Environmental Modelling & Software 54 (2014) 60–72

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

Development of a knowledge library for automated watershed modeling

Mateja Škerjanec [a,b,*], Nataša Atanasova [c,d], Darko Čerepnalkoski [e], Sašo Džeroski [e,f,g], Boris Kompare [a]

[a] University of Ljubljana, Faculty of Civil and Geodetic Engineering, Jamova cesta 2, 1000 Ljubljana, Slovenia
[b] Water Science Institute, Hajdrihova 28a, 1000 Ljubljana, Slovenia
[c] University of Algarve, Centre for Marine and Environmental Research (CIMA), Campus de Gambelas, 8005-139 Faro, Portugal
[d] International Centre for Coastal Ecohydrology (ICCE), Solar do Capitão Mor, EN125 Horta das Figuras, 8000-518 Faro, Portugal
[e] Jožef Stefan Institute, Department of Knowledge Technologies, Jamova cesta 39, 1000 Ljubljana, Slovenia
[f] Jožef Stefan International Postgraduate School, Jamova cesta 39, 1000 Ljubljana, Slovenia
[g] Centre of Excellence for Integrated Approaches in Chemistry and Biology of Proteins, Jamova cesta 39, 1000 Ljubljana, Slovenia

A R T I C L E I N F O

Article history:
Received 12 March 2013
Received in revised form
29 November 2013
Accepted 23 December 2013
Available online 19 January 2014

Keywords:
Watershed model
Hydrology
Nutrient loading
Knowledge library
Automated modeling

A B S T R A C T

In this work, we develop a library of components for building semi-distributed watershed models. The library incorporates basic modeling knowledge that allows us to adequately model different water fluxes and nutrient loadings on a watershed scale. It is written in a formalism compliant with the equation discovery tool ProBMoT, which can automatically construct watershed models from the components in the library, given a conceptual model specification and measured data. We apply the proposed modeling methodology to the Ribeira da Foupana catchment to extract a set of viable hydrological models. By specifying the conceptual model and using the knowledge library, two different hydrological models are generated. Both models are automatically calibrated against measurements and the model with the lower root mean squared error (RMSE) value is selected as an appropriate hydrological model for the selected study area.

Software availability

Name: Process-Based Modeling Tool (ProBMoT)
Developer and contact address: Darko Čerepnalkoski, Jožef Stefan Institute, Department of Knowledge Technologies, Jamova cesta 39, 1000 Ljubljana, Slovenia. Email: darko.cerepnalkoski@ijs.si
Year first available: 2012
Hardware requirements: x86 architecture, 1 GHz processor, 1 GB memory
Operating System: MS Windows (32 or 64 bit), GNU/Linux (32 or 64 bit)
Software requirements: Java 6 (32 or 64 bit)
Program language: Java
Program size: 15 MB
Availability: Available upon request from the authors

1. Introduction

Watershed modeling is recognized as a useful tool for evaluating the effects of land and water management practices on natural resources. It usually operates between various compartments, e.g., land and water, and involves many different disciplines, such as hydrology, agriculture, water management, and others. Thus, its complexity and transdisciplinary nature make it a part of the science of integrated environmental modeling (IEM, Laniak et al., 2013).

The environmental modeling community has been actively developing various watershed models, such as SWAT (Arnold and Fohrer, 2005), SPARROW (Schwarz et al., 2006) and GWLF (Haith and Shoemaker, 1987). These mostly differ in the way of conceptualizing the catchment, in the level of detail in describing catchment processes, in the specific mathematical formulations, and in the data requirements for simulation. Most models are only used by

* Corresponding author. University of Ljubljana, Faculty of Civil and Geodetic Engineering, Jamova cesta 2, 1000 Ljubljana, Slovenia. Tel.: +386 1 425 40 52; fax: +386 1 251 98 97.

E-mail addresses: mateja.skerjanec@fgg.uni-lj.si (M. Škerjanec), natasa.atanasova@icce-unesco.org (N. Atanasova), darko.cerepnalkoski@ijs.si (D. Čerepnalkoski), saso.dzeroski@ijs.si (S. Džeroski), boris.kompare@fgg.uni-lj.si (B. Kompare).

1364-8152/$ – see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.envsoft.2013.12.017

a few research groups, some gather a following within a region, and only a few become widespread.

Although it is often perceived to be easier to create a new model than to reuse an existing one, the benefits of model exchange and reuse have been largely recognized (Holzworth et al., 2010). From the user's perspective, the selection of a suitable (existing) model remains a difficult task, since various models can adequately fulfill the requirements. In other words, there is usually no single suitable model for a specific system. The model choice is guided by the modeling requirements, data availability, and modeler's skills.

To overcome this problem, a modeling concept was proposed that does not focus on searching for the best suitable model but rather on the most appropriate combination of modeling blocks (also termed modules or components) to represent the observed system (Leavesley et al., 2002; Argent, 2004). As a result larger software frameworks providing reusable components for building environmental models have been developed, e.g., the E2 (Argent et al., 2009), the Source IMS (Welsh et al., 2013) and the Object Modeling System (OMS, David et al., 2013). Such frameworks offer an infrastructure that supports inter-model communication. They usually comprise framework-dependent libraries containing reusable modules for simulating a variety of processes, where each single module represents a computable model of a part of the system. Within the selected environmental modeling framework, these executable components are coupled through the use of standard interfaces for data exchange (e.g., the OpenMI, Gregersen et al., 2007) in order to construct an integrated model of the observed system.

Libraries supporting environmental modeling frameworks usually focus on the collection and documentation of previously developed legacy models, considering each model as a single executable module. One of the most notable examples is the Library of Hydro-Ecological Modules (LHEM) introduced by Voinov et al. (2004). LHEM incorporates modules for the simulation of hydrological processes and nutrient cycling, along with other processes. Modules encoded in LHEM may be used either as stand-alone models to describe certain processes and ecosystem components, or may be put together into more complex structures by using the SME model building environment (Voinov et al., 1999).

Unlike the traditional component-based modeling approach, where each module represents a single executable model, we are introducing a new modeling approach that uses a declarative formalism for describing the system. Although we also decompose the modeling domain into several components, none of them represent an executable model by itself; rather, they represent the entities and processes involved in the domain of study. While entities correspond to the actors of the observed system, processes are used to define the relationships among them. Within the automated model generation procedure, all components are compiled together in order to produce a global model that encompasses all parts of the system.

In this work, we present (1) a domain-specific library that contains formalized watershed modeling knowledge and (2) a case study demonstrating the utility of the developed library for the extraction of viable watershed-scale hydrological models. At present, the knowledge in the library comprises hydrological processes, based on meteorological data, and nutrient loading processes, considering point and diffuse emission sources. Moreover, the library includes alternative formulations for the selected processes.

The library can be considered as an ontology, because it consists of organized and structured modeling knowledge. Similarly to other ontologies, it defines concepts (i.e., entities and processes) and the relationships between them. The taxonomies of entities and processes provide the inheritance (is-a) relation that is the essential part of an ontology. However, our domain library is much richer than the typical ontology. Besides listing the concepts and taxonomical relations between them, our library contains their properties (i.e., variables and constants for entities and equations for processes). The latter are the basic components that are put together to construct dynamical models in the form of systems of differential and/or algebraic equations.

The library can be used as a repository of modeling components when handcrafting semi-distributed watershed models running at a daily time step. The true usefulness of the library comes from its use by an automated modeling tool, such as Lagramge (Džeroski and Todorovski, 2003), HIPM (Todorovski et al., 2005), or the recently developed ProBMoT (Čerepnalkoski et al., 2012). These tools allow automatic induction of suitable models based on the libraries of domain-specific modeling knowledge and the measured data.

By developing a watershed library compliant with the ProBMoT, we are establishing a novel approach to automated modeling (AM) of watersheds that uses a combination of theoretical and data driven modeling. A similar approach, based on an aquatic ecosystem library (Atanasova et al., 2006), was successfully applied for lake food web modeling (Atanasova et al., 2011, 2008). However, to our knowledge, no such attempts have been made at a watershed scale, despite the similarity between the two modeling problems.

The paper is organized as follows. First, in Section 2, we briefly explain the watershed modeling domain. Next, in Section 3, we present the formalism for encoding the watershed modeling knowledge into the library and explain how the modeling task specification is included in the model induction procedure. In the following section (Section 4) we introduce ProBMoT. In Section 5, we present the model generation, calibration and validation for an experimental watershed. This is followed by a discussion (Section 6). Finally, conclusions and guidelines for further work are given in Section 7.

2. Watershed processes and modeling

When simulating the loadings of water quality constituents (sediment and nutrients) from watersheds, two basic groups of processes have to be taken into account: hydrological processes and constituent generation processes from various land use types, triggered by the water movement. The difference between precipitation and water losses (evapotranspiration, infiltration, percolation) results in the surface runoff and subsurface (groundwater) discharge. The generated surface and groundwater flows provoke soil erosion and constituent wash-off to various surface water recipients and ground water reservoirs. Besides the natural water cycle and related constituent loadings, we also have to consider human-generated water flows, such as septic effluents and other point sources (waste water treatment plants – WWTPs, industry discharges, and others) rich with nutrients.

Existing watershed models simulate these processes at different levels of detail. Physically based dynamic models (e.g., SWAT, Arnold and Fohrer, 2005; HSPF, Donigian et al., 1995) have a highly complex mass-balance structure and provide the best representation of the current understanding of watershed processes affecting pollution generation. However, the parameterization and calibration of this kind of models can be very difficult and time consuming. In contrast, the predominantly empirical steady state models (e.g., SPARROW, Schwarz et al., 2006; MONERIS, Behrendt et al., 1999) are compilations of expert knowledge and empirical relationships between the physiographic characteristics of the watershed and constituent loadings. Empirical models are conceptually simple and tend to be less expensive to implement compared to more physically based approaches. They commonly