



Jun 16th, 10:40 AM - 12:00 PM

# HydroShare: Advancing Collaboration through Hydrologic Data and Model Sharing

David G. Tarboton

*Utah State University, dtarb@usu.edu*

Ray Idaszak

*RENCI, University of North Carolina at Chapel Hill, rayi@renci.org*

Jeffery S. Horsburgh

*Utah State University, jeff.horsburgh@usu.edu*

Jeff Heard

*RENCI, University of North Carolina at Chapel Hill, jeff@renci.org*

Dan Ames

*Brigham Young University, dan.ames@byu.edu*

*See next page for additional authors*

Follow this and additional works at: <https://scholarsarchive.byu.edu/iemssconference>

 Part of the [Civil Engineering Commons](#), [Data Storage Systems Commons](#), [Environmental Engineering Commons](#), [Hydraulic Engineering Commons](#), and the [Other Civil and Environmental Engineering Commons](#)

Tarboton, David G.; Idaszak, Ray; Horsburgh, Jeffery S.; Heard, Jeff; Ames, Dan; Goodall, Jonathan L.; Band, Larry; Merwade, Venkatesh; Couch, Alva; Arrigo, Jennifer; Hooper, Richard; Valentine, David; and Maidment, David, "HydroShare: Advancing Collaboration through Hydrologic Data and Model Sharing" (2014). *International Congress on Environmental Modelling and Software*. 7. <https://scholarsarchive.byu.edu/iemssconference/2014/Stream-A/7>

This Event is brought to you for free and open access by the Civil and Environmental Engineering at BYU ScholarsArchive. It has been accepted for inclusion in International Congress on Environmental Modelling and Software by an authorized administrator of BYU ScholarsArchive. For more information, please contact [scholarsarchive@byu.edu](mailto:scholarsarchive@byu.edu), [ellen\\_amatangelo@byu.edu](mailto:ellen_amatangelo@byu.edu).

---

**Presenter/Author Information**

David G. Tarboton, Ray Idaszak, Jeffery S. Horsburgh, Jeff Heard, Dan Ames, Jonathan L. Goodall, Larry Band, Venkatesh Merwade, Alva Couch, Jennifer Arrigo, Richard Hooper, David Valentine, and David Maidment

# HydroShare: Advancing Collaboration through Hydrologic Data and Model Sharing

**David G Tarboton<sup>a</sup>, Ray Idaszak<sup>b</sup>, Jeffery S Horsburgh<sup>a</sup>, Jeff Heard<sup>b</sup>, Dan Ames<sup>c</sup>, Jonathan L Goodall<sup>d</sup>, Larry Band<sup>e</sup>, Venkatesh Merwade<sup>f</sup>, Alva Couch<sup>g</sup>, Jennifer Arrigo<sup>h</sup>, Richard Hooper<sup>h</sup>, David Valentine<sup>i</sup>, David Maidment<sup>j</sup>**

*a. Utah State University ([dtarb@usu.edu](mailto:dtarb@usu.edu), [jeff.horsburgh@usu.edu](mailto:jeff.horsburgh@usu.edu)), b. RENCI, University of North Carolina at Chapel Hill ([rayi@renci.org](mailto:rayi@renci.org), [jeff@renci.org](mailto:jeff@renci.org)), c. Brigham Young University ([dan.ames@byu.edu](mailto:dan.ames@byu.edu)), d. University of Virginia ([goodall@virginia.edu](mailto:goodall@virginia.edu)), e. University of North Carolina at Chapel Hill ([lband@email.unc.edu](mailto:lband@email.unc.edu)) f. Purdue University ([vmerwade@purdue.edu](mailto:vmerwade@purdue.edu)), g. Tufts University ([couch@cs.tufts.edu](mailto:couch@cs.tufts.edu)), h. CUAHSI ([jarrigo@cuahsi.org](mailto:jarrigo@cuahsi.org), [RHooper@cuahsi.org](mailto:RHooper@cuahsi.org)), i. San Diego Supercomputer Center ([valentin@sdsc.edu](mailto:valentin@sdsc.edu)), j. University of Texas at Austin ([maidment@mail.utexas.edu](mailto:maidment@mail.utexas.edu))*

**Abstract:** HydroShare is an online, collaborative system being developed for open sharing of hydrologic data and models. The goal of HydroShare is to enable hydrology researchers to easily discover and access hydrologic data and models, retrieve them to their desktop for local analysis and perform analyses in a distributed computing environment that may include grid, cloud or high performance computing. Users may also share and publish outcomes (data, results or models) into HydroShare, using the system as a collaboration platform. HydroShare is expanding the data sharing capability of the CUAHSI Hydrologic Information System by broadening the classes of data accommodated. HydroShare will take advantage of emerging social media functionality to enhance information about and collaboration around hydrologic data and models. One of the fundamental concepts in HydroShare is that of a resource. All content is represented using a Resource Data Model that has elements common to all resources as well as elements specific to the types of resources HydroShare will support. These will include different data types used in the hydrology community and models and workflows that require metadata on execution functionality. The HydroShare web interface and social media functions are being developed using the Django web application framework. A geospatial visualization and analysis component enables searching, visualizing, and analyzing geographic datasets. The integrated Rule-Oriented Data System (iRODS) is being used to manage federated data content and perform rule-based background actions on data and model resources, including the execution of models and workflows. This paper introduces the HydroShare functionality developed to date and elaborates on the representation of hydrologic data and models in this system as resources for collaboration.

**Keywords:** Data Sharing; Model Sharing; Information Model; Web Services.

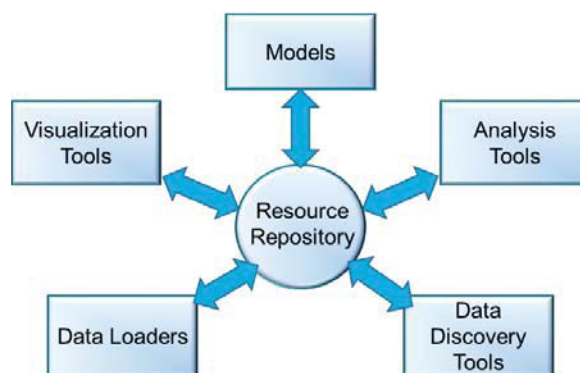
## 1 INTRODUCTION

Hydrologic information is collected by many individuals and organizations in government and academia for many purposes, including general monitoring of the condition of the water environment and specific investigations of hydrologic processes and environments. It is thus dispersed and heterogeneous. Advancing understanding in hydrology requires discovery, access to, and integration of data and information from multiple sources. It requires integrated modeling to codify and synthesize knowledge in a form that is testable through reproducible comparisons to data. It requires collaboration. Data and modeling information technology systems, or cyberinfrastructure, are required to address these problems and enhance the ability of hydrologic scientists to collaborate by sharing data and models. HydroShare is a system being developed to meet these needs. HydroShare will provide a community collaboration web site that enables users to easily discover and access data and models, retrieve them to a desktop computer or perform analyses in a distributed computing

environment that includes grid, cloud, or high performance computing model instances as necessary. We envision that HydroShare will enable more rapid advances in hydrologic understanding through collaborative data sharing, analysis, and modeling. Understanding will be advanced through the ability to integrate information from multiple sources. Outcomes (data, results, models) can then be published as new resources that can be shared with collaborators. Our goal is to make sharing of hydrologic data and models as easy as sharing videos on YouTube or shopping on Amazon.com.

The Consortium of Universities for the Advancement of Hydrologic Science Inc. (CUAHSI) is an organization representing 100+ universities and affiliated organizations, funded by the US National Science Foundation, to develop community infrastructure and services to advance hydrologic science. The CUAHSI Hydrologic Information System (HIS) (Tarboton et al., 2009) is a services-oriented-architecture established to support the sharing of hydrologic data. It is comprised of hydrologic databases and servers connected through web services as well as software for data publication, discovery and access that provides mechanisms for publishing, cataloging, discovering and accessing information using standardized web services. HIS supports the storing of point observations data using the Observations Data Model (ODM) (Horsburgh et al., 2008); sharing data through well-defined web services using a HydroServer (Conner et al., 2013; Horsburgh et al., 2010); and the discovery and integration of this information through the open source HydroDesktop client (Ames et al., 2012). HydroShare is a software development research project to expand and enhance the data sharing capability of the CUAHSI HIS by broadening the classes of data accommodated, expanding capability to include the sharing of models and model components, and taking advantage of emerging social media functionality to enhance information about and collaboration around hydrologic data and models.

The development of HydroShare is supported by the U.S. National Science Foundation (NSF) and involves eight U.S. universities, RENCi, and CUAHSI. HydroShare is still at an early stage of development. In this paper we describe the use cases that are driving the development of HydroShare. The objects that HydroShare works with are referred to as "resources". These can be datasets in any of the supported formats, models, or just generic resources comprising a file or files whose structure is not recognizable to and parseable by HydroShare. In this resource centric approach (Figure 1), tools or models perform actions (analyses) on resources in the resource repository that are in a standard format. This enables interaction among multiple models through the resource repository and allows a modeler to focus on modeling while taking advantage of standardized analysis, visualization loading and discovery tools.



**Figure 1.** Resource Repository Centric Paradigm for Modeling and Analysis.

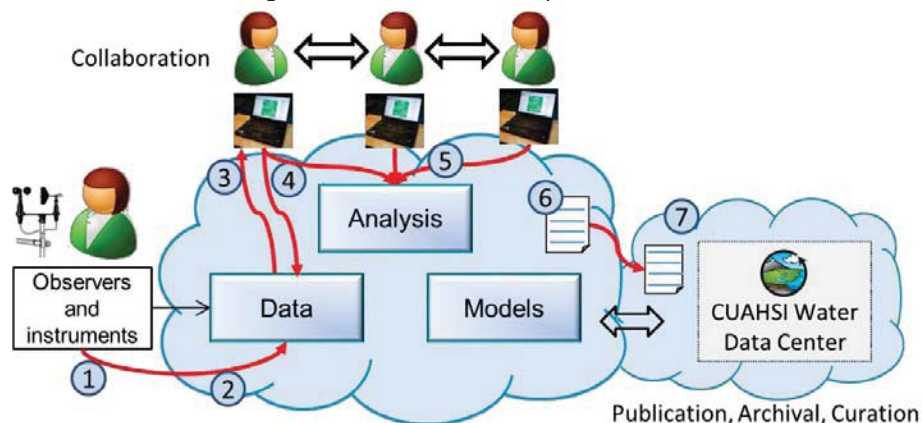
HydroShare is being developed using an open development model with community participation in code development managed in GitHub (<http://github.com/hydroshare/>). The development roadmap (<http://github.com/hydroshare/hydroshare2/wiki>) outlines the functionality anticipated for each release. It is challenging to commit to specific release dates for a project such as this, but release 1 that supports the basic resource sharing functionality is targeted for June 2014, with subsequent incremental releases during the second half of 2014 that add capability for additional resource types.

The way that data is structured can enhance or inhibit the analysis that it can support. The data model for representing resources, is thus critical, and we describe the resource data model we are using to structure the representation of data. We then give a brief overview of the architecture and framework being used to build HydroShare and conclude with remarks on how HydroShare will contribute to advances in hydrology through collaboration.

## 2 USE CASES

### 2.1 Collaborative Data Analysis and Publication Use Case

The first use case driving the development of HydroShare is that of collaborative data analysis and modeling (Figure 2) extending existing CUAHSI HIS data sharing functionality into a dynamic collaborative environment leading to the eventual archival publication of data.



**Figure 2.** Collaborative data analysis use case.

At (1) data are observed and then loaded (2). In CUAHSI HIS data is loaded into an ODM relational database on a HydroServer that publishes it using web services. Metadata is harvested by the HIS Central catalog, and supports geographic and context based data discovery. A HydroDesktop client user (3) discovers, downloads and analyzes the data, or uses it in a model. Steps 1 to 3 are supported by the existing CUAHSI HIS. HydroShare picks up from here allowing the user to next post the results (data and model) to HydroShare as resources, retaining provenance information on the original data source (4). This will be done through sharing features being added to HydroDesktop. HydroShare will also support direct entry of new resources. Upon ingestion, background actions parse metadata and enable analysis based on rules and policies. The user shares posted resources with colleagues (5), designating who has permission to access the resources. A group collaborates on refining the analysis, model or result. HydroShare tracks provenance supporting reproducibility and transparency. After iteration, the result is finalized and submitted for publication (6). At this point the resources produced (data, model, workflow, paper) are made immutable, access is opened and permanent persistent identifiers (e.g., DOIs) are assigned. The data may be moved to a permanent repository under the auspices of the CUAHSI Water Data Center (7).

### 2.2 Collaborative Integrated Modeling Use Case

Collaborative integrated modeling is another use case driving the development of HydroShare. HydroShare will support the sharing of models, scripts, and tools as resources, the pre-processing of model inputs based on nationally available and HydroShare resources, and the execution of models using HPC or cloud services.

#### 2.2.1 Model Metadata

Collaborative sharing and publication of models requires development of a resource content model for models and their inputs and outputs, as well as catalog functionality to support the discovery of

different model resources. Model resources will include a standard metadata template that will enable HydroShare users to discover, use, and visualize model resources within HydroShare. Publication of models as HydroShare resources will also enable users to create links with related resources, such as GIS data for topography and land use that were used to create that model resource. Morsy et al. (2014) presents model resource types in HydroShare and a team-based hydrologic modeling use case mapped to HydroShare model resource types with Dublin Core metadata properties. HydroShare developers are members of a number of Community Surface Dynamics Modeling System (CSDMS, 2010) working groups and are factoring elements from the CSDMS basic model interface (Peckham et al., 2013) and the CSDMS Standard Names (Peckham, 2014) into the design of model resource metadata and functionality to make HydroShare interoperable with CSDMS so as to take advantage of the model integration capability CSDMS has advanced.

### **2.2.2 Model Input Data Pre-processing**

Distributed watershed models require a variety of different spatial data from varied sources, including government agencies and individual researchers. Assembling and rectifying spatial data on soils, topography, hydrography, land cover and other key layers is one of the most time consuming and error prone steps in model development. HydroShare will support workflows to automate these procedures that, for example begin with identification of a watershed outlet, followed by the delineation of the stream network and watershed extent using information harvested directly from the National Hydrography Dataset. Then data is drawn from additional data sets from USGS, NRCS, EPA and other agencies while maintaining full provenance. The raw geographic information is rectified and processed to develop parameter files for a model such as RHESSys (Tague and Band, 2004), SWAT, and VIC. This reduces the pre-processing time and increases reproducibility, allowing a researcher to concentrate on higher level analysis and decision making.

### **2.2.3 Model Execution**

We will also provide gateway capability for access to high performance computing to execute models using remote cluster, grid, or cloud resources. Preliminary capability developed as part of SWATShare (Merwade et al., 2013) is being adapted for use in HydroShare, and extended to other models. A typical use case may include a user creating a HydroShare supported hydrologic model and submitting it to HydroShare for calibration using cloud services. Once the model is calibrated and verified, they will perform analyses, publish the findings in a journal, and then publish the model on HydroShare. Once the model is published on HydroShare, another user will discover the model and use it to answer a different research question by collaborating with the first user. During this collaborative work, both users will use HydroShare to run model simulations, visualize outputs, and store all the results as resources linked to the original model. Multiple instances of such shared collaborative models will also become ideal resources to use as educational tools for teaching hydrology in classrooms.

## **3 FACILITATING COLLABORATION**

At a technical level HydroShare will be a tool (or set of tools) that can be used for collaboration. However promoting the collaboration that we believe is necessary to accelerate hydrologic science is not only a technical problem but involves social elements related to the culture of sharing, rewards for sharing and building trust in the system (Bennett and Gadlin, 2012). The access control model for HydroShare is designed to be simple, transparent and open so that users of the system know who has access to their data. It is designed to give users the ability to protect their data, sharing only with trusted collaborators or groups in the initial stages, but then to make it easy to publish the data in a form where they can get credit through data citations (based on digital object identifiers) when it is ready for publication. Tools are designed to provide best practice hydrology value added functionality to encourage participation. Users will be attracted because of the functionality and value it provides directly to them. The social collaboration functionality involving commenting and rating has been designed to be simple and promote positive behavior. The metadata model for social collaboration



information has been designed so that statistical information on collaboration can be quantified to gauge the success of the system and inform and prioritize future developments.

#### 4 RESOURCE DATA MODEL

In HydroShare all content is persisted as a "resource". The Resource Data Model is based on the Open Archives Initiative's Object Reuse and Exchange (OAI-ORE) standard (Lagoze *et al.*, 2008), which is a standard for the description and exchange of aggregations of web resources. The OAI-ORE Abstract Data Model is described at <http://www.openarchives.org/ore/1.0/datamodel.html>. In this model, resources may be comprised of a single content file or an aggregation of multiple content files. Resources containing multiple content files may have a hierarchical file/directory structure. Each resource is described by "science" metadata, which is a separate unit of digital content that details the properties of the resource (i.e., resource level metadata). Each content file within a resource may be separately described by a "science" metadata document that is considered to be part of the resource content (i.e., content level metadata). Each resource is accompanied by "system" metadata that contains system level attributes of the resource, including time stamps, ownership, access control rules, etc. Persistent identifiers, access control, versioning, sharing, and cataloging for discovery are all managed at the resource level in HydroShare.

HydroShare will promote and develop value-added tools for a standard set of resource types. For the set of supported resource types, HydroShare will provide tools that enable users to open, visualize, convert, analyze, and otherwise manipulate the contents of resources conforming to the known types. HydroShare will not prevent users from uploading resources containing objects that are unknown to the system, but will not provide any value added functionality for those resources other than allowing users to upload them, describe them with metadata, set access control permissions on them, share them, and download them.

Following are the resource types planned for the initial HydroShare releases.

- **Generic Resource:** A generic resource is a package of one or more files for which HydroShare does not know the specific type. HydroShare will treat generic resources as opaque objects that can be created and shared, but do not have specific, value added functionality. Generic resources support sharing of content outside of the specific types listed below.
- **Time series:** Time series of hydrologic observations that conform to the information model (but not necessarily the formal schema) of ODM.
- **Referenced CUAHSI HIS data series:** A link to a HydroServer URL endpoint that represents a data series hosted on a CUAHSI HIS HydroServer.
- **Geographic features:** Points, lines, or polygons with their associated attributes.
- **Geographic raster:** Georeferenced grids containing datasets such as land cover, elevation, elevation derivatives.
- **Multidimensional Space/Time Data Sets:** Continuous space/time grids such as radar-based rainfall data.
- **Geochemistry/Sample-Based Observations:** Water quality and/or solid earth or other samples that may conform to the ODM2 information model (<https://github.com/UCHIC/ODM2>).

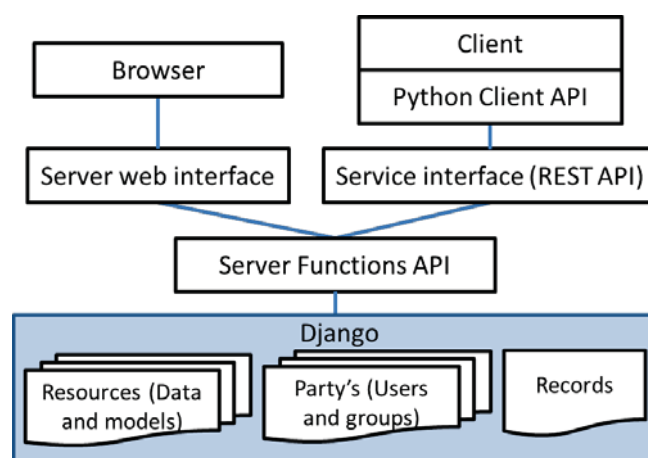
Other types of resources intended for later inclusion in HydroShare are composite resources, river geometry, and HydroDesktop project packages. These enable the representation of more complex and structured data and model content drawing upon the basic resource types above. To support modelling resource types to represent scripts, workflows, model programs, modules and components are under development. Documents (e.g. Microsoft Office or PDF) and tabular objects will support documentation and model inputs and outputs. HydroShare will also include the capability to reference resources held elsewhere and published using standard web services, thereby developing a capability for interoperability and collaboration involving these resources without requiring the data to be in HydroShare. Open Geospatial Consortium (OGC) compliant Web Services as well ESRI ArcGIS geospatial data services and THREDDS/OpenDap data services are under investigation here.

Where possible, HydroShare will encourage users to adopt the content data models for resource types that are supported and for which functionality has already been developed. The motivation for users will be that HydroShare will provide value added tools (e.g., visualization, processing, analysis,

transformation) for supported resource types, whereas HydroShare would treat unknown resource types as opaque objects with no such functionality provided.

## 5 ARCHITECTURE

The system will be designed using an architecture that separates the web application interface layer from the service layer, exposing as much of the functionality as possible through an application programmers interface to enable direct client access and interoperability with other systems (Figure 3).



**Figure 3.** HydroShare high level architecture

HydroShare is being built using a Python-based software stack and the Django Web Application Framework (Django, 2014). Django provides a high level web framework with RDBMS connectivity. Multiple Python libraries have been selected to support data manipulation and storage, content management, numerical analysis, RESTful APIs, search and discovery and visualization. iRODS (IRODS, 2010) provides federated resource storage replicated across multiple sites, a rule engine for policy driven data management via micro-services, scientific reproducibility and provenance, and execution of user workflows on remote HPC resources. The open source .Spatial based HydroDesktop software provides desktop client functionality including discovery and download for CUAHSI HIS time series and extensibility through its plugin architecture with plugins for R integration and modeling

## 6 CONCLUSIONS

We envision that HydroShare will enable more rapid advances in hydrologic understanding through collaborative data sharing, analysis, and modeling. HydroShare will provide a community collaboration site that enables users to easily discover and access data and models, retrieve them to a desktop computer or perform analyses in a distributed computing environment that includes grid, cloud, or high performance computing model instances as necessary. Understanding will be advanced through the ability to integrate information from multiple sources. Outcomes (data, results, models) can then be published as new resources that can be shared with collaborators. The provenance and metadata that HydroShare maintains will enhance reproducibility and transparency of the research conducted using HydroShare resources.

## ACKNOWLEDGMENTS

This work was supported by the National Science Foundation under collaborative grants OCI-1148453 and OCI-1148090 for the development of HydroShare (<http://www.hydroshare.org>). Any



opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

## REFERENCES

- Ames, D.P., Horsburgh, J.S., Cao, Y., Kadlec, J., Whiteaker, T., Valentine, D., 2012. HydroDesktop: Web services-based software for hydrologic data discovery, download, visualization, and analysis. *Environmental Modelling & Software* 37 146-156, <http://dx.doi.org/10.1016/j.envsoft.2012.03.013>.
- Bennett, L.M., Gadlin, H., 2012. Collaboration and Team Science: From Theory to Practice. *J Investig Med* 60(5) 768-775, <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3652225/>.
- Conner, L.G., Ames, D.P., Gill, R.A., 2013. HydroServer Lite as an open source solution for archiving and sharing environmental data for independent university labs. *Ecological Informatics* 18(0) 171-177, <http://dx.doi.org/10.1016/j.ecoinf.2013.08.006>.
- CSDMS, 2010. Community Surface Dynamics Modeling System, <http://csdms.colorado.edu/wiki/>.
- Django, 2014. Django - A high-level Python Web framework that encourages rapid development and clean, pragmatic design, <https://www.djangoproject.com/>.
- Horsburgh, J.S., Tarboton, D.G., Maidment, D.R., Zaslavsky, I., 2008. A Relational Model for Environmental and Water Resources Data. *Water Resour. Res.* 44 W05406, <http://dx.doi.org/10.1029/2007WR006392>.
- Horsburgh, J.S., Tarboton, D.G., Schreuders, K.A.T., Maidment, D.R., Zaslavsky, I., Valentine, D., 2010. Hydroserver: A Platform for Publishing Space-Time Hydrologic Datasets, 2010 AWRA Spring Specialty Conference Geographic Information Systems (GIS) and Water Resources VI. American Water Resources Association, Middleburg, Virginia, TPS-10-1: Orlando Florida, [http://www.awra.org/tools/members/Proceedings/1003conference/doc/abs/JefferyHorsburgh\\_7cb420e3\\_6602.pdf](http://www.awra.org/tools/members/Proceedings/1003conference/doc/abs/JefferyHorsburgh_7cb420e3_6602.pdf).
- IRODS, 2010. IRODS:Data Grids, Digital Libraries, Persistent Archives, and Real-time Data Systems, <https://www.irods.org/>.
- Lagoze, C., Van de Sompel, H., Johnston, P., Nelson, M., Sanderson, R., Warner, S., 2008. Open Archives Initiative Object Reuse and Exchange: ORE User Guide – Primer, <http://www.openarchives.org/ore/1.0/primer>.
- Merwade, V., Song, C., Zhao, L., Zhe, S., Rajib, M.A., 2013. SWATShare – A portal for sharing, publishing, and running SWAT models using XSEDE resources, Proceedings of the 2013 SWAT Users Conference and Workshops: Toulouse, France, July 2013, <http://water-hub.org/swatshare>.
- Morsy, M.M., Goodall, J.L., Bandaragoda, C., Castronova, A.M., Greenberg, J., 2014. Metadata for Describing Water Models, iEMSs Conference: San Diego, CA, June 15-19, 2014 (submitted).
- Peckham, S.D., 2014. The CSDMS Standard Names: Cross-Domain Naming Conventions for Describing Process Models, Data Sets and Their Associated Variables, iEMSs Conference: San Diego, CA, June 15-19, 2014.
- Peckham, S.D., Hutton, E.W.H., Norris, B., 2013. A component-based approach to integrated modeling in the geosciences: The design of CSDMS. *Computers & Geosciences* 53(0) 3-12, <http://dx.doi.org/10.1016/j.cageo.2012.04.002>.
- Tague, C.L., Band, L.E., 2004. RHESys: Regional Hydro-Ecologic Simulation System: An Object-Oriented Approach to Spatially Distributed Modeling of Carbon, Water, and Nutrient Cycling. *Earth Interactions* 8(19) 1-42, [http://dx.doi.org/10.1175/1087-3562\(2004\)8<1:RRHSSO>2.0.CO;2](http://dx.doi.org/10.1175/1087-3562(2004)8<1:RRHSSO>2.0.CO;2).
- Tarboton, D.G., Horsburgh, J.S., Maidment, D.R., Whiteaker, T., Zaslavsky, I., Piasecki, M., Goodall, J., Valentine, D., Whitenack, T., 2009. Development of a Community Hydrologic Information System, In: Anderssen, R.S., Braddock, R.D., Newham, L.T.H. (Eds.), 18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand and International Association for Mathematics and Computers in Simulation, pp. 988-994, July, [http://www.mssanz.org.au/modsim09/C4/tarboton\\_C4.pdf](http://www.mssanz.org.au/modsim09/C4/tarboton_C4.pdf).