Environmental Modelling & Software 61 (2014) 121–134

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

A comprehensive approach to evaluating watershed models for predicting river flow regimes critical to downstream ecosystem services

Seyoum Y. Gebremariam [a, e, *], Jay F. Martin [a], Carlo DeMarchi [b], Nathan S. Bosch [c], Remegio Confesor [d], Stuart A. Ludsin [e]

[a] Food, Agricultural and Biological Engineering Department, The Ohio State University, Columbus, OH 43210, USA
[b] Department of Earth, Environmental, and Planetary Sciences, Case Western Reserve University, Cleveland, OH 44106, USA
[c] Environmental Science, Grace College, Winona Lake, IN 46590, USA
[d] National Center for Water Quality Research, Heidelberg University, Tiffin, OH 44883, USA
[e] Aquatic Ecology Laboratory, Department of Evolution, Ecology, and Organismal Biology, The Ohio State University, Columbus, OH 43212, USA

A R T I C L E I N F O

Article history:
Received 10 October 2013
Received in revised form
2 July 2014
Accepted 3 July 2014
Available online

Keywords:
Watershed modeling
Eutrophication
Runoff
Great Lakes
Lake Erie
Allochthonous inputs
Non-point source pollution
Algal bloom

A B S T R A C T

Selection of strategies that help reduce riverine inputs requires numerical models that accurately quantify hydrologic processes. While numerous models exist, information on how to evaluate and select the most robust models is limited. Toward this end, we developed a comprehensive approach that helps evaluate watershed models in their ability to simulate flow regimes critical to downstream ecosystem services. We demonstrated the method using the Soil and Water Assessment Tool (SWAT), the Hydrological Simulation Program—FORTRAN (HSPF) model, and Distributed Large Basin Runoff Model (DLBRM) applied to the Maumee River Basin (USA). The approach helped in identifying that each model simulated flows within acceptable ranges. However, each was limited in its ability to simulate flows triggered by extreme weather events, owing to algorithms not being optimized for such events and mismatched physiographic watershed conditions. Ultimately, we found HSPF to best predict river flow, whereas SWAT offered the most flexibility for evaluating agricultural management practices.

1. Introduction

Many of the world's coastal and lake ecosystems that drain large agricultural watersheds are experiencing degraded water quality, including noxious algal blooms, hypoxia, and reduced water clarity (Cloern, 2001; O'Neil et al., 2012; Diaz and Rosenberg, 2008; Rabalais et al., 2009; Michalak et al., 2013). Watershed flow regimes have been shown to be drivers of such conditions by influencing nutrient runoff into the downstream environment (Donner et al., 2002; Vidon et al., 2009), and therefore need to be considered in nutrient mitigation or rehabilitation strategies (Royer et al., 2006; Scavia et al., 2014). Numerous factors interact to govern river outflows from the watershed, including topography, meteorology (e.g., precipitation, temperature), soil characteristics, and land-use practices and management (DeFries and Eshleman, 2004). Owing to the complexity of factors that control hydrologic processes, finding a way to reliably model flow regimes that are critical to stream ecology and downstream ecosystem services can be challenging. However, doing so is absolutely critical, if land-use planners and water-quality managers are to succeed in protecting downstream water bodies (DeFries and Eshleman, 2004; Royer et al., 2006).

To help research and management communities make well-informed choices regarding hydrology models, we describe a comprehensive approach to evaluate model performance in predicting river flow regimes critical to downstream ecosystem services. The approach was used to evaluate three commonly used

Abbreviations: SWAT, Soil and Water Assessment Tool; DLBRM, Distributed Large Basin Runoff Model; HSPF, Hydrology Simulation Program—Fortran; GOF, goodness-of-fit; SCS, Soil Conservation Service; HRU, hydraulic response unit; STATSGO, state soil geographic; USGS, U.S. Geological Survey; BMP, best management practice.

* Corresponding author. Aquatic Ecology Laboratory, Department of Evolution, Ecology, and Organismal Biology, The Ohio State University, Columbus, OH 43212, USA. Tel.: +1 208 596 5165; fax: +1 614 292 9448.
E-mail address: gebremariam.6@osu.edu (S.Y. Gebremariam).

http://dx.doi.org/10.1016/j.envsoft.2014.07.004
1364-8152/

watershed models, SWAT (version 528.0; Arnold et al., 1998), DLBRM (version 2004; Croley and He, 2005) and HSPF (version 12.0; Bicknell et al., 2001), in their ability to accurately quantify various flow-regime components of the Maumee River Basin, the largest watershed in the Great Lakes region of North America. We assessed the models in terms of (1) daily and monthly flow, (2) flood and low-flow pulse frequency, magnitude and duration, and (3) watershed response to extreme weather events. The models also were compared in terms of their ease of use. While our model comparison centers on the Maumee River watershed, our findings should have general application to other large watersheds and provide a better framework for future model assessment efforts.

2. Materials and methods

2.1. Performance assessment

Conducting performance evaluation of environmental models has attracted increased attention in recent years, as multiple models targeting one specific environmental problem have become more available. The answer to the question of which one of available models would better address a desired goal of modeling is not trivial and approaches to conduct performance tests may vary with modeling objectives (Jakeman et al., 2006; Bennett et al., 2013). Difficulty of multi-model testing increases with complexity of the models involved and it is usually a time-consuming task requiring knowledge of each model and input data preparation. While numerous guidelines to measuring model performance have been proposed in the literature, in this work an attempt was made to follow those proposed by Bennett et al. (2013). These include defining modeling goals, selecting performance criteria and developing methods for identifying systematic errors.

2.1.1. Modeling objectives

During the past decade, Lake Erie (USA–Canada), the smallest, shallowest, and most biologically productive lake of the North American Great Lakes, has experienced degraded water quality, including hypoxia (Hawley et al., 2006; Scavia et al., 2014) and harmful algal blooms (Stumpf et al., 2012; Michalak et al., 2013). These impairments have in large part been attributed to increased inputs of phosphorus-rich water from catchment basins (Burns et al., 2005; Rucinski et al., 2010; Scavia et al., 2014), including the Maumee River watershed, which is the largest watershed in the Lake Erie and Great Lakes basins. This watershed is dominated by agriculture (>70%: Lake, 1978; NRCS, 2005) and contributes roughly 48% of the phosphorus that enters western Lake Erie annually (Ohio EPA, 2010). Because Lake Erie provides numerous economically important ecosystem services to the region (e.g., fishing opportunities, drinking water supply, beach access) that depend on water quality, state, provincial, and Federal agencies have a strong interest in understanding how land-use practices and climate operate independently and interactively to influence inputs from the Maumee River watershed into downstream Lake Erie (Ohio EPA, 2010).

While tillage and fertilizer application practices have been implicated in the recent "re-eutrophication" of Lake Erie (Ohio EPA, 2010; Scavia et al., 2014), increased precipitation-driven river discharge also has been shown to play a dominant role. In fact, high river discharge from the Maumee River was found to be the primary driver of record-breaking inputs of phosphorus and sediment into western Lake Erie during 2007 (Richards et al., 2010) and the occurrence of the largest recorded harmful algal bloom in Lake Erie during 2011 (Michalak et al., 2013). Restoration of Lake Erie and its ecosystem services, as well as selection of watershed management strategies, require the use of watershed models, which necessitates testing of multiple hydrology/water quality models to select one that is suitable for quantifying flow regimes critical to the nutrient flux entering Lake Erie.

While numerous watershed models exist, including Annualized Agricultural Nonpoint Source (AnnAGNPS; Bingner et al., 2011), Areal Nonpoint Source Watershed Environment Response Simulation (ANSWERS-2000; Bouraoui and Dillaha, 1996), Hydrological Simulation Program – Fortran (HSPF; Bicknell et al., 2001), Soil and Water Assessment Tool (SWAT; Arnold et al., 1998), Distributed Large Basin Runoff Model (DLBRM; Croley and He, 2005), and MIKE SHE (Refsgaard and Storm, 1995), the ability of each to accurately model the flow regime of a river is likely to differ. Differences among models can arise for many reasons, including their differential use of algorithms to simulate overland flow and flow routing, how watersheds are disaggregated into spatial units, the time step used to calculate flow components, and dissimilarities in their ability to consider multiple watershed attributes (Borah and Bera, 2003; DeFries and Eshleman, 2004; Smith et al., 2004). Hydrology models also can vary greatly in their input data, availability of pre-processor and post-processor interfaces for data preparation and analysis, their capability to simulate changes in climate, land use and land cover, and their flexibility to allow specification of existing crop management practices (e.g., fertilizer application, tillage method, tile drainage, crop rotation). Additionally, models can vary in terms of availability of support, documentation and source code, and their ease of modification for further development.

Despite the widespread and growing use of watershed models to simulate water discharge and nutrient exports to downstream water bodies (Borah and Bera, 2003; Smith et al., 2004), a comprehensive approach to evaluate the performance of models for large watersheds is conspicuously lacking. Indeed, previous performance assessments of different models within a single watershed (e.g., Saleh and Du, 2004; Im et al., 2007; Nasr et al., 2007), as well as multiple models across multiple watersheds (e.g., Distributed Model Intercomparison Project (DMIP-1,2): Reed et al., 2004; Smith et al., 2004, 2012; Climate Impact Assessment Study: U.S. EPA, 2013), support this contention. Previous studies that have assessed the performance of models often only compared magnitudes of daily and (or) monthly predicted flow to observed data in terms of one or two goodness-of-fit statistics (e.g., Smith et al., 2004; Saleh and Du, 2004; Im et al., 2007; Nasr et al., 2007; U.S. EPA, 2013). Further, while the DMIP-1,2 project was comprehensive (comparing performance of 20 models across multiple watersheds) and showed that distributed models can supplement lumped models for operational flow forecasting, this effort focused on small (<2500 km^2) watersheds of less complexity in terms of land use and land cover and did not assess the capacity of models to describe flow-regime components important to driving conditions in downstream (receiving) ecosystems (Reed et al., 2004; Smith et al., 2004, 2012). For this reason, a major knowledge gap exists with respect to how well commonly used watershed models such as SWAT, HSPF, and DLBRM can simulate river flow regimes for any size of watershed. Thus, while classifying river flow into ecologically meaningful categories based on flow metrics (e.g., timing, frequency, duration, flashiness, and magnitude of river discharge) has become commonplace when assessing riverine inputs and their impact on downstream ecosystems services (Poff et al., 1997; Richter et al., 1996), how accurately numerical watershed models predict these flow metrics remains largely unknown, especially in large watersheds.

2.1.2. Selection of evaluation criteria

Models are assessed in terms of their ability to reproduce measurable behavior of the simulated variables; however, the criteria for performance assessment can be subjective or objective. Subjective criteria often involve visual inspections to determine whether temporal and spatial behaviors of the variable being modeled are reproduced, whereas objective criteria are quantitative scores computed from known statistical error estimates between simulated and observed behavior (Krause et al., 2005; Pushpalatha et al., 2012; Bennett et al., 2013). In hydrologic modeling the use of both methods is advocated (Boyle et al., 2000; Bennett et al., 2013) for determination of qualitative and quantitative assessment, both of which were applied in our work.

2.1.2.1. Statistical goodness-of-fit metrics. While numerous goodness-of-fit (GOF) criteria are available for model assessment, no one by itself is capable of fully characterizing performance of a model (Krause et al., 2005). Instead, each criterion has its own strength to depict certain aspects of a model that would not be obvious otherwise. For this reason, to assess agreement between observed and simulated daily flow, as well as monthly flow, we used a combination of GOF metrics (using R software's "gof" package, Zambrano-Bigiarini, 2012) that are commonly used for assessing model performance in hydrologic modeling. Metrics calculated included the Nash-Sutcliffe efficiency (NSE; Nash and Sutcliffe, 1970), coefficient of determination (R^2), percent bias (PBIAS; Gupta et al., 1999), relative index of agreement (rd; Willmott, 1981), mean absolute error (MAE), volumetric efficiency (VE) (Criss and Winston, 2008), and root mean square error (RMSE). These metrics were selected based on their merits to reflect different error estimates between simulated and observed data. For example, R^2 depicts the relationship between observed and simulated flow in terms of percent variance explained, whereas MAE explains the size of absolute error. PBIAS indicates if the model is under- or over-predicting observed behavior, whereas VE depicts a volumetric fraction explained by the model.

The NSE, which is the most widely used measure for assessing performance of hydrologic models, was calculated by:

$$\mathrm{NSE} = 1 - \left[\frac{\sum_{i=1}^{n}(O_i - S_i)^2}{\sum_{i=1}^{n}\left(O_i - \overline{O}\right)^2}\right]$$

where O_i is observed and S_i is simulated flow at the ith time step. $\overline{O}$ is the average value of observed flow during the calibration or validation period. The NSE ranges between $-\infty$ to 1, with a value of 1 indicating a perfect fit.

As with NSE, the R^2, which represents the proportion of variance in the observed data explained by the model, has been widely used to evaluate the predictive capability of hydrology models (Moriasi et al., 2007; Gassman et al., 2007; Donigian and Imhoff, 2009). It varies between 0 and 1, with 1 indicating that the variance in observed data is fully explained by the model.

PBIAS measures the average deviation of simulated data from observed data relative to observed data and is calculated as:

$$\mathrm{PBIAS} = \left[\frac{\sum_{i=1}^{n}(S_i - O_i)*100}{\sum_{i=1}^{n}(O_i)}\right]$$

PBIAS ranges from $-\infty$ to ∞, with 0 indicating no bias.