This is an Accepted Manuscript of an article published in Environmental Modelling and Software in April 2016, available online: DOI: 10.1016/ j.envsoft.2015.12.010

Using a Data Grid to Support Regional-Scale Hydrologic Modeling

Mirza M. Billah[1], Jonathan L. Goodall[2, 3, *], Ujjwal Narayan[4], Bakinam T. Essawy[2], Venkat Lakshmi[5], Arcot Rajasekar[6], Reagan W. Moore[6]

[1] *Department of Biological Systems Engineering, Virginia Tech, Blacksburg, Virginia*

[2] *Department of Civil and Environmental Engineering, University of Virginia, Charlottesville, Virginia*

[3] *Department of Civil and Environmental Engineering, University of South Carolina, Columbia, SC*

[4] *CMNS-Earth System Science Interdisciplinary Center, University of Maryland, College Park, MD*

[5] *Department of Earth and Ocean Sciences, University of South Carolina, Columbia, SC*

[6] *School of Library and Information Science, University of North Carolina, Chapel Hill, NC*

* *To whom correspondence should be addressed (E-mail: goodall@virginia.edu; Address: University of Virginia, Department of Civil and Environmental Engineering, PO Box 400742, Charlottesville, Virginia 22904; Tel: (434) 243-5019)*

Abstract

Modeling a regional-scale hydrologic system introduces major data challenges related to the access and transformation of heterogeneous datasets into the information needed to execute a hydrologic model. These activities are difficult to automate, making the reproducibility and extensibility of model simulations conducted by others difficult or even impossible. This study addresses this challenge by demonstrating how the integrated Rule Oriented Data Management Systems (iRODS) can be used to support workflow execution needed when using data-intensive models to simulate regional-scale hydrologic systems. Focusing on the Variable Infiltration Capacity (VIC) model as a case study, data preparation steps are written using micro-services and rules within iRODS. VIC and iRODS are applied to study hydrologic conditions in the Carolinas, USA during the period 1998-2007 to better impacts of drought within the region. The application demonstrates how automated workflows in iRODS can support hydrologic modelers to create more reproducible and extensible "end-to-end" model simulations.

Keywords: data management; workflows; hydrologic modeling

Software availability: Software is available by request to the corresponding author.

INTRODUCTION

Motivation

Application of regional-scale hydrologic models presents a number of challenges associated with handling and processing large datasets. The models are data intensive and require a significant amount of time and effort in order to transform available datasets into the form required by the hydrologic model (Leonard and Duffy, 2013). The data challenges also include collecting datasets from various data providers that have inconsistent data access protocols, file formats, and semantics (Horsburgh et al., 2014). The result is that the datasets required to setup, calibrate, and validate hydrologic models are not made available by data providers in a form that is ready for application directly into models, rather each model requires specific transformations of the available information before it can be used within a model (Leonard and Duffy, 2014). Due to the level of heterogeneity between data sources, these data preparation steps are difficult to automate and, with the exception of a few models with robust data preparation tools, often require significant manual intervention. Even for models with data preparation tools available, they are often tied to pre-specified data sources that may not be the best available information for a specific region or modeling objective. The end result is that modelers (i) consume significant time on tasks that could be automated and (ii) lack the ability to easily reproduce and extend past work completed by others to explore new scientific questions or water management goals.

Within the information and computer science communities, there has been work to create advanced data management and scientific workflow software that has the potential to address these challenges facing the hydrology community (Gil et al., 2007;

Ludäscher et al., 2006; Oinn et al., 2006). There has been uptake of these tools within some scientific communities, in particular bioinformatics, but with only a few exceptions (Fitch et al., 2011; Guru et al., 2009; Perraud et al., 2010; Piasecki and Lu, 2010), there has been minimal adoption of workflow tools for hydrologic modeling. This past work has clearly demonstrated the utility of workflow environments for implementing the data preparation tasks and allowing the software to coordinate and automate processing steps, as well as track the provenance of the datasets generated through the processing steps. The potential of this software to provide authenticated access to external data sources, along with procedures for automatically transforming data products to those required by a model, makes these software systems particularly well suited to automating hydrologic modeling workflows. However these workflow tools have challenges when used for hydrologic analysis and modeling that include but are not limited to the lack of common data models within hydrology (Perraud et al., 2010).

Building on this paper work, we postulate that a key reason for the lack of update of workflow environments within hydrology is the data-centric nature of hydrology, and the vast heterogeneity of datasets used within hydrologic models. For this reason, we believe that data management must play a central role within hydrologic workflows. Hydrologic data sources are distributed and maintained by different governmental and nongovernmental agencies. They are provided using different file types, structures, and semantics. We believe that centralizing and standardizing this data is not a sustainable long-term approach. Rather the long-term solution, we believe, will be a federated data grid approach with decentralized data management supplemented with server-side processing that allows for on-demand access to derived data produces required by specific hydrologic models. Thus, when we speak of workflows in this paper, we are describing processing pipelines that are able to operate on servers that house large data achieves and that can deliver to client machines, where hydrologic models are executed, derived data products at are much closer to the information required as input by the model.

This paper illustrates our concept by focusing on the DataNet Federation Consortium (DFC) grid, which is built using the data management system Integrated Rule-Oriented Data System (iRODS), and the regional-scale hydrologic model Variable Infiltration Capacity (VIC). These systems are briefly introduced in the background section that follows this introduction section. Then, in the design and implementation section, the approach for automating VIC data preparation and post-processing workflows using iRODS is presented. The workflows were demonstrated for an example of modeling drought in the Carolinas region of the United States. Finally, this paper concludes with a discussion of applying iRODS to support hydrologic modeling.

Background

The DFC grid (http://www.datafed.org) was built as part of an NSF-funded research project to provide storage and compute resources that allow for long-term access to the stored datasets. DFC is enabled by the iRODS, an open source, policy-based cyberinfrastructure developed by the Data Intensive Cyber Environments (DICE) group for distributed data management (Rajasekar et al., 2010b) (http://www.irods.org). It is used by a wide variety of end users including the bioinformatics community (Goff et al., 2011). Data management tasks or policies are implemented within iRODS as rules. Rules specify a sequence of lower-level micro services that are used to operate on datasets within the data grid. The user can specify the sequences of data collection, transformation, curation, preservation, and processing steps in a workflow within one or more rules that use micro-services developed by users or administrators. The system is capable of remote execution of workflows as well, meaning data preparation steps can be executed on remote servers rather than on a client machine, which is especially beneficial for large-datasets.

The Variable Infiltration Capacity (VIC) model is a large-scale hydrologic model that applies water and energy balances to simulate terrestrial hydrology at a regional spatial scale (Liang et al., 1996a). The scientific background of the model is summarized in the case study section, while here the model is presented from a data management perspective. The model requires several input datasets to be generated by applying multiple data processing steps before a simulation run can be executed (Figure 1). The datasets for preprocessing include precipitation, maximum and minimum temperature, wind speed, topography, soil, and vegetation information. Each dataset is processed using scripts that require the execution of data processing routines to generate new datasets used in the model simulation. Each data processing script performs a certain task that is a prerequisite to the following script in the workflow. Running these data collection and processing scripts currently requires significant manual intervention and time to complete. In addition, a large amount of data is generated during the procedure, introducing issues with storing intermediate datasets for reproducibility of model results as well as further analysis of key model inputs.

Study Objective and Scope

Given these data challenges in running VIC, which are not dissimilar from any other data-intensive hydrologic simulation model, and given the potential advantages of advanced data management systems such as iRODS, the objective of this study is to apply iRODS for automating the execution of VIC. The study is built on prior work where the VIC model was used for modeling water balances for South Carolina river basins (Billah and Goodall, 2011; Billah et al., 2015). VIC has also been applied by others for a number of climate conditions and basins to estimate several hydrologic variables with high accuracy (Abdulla et al., 1996; Lakshmi et al., 2004; Lohmann et al., 1998; Sheffield and Wood, 2007; Sheffield et al., 2004). However, there has been less work on the data challenges associated with running VIC. Therefore, this work addressed these data challenges by integrating VIC with iRODS in the DFC grid using federated data access and data processing for executing VIC for regional-scale hydrologic analysis.

The resulting system focused on developing data management workflows that automate pre and post-processing of data for VIC model and visualize the model results for state water managers for drought decision-making. The work demonstrated a case study for the drought in South Carolina during 1998 to 2002 (Badr, A. W., Wachob, A., Gellici, 2004) that had a significant effect on water resources across the Carolinas. The availability of detailed spatial and temporal information of hydrologic systems across the Carolinas, and in particular the ability to forecast future conditions, is a useful tool for water resources management. Soil moisture in particular is a difficult parameter to observe at state-level spatial scales (Sheffield et al., 2004), but it is an important indicator of drought at various scales (Lakshmi et al., 2004; Sheffield et al., 2004). The modeling

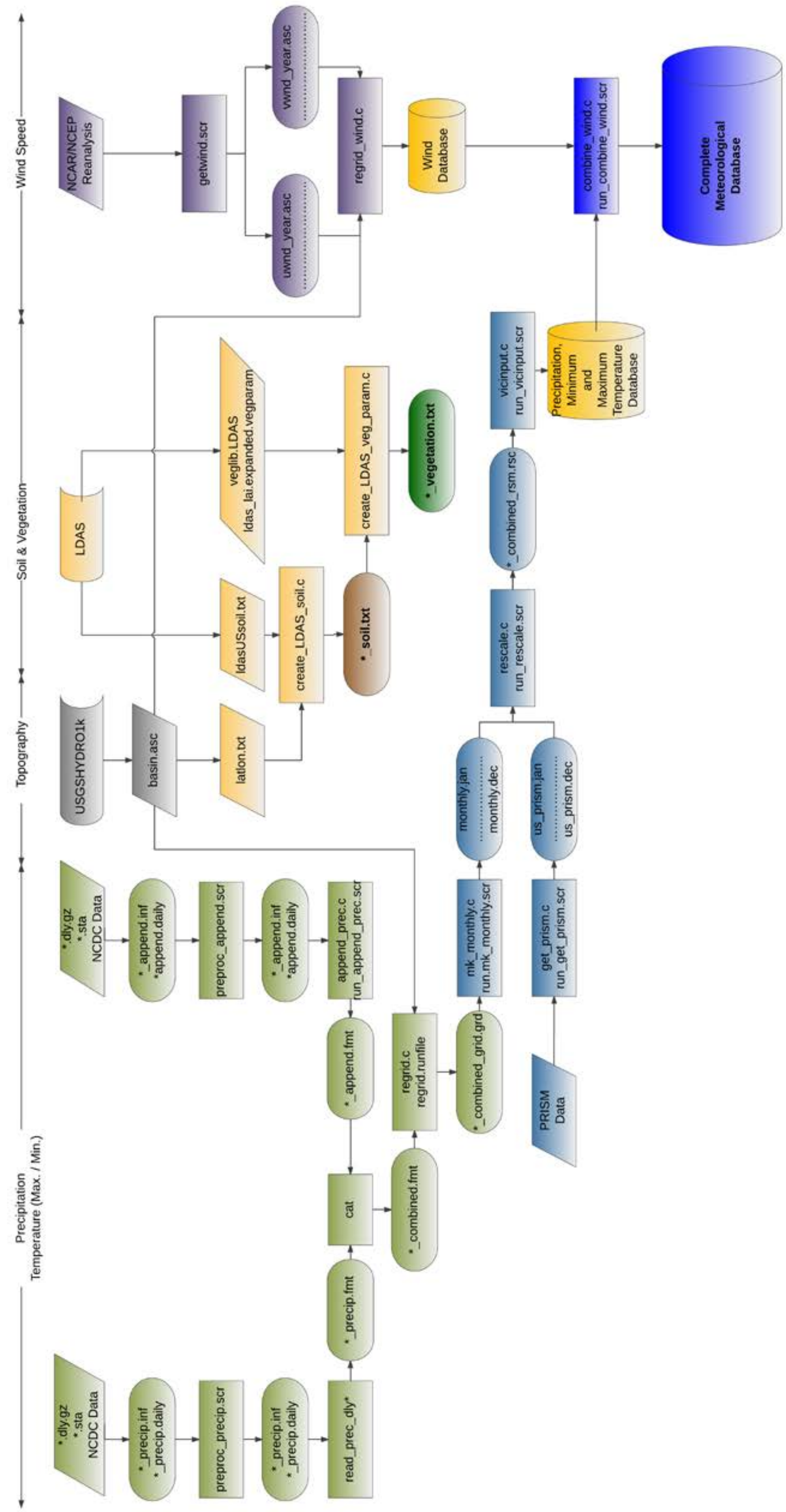

Figure 1: Data pre-processing steps for VIC model.

system created through this study provides estimates of soil moisture across the Carolinas for the period of 1998 to 2007. The analysis demonstrated insights of the methodological advancement in terms of accessibility, reproducibility, human intervention, time consumption, and storage management for the VIC model application for hydrologically-based drought analysis at regional spatial scales

SYSTEM DESIGN AND IMPLEMENTATION

The presentation of the system design and implementation is organized within two broad categories. The first category discusses the server-side application for workflow implementation and managing datasets, while the second category focuses on the client-side application for workflow execution using the server-side setup. Both the server and client-side tool are designed to manage datasets using iRODS for VIC model. Tasks associated with these categories are described in the following subsections.

Server-side Design and Implementation

The server-side application for iRODS is deployed on a remote server for hydrologic analysis that is part of the DFC federated grid. The application uses input datasets either from remote web services or a federated grid. The federated grid is a cyberinfrastructure that remotely connects several discipline specific resource servers. The server accommodates data processing sequences to produce output that can be transferred to a client.

The DFC-Hydrology Grid

The DFC-Hydrology Grid is deployed as part of the larger DFC (Figure 2), which is a project funded by the National Science Foundation (NSF) to support data collection, analysis, preservation, sharing, and publication of scientific data and models (http://www.datafed.org). The DFC- Hydrology grid consists of an iRODS server that communicates with the iRODS metadata catalog (iCAT) database located in the DFC-Federation Hub of the DFC grid. The server is administered by RENCI (Renaissance Computing Institute) in Chapel Hill, North Carolina. The server contains a Rule Engine (RE) that interprets rules (executed from the client-slide) using micro-services on the DFC-Hydrology data grid or the DFC-Federation Hub. The RE also connects to the catalog server in the DFC-Federation Hub and updates the iCAT database. The hydrology related data collected from various external sources are stored and preserved in the DFC-Hydrology server. The workflow can be implemented on any of the federated servers for data collection and transformation, however, transformed datasets are sent back to the client-side. These transformed datasets are replicated using RENCI operated resources within the DFC federated grid after completing data processing tasks in the client- side. For instance, the collection of precipitation data from a remote location is done by the DFC-Federation Hub and transformation is performed in the DFC-Hydrology server. The iRODS server-side application automates the process of accessing remote sources using different protocols (e.g., HTTP or FTP) for retrieving datasets. The iRODS server-side application is responsible for managing resource storage and federating these with other DFC grids, thereby providing seamless access to the data stored in DFC. These grids provide continuous support for data access and, in the case of connection failure, the federated grids provide data access from one or more other storage resources.

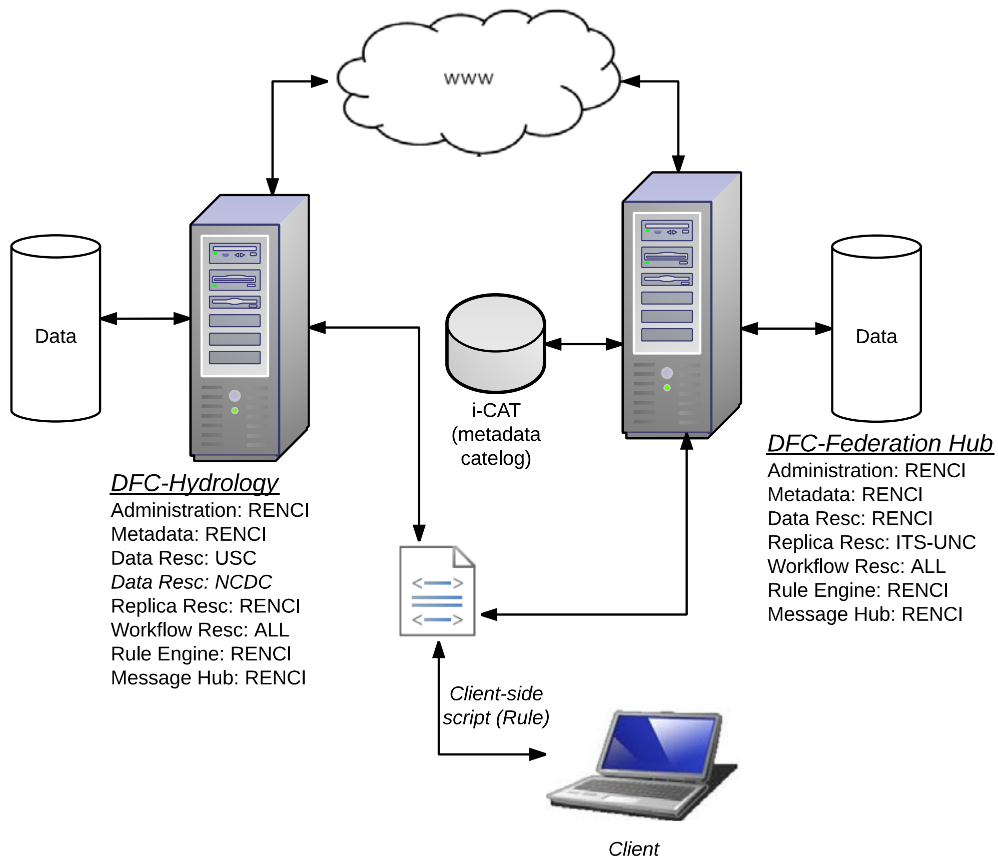

Figure 2: Schematic diagram of the NSF supported Data Federation Consortium (DFC) data management system showing the connections between the DFC-Hydrology with the DFC-Federation Hub.

Micro-service Implementation

Micro-services are the building blocks for implementing policy-based data management within DFC grid (Rajasekar et al., 2010a). A well-defined function in a micro-service performs a specific task as part of a distributed workflow system. A number of micro-services are available to automate data collection, processing, and storage in the DFC federated resource servers. These micro-services are primarily developed by system or application programmers and are compiled into the iRODS server code. The micro-services that are applied for this study are listed in Table 1. These micro-services are chained together to provide a higher-level of functionality to implement multiple tasks. Although the flexibility to chain a number of micro-services provides multiple ways to complete a series of tasks, iRODS applies priorities and validation conditions to select the best micro-service to complete a given task. Application of multiple micro-services in iRODS makes it possible to chain routines to perform multiple tasks within a single workflow.

Client-side Design and Implementation

Client-side application is used to execute data management workflow within the federated grid. This consists of several command-line utilities known as i-commands to manage datasets and data processing rules to execute data management workflow (Rajasekar et al., 2010a). These i-commands are used to download and upload data from/to the DFC grids and execute micro-services in the server-side. Furthermore, the client-side application provides opportunity to create and execute rules to gain access to

Table 1: Micro-services applied for VIC model application using iRODS.

No.	Micro-service	Purpose
1	msiExecCmd	Execute commands
2	msiCollCreate	Make data collection
3	msiDataObjCreate	Create data object
4	msiDataObjWrite	Write data object
5	msiDataObjClose	Close data object
6	msiAddSelectFieldToGenQuery	Make query for data using field
7	msiAddConditionToGenQuery	Make query for data using condition
8	msiExecGenQuery	Execute Query
9	msiGetValByKey	Extract value from query result
10	msiSplitPath	Get directory path
11	msiDataObjUnlink	Delete temporary file
12	msiRmColl	Remove data collection
13	msiGetSystemTime	Get time stamp

heterogeneous data sources, to process datasets into the formats required by models or scientific communities. The rule or action is a critical and fundamental component for iRODS. It provides a flexible mechanism to integrate external systems for specialized processing and metadata management (Hedges et al., 2009, 2007). The rules implement data processing steps as policies on the DFC federated grid and respond to various requests and conditions by integrating related micro-services from the server-side. These data processing rules, specifically for the VIC model, are generally categorized into two separate workflows: i) data pre-processing and ii) data post-processing. The pre-processing workflow is responsible for data collection and transformation into standard VIC format, while the post-processing workflow is used for model results visualization. Details of these categories are discussed in the following.

VIC Pre-processing Workflows

Data pre-processing involves collecting and transforming datasets from heterogeneous sources. For our purpose, data are collected from the United States Geological Survey (USGS), National Climatic Data Center (NCDC), National Center For Atmospheric Research (NCAR), National Centers for Environmental Prediction (NCEP), and Land Data Assimilation System (LDAS). These datasets are then processed in the DFC-Hydrology grid and stored in the DFC-Federation Hub. The metadata catalog in the DFC-Federation Hub is updated automatically while data are uploaded with recent information. This metadata catalog functions as an information center and enables the discovery of data that are stored into the grid.

Data processing workflows also include several data-specific processing rules that transform the collected datasets from the external sources into model readable inputs

(Figure 3). This data-specific rule is a combination of multiple step-based routines each of which performs a particular task. The step-based routines are simple, such as retrieving

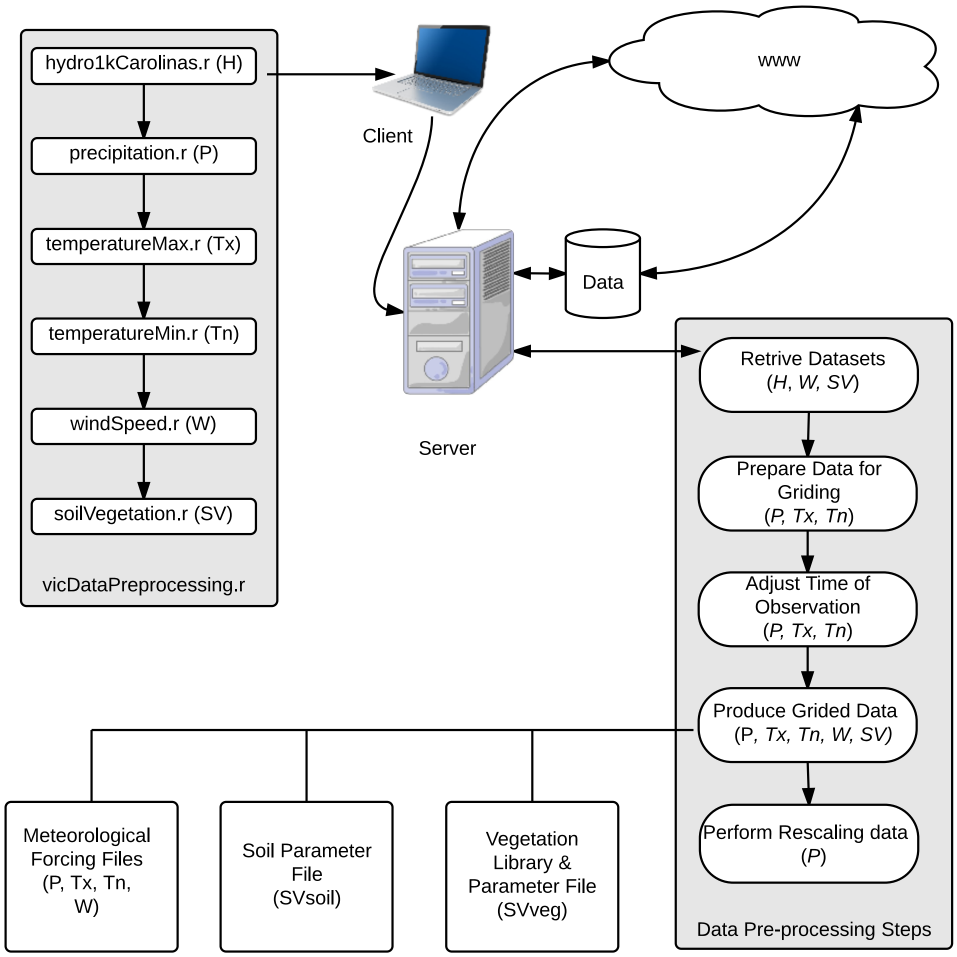

Figure 3: Model pre-processing workflows showing the major steps for transforming datasets to set up the VIC model for a specific study area. Rules are initiated from a client but executed on a server using micro-services.

data, preparing data for gridding, adjusting observation times, and transforming gridded and rescaled datasets. For our study, these step-based routines are integrated into data-specific rules used to complete a series of tasks from collection and transformation of the datasets into model inputs. Therefore, for each dataset, a separate rule is created to perform the data transformation tasks including the data collection from respective sources.

Table 2: Rules created within iRODS for VIC data pre-processing.

No.	Micro-service	Purpose
1	hydro1kCarolinas.r	Collect HYDRO1k and Climate data for the study area
2	precipitation.r	Collect and process daily precipitation data
3	temeperatureMax.r	Collect and process daily maximum temperature data
4	temeperatureMin.r	Collect and process daily minimum temperature data
5	windSpeed.r	Collect and process annual wind speed data
6	soilVegetation.r	Collect and process soil and vegetation data

The sequential implementation of the data-specific rules can be integrated into a single model-specific rule called *vicDataPreprocessing.r* to process complete data inputs for the VIC model. However, we have used six separate data-specific rules to process the datasets for VIC model for the period of 1998 to 2007 (Table 2). In the data-specific rules, step-based calculations are chained in such a way that data processing steps are not violated and perform only designated tasks (Figure 3). For example, preparation for data gridding is not executed without collecting and storing data in the DFC grid or rescaling is not performed before gridding the datasets. Also, not all of the data processing routines are performed for all the datasets. For instance, preparation for data processing is not executed for wind, soil and vegetation datasets because these data do not require grid preparation. Overall, data retrieval is performed by hydro1kCarolinas.r (H), windSpeed.r (W), and soilVegetation.r (SV), while precipitation.r (P), temperatureMax.r (Tx), and temperatureMin.r (Tn) implement respective workflows for data trans-formation using the retrieved climate data.

Data-specific rules are mostly inherent data processing scripts from the VIC that execute a series of routines that are associated with tasks (Table 2). For instance, we have retrieved climate data (precipitation, maximum and minimum temperature data) from the

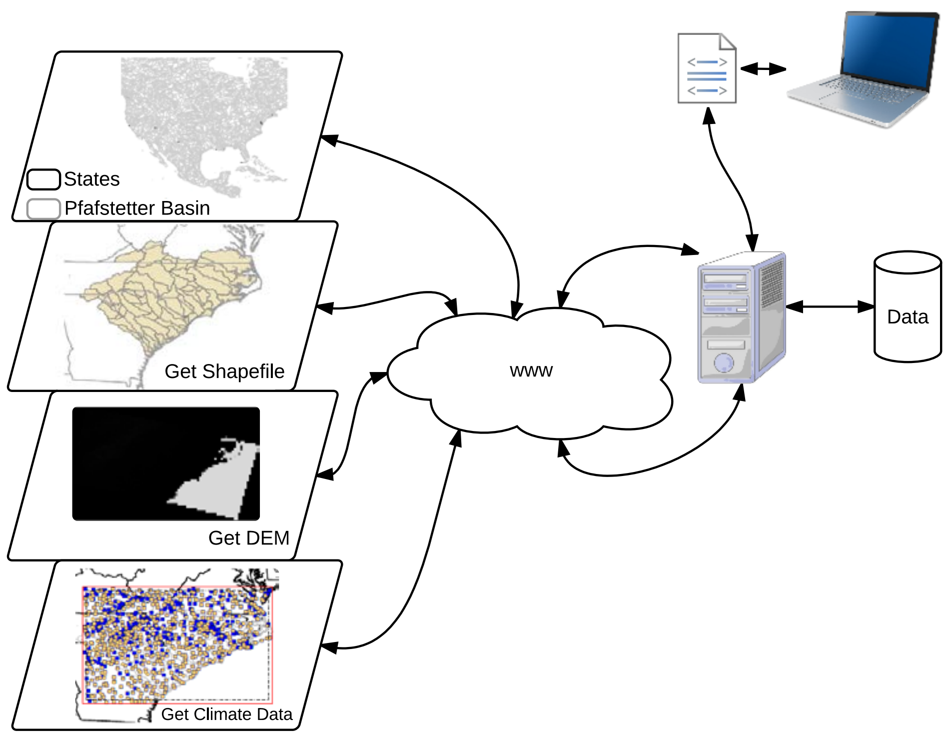

Figure 4: Data flow in the hydro1kCarolinas.r rule that extracts climate data from NCDC GHCND using HYDRO1K basin/DEM datasets to define a study region.

DFC federated grid by implementing a rule hydro1kCarolinas.r (Figure 4). This rule uses a series of micro-services in the DFC federated grid and executes a series of tasks sequentially to extract and register data over a defined area. We used the Pfafstetter basin numbering system as described in Furans and Olivera (2001) for defining study area from HYDRO1k basin and DEM datasets. The climate datasets for the defined study area were downloaded via FTP from the NCDC Global Historical Climatology Network (GHCND) database. While downloading the climate data, a buffer of 0.25° was considered around the defined study area to collect sufficient climate data. The climate data contained precipitation, maximum and minimum temperature, and wind speed datasets. The precipitation data was processed using the precipitation.r rule, which used the GHCND data downloaded through DFC server and converted the station specific datasets into gridded datasets with a spatial resolution of $1/8^{\circ}$. Similarly, temperatureMax.r and temperatureMin.r rules downloaded and converted station specific temperature values into gridded datasets with $1/8^{\circ}$ spatial resolution. Furthermore, we also collected wind speed, soil and vegetation data from their respective sources while executing respective data-specific rules in the DFC-Hydrology. The annual wind data were collected from NCAR/NCEP and processed to generate gridded datasets of $1/8^{\circ}$ resolution for the study area using windSpeed.r. The soilVegetation.r rule was applied to transform the LDAS soil and vegetation information into information required for the model.

VIC Post-processing Workflows

A VIC model simulation outputs hydrologic flux and state variables for a gridded discretization of the landscape. The hydrologic flux and state variables include evapotranspiration, soil moisture, baseflow, and snow depth. These variables are output into text files and require additional processing in order to visualize the model results and gain a better understanding of water movement within the system being studied. Workflows can play an important role in this regard because they are able to automate the steps required to transform model output data into more useful visualizations such as maps or graphs. Scientists can use workflows to more easily visualize the VIC model predictions and can share their approaches for creating model visualizations with other VIC modelers. Because rules are written in a simple scripting language, it is possible for multiple stakeholders to make use of and potentially modify data post-processing rules to create useful visualizations and analysis routines that summarize model predictions.

EXAMPLE APPLICATION

Study Area

The region used for the case study application covers all the major river basins in both North and South Carolina with a total area of 280,736 km^2 (108,393 mi^2) (Figure 5).

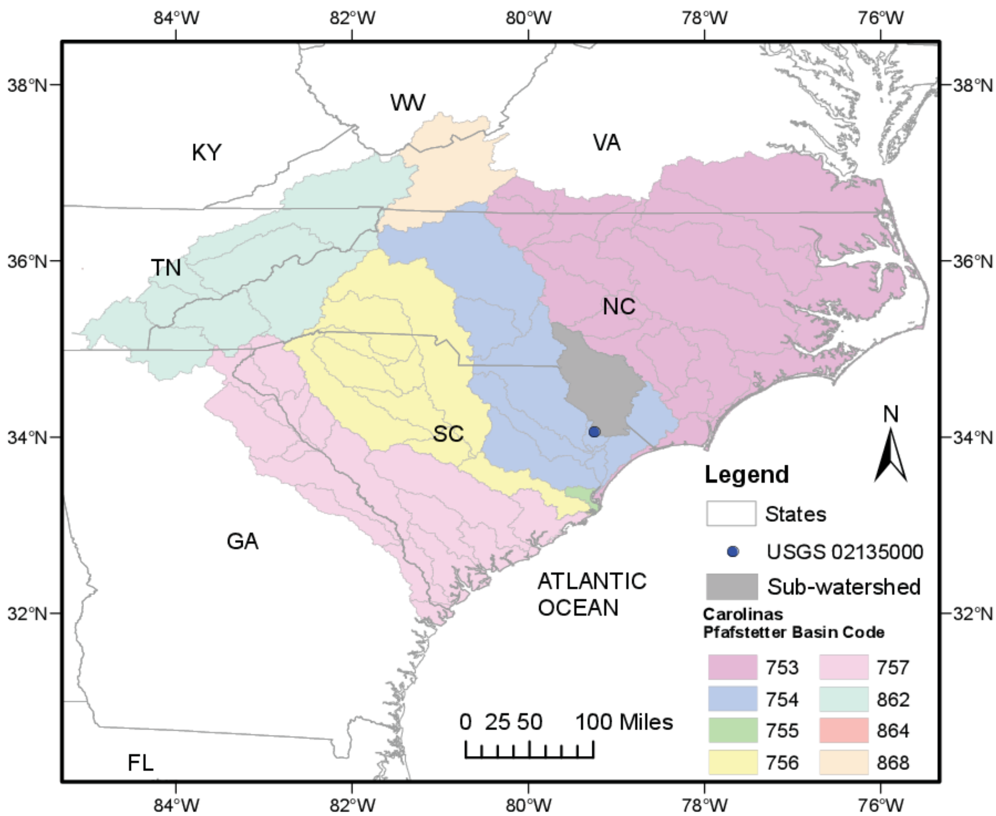

Figure 5: Study area with major river basins in North and South Carolina, USA. These subwatersheds were extracted using selected Pfafstetter Basin Code from the Hydro1K dataset. The sub-watershed and stream gage station used for the VIC model calibration are also shown in the Figure.

The Pfafstetter Basin Code system was used to define this “Carolinas” study region from the HYDRO1k basin dataset. The codes were incorporated in the hydro1kCarolina.r rule to extract the area during the automation of the workflow.

Model Setup

VIC is able to simulate the land surface portion of the hydrologic cycle by solving the full water and surface energy balance equations (Liang and Lettenmaier, 1994; Liang et al., 1996b) in the Carolinas on a daily time step. VIC was calibrated using the following seven parameters: variable infiltration curve (b), maximum base flow (D_{smax}), fraction of base flow where base flow occurs (D_s), fraction of maximum soil moisture content above which nonlinear base flow occurs (W_s), mid (d_2) and deep (d_3) soil layer depth, and minimum stomatal resistance (r_0) (Abdulla and Lettenmaier, 1997a, 1997b; Crow, 2003; Troy et al., 2008). The range investigated and the final values of the parameters applied in this study were described in (Billah et al., 2015) and a comparison of the monthly average streamflow for the calibrated VIC model and streamflow from the USGS streamflow station are provided in (Figure 6). A trial and error approach was followed for calibrating the model for a selected portion of the overall study region for different parameter values. This approach was used because the model execution time for the overall study area was too long to use the entire study region for calibration. We compared the VIC model prediction for streamflow at the Little Pee Dee River at Galivants Ferry stream gage station that is part of the USGS National Water Information System (NWIS) network (USGS 02135000; Figure 5) for the period of 1998 to 2007. This streamflow station has a drainage area of 7257 km^2 and includes portions of both North and South Carolina. This station was selected based on its available time series record and because it is on an unmanaged portion of a river network. Nash-Sutcliffe Efficiency (NSE) index between simulated streamflow using VIC routing scheme and observed streamflow for the station, which is the commonly used approach for evaluating VIC models, was used as an objective function in the calibration. The NSE index of the final calibration is 0.6, a value that is considered to be a satisfactory calibration by watershed-scale hydrologic modelers (Moriasi et al., 2007).

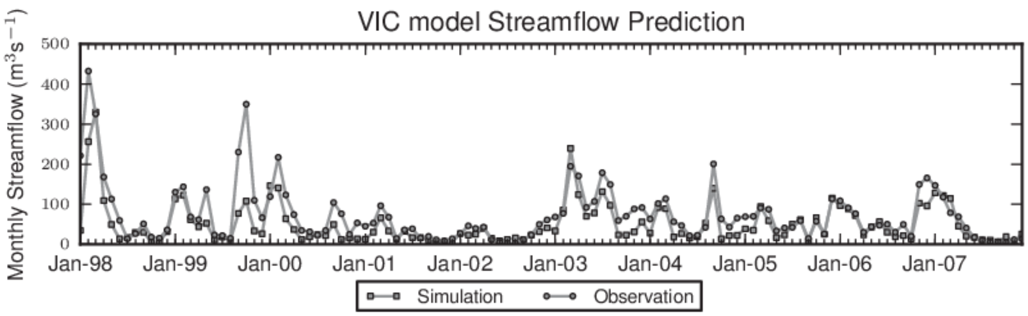

Figure 6: Streamflow comparison between the VIC model predictions and USGS observations. The comparison is performed at the USGS station Little Pee Dee River at Galivants Ferry, SC (Station Number 02135000).

Model Results

VIC model runs generate a number of grid-based hydrologic flux and state variable outputs. One of these outputs is soil moisture estimated for each grid cell in the simulation domain and for each soil layer within the model. Soil moisture is an important indicator of drought, so it was used as the example model output for creating a post-processing rule. This post-processing rule (soilmoisture.r) works by extracting model results from the various model output files, then summarizing the soil moisture over the study region for each soil layer. The rule results in a time-series plot of monthly soil moisture within the three VIC soil layers (Figure 7). The plot of soil moisture estimates provides understanding of the impact of drought on soil moisture, particularly the deep soil layer which is more sensitive to long term trends in water availability compared to the middle and upper soil layers. The plot represents the final product from the modeling work that would be used in a journal publication describing the work. Because generation of the plot can be easily repeated by re-running the soilmoisture.r rule, the results presented in the journal publication can be more easily reproduced and extended by other researchers.

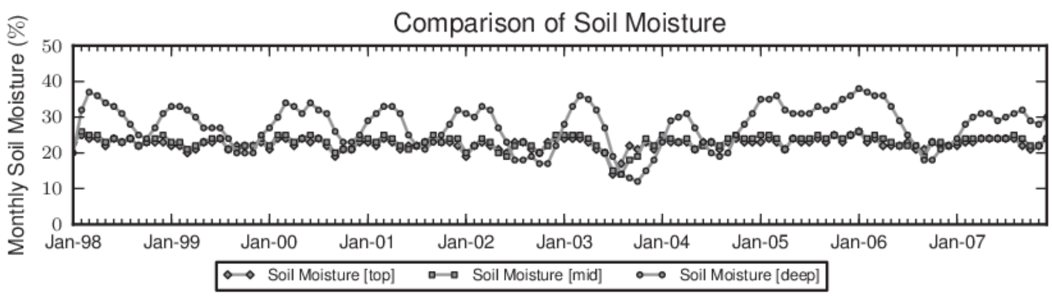

Figure 7: Comparison of monthly averaged soil moisture in the three soil layers predicted by the VIC model in the Carolinas for the periods of 1998 to 2007.

SUMMARY, DISCUSSION, AND CONCLUSIONS

The hydrologic modeling process can involve many steps from data access and transformation, to model setup, calibration, and validation, to analysis and visualization of model outputs. This entire “end-to-end’’ process involves some steps that are easily automated and others that often require intervention by expert modelers. The goal in this and related work is to automate those steps that are straightforward but tedious, while still allowing experts to guide the process and intervene when needed. Currently too many steps that can be automated are not. As a result, modelers are unable to focus on the important tasks that require their expertise and insights because time must be spent on more basic data gathering and transformation steps. Furthermore, the steps that could be automated are typically not thoroughly documented and, even if they are thoroughly documented, are time consuming to repeat. This makes independent reproducibility of model results, a requirement for scientific progress and water resource management objectives, difficult or even impossible.

This work addresses these challenges by leveraging the iRODS technology and the DataNet Federation Consortium (DFC) cyberinfrastructure to create workflows that automate pre and post-processing routines for the hydrologic model VIC. VIC is a widely used hydrologic model that is typically applied to large spatial regions to address questions related to water resource availability during periods of drought. The workflows developed for VIC include sufficient information to allow others to independently reproduce the model results, from raw data products to visualizations of model outputs used in publications. The workflows therefore act as a means for forced documentation of the steps used to create model input files including the provenance of data as it is transformed from the form provided by federal and academic data repositories to the form output by the models. In the case of VIC, and likely in the case of other hydrology applications as this work is extended to include other hydrologic models, the workflows were created by leveraging existing scripts written to complete specific data pre or post-processing tasks. The approach used provides a means for placing these scripts in larger workflows and removes the need to access and understand the original scripts to reproduce model results or to reuse the tools for a new study. In the latter case, modification is only required when selecting a new area of interest. The Pfafstetter Basin Codes are replaced in the shared workflow for hydrologic systems analysis workflow hydro1kCarolinas.r. However, all other workflows are used without modification to automate the remaining data processing steps, which effectively build a complete VIC model that is ready for simulation.

A key contribution of this work is demonstrating a methodological approach to assist in data-intensive hydrologic modeling. Using iRODS has advantages that include workflow automation, datasets curation and preservation, data replication, and workflow sharing for results reproduction. iRODS is flexible and robust and it was possible to extend the software to develop rules specific for the data processing workflows associated with running the VIC model. Because of the data grid concept used by iRODS, it was possible to design the workflows in such a way that both server-side and client-side applications could be leveraged to perform the data gathering and preparation steps. It was also possible to create shared workflows that emphasized the interoperability of the DFC federated grids for sharing rules and datasets among distributed users. For instance, we applied ecohydroworkflow (Miles, B., Band, 2013), a shareable workflow for data management for hydrologic models, to collect and register GHCND and HYDRO1k datasets from NCDC and USGS respectively in the DFC-Federation Hub. This implementation provided an example of how datasets collected for multiple models and applications can be easily transferred between grids within an iRODS federation using i-commands. The ability to execute data-specific rules on the client-side allowed for end-to-end data management, reduction of time and potential for human error during data processing, model prediction reproducibility, large data management, and opportunity to analyze model predictions before registering and sharing datasets within the DFC federated grid.

The workflows created within iRODS to enable end-to-end execution of model pre and post-processing steps are a key step to achieving reproducible hydrologic model runs. It was possible to chain all the data pre-processing routines by including all datasets in model-specific rule vicDataPreprocessing.r using micro-services in iRODS. Despite the opportunity to combine all data processing routines into a model-specific rule, the system was purposely designed in such a way that each dataset had its own sub-workflow to transform the raw data into model readable information. Doing so provided a level of granularity so that the data-specific rules could be later re-used within other hydrologic modeling applications in the DFC federated grid. Data post-processing workflows were also created to automate the tasks required to create visualizations and publication-quality figures based on model outputs.

By running data pre-processing, model execution, and data post-processing rules in sequence, it is possible to go from raw data from federal data repositories to figures summarizing model outputs that are used in journal publications and conference presentations. Having such capability would reduce human errors that could occur by not correctly performing a transformation step during a manual execution of the work. It would also free researchers to devote more time to enhancing, calibrating, and validating models, rather than on tedious steps required to set-up first iterations of the model. A longer term goal could be for researchers to publish such workflows that can be used to recreate publication figures as supplemental resources with the journal paper itself.

It is important to note the specific data challenges required for a hydrologic model application. First, VIC requires a large amount of data when applied to a region the size of the Carolinas, and of course even more data would be required for Continental scale model executions, which are not uncommon when applying VIC. The data used in the VIC model pre-processing steps included meteorological datasets at stations and on grids, topography datasets available as grids, and soil and vegetation datasets also available as grids. Transformation of these raw input data resulted in intermediate datasets with different spatial projections, filled gaps, and other modifications required before initiating the model simulation. Over the years, researchers have created scripts for completing many of these data pre-processing steps, and so it was relatively straight forward to wrap these scripts as iRODS rules using i-commands. One advantage of using iRODS rather than just running the scripts outside of iRODS is that iRODS provides a metadata server in the DFC-Federation Hub that is automatically updated with information tracking the provenance of the datasets during the transformation process. Also, the large datasets used in the VIC modeling exercise were automatically curated and preserved in the DFC federated grid for future uses and can be referenced through publications describing the analysis.

Creating data post-processing workflows provided the opportunity to visualize model results more rapidly and interactively as part of the modeling process. The data post-processing workflow combined results acquisition, cleaning and analyzing, and sharing findings to expose the relationships contained within the data. Visual observation of model results in the form anticipated for the final publication helped in understanding how changes to both the data pre-processing steps and the model execution steps impacted key model results. It is common that stakeholders will have unique interests for each model application, so creating general visualization tools will not always be possible. Therefore, creating workflows that leverage lower-level micro-services to visualize model results in customized ways is a powerful tool provided by iRODS. This process of creating visualizations is another time saving strategy that allows the modeler to focus on model specific tasks rather than technique approaches for transferring data between the model and a visualization system.

Questions remain regarding the level of reuse that will be practical when micro-services and rules are used across hydrologic simulation models. There are classes of hydrologic models that can be grouped based on the use cases considered when developing the model. VIC falls into the "macro-scale" class of hydrologic models, meaning it is typically applied to regional, continental, or even global scale hydrologic systems. Other hydrologic models focus on more local scale systems such as a single catchment. Different classes of hydrologic models will likely require different schemes for pre-processing tasks such as discretizing the landscape, and will likely make use of different raw datasets to setup and parameterize the model. A key challenge moving forward, therefore, will be to ensure a flexible workflow environment where new data access and transformation tools specific to certain models can be easily developed and shared within focused hydrologic communities.

ACKNOWLEDGMENTS

The authors wish to acknowledge the National Science Foundation (NSF) under the project DataNet Full Proposal: DataNet Federation Consortium (Award Number:094084.

REFERENCES

Abdulla, F. a., Lettenmaier, D.P., 1997a. Application of regional parameter estimation schemes to simulate the water balance of a large continental river. J. Hydrol. 197, 258–285. doi:10.1016/S0022-1694(96)03263-5

Abdulla, F. a., Lettenmaier, D.P., 1997b. Development of regional parameter estimation equations for a macroscale hydrologic model. J. Hydrol. 197, 230–257. doi:10.1016/S0022-1694(96)03262-3

Abdulla, F. a., Lettenmaier, D.P., Wood, E.F., Smith, J. a., 1996. Application of a macroscale hydrologic model to estimate the water balance of the Arkansas-Red River Basin. J. Geophys. Res. 101, 7449. doi:10.1029/95JD02416

Badr, A. W., Wachob, A., Gellici, J.A., 2004. South Carolina, South Carolina Water Plan. Second Edition. South Carolina Department of Natural Resources. Land, Water and Conservation Division, Columbia, SC.

Billah, M.M., Goodall, J.L., 2011. Annual and interannual variations in terrestrial water storage during and following a period of drought in South Carolina, USA. J. Hydrol. 409, 472–482. doi:10.1016/j.jhydrol.2011.08.045

Billah, M.M., Goodall, J.L., Narayan, U., Reager, J.T., Lakshmi, V., Famiglietti, J.S., 2015. A methodology for evaluating evapotranspiration estimates at the watershed-scale using GRACE. J. Hydrol. 523, 574–586. doi:10.1016/j.jhydrol.2015.01.066

Crow, W.T., 2003. Multiobjective calibration of land surface model evapotranspiration predictions using streamflow observations and spaceborne surface radiometric temperature retrievals. J. Geophys. Res. doi:10.1029/2002JD003292

Fitch, P., Perraud, J.-M., Cuddy, S., Seaton, S., Bai, Q., Hehir, D., Sims, J., Merrin, L., Ackland, R., Herron, N., 2011. The Hydrologists Workbench: more than a scientific workflow tool, in: Proceedings, Water Information Research and Development Alliance Science Symposium.

Furnans, J., Olivera, F., 2001. Watershed Topology - The Pfafstetter System. Esri User Conf. Vol. 21.

Gil, Y., Deelman, E., Ellisman, M., Fahringer, T., Fox, G., Gannon, D., Goble, C., Livny, M., Moreau, L., Myers, J., 2007. Examining the Challenges of Scientific Workflows. Computer (Long. Beach. Calif). 40, 24–32. doi:10.1109/MC.2007.421

Goff, S.A., Vaughn, M., McKay, S., Lyons, E., Stapleton, A.E., Gessler, D., Matasci, N., Wang, L., Hanlon, M., Lenards, A., Muir, A., Merchant, N., Lowry, S., Mock, S., Helmke, M., Kubach, A., Narro, M., Hopkins, N., Micklos, D., Hilgert, U., Gonzales, M., Jordan, C., Skidmore, E., Dooley, R., Cazes, J., McLay, R., Lu, Z., Pasternak, S., Koesterke, L., Piel, W.H., Grene, R., Noutsos, C., Gendler, K., Feng, X., Tang, C., Lent, M., Kim, S.-J., Kvilekval, K., Manjunath, B.S., Tannen, V., Stamatakis, A., Sanderson, M., Welch, S.M., Cranston, K.A., Soltis, P., Soltis, D., O'Meara, B., Ane, C., Brutnell, T., Kleibenstein, D.J., White, J.W., Leebens-Mack, J., Donoghue, M.J., Spalding, E.P., Vision, T.J., Myers, C.R., Lowenthal, D., Enquist, B.J., Boyle, B., Akoglu, A., Andrews, G., Ram, S., Ware, D., Stein, L., Stanzione, D., 2011. The iPlant Collaborative: Cyberinfrastructure for Plant Biology. Front. Plant Sci. 2, 34. doi:10.3389/fpls.2011.00034

Guru, S.M., Kearney, M., Fitch, P., Peters, C., 2009. Challenges in using scientific workflow tools in the hydrology domain, in: 18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation. Cairns, Qld., pp. 3514–3520.

Hedges, M., Blanke, T., Hasan, A., 2009. Rule-based curation and preservation of data: A data grid approach using iRODS. Futur. Gener. Comput. Syst. 25, 446–452. doi:10.1016/j.future.2008.10.003

Hedges, M., Hasan, A., Blanke, T., 2007. Management and preservation of research data with iRODS. Proc. ACM first Work. CyberInfrastructure Inf. Manag. eScience CIMS 07 17–22. doi:10.1145/1317353.1317358

Horsburgh, J.S., Tarboton, D.G., Hooper, R.P., Zaslavsky, I., 2014. Managing a community shared vocabulary for hydrologic observations. Environ. Model. Softw. 52, 62–73. doi:10.1016/j.envsoft.2013.10.012

Lakshmi, V., Piechota, T., Narayan, U., Tang, C., 2004. Soil moisture as an indicator of weather extremes. Geophys. Res. Lett. 31, 2–5. doi:10.1029/2004GL019930

Leonard, L., Duffy, C.J., 2013. Essential Terrestrial Variable data workflows for distributed water resources modeling. Environ. Model. Softw. 50, 85–96. doi:10.1016/j.envsoft.2013.09.003

Leonard, L., Duffy, C.J., 2014. Automating data-model workflows at a level 12 HUC scale: Watershed modeling in a distributed computing environment. Environ. Model. Softw. 61, 174–190. doi:10.1016/j.envsoft.2014.07.015

Liang, X., Lettenmaier, D.P., 1994. A simple hydrologically based model of land surface water and energy fluxes for general circulation models. J. Geophys. Res. 99, 14,415–14,428. doi:10.1029/94JD00483

Liang, X., Lettenmaier, D.P., Wood, E.F., 1996a. One-dimensional statistical dynamic representation of subgrid spatial variability of precipitation in the two-layer variable infiltration capacity model. J. Geophys. Res. 101, 21403. doi:10.1029/96JD01448

Liang, X., Wood, E.F., Lettenmaier, D.P., 1996b. Surface soil moisture parameterization of the VIC-2L model: Evaluation and modification. Glob. Planet. Change 13, 195–206. doi:10.1016/0921-8181(95)00046-1

Lohmann, D., Raschke, E., Nijssen, B., Lettenmaier, D.P., 1998. Regional scale hydrology: I. Formulation of the VIC-2L model coupled to a routing model. Hydrol. Sci. J. 43, 131–141. doi:10.1080/02626669809492107

Ludäscher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger, E., Jones, M., Lee, E.A., Tao, J., Zhao, Y., 2006. Scientific workflow management and the Kepler system. Concurr. Comput. Pract. Exp. 18, 1039–1065. doi:10.1002/cpe.994

Miles, B., Band, L., 2013. EcohydroWorkflowLib. <http://pythonhosted.org/ecohydroworkflowlib/index.html> (verified 05.29.2013).

Moriasi, D.N., Arnold, J.G., Liew, M.W. Van, Bingner, R.L., Harmel, R.D., Veith, T.L., 2007. M e g s q a w s 50, 885–900.

Oinn, T., Greenwood, M., Addis, M., Alpdemir, M.N., Ferris, J., Glover, K., Goble, C., Goderis, A., Hull, D., Marvin, D., Li, P., Lord, P., Pocock, M.R., Senger, M., Stevens, R., Wipat, A., Wroe, C., 2006. Taverna: lessons in creating a workflow environment for the life sciences. Concurr. Comput. Pract. Exp. 18, 1067–1100. doi:10.1002/cpe.993

Perraud, J., Fitch, P.G., Bai, Q., 2010. Challenges and Solutions in Implementing Hydrological Models within Scientific Workflow Software. AGU Fall Meet. Abstr. -1, 06.

Piasecki, M., Lu, B., 2010. Development of a Hydrologic Modeling Platform Using a Workflow Engine, in: AGU Fall Meeting Abstracts. p. 1239.

Rajasekar, A., Moore, R., Hou, C.-Y., Lee, C. a., Marciano, R., de Torcy, A., Wan, M., Schroeder, W., Chen, S.-Y., Gilbert, L., Tooby, P., Zhu, B., 2010a. iRODS Primer: Integrated Rule-Oriented Data System, Synthesis Lectures on Information Concepts, Retrieval, and Services. doi:10.2200/S00233ED1V01Y200912ICR012

Rajasekar, A., Moore, R., Wan, M., Schroeder, W., 2010b. Policy-based Distributed Data Management Systems. JODI J. Digit. Inf. 11, 1–16.

Sheffield, J., Goteti, G., Wen, F., Wood, E.F., 2004. A simulated soil moisture based drought analysis for the United States. J. Geophys. Res. D Atmos. 109, 1–19. doi:10.1029/2004JD005182

Sheffield, J., Wood, E.F., 2007. Characteristics of global and regional drought, 1950-2000: Analysis of soil moisture data from off-line simulation of the terrestrial hydrologic cycle. J. Geophys. Res. Atmos. 112, 1–21. doi:10.1029/2006JD008288

Troy, T.J., Wood, E.F., Sheffield, J., 2008. An efficient calibration method for continental-scale land surface modeling. Water Resour. Res. 44, 1–13. doi:10.1029/2007WR006513