Ecohydrology Models without Borders?

Using Geospatial Web Services in EcohydroLib Workflows in the United States and Australia

Brian Miles[1] and Lawrence E. Band[1,2]

[1] Institute for the Environment, University of North Carolina, Chapel Hill, USA
[2] Department of Geography, University of North Carolina, Chapel Hill, USA
brian_miles@unc.edu

Abstract. Ecohydrology models require diverse geospatial input datasets (e.g. terrain, soils, vegetation species and leaf area index), the acquisition and preparation of which are labor intensive, yielding workflows that are difficult to reproduce. EcohydroLib is a software framework for managing spatial data acquisition and preparation workflows for ecohydrology modeling, while automatically capturing metadata and provenance information. The goal of EcohydroLib is to enable water scientists to spend less time acquiring and manipulating geospatial data and more time using ecohydrology models to test hypotheses, while making it easier for models to be shared and scientific results to be reproduced. This increased reproducibility, ease of sharing, and researcher productivity can enable both model inter comparison of interest within a country, and site inter comparison of interest across national borders. Currently, EcohydroLib allows modelers to work with geospatial data stored locally as well as high spatial resolution U.S. national spatial data available via web services, for example 30-meter digital elevation model and land cover data, and 1:12,000 scale soils data. While researchers working in watersheds outside the U.S. can use EcohydroLib, they must manually download data for their study areas before these data can be imported into EcohydroLib workflows. Though national agencies in the U.S. and Australia offer some datasets via web services, with a few exceptions these are either lower resolution datasets or data made available via Open Geospatial Consortium (OGC) Web Map Service (WMS) interfaces of use primarily for cartography, rather than via OGC Web Coverage Service (WCS) or Web Feature Service (WFS) interfaces needed for integration with numerical models. In this paper we explore: (1) availability of high-resolution national geospatial data web services in the United States and Australia; and (2) integration of Australian web services with EcohydroLib.

Keywords: hydroinformatics, workflows, ecohydrology modeling, RHESSys.

1 Introduction

Researchers working in the interdisciplinary field of ecohydrology are concerned with the cycling of energy, carbon, water, and nutrients through coupled climate-soil-vegetation systems (Rodríguez-Iturbe 2000), and with the interaction between water

R. Denzer et al. (Eds.): ISESS 2015, IFIP AICT 448, pp. 311–320, 2015.

cycling and the ecological community. Ecohydrology models require diverse geospatial datasets (e.g. terrain, soils, vegetation species and leaf area index), the acquisition and preparation of which are labor intensive, yielding workflows that are difficult to reproduce. When applied to sites with complex terrain, for example mountainous forested ecosystems or urbanized ecosystems, high spatial resolution data (<= 30-m) are needed to accurately simulate hydrologic processes (Band 1993, Band et al. 2005).

Environmental modeling can incorporate information from diverse sources and relies heavily on cyberinfrastructure (CI; e.g. hardware, software, sensors, networks, and the human and social capital necessary to make use of these) to assist in the collection, analysis, and visualization of observed and model output data. The increasing use of cyberinfrastructure to carry out complex modeling, data analysis, and visualization tasks has led to the adoption of workflow systems of increasing complexity (Deelman et al. 2009). Workflow construction tools (e.g. Cyberintegrator, Kepler, VisTrails) enable scientists to create workflows by combining a series of services needed to complete a set of tasks (e.g. data preparation, analyses and modeling, visualization; Goble et al. 2010). Over the past decade, workflow systems have been gaining use across the sciences (e.g. high-energy physics, biological science as well as climate science). More recently, domain and information scientists have turned their attention to improving cyberinfrastructure for geoscience workflows in general (cf. U.S. National Science Foundation EarthCube; http://www.earthcube.org/) and water science in particular. In an analysis of Australia's Water Information Research and Development Alliance (WIRADA), Plale (2012) identifies the benefits of adopting workflow systems to carry out tasks common to water sciences, and considers how amenable these tasks are to representation in scientific workflow systems. The tasks include: (1) data discovery; (2) data cleaning and formatting; (3) data ingest; (4) data assimilation and forecast model execution; and (5) data analysis. The challenge for workflow systems in geosciences is to enable easy: (1) workflow creation, including debugging; (2) validation of workflows; (3) running of workflows across workflow systems; (4) results visualization; (5) results publishing; (6) sharing and reuse of workflows among scientists in a community and across disciplines to allow results to be reproduced (Duffy et al. 2012; Guo et al. 2012).

EcohydroLib (https://github.com/selimnairb/EcohydroLib) is a software framework for managing spatial data acquisition and preparation workflows for ecohydrology modeling, and provides library code and workflow commands for acquiring, manipulating, and managing geospatial data needed to run a variety of ecohydrology models (e.g. RHESSys, SWAT, VIC). Workflow steps common across models can be performed using EcohydroLib commands, for example identifying a study area, or acquiring terrain or soils data via web services (Fig. 1).

Modelers can then use model-specific workflow commands built on top of EcohydroLib to transform input data into formats appropriate for direct use by a particular model. In this way, EcohydroLib increases the amount of code devoted to data acquisition that can be shared across models, reducing duplication of programming effort, and facilitating model and site inter comparison. EcohydroLib workflows are composed of loosely coupled commands for performing geospatial data acquisition and preparation; data acquired by these commands are stored in a directory on

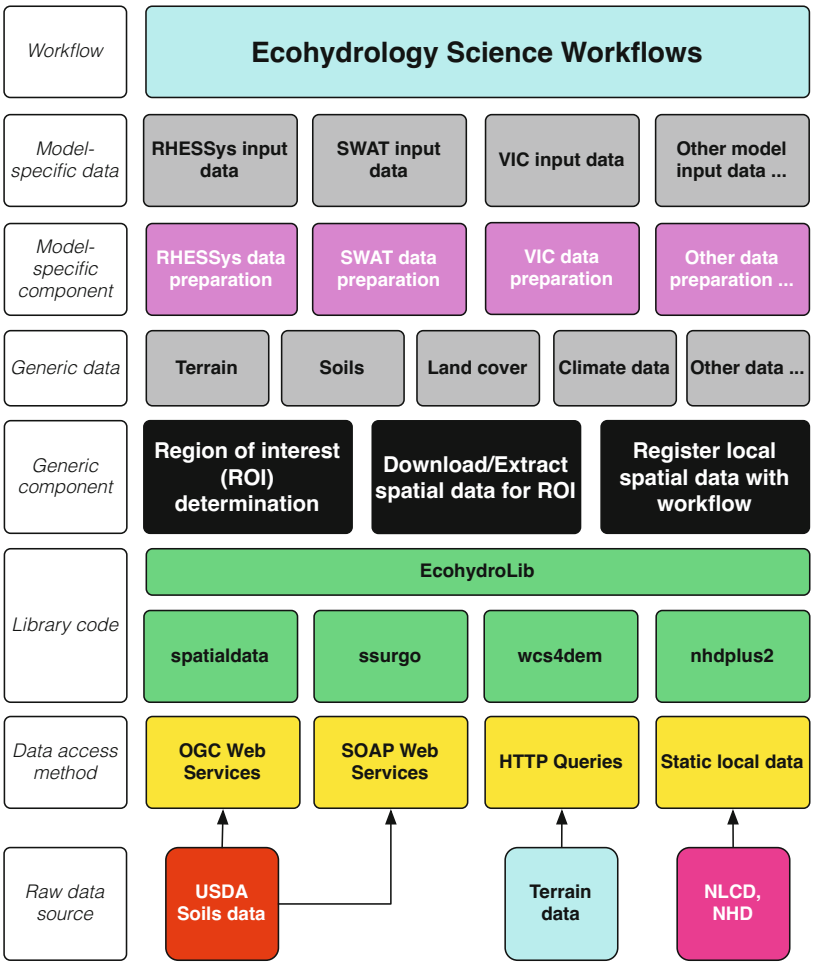

Fig. 1. Illustration of layered architecture of EcohydroLib. Read from the bottom, raw data sources are accessed using web services (yellow), or as static data stored on local files systems (yellow), using library code (green) in EcohydroLib. Generic data acquisition tools (black) are built on top of EcohydroLib library code. These tools yield data in generic formats not specific to any model (grey). Model-specific tools (magenta) use these generic data to produce model-specific input data parameterizations (grey, above magenta). The model specific data make possible ecohydrology modeling science workflows (cyan) whose goal is to answer particular science and management questions.

local disk specific to each project (i.e. the project directory). These workflows are orchestrated via a metadata store in the project directory. Workflow commands are built using task-oriented APIs defined in the package *ecohydrolib*. These commands provide tools for downloading and manipulating geospatial data needed to run ecohydrology models, information such as: digital elevation model (DEM), soils, land cover. The metadata store, essentially a key-value store (e.g. a dictionary), is used to

orchestrate a series of workflow commands used to prepare data for an ecohydrology model. The metadata contain information related to the study area (e.g. bounding box coordinates, spatial reference, spatial resolution), provenance information for each spatial data layer imported, and a processing history that records the order in which commands were run as well as the parameters used to invoke each command. Currently, EcohydroLib allows modelers to work with geospatial data stored locally as well as high spatial resolution U.S. national spatial data available via web services, for example 30-meter digital elevation model and land cover data, and 1:12,000 scale soils data. While researchers working in watersheds outside the U.S. can use EcohydroLib, they must manually download data for their study areas before these data can be imported into EcohydroLib workflows.

Compared to systems for manually downloading geospatial data, web services can make it easier to acquire and integrate such data into geoscience workflows. Manual-download workflows typically require a modeler to use a web browser-based GIS tool to define the region of interest and choose a data product to download before being able to download data, perhaps after a delay of minutes to hours while the data are fetched from offline storage or otherwise processed. Once downloaded, these data may consist of several image tiles that need to be joined together before being usable in an ecohydrology model. Instead, web services can offer more-or-less instant access to seamless geospatial data covering the entire study area. Further, when accessed via workflow tools that store study area information (e.g. geographic bounding box, spatial reference) such as EcohydroLib, users are not required to specify spatial information each time within a given workflow that they acquire datasets via web services. Lastly, compared to manual-download systems, web services make it possible to acquire geospatial data in spatial reference systems and resolutions specific to each workflow, with conversions being done on the server side, reducing the work required of the modeler. It must be noted that both web services and manual-download data acquisition workflows require the modeler to take care when combining geospatial datasets to ensure that the data being combined are of commensurate spatial scales or resolutions as well as being temporally and semantically compatible.

To be of use in numerical ecohydrology models, web services must provide data as Open Geospatial Consortium (OGC) Web Coverage Service (WCS) end points (http://www.opengeospatial.org/standards/wcs) or OGC Web Feature Service (WFS; http://www.opengeospatial.org/standards/wms), rather than as OGC Web Map Service (WMS; http://www.opengeospatial.org/standards/wms), which are of use primarily for cartography applications. While national agencies in the U.S. and Australia are offering some datasets via web services, these are either lower resolution datasets or are offered as WMS end points. In this paper we explore: (1) availability of high-resolution national geospatial data web services in the United States and Australia; and (2) integration of Australian web services with EcohydroLib.

2 Availability of National Spatial Data Web Services in Australia and the United States

Geospatial datasets needed to parameterize process-based ecohydrology models include: digital elevation model (DEM); land cover; soil surface texture; and vegetation leaf area index. Some of these data are available via web services interfaces as national coverages

for both Australia and the United States (Table 1). In Australia, all data listed are provided by Geoscience Australia (GA). Currently, Australia has national-scale 1-second (~30-m) spatial resolution DEM data available via a WCS interface, though these data date from the 2000 Shuttle Radar Topography Mission. There is a relatively coarse 250-m land cover data dataset for Australia, however at present the GA only offers a WMS web service for these data, not WCS. Soils data (e.g. Australian Soil Resource Information System; ASRIS) do not appear to be available via web services at this time, though a web browser-based download tool is provided by CSIRO (the Commonwealth Scientific and Industrial Research Organisation) and the Department of Agriculture, Fisheries and Forestry (http://www.asris.csiro.au/).

Table 1. U.S. and Australia national geospatial data of use in ecohydrology modeling available via web services interfaces. AU = Australia, US = United States, GA = Geoscience Australia, USGS = U.S. Geological Survey, NLCD = National Land Cover Database, ORNL DAAC = Oak Ridge National Laboratory Distributed Active Archive Center for Biogeochemical Dynamics, NASA = National Aeronautics and Space Administration, GMU = George Mason University, USDA = U.S. Department of Agriculture. More info on GA web services: http://www.ga.gov.au/data-pubs/web-services/ga-web-services USGS web services: http://viewer.nationalmap.gov/example/services/serviceList.html ORNL web services: http://webmap.ornl.gov/wcsdown/index.jsp GeoBrain WCS4DEM web service: http://geobrain.laits.gmu.edu/wcs4dem.htm SSURGO web services: http://sdmdataaccess.nrcs.usda.gov

Country	Dataset	Spatial resolution	Web service type	Hosting organization	Notes
AU	Land cover	N/A	WMS	GA	
AU	DEM	9-second (~250-m)	WCS, WMS	GA	
AU	DEM (SRTM)	1-second (~30-m)	WCS, WMS	GA	
AU	LiDAR	20-cm	WCS	GA	Murray Darling basin only
AU	DEM (LiDAR)	1-m	WCS	GA	Murray Darling basin only
US	NLCD (1992-2011)	N/A	WMS	USGS	
US	NLCD (2011)	30-m	WCS	USGS	
US	NLCD (1992-2011)	30-m	WCS	ORNL DAAC / NASA	

Table 1. (*continued*)

US	**DEM (SRTM, NED)**	**30-m**	**WCS**	**GeoBrain GMU / NASA**
US	Soil Survey Geographic Database (SSURGO)	~1:12,000	WFS, SOAP	USDA

In the U.S., DEM data are available via WCS4DEM, a WCS service provided by the GeoBrain project at George Mason University and funded by NASA (Table 1); both high-resolution U.S. national data (National Elevation Dataset, NED; SRTM) as well as global DEM data (e.g. GLSDEM, GTOPO) coverages are available via WCS4DEM. High spatial resolution (30-m) land cover data for the continental U.S. are available over a number of years (1992, 2001, 2006, and 2011) as part of the National Land Cover Database (NLCD), which is available for download via WCS interfaces from the NASA Distributed Active Archive (DAAC) for Biogeochemical Dynamics center at Oak Ridge National Laboratory (ORNL) for years 1992, 2001, and 2006, as well as through the U.S. Geological Survey for 2011 (Table 1). Finally, U.S. soils data are available via WFS and Simple Object Access Protocol (SOAP) web services interfaces provided by the U.S. Department of Agriculture (USDA).

Unfortunately, at present there are no national-scale high spatial resolution vegetation leaf area index (LAI) data products available for either the U.S. or Australia via WCS interfaces. However, it is possible to derive 30-m peak LAI data from Landsat satellite data as well as LAI phenology (e.g. green-up, senescence) from MODIS satellite data, which are available globally. ORNL DAAC offers seasonal and yearly Landsat mosaics for 2008 and 2009 via WCS (http://webmap.ornl.gov/wcsdown/dataset.jsp?ds_id=111112) from the USGS Web-enabled Landsat Data (WELD) project (http://weld.cr.usgs.gov/). These data could serve as a data source for a prototype WCS-enabled peak LAI service.

3 Integrating EcohydroLib with U.S. and Australian National Spatial Data Web Services

When EcohydroLib 1.0 was released in July 2013, it provided the ability to acquire the following geospatial datasets from remote web services: U.S. and global DEM data from WCS4DEM, NLCD 2006 data from ORNL DAAC, and U.S. surface soil texture data from USDA SSURGO. Here we describe an additional module recently developed for EcohydroLib 1.20 to provide access to national-scale 1-second (~30-m) spatial resolution SRTM-based DEM data provided by Geoscience Australia. Geoscience Australia (GA) provides 1-second spatial resolution DEM data for all of mainland Australia and Tasmania based on the 2000 Shuttle Radar Topography Mission data. These data are offered in three forms: unsmoothed data (DEM); smoothed data

(DEM-S; with noise removed); and a hydrologically enforced version of the smoothed data (DEM-H; which includes hydrologic flowpaths derived from both SRTM data and mapped streams). The hydrologically enforced data are deemed to be suitable for watershed delineation (Geoscience Australia 2011). EcohydroLib provides access to all three datasets using the *GetGADEMForBoundingBox* command (for a full description EcohydroLib, including installation instructions and tutorials, see: https://github.com/selimnairb/RHESSysWorkflows). The EcohydroLib workflow steps required to acquire GA DEM data are summarized in Fig. 2.

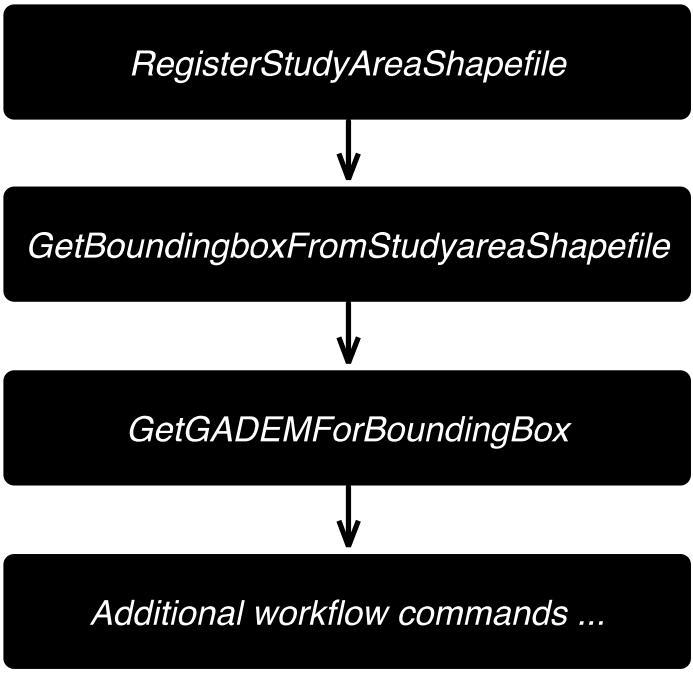

Fig. 2. Initial EcohydroLib workflow steps including acquisition of DEM data from Geoscience Australia WCS web service interface

The first workflow step, *RegisterStudyAreaShapefile*, defines the spatial extent of the study area using a single polygon encoded in an ESRI shapefile file format. This shapefile, accessed locally via a disk mounted on the computer where EcohydroLib is running, is copied into the project directory specified to *RegisterStudyAreaShapefile*; as it is copied, the shapefile is reprojected into the WGS84 geographic coordinate system (EPSG:4326). This reprojected shapefile can be visualized using a GIS such as QGIS.

Once the study area is defined, the bounding box of the study area must be derived from the study area polygon using the *GetBoundingboxFromStudyareaShapefile* command, which stores the bounding box in the *metadata.txt* file stored in the project directory. The bounding box (a.k.a. minimum bounding rectangle) is defined by two coordinate pairs that represent the upper right (e.g. northeast) and lower left

(e.g. southwest) corners of a rectangle that circumscribes the study area polygon. With the bounding box defined, it is possible to obtain GIS data for the study area via web services; by default *GetBoundingBoxFromStudyareaShapefile* will slightly buffer the bounding box rectangle to help ensure that any data downloaded provide sufficient coverage of the study area watershed.

The *GetGADEMForBoundingBox* command is used to download the 1-second DEM data from web services provided by Geoscience Australia. The type of DEM to be downloaded (e.g. DEM, DEM-S, or DEM-H) can be specified using the –d (a.k.a. –demType) option. A sample DEM for an example study area is shown in Fig. 3. By default, *GetGADEMForBoundingBox* will reproject the DEM into the UTM (WGS84) zone appropriate for the centroid of the study area; the native resolution of the DEM will be maintained. Both the spatial reference and resolution of the DEM can be specified as optional arguments to *GetGADEMForBoundingBox*. With the DEM data in hand, subsequent EcohydroLib commands can be run to acquire and manipulate data needed to run ecohydrology models (Fig. 2). All subsequent data imported into the project (either acquired via web services, or imported from locally stored data) will be resampled to the spatial reference system and resolution of the DEM; the user can choose any re-sampling method supported by GDAL, or can disable raster re-sampling on import. When working in Australia, land cover, soils, and other necessary geospatial data would need to be downloaded by hand before being imported into an EcohydroLib project (e.g. via the *RegisterRaster* command), given the lack of suitable web services for these data (see above).

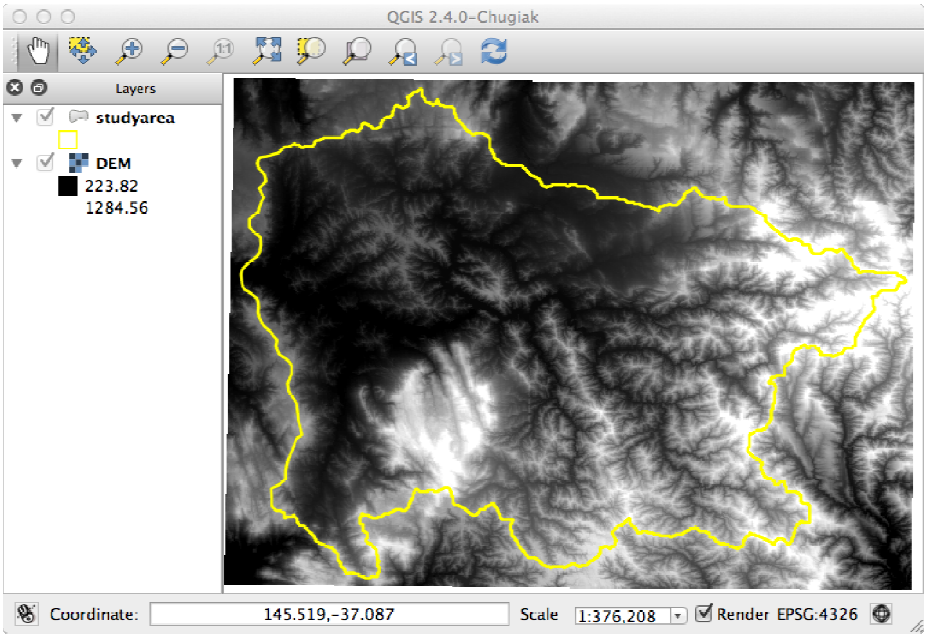

Fig. 3. QGIS visualization of Geoscience Australia DEM acquired for bounding box of study area defined by watershed shapefile

4 Discussion and Conclusion

In this paper we have described common high spatial resolution (<=30-m) geospatial data needed for parameterizing ecohydrology models, and listed known OGC web services-based interfaces for these data for the United States and Australia. We have also described a new module for EcohydroLib that provides access to high spatial resolution (~30-m) national-scale Australian DEM data accessible via OGC WCS interfaces provided by Geoscience Australia. While this new module does not address all data gaps for EcohydroLib users working in Australia, it is a useful first step toward providing easy access to geospatial data needed for parameterizing ecohydrology models for Australian catchments. Once other necessary geospatial datasets are available via suitable OGC web services interfaces (e.g. high resolution land cover raster data via WCS, and soils raster or vector data via WCS or WFS), these can be similarly integrated into EcohydroLib, providing similar ease of access currently afforded to users working in U.S. watersheds. In the mean time, EcohydroLib can still be of benefit when applied to Australia due to its ability to improve metadata and provenance information capture, even for datasets downloaded manually. Further, it is our hope that this work will show the benefits to the water science community when providers of national geospatial data make these data available via OGC WCS and WFS web services interfaces required for integration with numerical models, rather than cartography-oriented WMS services. Such web services, when integrated with tools like EcohydroLib, hold the potential to enable transformative water science by improving scientific reproducibility and researcher productivity while making model and site inter comparisons easier to achieve.

References

1. Band, L.: Effect of land surface representation on forest water and carbon budgets. Journal of Hydrology 150(2-4), 749–772 (1993), doi:10.1016/0022-1694(93)90134-u
2. Band, L.E., Cadenasso, M.L., Grimmond, C.S., Grove, J.M., Pickett, S.T.A.: Heterogeneity in urban Ecosystems: Patterns and Processes. In: Lovett, G.M., Turner, M.G., Jones, C.G., Weathers, K.C. (eds.) Ecosystem Function in Heterogeneous Landscapes, New York, pp. 257–278 (2005), http://dx.doi.org/10.1007/0-387-24091-8_13
3. Deelman, E., et al.: Workflows and e-Science: An overview of workflow system features and capabilities. Future Generation Computer Systems 25(5), 528–540 (2009)
4. Duffy, C., et al.: Designing a Road Map for Geoscience Workflows. Eos 93(24), 225–226 (2012), http://www.agu.org/pubs/crossref/2012/2012EO240002.shtml
5. Geoscience Australia, Metadata for SRTM-derived 1 Second Digital Elevation Models Version 1.0 (2011), http://www.ga.gov.au/metadata-gateway/metadata/record/gcat_72759
6. Goble, C.A., et al.: myExperiment: a repository and social network for the sharing of bioinformatics workflows. Nucleic Acids Research 38(Web Server), W677–W682 (2010)
7. Guo, D., et al.: Scientific workflow challenges. In: WIRADA Science Symposium Proceedings, Melbourne, Australia, August 1-5, pp. 54–60 (2012)

8. Plale, B.: The challenges and opportunities of workflow systems in environmental research. In: WIRADA Science Symposium Proceedings, Melbourne, Australia, August 1-5, pp. 48–53 (2012)
9. Rodríguez-Iturbe, I.: Ecohydrology: a hydrologic perspective of climate-soil-vegetation dynamics. Water Resources Research 36(1), 3–9 (2000)