Environmental Modelling & Software 52 (2014) 51–61

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Data visualization and analysis within a Hydrologic Information System: Integrating with the R statistical computing environment

Jeffery S. Horsburgh*, Stephanie L. Reeder

Utah Water Research Laboratory, Utah State University, 8200 Old Main Hill, Logan, UT 84322-8200, USA

ARTICLE INFO

Article history:
Received 14 May 2013
Received in revised form
11 October 2013
Accepted 15 October 2013
Available online 13 November 2013

Keywords:
Data analysis
Data management
CUAHSI HIS
Hydrologic observations
R

ABSTRACT

This paper presents a prototype software system for visualization and analysis of hydrologic data that provides interoperability between the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) Hydrologic Information System (HIS) and the R statistical computing environment. By linking these two systems within a single desktop software application, an integrated hydrologic data management and analysis environment has been created that simplifies the process used by scientists and engineers to find, access, organize, and analyze the hydrologic data needed in modeling and managing hydrologic and environmental systems. The implementation of this work is a software plug-in for the CUAHSI HIS HydroDesktop software system called HydroR. We describe the design, graphical user interface, and implementation of the HydroR plug-in. An example application of HydroR is presented in which total suspended solids concentrations are modeled for the Little Bear River using a regression developed from turbidity and total suspended solids observations downloaded from the CUAHSI HIS using HydroDesktop. Finally, we conclude with a summary of our experience in developing interoperability between HIS and R and suggest future developments that can extend the capabilities we have developed.

Software availability

Name of software: HydroR plug-in for HydroDesktop
Developers: Jeffery S. Horsburgh and Stephanie L. Reeder
Contact: jeff.horsburgh@usu.edu or stephanie.reeder@usu.edu
Year first available: 2012
Hardware required: PC running Microsoft Windows
Software required: HydroDesktop – available at http://hydrodesktop.codeplex.com R Version 2.14.0 and up
Software availability: all source code and documentation for the HydroR plug-in for HydroDesktop, the HydroR package for R, and the main HydroDesktop software application can be accessed at http://hydrodesktop.codeplex.com.
Cost: Free. Software and source code are released under the New Berkeley Software Distribution (BSD) License, which allows for liberal reuse of the software and code.

1. Introduction

Scientists and engineers who study and manage hydrologic and environmental systems require the ability to discover, access, manage, and analyze data from multiple sources, projects, and research efforts. It is important, therefore, to consider approaches for accomplishing these disparate tasks. While there are many examples of data management and analysis systems as separate tools (e.g., R and Matlab), there are fewer examples of data discovery systems capable of accessing data from many different sources and an even smaller number of integrated systems capable of handling all of these activities. Indeed, some scientists report spending as much as 50–80% of their time in the discovery, access, and management activities (Bandaragoda et al., 2006), leaving less of their time available for the necessary visualization, analysis, and modeling activities needed to investigate complex hydrologic processes. This imbalance is likely due, in part, to the lack of interoperable tools for integrating all of these activities.

The recent development of standards for sharing and accessing observational data within the hydrology community, along with the availability of robust, freely available data visualization and analysis software suggest a path forward. This paper focuses on two existing software technologies, the Consortium of Universities for

* Corresponding author. Tel.: +1 435 797 2946; fax: +1 435 797 3663.
E-mail address: jeff.horsburgh@usu.edu (J.S. Horsburgh).

1364-8152/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.envsoft.2013.10.016

the Advancement of Hydrologic Science, Inc. (CUAHSI) Hydrologic Information System (HIS) and the R statistical computing environment. We explored how these two technologies can be combined to form an integrated hydrologic data discovery, management, visualization, and analysis environment. The CUAHSI HIS was developed with the goal of enhancing access to hydrologic data (Maidment, 2008; Tarboton et al., 2009). The R statistical computing software, on the other hand, was developed as a language and environment for statistical computing and graphics. Developed by very different communities, there was no mechanism for using the CUAHSI HIS and R together.

The objective of this research was to explore how interoperability between the data discovery, access, and management capabilities of the CUAHSI HIS and the existing data analysis, visualization, and modeling capabilities of other software packages could be achieved, using R as a specific example. The CUAHSI HIS and R are complementary software technologies. The strengths of CUAHIS HIS are in supporting data discovery, access, and management, whereas the strengths of R lie in data analysis, visualization, and modeling. Using the knowledge we gained, we developed a prototype software application called HydroR as a plug-in to the CUAHSI HydroDesktop software system (Ames et al., 2012). The combination of HydroDesktop and HydroR enables users to search for, download, and store hydrologic datasets in a local data repository database and then access those datasets directly from R for visualization and analysis via a tightly integrated R Console and scripting window.

In Section 2 we provide further background on the CUAHSI HIS and R. Next, we present our approach for integrating the CUAHSI HIS and R and the challenges that we overcame. We then present HydroR as a prototype application that provides the ability to create visualizations and analyses in R that use HIS data. An example is then used to demonstrate how these systems can be applied in the discovery, access, visualization, and analysis of hydrologic data. Finally we conclude with a summary of the research results.

2. Background

2.1. CUAHSI HIS

The CUAHSI HIS was developed to bring together hydrologic observations from multiple sources across the United States and internationally into a uniform, standards-based, service-oriented environment where heterogeneous data can be seamlessly integrated for advanced computer-intensive analysis and modeling (Tarboton et al., 2009, 2011). CUAHSI HIS includes three main components that together enable publishing and accessing hydrologic data collected at point locations (e.g., time series of observations from stream gages, water quality monitoring sites, weather stations, etc.). The first is HydroServer, which is a software stack for publishing hydrologic observations (Horsburgh et al., 2009, 2010). HydroServer relies on a relational database schema called the Observations Data Model (ODM) (Horsburgh et al., 2008) for storing observational time series. Data within an ODM database are published using a standard web service interface called WaterOneFlow that exposes data from the database in a standard XML schema called Water Markup Language (WaterML) (Zaslavsky et al., 2007; Taylor, 2012). HIS Central, which is the second component, is a web service registry and metadata catalog that provides search services across all of the federated HydroServer instances. Metadata from registered WaterOneFlow web services are regularly harvested and cataloged at HIS Central, and the metadata catalog is exposed via a web service search interface that can be called by client applications to search the metadata catalog.

The third component of HIS is HydroDesktop (Ames et al., 2012). HydroDesktop is a desktop client software application that enables users to search the HIS Central metadata catalog using spatial, temporal, and keyword-based criteria and then access discovered data by making web service calls against HydroServers. HydroDesktop uses a local data repository database that has a schema similar to ODM to store datasets downloaded from HydroServers on a user's local computer. The HydroDesktop graphical user interface (GUI) is built upon open-source geographic information systems (GIS) software called DotSpatial (Cao and Ames, 2012; http://dotspatial.codeplex.com) and allows for extension of the core HydroDesktop functionality through development of plug-ins. Plug-ins can be developed using the C# programming language (Microsoft .NET Framework 4.0) using a HydroDesktop plug-in interface standard. HydroR was developed as a plug-in that adds data visualization and analysis capabilities to HydroDesktop.

2.2. Scientific data analysis in R

R was originally developed for statistical computing. Over the past decade, it has seen tremendous growth as a scientific computing environment for several reasons. First, R is both a programming language and a software environment that sustains a wide variety of computational techniques (linear and nonlinear modeling, time series analysis, statistical tests, and others) (Soetaert and Meysman, 2012). The R software environment has a base set of functionality that helps in performing a number of computing tasks. R includes a data handling and storage facility, a suite of operators for calculations on arrays and matrices, a large collection of tools for data analysis, and graphical facilities for data analysis and display. Additionally, R is open source and highly extensible. Users can improve the base code of R and can also write extensions, or packages, that add functionality to the base implementation (R Development Core Team, 2010). Another factor that has contributed to the success of R is the existence of the Comprehensive R Archive Network (http://CRAN.R-project.org/), which is a single web repository that archives a wide range of extension packages for R. These packages are formally quality controlled and documented, and the existence of this extensive library of packages means that package developers can build upon existing packages, or include them by reference so they do not have to copy or re-develop needed functionality.

Although R was originally developed for statistical computing, its scope has been extended to the field of environmental data visualization, analysis, and modeling. Scientific computing using R ranges from users who develop custom R scripts for specific analyses, to developers who write entire R packages with functionality that includes specific visualizations and analyses. For example, Pino-Mejías et al. (2010) implemented data mining models for the prediction of potential habitats for the oak forest type in Mediterranean areas of southern Spain using specific functions available within R. Petzoldt and Rinke (2007) describe an object-oriented framework for ecological modeling in R and implemented as an R package that provides an open structure to implement, simulate, and share ecological models. Carslaw and Ropkins (2012) developed an R package called "openair" for analysis of air pollution measurement data, and Soetaert and Meysman (2012) developed a set of R packages for modeling reactive transport of water quality constituents in rivers.

Additionally, there are more advanced software development efforts that encapsulate R or R functions within other software. A number of GUI tools are available that include interactive development environments (IDEs) or script editors that aim to provide feature-rich environments to edit R scripts and code. Examples include Notepad++/NppToR (http://npptor.sourceforge.net/) and