. Maidment, and I. Zaslavsky, 2011: Components of an environmental observatory information system. *Comput. Geosci.*, **37**, 207–218, doi:10.1016/j.cageo.2010.07.003.
- Hudson River Natural Resource Trustees, 2013. PCB contamination of the Hudson River ecosystem. NOAA Rep., 38 pp. [Available online at www.fws.gov/contaminants/restorationplans/HudsonRiver/docs/Hudson%20River%20Status%20Report%20Update%20January%202013.pdf.]
- Jeong, S., Y. Liang, and X. Liang, 2006: Design of an integrated data retrieval, analysis, and visualization system: Application in the hydrology domain. *Environ. Modell. Software*, **21**, 1722–1740, doi:10.1016/j.envsoft.2005.09.007.
- Kanwar, R., U. Narayan, and V. Lakshmi, 2010: Web service based hydrologic data distribution system. *Comput. Geosci.*, **36**, 819–826, doi:10.1016/j.cageo.2007.10.017.
- Lebo, T., S. Sahoo, and D. McGuinness, cited 2013: PROV-O: The PROV ontology. [Available online at www.w3.org/TR/2013/REC-prov-o-20130430/.]
- Maidment, D. R., 2008: Bringing water data together. *J. Water Resour. Plann. Manage.*, **134**, 95–96, doi:10.1061/(ASCE)0733-9496(2008)134:2(95).
- Maier, D., V. M. Megler, A. M. Baptista, A. Jaramillo, C. Seaton, and P. J. Turner, 2012: Navigating oceans of data. *Scientific and Statistical Database Management*, A. Ailamaki, and S. Bowers, Eds., Vol. 7338, *Lecture Notes in Computer Science*, Springer, 1–19, doi:10.1007/978-3-642-31235-9_1.
- Megler, V. M., 2013: Taming the metadata mess. *Proc. 29th Int. Conference on Data Engineering Workshops*, Brisbane, Australia, IEEE, 286–289, doi:10.1109/ICDEW.2013.6547465.
- , and D. Maier, 2011: Finding haystacks with needles: Ranked search for data using geospatial and temporal characteristics. *Scientific and Statistical Database Management*, J. Bayard Cushing, J. French, and S. Bowers, Eds., Vol. 6809, *Lecture Notes in Computer Science*, Springer, 55–72, doi:10.1007/978-3-642-22351-8_4.
- NOAA, 2009: Integrated Ocean Observing System data management and communications concept of operations, version 1.5. NOAA Rep., 94 pp. [Available online at www.ioos.noaa.gov/library/dmac_cops_v1_5_01_09_09.pdf.]
- Piasecki, M., and B. Beran, 2009: A semantic annotation tool for hydrologic sciences. *Earth Sci. Inf.*, **2**, 157–168, doi:10.1007/s12145-009-0031-x.
- Ramachandran, R., S. A. Christopher, S. Movva, X. Li, H. T. Conover, K. R. Keiser, S. J. Graves, and R. T. McNider, 2005: Earth science markup language: A solution to address data format heterogeneity problems in atmospheric sciences. *Bull. Amer. Meteor. Soc.*, **86**, 791–794, doi:10.1175/BAMS-86-6-791.
- Ruberg, S. A., and Coauthors, 2008: Societal benefits of the Real-Time Coastal Observation Network (ReCON): Implications for municipal drinking water quality. *Mar. Technol. Soc. J.*, **42**, 103–109, doi:10.4031/002533208786842471.
- Strayer, D. L., 2012: *The Hudson Primer: The Ecology of an Iconic River*. University of California Press, 224 pp.
- Tarboton, D. G., and Coauthors, 2009: Development of a community hydrologic information system. *Proc. 18th World IMACS Congress and MODSIM09 Int. Congress on Modelling and Simulation*, Cairns, Australia, Modelling and Simulation Society of Australia and New Zealand and International Association for Mathematics and Computers in Simulation, 988–994. [Available online at http://mssanz.org.au/modsim09/C4/tarboton_C4.pdf.]
- Taylor, P., 2012: OGC WaterML 2.0: Part 1—Time series. Open Geospatial Consortium Discussion Paper OGC 10-126r3, 149 pp. [Available online at www.opengis.net/doc/IS/waterml/2.0.]
- Vörösmarty, C. J., P. Green, J. Salisbury, and R. B. Lammers, 2000: Global water resources: Vulnerability from climate change and population growth. *Science*, **289**, 284–288, doi:10.1126/science.289.5477.284.
- Zaslavsky, I., D. Valentine, and T. Whiteaker, 2007: CUAHSI WaterML. Open Geospatial Consortium Discussion Paper OGC 07-041r1, 88 pp. [Available online at http://portal.opengeospatial.org/files/?artifact_id=21743.]
- Zheng, J., P. Wang, E. W. Patton, T. Lebo, J. S. Luciano, and D. L. McGuinness, 2011: A semantically-enabled provenance-aware water quality data portal. *Proc. Environmental Information Management Conf.*, Santa Barbara, CA, University of California, Santa Barbara, 151–156.