International Environmental Modelling and Software Society (iEMSs)
7th Intl. Congress on Env. Modelling and Software, San Diego, CA, USA,
Daniel P. Ames, Nigel W.T. Quinn and Andrea E. Rizzoli (Eds.)
http://www.iemss.org/society/index.php/iemss-2014-proceedings

Web Applications that Share Level-12 HUC Data and Models of the CONUS

Lorne Leonard[a] , Christopher J Duffy[a]
(a) Department of Civil & Environmental Engineering, The Pennsylvania State University, United States of America *{ lnl3@psu.edu, cxd11@psu.edu }*

Abstract: HydroTerre web applications and services provide the Essential Terrestrial Variable (ETV) datasets to create common hydrological models anywhere in the continental United States (CONUS). This service allows web users to download data for their own purposes in their own computing environment. The datasets are provided using standard Geographic Information System formats and the data transformation is dependent on the users own needs, goals, and computing environment. In this article, we demonstrate three web applications and web services that share data and models with users via web interfaces, which automate the data-transformations for United States Geological Survey level-12 Hydrological Unit Codes (HUC-12) in order to be consumed in hydrological models. Penn State Integrated Hydrological Model (PIHM) is demonstrated here, but the workflows serve as a template for other models to adapt and become new services. The emphasis of this article is to demonstrate the feasibility of sharing both model input and output data via web interfaces that capture users' provenance. Capturing provenance serves as a new data resource to share amongst modelers, which we believe will improve modeling reproducibility and shift the focus from data preparation to data analysis. We want to demonstrate that workflows empower modelers to create hydrological models rapidly anywhere in the CONUS and the interface to do so is a critical part of the process. To explain the interface, an explanation of both hardware and software architecture is required, as the way they are coupled is critical for performance and constrain the interfaces that execute workflows in a distributed computing environment.

***Keywords**:* Workflows; HydroTerre; Web services.

1 INTRODUCTION

HydroTerre (www.hydroterre.psu.edu) provides geospatial data sources to support distributed hydrologic models at the HUC-12 scale within the CONUS as described by Leonard & Duffy (2013). The HUC-12 is a sub watershed typically ranging from 10 to 40,000 acres in size. The geospatial data includes soils, land-cover, elevation, stream networks, and time-series North American Land Data Assimilation System NLDAS (2011) climate forcing. In this article, we briefly explain the system design that supports these web services followed by a description of three services to gain access to and share data. The first service describes how web users can download an individual bundle of ETV data for any CONUS HUC-12. The second service explains how a web user can select all upstream HUC-12s based on one selected HUC-12 to determine disk and computation needs for the PIHM model. The final service demonstrates how a web user can select HUC-12s by an administration region to retrieve an ETV data bundle and create PIHM input and output datasets.

1.1 Scope

This article focuses on the web user interfaces that share HUC-12 data and models within the CONUS to illustrate the importance of provenance to improve confidence in hydrological modeling. The reader is referred to Leonard & Duffy (2013) for an explanation of the ETV data workflows that support these web interfaces. Leonard & Duffy (2014a) will provide the reader with a detailed explanation of the data-model workflows that the interfaces control and execute. The second service relies on a depth-first graph derived from the National Hydrography Dataset (NHD) provided by the United States Geological Survey (USGS) and is explained in detail at Leonard & Duffy (2014b). Here

we are focusing on the interface that enables users to download upstream HUC-12 bundles for their own modeling needs. It is beyond the scope of this article to discuss hydrological model calibration and the validity of the model results. Due to administrative restrictions, executing model workflows discussed in the article may not be available to the public at times, but ETV data workflows are available. In addition, the data workflows presented here are restricted to one HUC-12 selected by the web user.

2 SYSTEM DESIGN

The web interfaces described in this article control workflows and web services are supported on a distributed computing environment. The data-model workflows are implemented in a three-tier hardware layer system (Figure 1), (1) web interface, (2) data support and (3) model development. Web application interfaces remove the complexity of consuming service-oriented architecture and communication methods between these three tiers that depend on Simple Object Access Protocol as defined by World Wide Web Consortium (2014a), Representational State Transfer described in Fielding & Taylor (2002), Web Services Description Language as defined by World Wide Web Consortium (2014b), and Database Markup Language by Microsoft (2014a). The hardware has been organized to support large volumes of data required to support data-model workflows anywhere in the CONUS. The data-model processing is distributed within the data-tier of the HydroTerre system, and the hydrological modeling is distributed to other High Performance Computing (HPC) systems to compute PIHM models. Service-oriented architecture that is efficient and robust is critical to support the rapid prototyping and distribution of the data-model workflows as is demonstrated in service three.

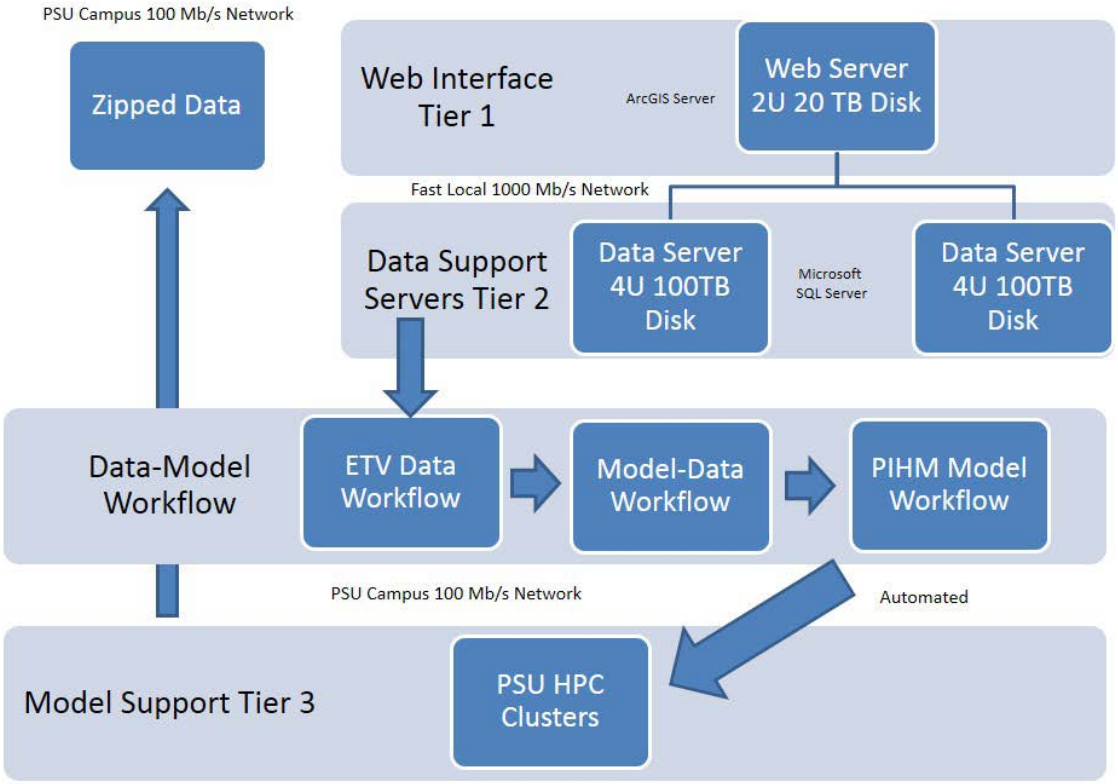

Figure 1. Tier-one supports web applications, tier-two supports the data services, and tier-three supports the model development. Both tiers one and two support the data-model workflows.

2.1 Hardware and Administration Layers

The web interface tier hosts the web applications and services. ArcGIS server software by ESRI (2014) was used for serving spatial data and because it contains

development kits that support both GIS web applications and SQL server databases Microsoft (2014b) for data storage and querying. The web server disk is partitioned into three components for maximum performance; the windows server operating system, GIS software that performs select and query operations, and the third partition for workflow results that are temporarily stored for users to download via a web link. Two 100-terabyte data support servers constitute the second tier, with each server containing multiple processors. These servers contain thousands of datasets that form the 30-year NLDAS-2 climate forcing dataset. Microsoft SQL server is used to store and query datasets in the second tier. Components of the data workflows that retrieve the forcing data are implemented on this tier and form the first layer of the distributed computing system. All data retrieving and processing components are executed in parallel for maximum performance that range from minutes to hours depending on the catchment size. The reader is referred to Leonard & Duffy (2013) for further details about compute times and data sizes of ETV datasets.

The web and data tiers are tightly connected via a private fast network router to minimize performance lost when retrieving datasets between servers. The data-model workflows reside on these two tiers and are explained further in Leonard & Duffy (2014a). Both ETV and data-model workflow results, which are compressed and zipped, reside on the web interface tier. Web users gain access to these results via a public network connection. The model support tier also gains access to the zipped data using the same public network. The model is executed automatically after the data-model transformation that transforms the ETV data into model input files using a HPC environment. Using a specified PIHM account, a custom PIHM dispatcher application runs continuously and uses web services to retrieve PIHM jobs using Penn State Universities CyberSTAR (2014) cluster. In this mode, the user does not need to login to the compute environment and the PIHM models are automatically dispatched to the compute nodes.

Job management is achieved via the data tier, with all compute nodes accessing job tasks from this tier, via user-project databases. Statistics about data and the model performance are returned to the web tier database to inform the user of the workflow status. Access to the model results is automatically sent back to the web interface tier. The computer names are recorded as part of provenance data when a PIHM job is executed. Thus, the exact workflow can easily be replicated on the exact distributed computer. Having the ability to replicate the entire workflow is extremely useful for debugging purposes when trying to determine if data is being lost or corrupted via network connections. Another issue for debugging is determining whether hardware and/or operating systems are causing problems with the model jobs. To the reader this may appear unnecessary, but when hundreds of thousands of model jobs are running within different types of HPC environments, having this information is valuable to understand failure. Hardware provenance is important, as is data and software provenance in order for web users to replicate all the steps when constructing and reconstructing shared hydrological models. The web interface handles these tasks for the user, making the distributed hardware and software implementation appear as one identity.

3 SERVICES TO SHARE DATA AND MODELS

HydroTerre has been designed to create and model HUC-12 watersheds in the CONUS as rapidly as possible. To do so requires HPC hardware and the software architecture has to be designed and developed to take advantage of such hardware. Thus, all HydroTerre software components need to be tuned to run in parallel, be robust and efficient, and able to take advantage of memory, central processing units, and disk speeds. In addition, how data is kept and structured is a critical part of the strategy to make sharing data and models efficient with web users. These are tasks that must be considered when designing the web interface for sharing data and models with users. This section describes three services for users to gain access to data. The first is providing rapid access to any individual CONUS HUC-12 ETV data bundle that can be used for any GIS and modeling purposes. The second explains how a user can select all upstream HUC-12s from a selected HUC-12 and calculate data requirements for PIHM. The final service demonstrates how a web user can select HUC-12s and go through the entire process from ETV data bundle to data-model transformation and executing the PIHM model on a distributed computing environment.

3.1 Service 1: ETV data bundle for individual level-12 HUCs

The ETV workflow is a service layer with the sole purpose of retrieving data rapidly for any HUC-12 in the CONUS with minimal interaction from the user. The web application interface consumes the ETV service layer by asking the user to select a HUC-12, provide an email address, and specify the forcing period (http://hydroterre.psu.edu/Development/HydroTerre_National/HydroTerre_National.aspx). The application code is responsible for retrieving and visualizing geospatial data using ArcGIS Server software. The application stores the provenance of what data services are being called, the explicitly defined user parameters (for example forcing period), and the implicitly defined parameters (for example the bounding box of the selected HUC-12). The bounding box is then used to clip and extract soil, geology, land cover, and elevation from the national datasets as described in Leonard & Duffy (2013) to produce GeoTIFFs and text files. When possible, existing ArcGIS tools are used to do common GIS operations such as select and clip if the performance is adequate and the tools take advantage of the HPC environment. However, new tools are needed, for example how to handle the distributed forcing datasets that consist of millions of tables on the data tier. The forcing tool is responsible for identifying which forcing cells overlap the HUC-12, then in parallel, query the forcing variables (for example precipitation and temperature) and generate an XML file for easy sharing. These processes take place on tiers one and two outlined in Figure 1. Once the data workflow is completed, the user is notified via email where the zipped file can be downloaded. On average, a web user gains access to a one-year forcing period in less than two minutes.

3.2 Service 2: ETV data bundles for all upstream level-12 HUCs

The second service builds upon the first, with a web service that determines disk storage needs upstream of a selected HUC-12 from any CONUS HUC-12 using a depth-first hierarchy. The selection hierarchy is an important step towards database driven large-scale hydrological modeling on a distributed computing environment using HPC. The web interface executes a data workflow, on the data tier, to create a depth-first graph built on the fly using the selected HUC-12 as the root node. Access to the HUC-12 selection hierarchy is available via a web browser at the following address http://www.hydroterre.psu.edu/Development/HUC12_hierarchy. The steps to start this service are similar to service one and the processes take place on tiers one and two, with the web application storing user provenance for reproducibility purposes. Within seconds to less than five minutes for the largest catchment (Mississippi scale), an xml selection hierarchical data file is generated. We use this service to demonstrate the importance of how data is stored and structured, which is critical for performance and sharing data. Briefly, NHD stores one HUC designation key per HUC object in shape files, which is a fundamental problem when many HUC-12s have more than one designation. Using a shape file makes it difficult to check that HUC designations are adjacent, and to select all the geometry objects at the Mississippi scale takes tens of minutes. By extracting the HUC key, neighboring HUC-12 keys and the designation data from the shape file into a database, the issue of identifying missing HUC-12 keys, and designations that were not adjacent became evident within seconds. The reader is referred to article Leonard & Duffy (2014b) regarding further technical issues resolved to generate the depth-first graphs. The selection hierarchy generated is an xml file due to its flexibility of sharing data independent of software and operating systems.

The intention of this service is to understand what resources are needed to distribute hydrological modeling in cloud and other HPC environments using HUC-12 configurations. A simple example is shown in Figure 2.

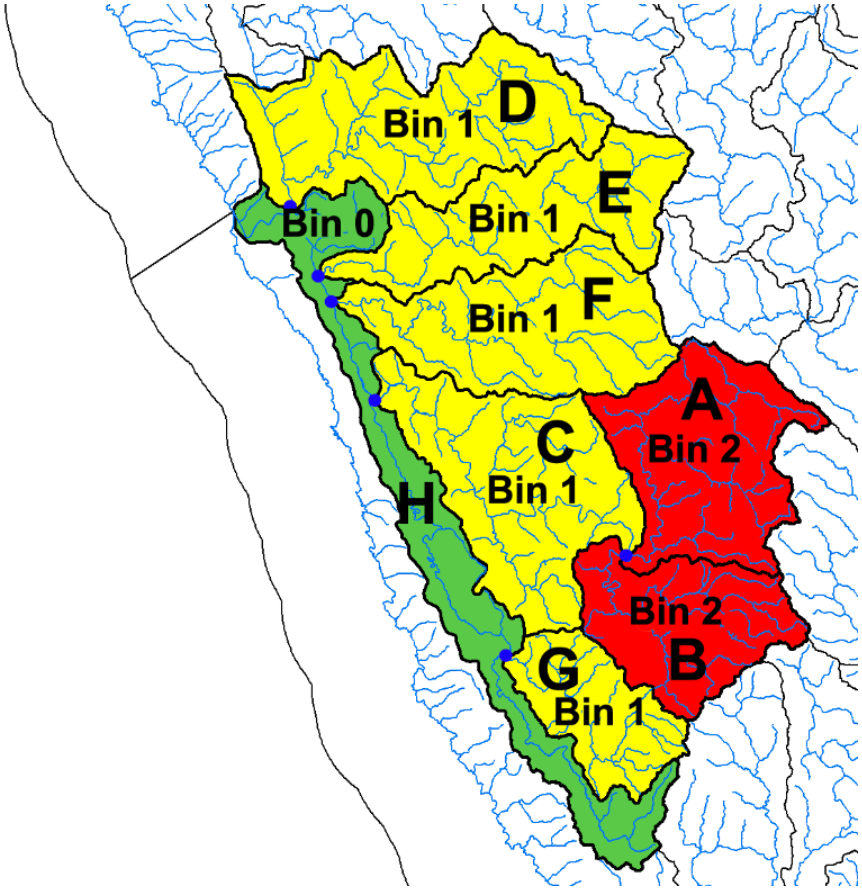

Figure 2. Demonstrating traversing HUC-12 graph geodesic hierarchy with PIHM. In this simple example, HUC-12s A and B are simulated first. Both A and B feed into C. Then, bin with graph geodesics 1 are simulated (D,E,F,C,G). Finally, graph geodesic 0 (H) is simulated which is fed by graph geodesic 1 (D,E,F,C,G). This process is repeated per time step in the simulation model. As HUC-12s are relatively small to simulate in this example, each bin graph geodesic could be easily simulated on 3 computers; 1 computer simulating each graph geodesic.

There are three graph geodesics (0, 1, 2), thus there are three “HUC12” bins. Each HUC-12 polygon is treated as a graph node and the intention is to graph the shortest paths between HUC-12s. In this example, there are no closed basins. Bin 2 contains HUC-12s (A 180101090103, B 180101090102). Bin 1 contains (D 180101090107, E 180101090106, F 180101090105, C 180101090104, and G 180101090101). Bin 0 contains the root node (H, 180101090108). With PIHM, time step 0 with bin 2 HUC-12s is simulated first. Then catchment boundary and stream data, passed as PIHM output files, are fed to the HUC-12s in bin 1. In particular, A to C, B to C. All the HUC-12s in bin 1 are simulated at time step 0. Data generated from HUC-12s in bin 1 are fed to the HUC-12s in bin 0. In particular, D to H, E to H, F to H, C to H, and G to H. This process is repeated for the entire simulation. All the connections between HUC-12s are listed in each “HUC12” object, which is necessary to modify PIHM input files to specify boundary conditions at the stream connection locations. Thus, with naïve parallelization, only three computers are necessary to simulate with PIHM on all eight HUC-12s. Before executing this simulation, one would expect the web application to calculate the amount of disk storage required. On average, 10 gigabytes of data are required as PIHM inputs per HUC-12 for 30-year duration. Thus, in this example the software package can quickly determine that 8 HUC-12s will require 80 gigabytes of storage. The same process is necessary for PIHM model outputs, 30 gigabytes of output per HUC-12 are needed, totaling 240 gigabytes. Thus, a total storage of 320 gigabytes of storage are required for all eight HUC-12s. This is a trivial example, but the web interface strategy to share the xml graph hierarchy of any CONUS HUC-12 simplifies the process of understanding data and model needs for larger catchment studies by simply selecting one HUC-12.

3.3 Service 3: Consuming individual ETV data to create PIHM input and output datasets

Services one and two demonstrate how web users can retrieve ETV data applicable to many types of models. This section describes transforming the ETV datasets to be consumed within PIHM and focuses on the interface to enable users to gain access to the ETV data and the PIHM input and output datasets. The first step is to have the user select HUC-12s of interest, as shown in services one and two, but additionally by administration layers such as by county and state boundaries. Once the user has created a selection list (Figure 3a), the next step is to define a data workflow (Figure 3b) for the selection list. In this prototype, the user can select to create stream networks using Tarboton's TauDEM (2011), or they can use NHD stream networks. The user can define data-model workflow parameters by clicking on the interface button highlighted in Figure 3c to reveal the user interface control to modify parameters. After defining the data-model workflow properties, the user can select which PIHM workflow version they wish to use, and which HPC resource to use (Figure 3d). The user can define and control PIHM's calibration and parameter inputs by clicking on the interface button highlighted in Figure 3e to reveal a user interface control for full customization. Any changes to data and model workflows will be applied to the selected HUC-12 selection list highlighted in Figure 3a.

Assuming the user has valid credentials, the user initiates the process by clicking on the submit model button (Figure 3f) which adds the project objects to the workflow submission list (Figure 3g). The project object indicates the users' email, the project name, the HUC identification, and when the job was added to the submission list. The user can investigate all the workflow settings by clicking on the appropriate buttons highlighted in Figure 3h. The status of the workflows (Figure 3i) are indicated to the user with four colors; white indicates the workflow was not requested by the user, orange indicates the workflow has started, green indicates the workflow succeeded, and red indicates the workflow failed.

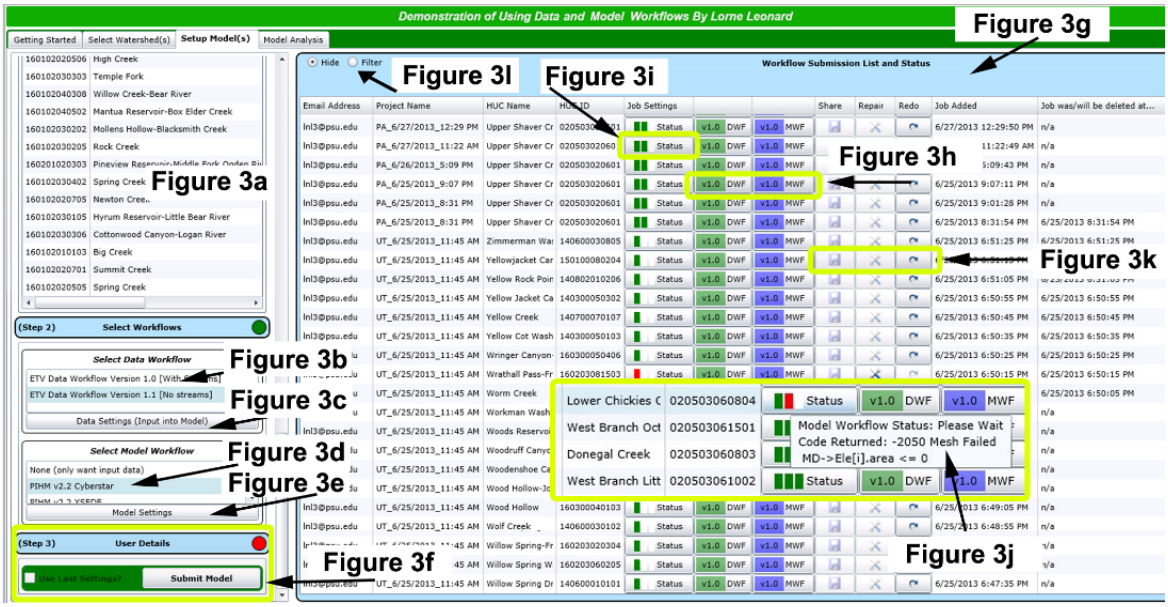

Figure 3. Demonstration of using data-model workflows to start a PIHM HUC-12 model quickly in the CONUS.

When one workflow has succeeded, the next workflow starts on the next available computing environment. The ETV and data-model workflows execute on the HydroTerre distributed computing system (Figure 1), while the PIHM models are distributed on the user selected HPC resources. If the workflow has failed, the reason for failure is available to the user by clicking on the status button (Figure 3i) which reveals the dialog (Figure 3j) using error codes that reveal to the modeler the source of the problem. This provides the user valuable information about ways to fix the problem. The main reason for failure using this prototype is poor meshes, which requires the user to modify catchment simplification parameters. However, this prototype web application simplifies the process of creating a new model, by the user selecting an existing project, and cloning the project (Figure 3k). Then, the user changes parameters (Figure 3c and 3e) to resubmit the workflow processes (Figure 3f). For details about the workflows, the reader is referred to Leonard (2014a).

The submission list embraces the provenance data associated with all the workflows making it feasible and easy to reproduce an entire workflow process. These workflow processes are accessible to other users who can filter (Figure 3l) HUC-12s of interest from other modelers. All the parameters chosen by users, explicitly and or implicitly, are kept in a user project database that allows users to query other modeler's choices involved with any CONUS HUC-12 that are used to populate the interface shown in Figure 3. Unfortunately, model results are not stored permanently, due to the large amount of disk resources that would be required. To store one climate norm (30-years of forcing) of PIHM model results for all CONUS HUC-12s requires three petabytes of disk storage. However, by storing the provenance steps only, the need for large amounts of disk storage is reduced. Thus, this prototype can store thousands of workflow settings per HUC-12 across the CONUS, providing a new resource for modelers using the system to gain insight to start a new model study and to download a refined model.

4 CONCLUSIONS AND RECOMMENDATIONS

In this article, we describe three web applications and services that implement data-model workflows designed to rapidly share both data and models anywhere of the continental USA (CONUS). The first application enables users to download Essential Terrestrial Variables for a single level-12 Hydrological Unit Code anywhere in the CONUS applicable to most hydrological models. The second application is a web service that generates an xml file with the upstream depth-first HUC-12 graph for users to estimate data and model storage needs. The third application demonstrates how combining workflows can be used to share ETV data, PIHM input data, and PIHM model results for any CONUS HUC-12. The third application also demonstrates how workflows create provenance data ready to be shared amongst modelers to start a new PIHM hydrological model quickly. The web interfaces are built upon data and model workflows that are executed in a distributed compute environment to take advantage of different hardware and software configurations.

We have presented an important step towards eliminating hurdles involved with using physics based models, such as PIHM, in a High Performance Computing (HPC) environment by seamless allocation of resources with minimal interaction from the user through web-based workflows, shared software, data and HPC resources. The next phase is to scale from level-12 Hydrological Unit Codes to major river basins using the sharing services discussed in this article with ease and provenance. However, we foresee that beyond HUC-12 scales the feasibility of sharing data and model results will be difficult without visualization workflows to calibrate and interrogate hydrological model results. Coupling data-model, model, and visualization workflows is an important step towards providing numerical watershed predictions as a product in the form of a dynamic watershed atlas that provides surface and groundwater budgets from model simulations with data and analysis provenance. To achieve these goals, we need robust interfaces to execute, reproduce, and share provenance amongst modelers and stakeholders.

REFERENCES

CyberSTAR, 2014. A Scalable Terascale Advanced Resource for Discovery through Computing. http://www.ics.psu.edu/infrast (last accessed 01.01.14.).

ESRI, 2014. ArcGIS Server. http://www.esri.com/ (last accessed 01.01.14.).

Fielding, R., Taylor, R., 2002. Principled design of the modern Web architecture. ACM Transactions on Internet Technology. vol. 2, (2), 115-150.

Leonard, L., Duffy, C.J., 2013. Essential Terrestrial Variable data workflows for distributed water resources modeling. Environ. Modell. Softw. 50, 85-96.

Leonard, L., Duffy, C.J., 2014a. Automating Data-model Workflows at a level-12 HUC scale in a Distributed Computing Environment. In: International Environmental Modelling and Software Society (iEMSs) 7th International Congress on Environmental Modelling and Software San Diego, California, USA, D.P. Ames, N. Quinn (Eds.).

Leonard, L., Duffy, C.J., 2014b. HydroTerre: Selecting Up-stream level-12 HUCS using Depth-First Graphs anywhere in the Continental USA**.** In: 11th International Conference on Hydroinformatics, HIC 2014, New York City, USA.

Microsoft, 2014a. Database Markup Language, DML. http://msdn.microsoft.com/en-us/library/bb399400%28v=vs.110%29.aspx (last accessed 01.01.14.).

Microsoft, 2014b. SQL Server. https://www.microsoft.com/en-us/sqlserver (last accessed 01.01.14.).

NLDAS, 2011. North American Land Data Assimilation System, NLDAS. http://ldas.gsfc.nasa.gov/nldas/NLDAS2forcing.php (last accessed 08.01.11.).

Tarboton, D. G., 2011. TauDEM Hydrology Research Group. http://hydrology.usu.edu/taudem/taudem5.0/index.html (last accessed 02.03.11.).

World Wide Web Consortium, 2014a. Simple Object Access Protocol, SOAP. http://www.w3.org/TR/soap12-part1 (last accessed 01.01.14.).

World Wide Web Consortium, 2014b. Web Services Description Language, WSDL. http://www.w3.org/TR/wsdl (last accessed 01.01.14.).