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

ARTICLE INFO

Article history:
Received 10 October 2013
Received in revised form
2 July 2014
Accepted 3 July 2014
Available online 5 August 2014

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

ABSTRACT

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.

The rd ranges between $-\infty$ and 1, with a value of 1 indicating perfect agreement between observed and simulated data. It is defined as:

$$\mathrm{rd} = 1 - \frac{\sum_{i=1}^{n}\left(\frac{O_i - S_i}{O_i}\right)^2}{\sum_{i=1}^{n}\left(\frac{|S_i - \overline{O}| + |O_i - \overline{O}|}{\overline{O}}\right)^2}$$

MAE calculates the absolute deviation between S_i and O_i and is defined as:

$$\mathrm{MAE} = n^{-1}\sum_{i=1}^{n}|S_i - O_i|$$

where n is the number of modeled time intervals (e.g., years, seasons, days) in the calibration or validation dataset. MAE ranges from 0 to ∞, with a value of 0 signifying a perfect fit.

Volumetric efficiency (VE) is a modified form of MAE in which the absolute deviation is normalized by total sum of observed flow and is defined as:

$$\mathrm{VE} = 1 - \frac{\sum_{i=1}^{n}|S_i - O_i|}{\sum_{i=1}^{n}O_i}$$

VE ranges from 0 to 1 (perfect agreement) and represents matching volumetric fraction of discharge.

The root mean square error (RMSE), which shows optimal fit when value equals zero is defined as:

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n}(O_i - S_i)^2}$$

2.1.2.2. Visual methods. The use of graphs and plots of simulated and observed data provides a visual opportunity to determine if systematic errors are present or if temporal behaviors of observed data were captured by the model. In this regard, we used scatter plots, monthly and daily hydrographs, as well as flow duration curves during model calibration and validation steps. We followed the hybrid approach proposed by Boyle et al. (2010), in which auto-calibration schemes were complemented with graphical methods to enhance calibration of the models. The result of this is provided in Appendix A and Tables A.1–A.3.

2.2. Study area

The Maumee River Basin, most of which was largely comprised of the historic Great Black Swamp, is the largest watershed in the Great Lakes Basin with a drainage area of 16,480 km^2 (Lake, 1978). The basin is relatively flat with ~50% of it having a slope less than 0.5% and greater than 80% of it having a slope of less than 2%. Over 90% of this watershed's cropland is drained through surface ditches and subsurface drains because poorly drained soils are the norm (Lake, 1978; NRCS, 2005). Agriculture dominates land use (73%), with forest cover (8%) and urban land (including road development; 10%) being minimal. Major crops grown within the basin include corn, soybean, and wheat, with a small fraction of cropland devoted to oats and hay (Richards et al., 2002; NRCS, 2005).

In terms of hydrology, the Maumee River Basin consists of seven 8-digit hydrologic unit code watersheds drained by a number of major rivers, including the St. Joseph, St. Marys, Blanchard, Auglaize, and Tiffin, which are tributaries of the Maumee River (Fig. 1). Formed at the confluence between the St. Joseph River and St. Marys River near Fort Wayne, IN, the Maumee River meanders northeast through mainly agricultural fields for about 240 river km at an average slope of 0.25 m/km before it discharges into Maumee Bay, located near Toledo, OH at the western end of Lake Erie (Fig. 1).

2.3. Synopsis of the watershed models

2.3.1. DLBRM

The DLBRM was created in 2004 from the Large Basin Runoff Model, which is a tank model that was developed by the National Oceanic and Atmospheric Administration (NOAA)'s Great Lakes Environmental Research Laboratory (GLERL) to estimate watershed runoff into the Laurentian Great Lakes (Croley, 2002; Croley and He, 2005). The DLBRM subdivides the watershed into a grid of square cells (typically 1 km^2 in size) and runs continuously at a daily time step. Each grid cell has a three-layer hypothetical reservoir representing the upper soil zone, lower soil zone, and groundwater zone flow regimes (Croley and He, 2005). The DLBRM has been applied to 34 watersheds in the Great Lakes Basin (DeMarchi et al., 2009, 2011), but has rarely been applied to watersheds outside this region (He et al., 2013). Hydrologic process components in DLBRM include, snow melt, evaporation, overland flow, infiltration, groundwater, and lateral flow. The DLBRM has only 14–16 parameters that need to be optimized globally during calibration, which is less than HSPF and SWAT.

2.3.2. HSPF

The HSPF model is the direct outgrowth of the first hydrology model (the Stanford Model), which was developed during the 1960s. HSPF has been developing since this time through sustained financial support from the U.S. Environmental Protection Agency to simulate hydrology and point-source and nonpoint-source pollution in a partially lumped manner (Bicknell et al., 2001). In the HSPF model, the catchment is divided into sub-watersheds (reaches) based on drainage areas of the tributaries that are partitioned into pervious and impervious land segments on the basis of land use and meteorological conditions. The primary modules of the HSPF model include PERLND, IMPLND, and RCHRES, which are used to simulate hydrologic and transport process from pervious surfaces, impervious surfaces, and streams/reservoirs, respectively. This model has the flexibility to have 1 min to 24 h time steps, with the temporal resolution depending on availability of weather data as input. For simulation of hydrology and water quality, the HSPF model is

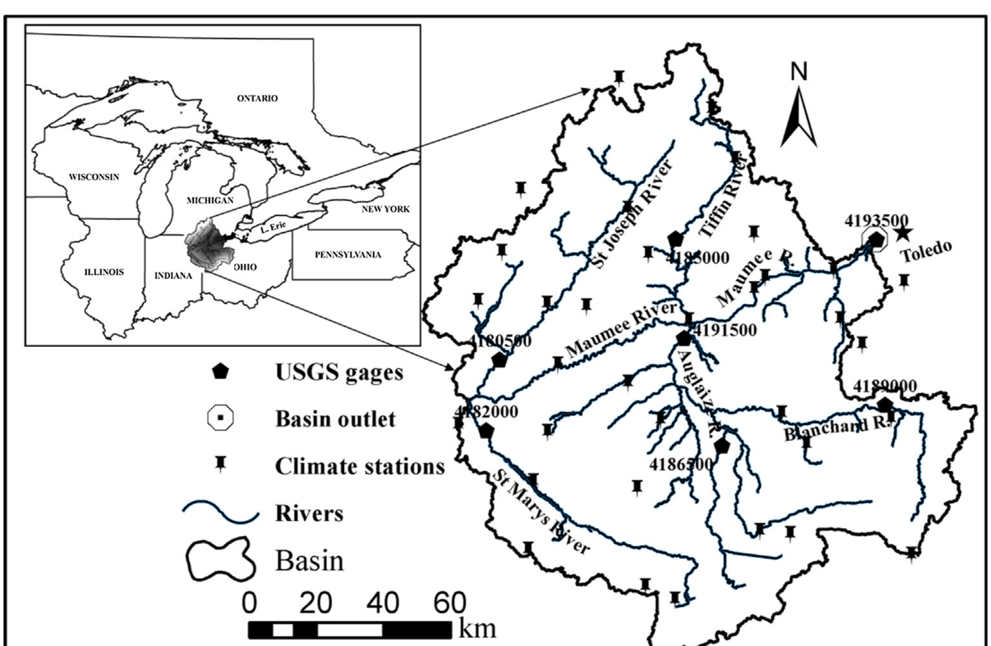

Fig. 1. The Maumee River Basin, which spans 26 counties in MI, IN, and OH (USA) and drains into western Lake Erie near Toledo, OH.

meteorologically data intensive. Unlike the DLBRM, the HSPF model has been applied widely to watersheds in the USA and in many other parts of the world (Donigian, 2000). It requires calibration of a large number of parameters to allow correct performance of its numerous process components, including surface runoff, lateral flow, evaporation, groundwater, and percolation. In addition to parameters that need to be adjusted via calibration, the HSPF model has a number of parameters that have to be estimated from geographic datasets (Saleh and Du, 2004).

2.3.3. SWAT

Similar to the HSPF model, SWAT is a partially lumped watershed model. It was developed by the U.S. Department of Agriculture's (USDA) Agricultural Research Service (ARS) for simulating flow, sediment, nutrients, and pesticide runoff from agricultural watersheds (Arnold et al., 1998). Having evolved out of various hydrology, crop growth, and chemical transport models that were developed by ARS during the past 30 years (White et al., 2010), SWAT has primarily been used to assess the benefits of conservation practices aimed at reducing sediment, nutrient, and chemical runoff from agricultural fields (Arnold et al., 2011). For this reason, SWAT offers a wide range of alternatives for simulating conventional and conservation farming practices. In a manner similar to that of HSPF, SWAT's sub-watersheds are divided into hydrological response units (HRUs) that have unique combinations of slope, land use, and soil type within the sub-basin, and form the basic land segment for computing flow and transport. The HRUs in SWAT correspond to pervious/impervious segments in the HSPF model and grid cells in the DLBRM. While SWAT commonly runs at a daily (24 h) time step, it can run at sub-daily time steps, if sub-daily weather data are available. Different from the HSPF model and the DLBRM, SWAT also is capable of producing weather data from climatic averages at times when records are unavailable. It also has fewer parameters that need to be manually estimated compared to the HSPF model. Process components for estimating flow in the SWAT include overland flow, lateral flow, shallow ground water return flow, percolation, evaporation, and channel transmission.

2.3.4. Comparison of model attributes

In general, distributed watershed-hydrology models are based on digital elevation models and require weather, soil, and land-use data as input to derive physical parameters needed for water-balance calculations, transport, and flow routing. Differences and similarities in terms of process feature, input/output, and software structure are summarized for the three models in Table 1.

2.4. Input data preparation

A number of necessary input datasets (Table 2) were collected and prepared as required by each model. Climate and stream gage data were split over three periods for model initialization (1992–1994), calibration (1995–2002), and validation (2003–2009) purposes. When required, the input data were further processed in ArcGIS software so as to properly set up the models. Identical weather stations (Fig. 1) and weather data were used to drive the models, except that weather data for the HSPF model were at an hourly time-step in contrast to daily data used for the DLBRM and SWAT. Using different temporal resolutions allowed the models to run at time steps at which they are commonly used.

2.5. Watershed discretization

The GIS data were re-projected to the Albers-Equal-Area conic NAD83 projection before watershed discretization. For SWAT, the ArcSWAT interface (version 2009) was used to divide the watershed into 71 sub-watersheds. By implementing percentage threshold criteria of 5, 10, and 10 for land use, soil, and slope, respectively, we created 956 HRUs without substantially distorting the original land-use proportions. The BASINS interface (version 4.0) was used to setup the HSPF model. To avoid variation in watershed and sub-watershed delineation becoming a confounding factor for potential differences between SWAT and HSPF, we imported stream reach and sub-watershed layers from SWAT into BASINS to create the HSPF model. Doing so resulted in 71 stream reaches and 828 land segments in the HSPF model. For the DLBRM, the watershed was discretized into 16,395 1 km^2 grid cells using the AvDLBRM interface (version 2.0). For each model, the final watershed outlet was selected to be Waterville, OH. We chose this location because it is the closest USGS gaging station (#04193500) to where water from the Maumee River flows into Lake Erie (~20 km upstream; Fig. 1).

2.6. Crop management

To account for agricultural activities in the watershed, common farming practices were implemented in the SWAT and HSPF model for land-use classes that were

Table 1
Selected attributes of the Distributed Large Basin Runoff Model (DLBRM), the Hydrological Simulation Program–Fortran (HSPF) model, and the Soil and Water Assessment Tool (SWAT).

Feature	DLBRM (Croley, 2002)	HSPF (Bicknell et al., 2001)	SWAT (Arnold et al., 2011)
Weather data	Daily precipitation, temperature (minimum, maximum), monthly solar radiation	Hourly precipitation, air temperature, solar radiation, cloud cover, wind, dew point, potential evapotranspiration	Daily precipitation, air temperature (minimum, maximum), solar radiation, relative humidity, wind
Weather generator	NA	NA	WXGEN
Preprocessor	AvDLBRM (not public)	BASINS	ArcSWAT
Graphic user interface	NA	WinHSPF	SWAT Editor
Calibration module	DLBRM auto-calibration	PEST auto-calibration	SWAT-CUP auto-calibration
Flow calibration parameters	15	20–25 parameters typically lead to optimal results (but the number of parameters can increase depending on the number of processes activated and ability to estimate parameters from watershed data)	20–30 parameters typically lead to optimal results (but the number of parameters can be numerous depending on whether parameters were varied within HRUs)
Postprocessor	Output visualizer (not public)	GenScn	VizSWAT (commercial)
Overland flow and infiltration	Empirical equation based on wetted contributing area proportion	Infiltration is calculated using Philip's equation, whereas overland flow is computed with Chezy-Maning equation using hourly precipitation	SCS curve number method with daily precipitation or Green–Ampt Mein–Larson infiltration equation with sub-daily time step precipitation
Water routing	Linear reservoir routing method	Storage routing or kinematic wave method	Variable storage or the Muskingum routing method
Channel geometry	Not applicable	User-defined	Trapezoidal
Shallow aquifer	Yes	Yes	Yes
Deep aquifer	Yes	Yes	Yes
Tile drainage	No	No	Yes
Snow melt	Degree-day method	Energy balance or degree-day method	Degree-day method
Smallest spatial feature	User-defined square cells	Land use and climate stations-based lumped land segments (impervious, pervious)	Land use, soil- and slope-based lumped HRUs
Simulation time step	Daily	Sub-daily to daily	Sub-daily or daily
Predictable water quality variables	None	Sediment, inorganic suspended sediment, pathogens, biological oxygen demand, dissolved oxygen, pH, alkalinity, pesticides, inorganic nitrogen, nitrite, ammonia, nitrate, orthophosphate, inorganic phosphorus, conservative tracers, carbon dioxide, inorganic carbon, zooplankton, phytoplankton, benthic algae, organic carbon, and fecal coliform	Sediment, organic nitrogen, nitrate, organic phosphorus, mineral phosphorous, ammonium, nitrite, algae as chlorophyll *a*, conservative metals, fecal coliform, biological oxygen demand, dissolved oxygen, and pesticides

Table 2
Description of input data used for developing the Maumee River watershed-hydrology models (i.e., the SWAT, HSPF model, and DLBRM).

Data	Source	Resolution
Topography	USGS Digital Elevation Model data[a]	30 × 30 m
Land use	USGS 2001 National Land Cover data[a]	30 × 30 m
Hydrography	USGS National Hydrography data[a]	30 × 30 m
Soil	USDA STATSGO Soil data[b]	225 × 200 m
Weather	NOAA National Climatic Data Center[c]	
Stream gage	USGS (Station # 04193500; see Fig. 1)[d]	

[a] http://viewer.nationalmap.gov/viewer/.
[b] http://websoilsurvey.sc.egov.usda.gov/App/HomePage.htm.
[c] http://www.ncdc.noaa.gov/.
[d] http://waterdata.usgs.gov/nwis.

designated as "cultivated crops" in the USGS land-use map. The annual cycle included conservation tillage implemented on 20 April, crops planted along with fertilizer application during the first week of May, and harvest of crops during mid-October based on published data for the watershed (USDA, 2010). To determine fertilizer application rates, we first calculated the average areal proportion of major crops (corn, soybean, and winter wheat) from 2007 to 2011 crop-layer data (Han et al., 2012). We then applied fertilizer at an area-weighted rate determined from agronomic rate data recommended for corn, soybean, and winter wheat. Tile drainage was simulated in SWAT but not in the HSPF and DLBRM. Unlike SWAT and HSPF, the DLBRM does not support specification of crop management inputs such as crop rotation, tillage operation, planting, and harvest dates, although a new version that incorporates these details is under development (C. DeMarchi, personal communication).

2.7. Calibration and validation

The models were calibrated to optimize their performance in simulating daily river flows recorded at the basin outlet (i.e., Waterville station; Fig. 1). Following the method of Morris (1991), where perturbation of one parameter is used to measure changes in model output, we altered only the most sensitive model parameters for the SWAT and HSPF model so as to minimize time and effort required for their calibration. Because so few parameters required calibration in the DLBRM (Table 1), all parameters were calibrated. Parameter optimization was carried out following commonly used approaches recommended for each model, as is briefly described below. Calibration parameter ranges were defined based on recommended ranges in software manuals and publications (Bicknell et al., 2001; Arnold et al., 2011), as well as the peer-reviewed literature (Arabi et al., 2007). In choosing parameter values, we accounted for physiographic conditions, agricultural operations, and soil hydrology classification groups within the Maumee River Basin. The HSPF model was manually calibrated by changing parameters within recommended ranges until an optimum fit between observed and predicted daily flow was achieved following the method outlined in Duda et al. (2012) using NSE (Nash and Sutcliffe, 1970) as a fit criterion. SWAT was calibrated using SWAT-CUP's SUFI-2 optimization software (Abbaspour et al., 2007) with NSE as an objective function. The DLBRM was parameterized using its own global optimization auto-calibration algorithm (minimizing RMSE), from which grid-cell level parameters were calculated on the basis of local features such as land use, slope, and soil characteristics. For SWAT, the base-flow parameter was estimated using automated base-flow separation software (Arnold and Allen, 1999). For HSPF and DLBRM, base flow was adjusted as part of the calibration process.

After calibration, the models, with their new calibration parameters and a new set of meteorological data, were assessed for their ability to predict flow outside of the calibration period both at the Maumee River Basin outlet (Waterville station) and six other interior USGS gaging stations to which the models were not calibrated (see Fig. 1). These interior USGS gaging stations are located at outflow of major tributaries, including the St. Joseph and St. Marys (near Fort Wayne, IN), the Auglaize (near Defiance, OH and Fort Jennings, OH), the Blanchard (near Findlay, OH), and the Tiffin (at Stryker, OH). Calibration parameters for each model are listed in Appendix A and Tables A.1–A.3.

2.8. Identification and classification of flow regimes

We explored the ability of the models to simulate flows at daily and monthly time scales during 1995–2009, as different applications might require reliable data at different temporal scales. We also assessed the ability of models to simulate several aspects of the flow regime that might be of ecological importance during this same time period. These included the 3 d maximum and 30 d minimum median flows, the median fall and rise rates, zero flow, and the frequency of high- and low-flow pulses. To do so, we used the Nature Conservancy's (2009) Indicators of Hydrologic Alteration (IHA) software (version 7.1). All flows that exceeded 75% of daily flows for the water year were classified as flood pulses, whereas all flows below 25% of daily flows were classified as low-flow pulses. In addition, flow–duration curves were plotted to compare prediction trends of the models.

We also conducted analyses of single extreme events and the capacity of the models to predict them. While none of the three models was exclusively designed for simulating a single extreme flow event of short duration, their capacity to predict and characterize watershed responses to intense storms has attracted interest because such events often contribute most of the annual pollutant loads leaving the watershed (David et al., 1997; Borah et al., 2007). Such ability also is of interest because the frequency of extreme rain events is projected to increase in the Great Lakes region (Angel and Kunkel, 2010; U.S. EPA, 2013). We searched the daily weather data along with measured flow data for extreme events that occurred during 1995–2009 and identified four to investigate: Event #1 — heavy precipitation from multiple winter storms in 1999; Event #2 — exceedingly wet conditions characterized by heavy rain spanning multiple days during early spring of 2009; Event #3 — a summer storm following a prolonged dry spell in 2007; and Event #4 — prolonged dry conditions during autumn of 1999. Events #1 and #2 targeted wet weather conditions that produced a daily average flow rate that exceeded 2000 m^3/s; such events ($n = 4$ total in our datasets) usually occurred during January through March, when precipitation is accompanied by snow melt or sequential heavy storms. Events #3 and #4 focused on dry periods that resulted in low daily average flows that were less than 20 m^3/s; such low average daily flows were observed commonly during summer and autumn.

2.8.1. Parametric statistical analysis

Inferential parametric statistical tests were conducted to further compare data. Prior to testing, data were checked to verify assumptions of independence, homogenous variances, and normality. Stationarity and autocorrelation were respectively tested using Augmented Dickey–Fuller and Ljung–Box tests, whereas normality and homogeneity of variance were tested with Shapiro–Wilks and Levene's tests, respectively. Afterwards, we tested for differences in mean of 3 d maximum and 30 d minimum median flow, frequency of flood and low-flow pulses, and flow rise and fall rates between model-predicted and observed data using one-way ANOVA. The alpha level for all tests was set to 0.05.

2.8.2. Software technical preferences

Following the methodology proposed by Cunderlik (2003), we assessed the models in terms of their ease of use regarding model setup (e.g., watershed delineation) and calibration, as well as their ability to specify agricultural management practices and potentially assess the impact of climate and land-use change. We also commented on the ability to access and modify the source code of models, which might be important when trying to integrate a model with other models simulating biophysical processes (e.g., coastal or lake ecosystem hydrodynamic and lower food web models). Model preference criteria (Table 5) were identified and assigned a score between 0 and 5 (5 being most preferred; *sensu* Cunderlik, 2003) based on experience gained while running the models.

3. Results

3.1. Daily and monthly flow

3.1.1. Graphical methods

Overall, each model performed well at simulating the magnitude of observed daily flows (Fig. 2), as well as replicating the timing of daily (Fig. 3a) and monthly (Fig. 3b) peak flows. The slope of the least-squares regression line between simulated and observed flow data (Fig. 2) was close to that of the slope of line of perfect agreement (i.e., a 1:1 line) for each model, with R^2 values ranging between 0.78 and 0.85. The slight departure of the slopes of the regression lines from unity (not statistically significant), however, indicates that the DLBRM and the SWAT somewhat under-predicted higher flows (slope >1), whereas the HSPF model slightly over-predicted them (slope <1). A comparison of daily flow hydrographs (Fig. 3a) also showed that each model could reliably estimate precipitation-driven peaks in flow, as the peaks in precipitation and flow were closely matched in time. A side-by-side comparison of daily (Fig. 3a) and monthly (Fig. 3b) hydrographs also showed that each model better predicted peak flows at monthly time scales compared to daily time scales. However, we found no obvious difference in each model's ability to predict the magnitude and timing of the daily versus monthly peaks.

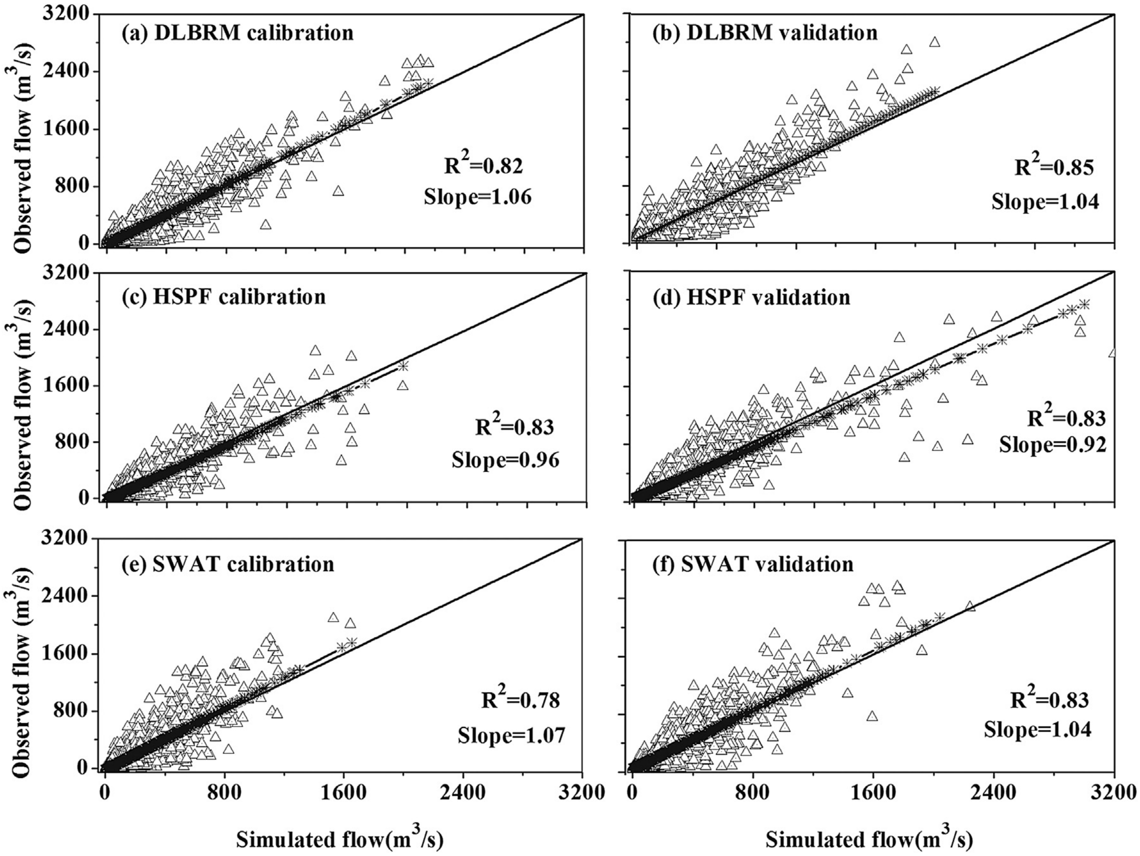

Fig. 2. Scatterplots of simulated versus observed daily Maumee River flows (Δ) for the DLBRM (a,b), HSPF model (c,d), and the SWAT (e,f) for calibration years (1995–2002; left panels) and validation years (2003–2009, right panels) at the watershed outlet (Waterville, Ohio; USGS station 04193500). The solid black line is a 1:1 line. The line of asterisks (*) represents the least-squares regression line, with its slope and R^2 shown in each panel.

3.1.2. Goodness-of-fit (GOF) metrics

Based on previously established model evaluation criteria, each of the models would be considered "very good" or "excellent" (Moriasi et al., 2007; Donigian and Imhoff, 2009) at simulating both daily and monthly flows at the Maumee River watershed outlet. For monthly data, for example, these criteria include a NSE of ≥ 0.5 accompanied by a PBIAS of ± 25 (Moriasi et al., 2007) or a $R^2 \geq 0.70$ accompanied by a PBIAS of ± 15–20 (Donigian and Imhoff, 2009). All of our models met (and in fact greatly surpassed) these minimum criteria not only for monthly flows for both the calibration and validation periods, but daily ones as well (Table 3). For example, all NSE values ranged between 0.76 and 0.91, all PBIAS values ranged between −0.5% and 6.4%, and all R^2 values ranged between 0.78 and 0.91. Additionally, no single model had the best values for our entire suite of GOF statistics. For example, the VE and rd values for the HSPF model were consistently higher (i.e., better) than for the DLBRM or SWAT for both daily and monthly flow data during the calibration and validation periods. Likewise, the MAE values were consistently lower (i.e., better) for the HSPF model than for the other models during both periods at both time steps (Table 3). However, PBIAS values for the HSPF model were intermediate to the DLBRM and SWAT.

As a further model validation step, we assessed the ability of each model to simulate observed monthly flows during the

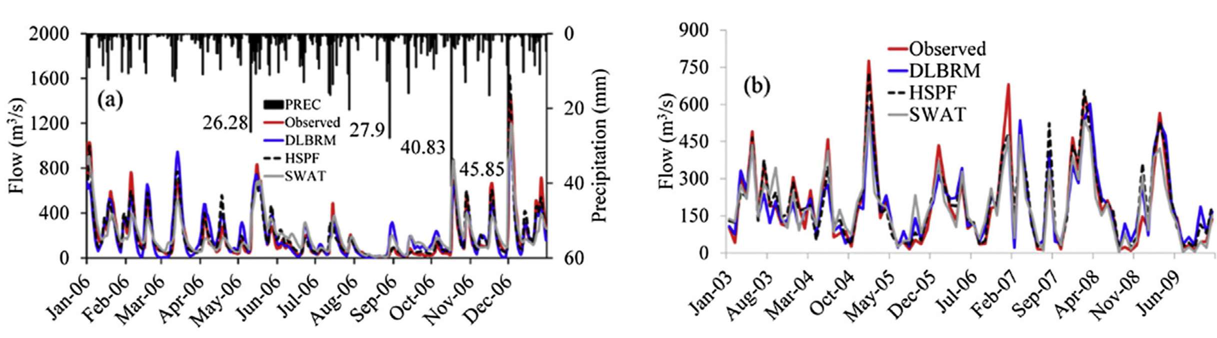

Fig. 3. Representative hydrographs of daily flow for 2006 (a) and monthly flow during the validation years (2003–2009) (b) for the Maumee River as predicted by the DLBRM, HSPF model, and SWAT at the watershed outlet (Waterville, Ohio; USGS station 04193500). The observed flow hydrograph is shown by the red line in both panels, whereas inverted bar graphs in (a) are average daily watershed precipitation values corresponding to the hydrographs. Numbers in (a) are values of major precipitation peaks in mm. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Goodness-of-fit metrics describing the ability of the DLBRM, HSPF model, and SWAT to predict observed daily and monthly Maumee River flows at the watershed outlet (Waterville, Ohio; USGS station 04193500) during calibration (1995–2002) and validation (2003–2009) years.

Model	Metric	Calibration		Validation	
		Daily	Monthly	Daily	Monthly
DLBRM	NSE	0.82	0.86	0.85	0.87
	R^2	0.82	0.87	0.85	0.88
	PBIAS (%)	6.40	6.40	0.50	0.60
	rd	0.23	0.10	0.67	0.65
	MAE (m^3/s)	65.87	39.74	72.63	45.39
	VE	0.57	0.74	0.63	0.77
	RMSE	102.04	52.36	117.13	61.86
HSPF	NSE	0.83	0.91	0.82	0.91
	R^2	0.83	0.91	0.83	0.91
	PBIAS (%)	4.50	4.40	5.00	5.20
	rd	0.81	0.81	0.74	0.85
	MAE (m^3/s)	53.50	31.19	64.47	36.33
	VE	0.65	0.80	0.67	0.81
	RMSE	97.44	42.19	126.7	52.51
SWAT	NSE	0.76	0.85	0.82	0.86
	R^2	0.78	0.85	0.83	0.88
	PBIAS (%)	1.40	1.30	−0.50	−0.40
	rd	0.62	0.68	0.56	0.75
	MAE (m^3/s)	63.37	39.00	74.13	48.37
	VE	0.59	0.75	0.62	0.75
	RMSE	115.81	54.01	129	65.25

validation period (2003–2009) by computing GOF metrics (NSE and PBIAS) at six interior USGS gages located within the Maumee River watershed (Table 4; see Fig. 1). Importantly, each model was calibrated to predict flows at the watershed outlet (Waterville, Ohio; USGS station 04193500), not these interior stations. With exception of simulated flows from SWAT at the Tiffin River gage, which is located in the far upstream location of the Maumee River watershed, all models performed in the "very good" to "excellent" range (Moriasi et al., 2007), with NSE values exceeding 0.5 and PBIAS values within ±25% (Table 4).

3.2. Flood and low-flow pulse frequency and magnitude

The models varied in their ability to replicate the frequency and magnitude of flood and low-flow pulses, with no model proving to be superior to the others at predicting both high and low flow events. While the predicted average frequency of flood pulses did not differ from observed data (12 per year) for any model, simulated average low-flow pulse frequencies did differ for all models (all $p < 0.05$). Average low-flow frequencies (9 per year) were under-predicted by the HSPF model (7 per year) and SWAT (7 per year), but over-predicted by the DLBRM (12 per year). The models also varied in their ability to predict the magnitude of low flows. For example, the DLBRM predicted up to 31 d of zero-flow events per year when none was observed (the Maumee River is perennial). By contrast, the HSPF model and SWAT never predicted zero-flow days. The models also varied in their ability to accurately predict median values of 3 d maximum and 30 d minimum flows during each year (Fig. 4). While the 3 d maximum flow was accurately predicted by the HSPF model during each year, it was under-predicted by the DLBRM ($p < 0.06$) and especially by the SWAT ($p < 0.006$). More specifically, the DLBRM and SWAT under-predicted the median value of the 3 d maximum flows by up to 43% (in 2004) and 51% (in 2002), respectively (Fig. 4b). By contrast, the 30 d minimum flow in each year was closely predicted by the DLBRM and SWAT but over-predicted by the HSPF model at a statistically significant level (both $p \leq 0.02$). In fact, the HSPF model over-predicted the median 30 d minimum flow by as much as 165% (2002–2003; Fig. 4a).

Inspection of flow–duration curves (Fig. 5) further highlights that all models were better at predicting high flows than low flows. For example, curves from simulated data closely matched the observed data for exceedance probabilities of <70% for both the calibration (Fig. 5a) and validation (Fig. 5b) years. However, for flows with exceedance values >90% (i.e., low flows), obvious differences in predictive capabilities were evident. In particular, during low- and base-flow conditions, the DLBRM tended to under-predict observed flows, whereas the HSPF model tended to over-estimate flows, with the SWAT being intermediate (Fig. 5). Further, the bias exhibited by each model was <7% and <15% when predicting discharge that was exceeded 10% of time (Q_{10}) during calibration and validation years, respectively (i.e., little bias existed when predicting high flows). By contrast, flows that were exceeded 90% of the time (i.e., low flows Q_{90}) were either severely under-predicted (e.g., bias of 50% by the DLBRM) or over-predicted (e.g., bias of 200% by the HSPF model) during both the calibration and validation periods (Fig. 5).

Analysis of the rise and fall rates (i.e., the rate at which simulated flows changed from day to day) during 1995–2009 also revealed differences among models (Fig. 6). For example, the DLBRM consistently predicted significantly greater rise and fall rates than was observed ($p < 0.001$), which might lead one to conclude that the Maumee River is "flashier" (e.g., quickly floods, then recedes) than it really is. Specifically, with exception of 2006, the DLBRM consistently predicted rise rates that exceeded observed ones, with the difference being >400% at times (Fig. 6a). Likewise, the median fall rate simulated by the DLBRM (Fig. 6b) was positively biased, although by a lesser degree (maximum positive bias was <200%). While simulated data from the SWAT and HSPF model also showed positive bias in terms of rise and fall rates, the median rise rate ($p > 0.09$) and median fall rate ($p > 0.1$) for these models did not significantly differ from the observed data.

3.3. Extreme events

3.3.1. Winter storm accompanied by snow melt (Event #1)

The first extreme wet weather event occurred during 2–3 January 1999. During this 2 d period, a storm produced about

Table 4
Statistics describing the ability of the DLBRM, HSPF model, and SWAT to predict observed monthly tributary flows at six USGS gage stations located within the Maumee River watershed during 2003–2009. Each model was calibrated to predict flows at the watershed outlet (Waterville, Ohio; USGS station 04193500), not these interior stations. Reported are the NSE and PBIAS (in parentheses).

USGS Gage	Tributary (Gage location)	Drainage Area (km^2)	Model		
			DLBRM	HSPF	SWAT
4191500	Auglaize (Defiance, OH)	6004	0.77 (−12.9)	0.87 (−3.4)	0.70 (0.6)
4186500	Auglaize (Fort Jennings, OH)	860	0.74 (−7.8)	0.86 (−10.4)	0.62 (−5.3)
4189000	Blanchard (Findlay, OH)	896	0.70 (−16.5)	0.65 (−23.0)	0.53 (4.2)
4182000	St. Marys (Fort Wayne, OH)	1974	0.79 (14.2)	0.90 (14.6)	0.69 (−14.4)
4185000	Tiffin (Stryker, OH)	1062	0.79 (−0.9)	0.77 (1.7)	0.28 (30.3)
4180500	St. Joseph (Fort Wayne, OH)	2745	0.84 (−3.4)	0.77 (14.4)	0.75 (2.1)

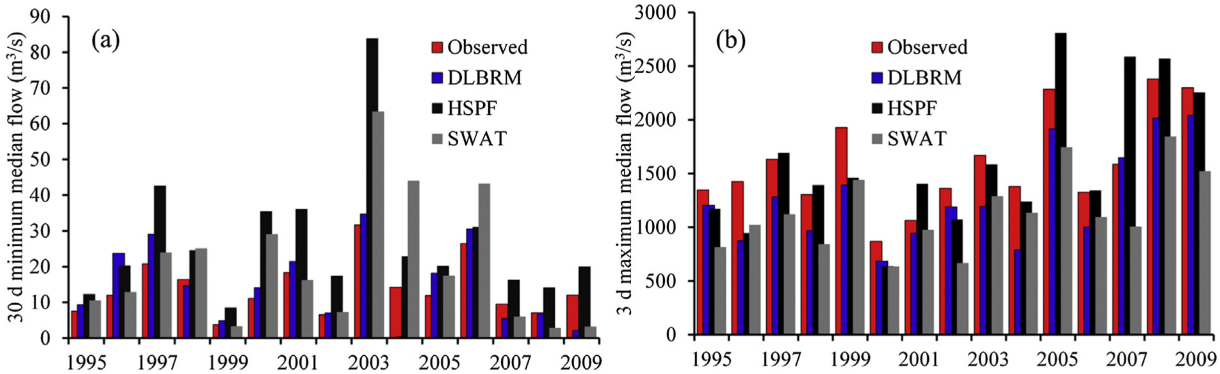

Fig. 4. Observed and simulated median values of 30 d minimum (a) and 3 d maximum (b) flows for the Maumee River for each year, 1995–2009. Observed data were measured at outlet of the Maumee River watershed (Waterville, Ohio; USGS station 04193500), whereas predicted data were generated by three watershed models: the DLBRM, HSPF, and SWAT.

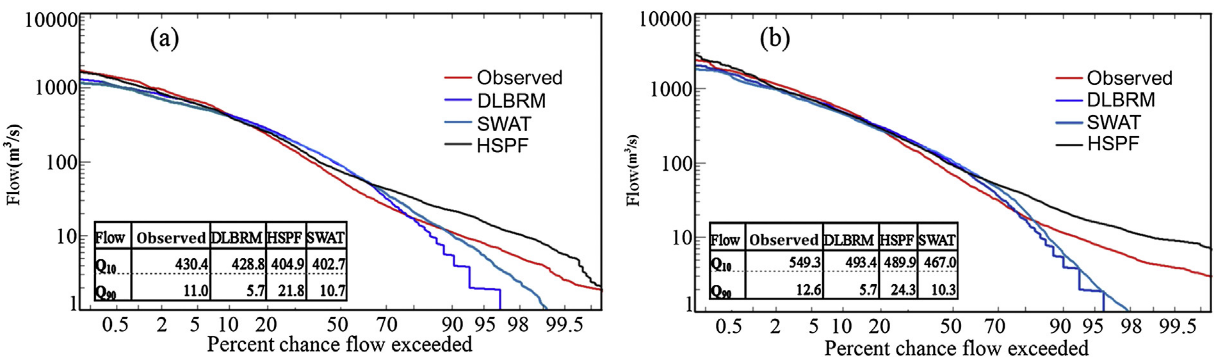

Fig. 5. Observed and simulated flow–duration curves for the Maumee River during calibration (a, 1995–2002) and validation (b, 2003–2009) years. Also reported are Q_{10} (i.e., high flows that were only exceeded 10% of the time in the dataset; m^3/s) and Q_{90} values (i.e., low flows that were exceeded 90% of the time in the dataset; m^3/s) for each period. Observed data were measured at outlet of the Maumee River watershed (Waterville, Ohio; USGS station 04193500), whereas predicted data were generated by three watershed models: the DLBRM, HSPF, and SWAT.

30 mm of precipitation, yet river discharge remained at base flow primarily because freezing temperatures prevented rain and snow melt. All models were able to closely predict this low winter flow (Fig. 7a, leftmost arrow). However, on 22 January 1999 another winter storm resulted in 20 mm of rain during which time both watershed average maximum and minimum temperatures rose from below freezing to 9.8 °C and 1.2 °C, respectively. This sudden temperature increase initiated snow melt that was accompanied by heavy rainfall, which triggered a daily average flow rate of 2087 m^3/s and a peak discharge of 2178 m^3/s (Fig. 7a, rightmost arrow). The DLBRM, HSPF model, and SWAT severely under-predicted the daily average flow rate triggered by this event by 28.3%, 33.4%, and 26.7%, respectively.

3.3.2. Continuous multiple-day early spring precipitation (Event #2)

The second event consisted of a rainy period during 6–11 March 2009, which resulted in a total watershed average precipitation of about 90 mm and triggered an average daily flow of 2009 m^3/s and a peak flow rate of 2758 m^3/s on 11 March (Fig. 7b). This storm

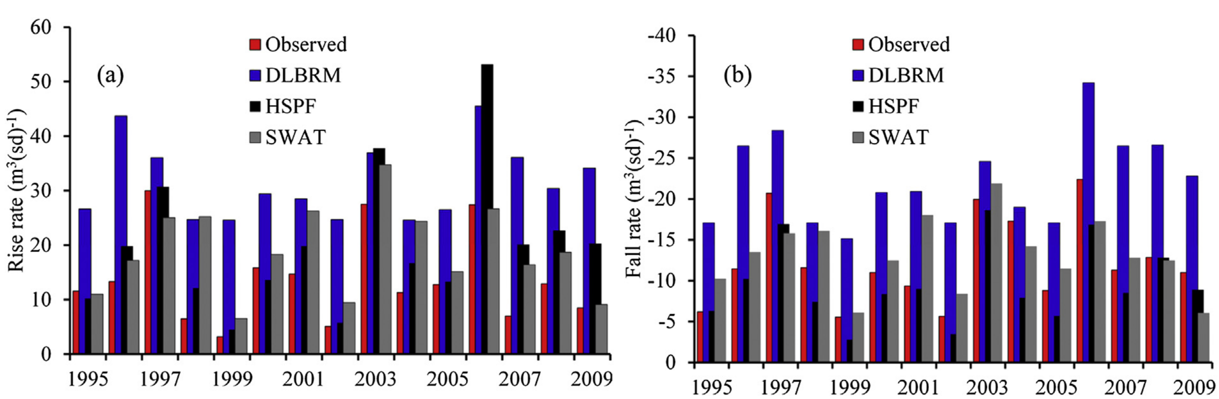

Fig. 6. Observed and simulated median Maumee River flow rise (a) and fall (b) rates during 1995–2009. Observed data were measured at outlet of the Maumee River watershed (Waterville, Ohio; USGS station 04193500), whereas predicted data were generated by three watershed models: the DLBRM, HSPF, and SWAT.

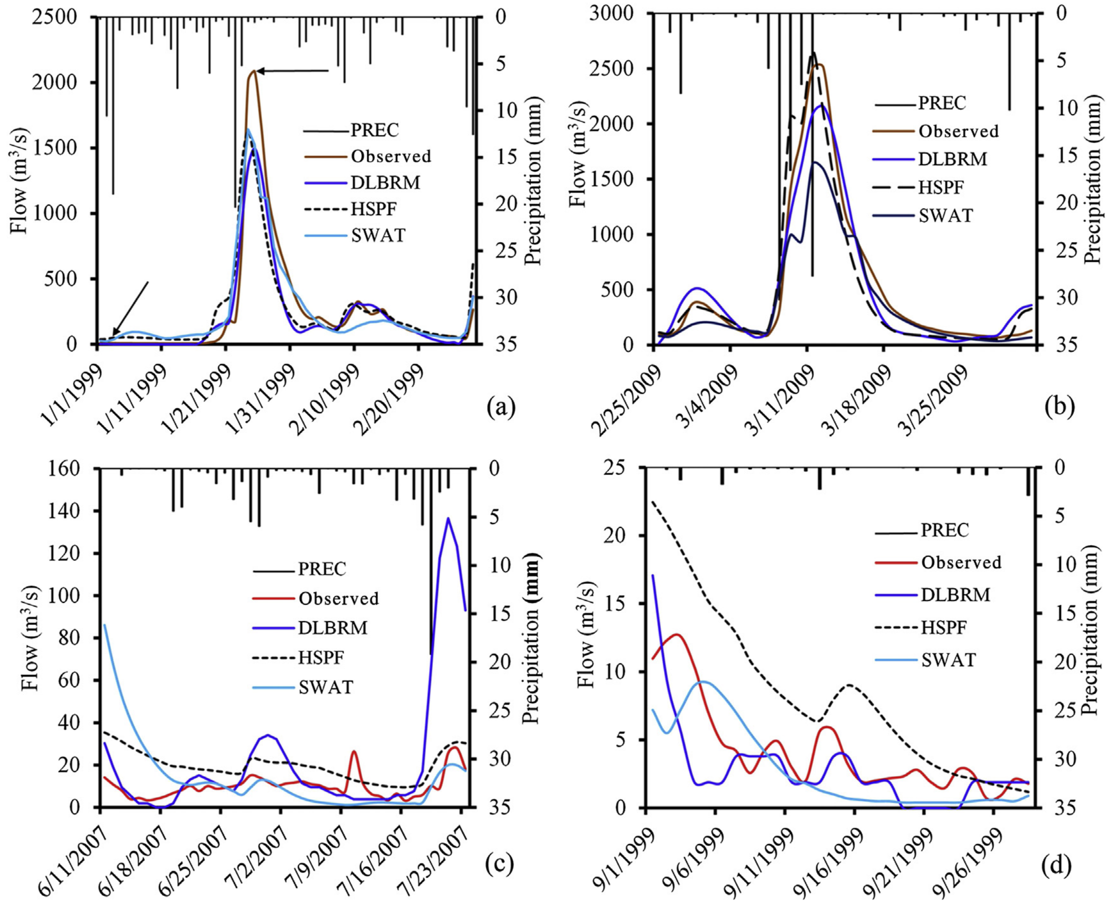

Fig. 7. Observed and simulated Maumee River flows during select "extreme" events: multiple winter storms accompanied by snow melt (a); multiple-day early spring precipitation (b); summer storm following multiple dry spells (c); and prolonged dry condition in autumn (d). Observed precipitation (PREC) is show on the top of each panel. Observed data were measured at outlet of the Maumee River watershed (Waterville, Ohio; USGS station 04193500), whereas predicted data were generated by three watershed models: the DLBRM, HSPF, and SWAT.

system produced precipitation continuously with the largest amounts being produced on March 8 (30 mm) and on March 11 (28 mm). The HSPF model slightly over-predicted average daily flow (by 6.5%), whereas the DLBRM and SWAT under-predicted flow (by 17% and 35%, respectively).

3.3.3. Summer storm following extended dry period (Event #3)

The third event occurred following prolonged dry weather during June through July 2007 that resulted in sustained low flows (mostly < 15 m^3/s). Multiple dry spells between episodic rain events (≤6 mm) resulted in persistent low Maumee River flows during this period. This low-flow condition was followed by a precipitation event that produced 19 mm of rain on 19 July that resulted in an observed flow rate of 28 m^3/s at the Maumee River watershed outlet (Fig. 7c). The HSPF model and SWAT did a reasonable job of simulating daily flow during this time, being within 10% of the observed flow. However, the DLBRM showed a large bias, over-predicting the observed flow rate by >300% (Fig. 7c).

3.3.4. Prolonged dry condition in autumn (Event #4)

Discharge from the Maumee River typically is low during late summer and early autumn, owing to infrequent precipitation events of small magnitude and high evaporation rates. To illustrate the ability of the models to replicate base-flow conditions during this time (i.e., when flows are <10 m^3/s for a prolonged period), we examined their performance during September 1999 (Fig. 7d). During this period, episodic small precipitation events (<4 mm) occurred, but flows remained low. Similar to previous results, the DLBRM mostly under-predicted flows, even predicting no flow at times (Fig. 7d). By contrast, the SWAT over-predicted flow at the outset, but under-predicted it later, with the HSPF model over-predicting flow the entire time (often by 100–200%; Fig. 7d).

3.4. Ease of use

We ranked the models based on a variety of criteria (following Cunderlik, 2003) related to their ease of use (Table 5). Each of the factors used in this assessment was carefully considered and scored based on evaluations made while conducting this work. Overall, we found the SWAT to be the most robust, user-friendly model. Based on the suite of criteria presented in Table 5, SWAT scored highest (sum of all = 54 points), followed by the HSPF model (36 points), and then the DLBRM (29 points).

3.4.1. Documentation and technical support

Availability of technical documentation plays a significant role in successful watershed model development. The documentation available for the SWAT was much better than for the HSPF model, owing to a large available literature and the existence of detailed user manuals (http://swat.tamu.edu/software/swat-model/). A manual for the DLBRM has yet to be developed. Additionally, the SWAT and HSPF model have user forums from which technical advice can be garnered, whereas little online help was available for

Table 5
Attributes compared among three Maumee River watershed-hydrology models (the DLBRM, HSPF model, and SWAT) used in this study. Assessment methods followed that of Cunderlik (2003).

Criteria	Score range		Scores		
	0	5	DLBRM	HSPF	SWAT
Documentation	None	Widely available	1	4	5
Technical support	Unavailable	Widely available	0	3	5
Watershed delineation for large basins	Difficult	Easy	3	3	5
Auto-calibration procedures	Difficult	Easy	4	2	5
Critical risk & source area identification	Impossible	Easy	5	2	3
Ability to vary annual land-use & land-cover data	Impossible	Easy	0	0	5
Cropping practice specification & implementation	Impossible	Many options (easy)	0	4	5
Cropland BMP specification & implementation	Impossible	Many options (easy)	0	4	5
Climate change assessment	Impossible	Easy	3	4	5
Data file management	Difficult	Easy	5	5	1
Compatibility with 64 bit Windows	Incompatible	Highly compatible	5	2	5
Source-code modification	Challenging	Easy	3	3	5

the DLBRM. SWAT users also can easily obtain support from SWAT developers, which is not the case for the HSPF model and the DLBRM.

3.4.2. File management and software compatibility

The models also varied in terms of ease of file management, compatibility with operating systems, and the potential to integrate with other software programs. SWAT posed file management problems as thousands of input files were required to model the Maumee River watershed in this study, which was not the case for the HSPF model and the DLBRM. In terms of compatibility, the HSPF model was found to be the least compatible because all of its pre-processing (WMDUtil) and post-processing interfaces (GenScn) only worked on 32-bit computers running Windows XP. The stable versions of its main GUI interfaces, BASIN and WinHSPF, also only run on Windows XP with Beta versions running on Windows 7 only recently released (in 2013). Source-code modification also was challenging for the HSPF model and DLBRM primarily because of lack of documentation pertaining to code structure and subroutines.

3.4.3. Model setup and calibration

When customizing our three models for use in the Maumee River watershed, a considerable amount of time was spent setting up the model and calibrating it, as successful parameterization is critical to robust predictions of the flow regime. In terms of model setup and calibration, we found that the SWAT was better than the DLBRM and HSPF model for several reasons. Most notably, the SWAT model could be easily discretized into smaller watershed units, its auto-calibration software could optimize a large number of parameters (with an option to choose objective functions), and it has a user-friendly interface (ArcSWAT). By contrast, watershed discretization was challenging for the HSPF model primarily because of the large size of the Maumee River watershed; previous applications of the HSPF model focused on much smaller watersheds (e.g., Demissie et al., 2007). We also found the HSPF model's auto-calibration feature to not be useful in the large Maumee River watershed. Because this feature can only be used when the number of land segments is small (<200 segments), we were forced to manually calibrate the HSPF model. Likewise, although the DLBRM has a user-friendly auto-calibration program, the interface for setting up the DLBRM was less friendly than the SWAT or HSPF model.

3.4.4. Use for scenario testing

Watershed models are increasingly being used to help assess the impacts of changing land use/cover and climate on the flow of water and water constituents, including nutrients, sediments, and pollutants (e.g., Bosch et al., 2013; Scavia et al., 2014). To successfully predict hydrologic and water-quality processes under changing climatic conditions, land-use practices, and land-cover regimes, ability to vary land use/cover and climate data between simulation years is required (DeFries and Eshleman, 2004). Of the three models explored herein, only the SWAT allows the input of annually varying land-use and land-cover GIS layers (Pai and Saraswat, 2011). Additionally, while the HSPF model allows simulation of a similar number of crop management practices and best management practices (BMPs) also available in the SWAT, these practices can be implemented in the SWAT in a simpler, more mechanistic way than in the HSPF model (Radcliffe and Lin, 2007). Further, in contrast to the HSPF model, the SWAT also can simulate crop growth, which allows scenario analysis for impact of crop type and associated management at different stages of plant growth. Capabilities to simulate crop management and changes in land-use are unavailable in the DLBRM. Likewise, while each model can simulate hydrology under future climates, SWAT allows easier importing of climate projection data into the model after calibration and validation than the HSPF model or the DLBRM. Additionally, SWAT offers a better alternative than the HSPF model or the DLBRM for assessing the impact of climate change on watershed hydrology, owing to its capability to simulate plant growth and impacts of vegetation on hydrology under varying concentrations of atmospheric carbon dioxide (U.S. EPA, 2013).

Agencies also have expressed interest in identifying critical pollutant source areas (Croley et al., 2008) that disproportionately affect downstream ecosystems. With this knowledge, appropriate management practices and policies potentially can be implemented to mitigate pollution inputs. One simple way to identify the critical source of pollutants within a watershed is to spatially locate the smallest land segment used for flow computation before it is routed to channels. This strategy is much more easily implemented using the DLBRM than the HSPF model or SWAT because these latter two models depend on lumped land segments for basic calculations. Further, this approach is more likely to be successful using the SWAT than the HSPF model because the HRUs in the SWAT can be more easily spatially identified within subwatershed boundaries as compared to the lumped land segments used in the HSPF model.

4. Discussion

4.1. Daily and monthly flow prediction accuracy

All three models performed quite well in simulating daily and monthly flows within the Maumee River watershed. Based on NSE and PBIAS, which have been suggested as useful metrics to assess watershed-hydrology model performance (ASCE Task Committee, 1993; Moriasi et al., 2007), our models performed in line with or in some cases better than previously published models during both the calibration and validation periods; NSE and PBIAS for daily and monthly flows ranged 0.76–0.91 and −0.5%–6.4%, respectively. For example, in previous applications of the SWAT, the median daily NSE value was 0.6 (range = [−0.23, 0.93], n = 68) and 0.54 (range = [−0.16, 0.87], n = 47) for the calibration and validation periods, respectively (Gassman et al., 2007). Similarly, in a different application of the SWAT to 20 U.S. watersheds, a median NSE value of 0.74 over the entire calibration and validation period was found at the daily and monthly time scale (U.S. EPA, 2013). In yet another application of the SWAT to six different Great Lakes watersheds that included the Maumee River Basin, the NSE and PBIAS ranged 0.7–0.95 and −11–11, respectively, for both calibration and validation periods (Bosch et al., 2011). In terms of HSPF model and DLRBM performance, both outperformed applications found in previously published studies. For example, Demissie et al. (2007) applied the HSPF model to multiple watersheds within the Illinois River basin, reporting daily NSE values ranging from 0.32 to 0.75 (as compared to ≥0.82 herein). Similarly, DeMarchi et al. (2011) applied the DLBRM to multiple watersheds within the Great Lakes region, reporting daily NSE values that ranged 0.09–0.71 (as opposed to ≥0.82 herein).

In an attempt to assess whether these findings held at smaller spatial scales, we tested the ability of the SWAT, HSPF model, and DLBRM to simulate Maumee River flows at six USGS gages located within the watershed (as opposed to the Maumee River Basin outlet). Unlike a similar validation attempt conducted as part of the DMIP project (Reed et al., 2004), the performance of our three models at uncalibrated interior gages did not vary with gage drainage area. In general, our models reliably simulated flows of the six gages (NSE = [0.53, 0.90] and PBIAS = [−23, 14.6]) with the only exception being the performance of the SWAT at the Tiffin River gage (NSE = 0.28 and PBIAS = 30.3), which is located far upstream relative to the Basin outlet. The poor performance of the SWAT at the Tiffin River gage points to an advantage that fully distributed models (e.g., DLBRM) appear to have over lumped models (e.g., the SWAT and HSPF model) in simulating flows at interior locations, especially those found far upstream of the calibration site. Indeed, an advantage that distributed models have over lumped models is simulating flows anywhere within the watershed, without the need to calibrate more than at the basin outlet (Smith et al., 2004).

Although the models were calibrated using two different objective functions (SWAT and HSPF: NSE, DLBRM: RMSE), differences in the type of objective function did not appear to influence the performance of the models primarily because both objective functions involve second-order error computation. This is evident from the fact that the NSE value that we calculated for the DLBRM for calibration years was comparable to those obtained for the HSPF model and SWAT through calibration. Similarly, the RMSE values calculated for the HSPF model and SWAT were comparable to the value obtained for the DLBRM via calibration. Moreover, the SWAT and HSPF model, which were calibrated with the same objective function, displayed differing patterns in predicting high and low flows, indicating minimal influence of the calibration objective function. However, while the performance of the models was nearly equal when compared in terms of the most commonly used metrics (NSE and PBIAS), consideration of other recently recommended metrics (Willmott and Matsuura, 2005; Criss and Winston, 2008) could help discern subtle differences in predicting overall daily and monthly flow at watershed outlet. For example, if we compare the models in terms of MAE, we observe reduced model performance during the validation years as compared to calibration years for all three models, which makes intuitive sense given that the parameter values are not "trained" to the data during the validation phase. Also, we found MAE to be 16–19% lower for the HSPF model than for the DLBRM or SWAT at a daily time step during calibration, whereas at a monthly time scale, MAE was 20–25% lower for the HSPF model relative to the other two models for both the calibration and validation years. These results suggest that the HSPF model offers a more unbiased and reliable predictor of average Maumee River flows than either the SWAT or DLBRM.

Similar to our findings, previously published studies outside of the Great Lakes Basin have found the HSPF model to outperform the SWAT when simulating average daily flows (Saleh and Du, 2004; Im et al., 2007; Lian et al., 2007; Nasr et al., 2007). For example, Lian et al. (2007) found the HSPF model to outperform the SWAT when estimating average daily flow in the Illinois River watershed, when NSE was used as a comparative metric (NSE = 0.80 for HSPF versus 0.67 for SWAT). In the upper North Bosque River watershed (TX), the HSPF model outperformed the SWAT during both calibration (NSE = 0.72 versus 0.17, respectively) and validation (NSE = 0.70 versus 0.62, respectively) (Saleh and Du, 2004). We suggest that the lower performance of the SWAT in our study (and these other studies) is due to the use of a curve number in the SWAT that does not account for rainfall intensity. A similar conclusion was suggested when the SWAT produced lower NSE values than the HSPF model in a recent study conducted by U.S.EPA (2013), where both models were applied to five different watersheds.

4.2. *Flood and low-flow pulse frequency/magnitude prediction capability*

Despite each model's solid performance at predicting average daily and monthly flows, the SWAT, HSPF model, and DLBRM were somewhat limited in their ability to predict flood and low-flow conditions that could be critical to understanding the distribution and composition of stream biota, including macrophytes, invertebrates, and fish (Poff and Ward, 1989), as well as fisheries production in Lake Erie. For example, recruitment of Lake Erie's two most important recreational and commercial fishes (walleye *Sander vitreus* and yellow perch *Perca flavescens*) have been shown to depend on Maumee River discharge during the larval production period in spring, with low flows benefiting survival of larval walleye in the Maumee River (Mion et al., 1998) and high flows benefiting larval yellow perch survival through open-lake river-plume formation (Reichert et al., 2010). Herein, we found that the DLBRM and SWAT were more limited in their ability to simulate flood flows than the HSPF model, whereas the HSPF model and the DLBRM were less able to simulate low flows than the SWAT. Similar weaknesses have been reported for the SWAT (Lian et al., 2007; Srinivasan et al., 2005; Borah et al., 2007) and HSPF (Lian et al., 2007). Withstanding these problems, accurate predictions of flood-pulse frequencies by the models suggest their potential for using them for flood prediction, if they benefit from further enhancement to allow improved prediction of flood magnitudes (especially by the SWAT and DLBRM).

4.3. *Extreme-event performance*

The inability of all three models to accurately simulate flows during extreme events is an important limitation because numerous studies have demonstrated that a large fraction of pollutant (including nutrient) export from the watershed can occur during a few high flow events generated by snowpack melt (Steinheimer et al., 1998) and heavy storms (Vanni et al., 2001; Royer et al., 2006). For example, sequential heavy storms that triggered flows in excess of 2000 m^3/s caused excessive export of dissolved reactive phosphorus from the Maumee River watershed, which in turn, led to a record-setting algal bloom in Lake Erie during 2011 (Michalak et al., 2013). Because harmful algal blooms in Lake Erie have been shown to be positively related to Maumee River discharge (Stumpf et al., 2012), the need to correctly simulate extreme weather-related discharges is critical to the prediction and perhaps mitigation of this water-quality problem. Likewise, extremely low flows also need to be accurately simulated because they can result in high in-stream nutrient storage that can account for higher than expected in-stream biological production or downstream nutrient pulses (Hall et al., 2009). While our findings indicate a need to further improve each model's ability to predict extreme flows, they suggest that caution should be taken when selecting a watershed-hydrology model and interpreting its results.

All three of our models showed some inability to accurately simulate flows during extreme events within the Maumee River Basin. The largest discrepancy related to extreme weather was found in the DLRBM's prediction of flow from a summer storm event following a prolonged dry condition (see Fig. 7c). One likely reason for such large bias with the DLBRM is its runoff calculation mechanism, which is based on the partial-area concept (Croley, 2002). Using partial-area hydrology approaches limits the user's ability to correctly identify areas within the watershed that contribute to runoff from those contributing to groundwater recharge (VanDeGriend and Engman, 1985). After a prolonged dry period, one should expect that any precipitation should first replenish soil moisture and shallow groundwater before runoff occurs (as was evident from observed flow and predictions made by the SWAT and HSPF model). By contrast, the DLBRM, which predicted lower flows during dry spells preceding the storm (see Fig. 7c), also predicted a flow that was three times greater than the observed flow. This disparity indicates that the model incorrectly calculated the fraction of area contributing to overland flow. This result also was obvious from simulated fall and rise rates in which the model incorrectly portrayed the Maumee River as being “flashy”. Because partial-area hydrology seems inappropriate for watersheds such as the Maumee River, which are predominantly flat and characterized by poorly drained soils (Lake, 1978), our findings highlight the need to closely consider the match between hydrogeomorphology of the watershed and the underlying runoff generation mechanism when selecting a model.

Overall, all three models were better at simulating extreme wet conditions than extreme dry conditions for the Maumee River Basin, with the HSPF model showing marked improvement over the SWAT. One possible reason for the better performance of the HSPF model relative to the SWAT could be related to the temporal resolution of the input data. Hourly time step and precipitation data are used in HSPF model, which seemingly would allow the model to better partition precipitation between infiltration and overland flow relative to the SWAT, which lumps precipitation to a single daily value via the SCS curve number method. Results also demonstrated that the capacity of all three models to accurately predict flow was undermined by the concurrent occurrence of snow melt and rainfall (see Fig. 7a); Maumee River flow was underestimated in each case. Thus, while the SWAT, HSPF model, and DLBRM all have snow-melt routines, a need to better simulate the joint occurrence of snow melt and rainfall exists.

4.4. Ease of use

While ability to accurately simulate flow conditions is likely to be the most important factor to consider when choosing a watershed-hydrology model, consideration of other factors is recommended when no single model substantially outperforms all other models in terms of simulating flow conditions (as is the case here). Additional factors include the time required to setup and calibrate the model, the level of user support that is available from online sources (e.g., user forums) or written documentation (e.g., manuals), the file management system and compatibility with computer operating systems, the ability to use the model in a scenario-testing mode (e.g., to explore the impacts climate and land-use change), and the ability to access and modify the source code, which might be important when trying to integrate a model with other models simulating biophysical processes. In our evaluation of these and other factors (see Table 5), we found the SWAT excelled relative to the HSPF model and DLRBM in all factors except the file management system and the ability to potentially identify areas within the watershed that are contributing the flow of water or pollutants (e.g., pesticides, nutrients). Thus, for users who seek to develop a comprehensive watershed model to quantify water quantity and quality, evaluate management practices to reduce nutrient flux to Lake Erie, and (or) assess future climate change impacts on watershed hydrology and pollutant export, SWAT appears to offer the best combination of accuracy, user-friendliness, and flexibility. This conclusion is supported by a recent study (US E.P.A., 2013) that compared the HSPF model and SWAT across five U.S. watersheds.

5. Conclusions

1. Comprehensive evaluation of the models over a wide range of flow regimes, and spatial and temporal scales, provided a greater opportunity to assess limitations of the models than conventional comparison with one or two GOF indices at single gaging station. We recommend the use of this approach for future model assessment and the use of multiple gages and GOF metrics when comparing models.
2. Based on widely used assessment criteria (i.e., NSE, PBIAS, and R^2), the SWAT, HSPF model, and DLBRM simulated average daily and monthly Maumee River flows with an acceptable and equal degree of accuracy. The models also were capable of reliably predicting flows within the Maumee River Basin based on watershed outlet calibration. However, based on other GOF metrics that have been recently recommended for assessing model performance (e.g., MAE, VE), we conclude that the HSPF model is better at predicting flows within the Maumee River watershed than the SWAT and DLBRM.
3. Analysis of the frequency and duration of flows indicated that each model showed a greater limitation in predicting low-flow conditions than flood conditions for the Maumee River Basin. The DLBRM was least reliable in accurately predicting low flows and flow fall and rise rates, suggesting a mismatch between the hydrogeomorphology of the watershed and the DLBRM's underlying runoff generation mechanism. Flood pulses and extreme wet events were best predicted by the HSPF model.
4. Because performance differences among the models were relatively minor, when other model characteristics were included in the evaluation, we concluded that the SWAT is the best of the three models for applications to the Maumee River watershed. Advantages of the SWAT leading to this conclusion include ease of model use, ability to represent agricultural BMPs, availability of documentation and technical support, compatibility with current operating systems, ease of conducting climate change assessments, and ease of source-code modification.

Acknowledgments

Funding for this research was provided by the NSF Dynamics of Coupled Natural and Human Systems Program (BCS-1114934) and the Ohio Sea Grant College Program (R/ES 60043508).

Appendix A. Supplementary material

Supplementary material related to this article can be found at http://dx.doi.org/10.1016/j.envsoft.2014.07.004.

References

Abbaspour, K.C., Yang, J., Maximov, I., Siber, R., Bogner, K., Mieleitner, J., Zobrist, J., Srinivasan, R., 2007. Modelling hydrology and water quality in the pre-alpine/alpine Thur watershed using SWAT. J. Hydrol. 333 (2–4), 413–430.

Angel, J.R., Kunkel, K.E., 2010. The response of Great Lakes water levels to future climate scenarios with an emphasis on Lake Michigan-Huron. J. Great Lakes Res. 36, 51–58.

Arabi, M., Frankenberger, J.R., Engel, B.A., Arnold, J.G., 2007. Representation of agricultural conservation practices with SWAT. Hydrol. Process. 22 (16), 3042–3055. http://dx.doi.org/10.1002/hyp.6890.

Arnold, J.G., Allen, P.M., 1999. Automated methods for estimating base flow and groundwater recharge from stream. J. Am. Water Resour. Assoc. 35, 411–424.

Arnold, J.G., Srinivasan, R., Muttiah, R.S., Williams, J.R., 1998. Large-area hydrologic modeling and assessment: Part I. Model development. J. Am. Water Resour. Assoc. 34 (1), 73–89.

Arnold, J.G., Kiniry, J.R., Srinivasan, R., Williams, J.R., Haney, E.B., Neitsch, S.L., 2011. Soil and Water Assessment Tool User's Manual, Version 2009. Texas Water Resource Institute Technical Report No. 365.

ASCE Task Committee on Definition of Criteria for Evaluation of Watershed Models of the Watershed Management Committee, Irrigation and Drainage Division, 1993. Criteria for evaluation of watershed models. J. Irrig. Drain. Eng. 119 (3), 429–442.

Bennett, N.D., Croke, B.F.W., Guariso, G., Guillaume, J.H.A., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T.H., Norton, J.P., Perrin, C., Pierce, S.A., Robson, B., Seppelt, R., Voinov, A.A., Fath, B.D., Andreassian, V., 2013. Characterising performance of environmental models. Environ. Model. Softw. 40, 1–20.

Bicknell, B.R., Imhoff, J.C., Kittle Jr., J.L., Jobes, T.H., Donigian Jr., A.S., 2001. Hydrological Simulation Program – Fortran (HSPF): User's Manual for Release 12. USEPA, Athens, Ga.

Bingner, R.L., Theurer, F.D., Yuan, Y., 2011. AnnAGNPS technical processes documentation, Version 5.2. Available online at: ftp://199.133.90.201/pub/outgoing/AGNPS/AGNPS_Web_Files/pdf_files/Technical_Documentation.pdf (accessed 07.07.13.).

Borah, D.K., Bera, M., 2003. Watershed-scale Hydrologic and non-point source pollution models: review of mathematical bases. Trans. ASAE 46 (6), 1553–1566.

Borah, D.K., Arnold, J.G., Bara, M., Krug, E.C., Liang, X., 2007. Storm event and continuous hydrologic modeling for comprehensive and efficient watershed simulations. J. Hydrol. Eng. 12 (6), 605–616.

Bosch, N.S., Allan, J.D., Dolan, D.M., Han, H., Richards, R.P., 2011. Application of the Soil and Water Assessment Tool for six watersheds of Lake Erie: model parameterization and calibration. J. Great Lakes Res. 37, 263–271.

Bosch, N.S., Allan, J.D., Selegean, J.P., Scavia, D., 2013. Scenario-testing of agricultural best management practices in Lake Erie watersheds. J. Great Lakes Res. 39, 429–436.

Bouraoui, F., Dillaha, T., 1996. ANSWERS-2000: runoff and sediment transport model. J. Environ. Eng. 122 (6), 493–502.

Boyle, D.P., Gupta, H.V., Sorooshian, S., 2000. Toward improved calibration of hydrologic models: combining the strengths of manual and automatic methods. Water Resour. Res. 36 (12), 3663–3674.

Burns, N.M., Rockwell, D.C., Bertram, P.E., Dolan, D.M., Ciborowski, J.J.H., 2005. Trends in temperature, Secchi depth, and dissolved oxygen depletion rates in the Central Basin of Lake Erie, 1983–2002. J. Great Lakes Res. 31 (Suppl. 2), 35–49.

Cloern, J.E., 2001. Our evolving conceptual model of the coastal eutrophication problem. Mar. Ecol. Prog. Ser. 210, 223–253.

Criss, R.E., Winston, W.E., 2008. Do Nash values have value? Discussion and alternate proposals. Hydrol. Process. 22, 2723–2725.

Croley II, T.E., 2002. Large Basin Runoff Model. In: Singh, V., Frevert, D., Meyer, S. (Eds.), Mathematical Models of Large Watershed Hydrology. Water Resources Publications, Littleton, Colorado, pp. 717–770.

Croley II, T.E., He, C., 2005. Distributed-parameter large basin runoff model. I: Model development. J. Hydrol. Eng. 10 (3), 173–181.

Croley II, T.E., Raikow, D.F., He, C., Atkinson, J.F., 2008. Hydrological resource sheds. J. Hydrol. Eng. 13 (9), 873–885.

Cunderlik, J.M., 2003. Hydrologic Model Selection for the CFCAS Project: Assessment of Water Resources Risk and Vulnerability to Changing Climatic Conditions. Report No. 046 (accessed 24.07.13.). The University of Western Ontario. Available at: http://www.eng.uwo.ca/research/iclr/fids/publications/cfcas-climate/reports/Report%20I.pdf (accessed 07.07.13.).

David, M.B., Gentry, L.E., Kovacic, D.A., Smith, K.M., 1997. Nitrogen balance in and export from an agricultural watershed. J. Environ. Qual. 26 (4), 1038–1048.

DeFries, R., Eshleman, K.N., 2004. Land-use change and hydrologic processes: a major focus for the future. Hydrol. Process. 18, 2183–2186.

DeMarchi, C., Croley II, T.S., He, C., Hunter, T.S., 2009. Application of a distributed watershed hydrology and water quality model in the Great Lakes basin. In: Proceedings of the 7th International Symposium on Ecohydraulics 2009, Concepcion, Chile, January 12–16, 2009.

DeMarchi, C., Xing, F., Croley II, T.S., He, C., Wang, Y., 2011. Application of a Distributed Large Basin Runoff Model to Lake Erie: model calibration and analysis of parameter spatial variation. J. Hydrol. Eng. 16 (3), 193–202.

Demissie, M., Singh, J., Knapp, H.V., Saco, P., Lian, Y., 2007. Hydrologic Model Development for the Illinois River Basin Using Basins 3.0. Illinois State Water Survey Center for Watershed Science, ISWS Contract Report 2007-03.

Diaz, R.J., Rosenberg, R., 2008. Spreading dead zones and consequences for marine ecosystems. Science 321 (5891), 926–929.

Donigian Jr., A.S., 2000. Bibliography of HSPF and Related References. AQUA TERRA Consultants, Mountain View, Cal.

Donigian Jr., A.S., Imhoff, J.C., 2009. Evaluation and performance assessment of watershed models. WEF TMDL 2009, August 9–12, 2009. Minneapolis, MN. In: WEF Specialty Conference Proceedings on CD-ROM.

Donner, S.D., Coe, M.T., Lenters, J.D., Twine, T.E., Foley, J.A., 2002. Modeling the impact of hydrological changes on nitrate transport in the Mississippi river basin from 1955 to 1994. Global Biogeochem. Cyc. 16 (3), 1043.

Duda, P.B., Hummel, P.R., Donigian Jr., A.S., Imhoff, J.C., 2012. BASINS/HSPF: model use, calibration and validation. Trans. ASABE 55 (4), 1523–1547.

Gassman, P.W., Reyes, M.R., Green, C.H., Arnold, J.G., 2007. The soil water and assessment tool: historical development, applications, and future research directions. Trans. ASABE 50 (4), 1211–1250.

Gupta, H.V., Sorooshian, S., Yapo, P.O., 1999. Status of automatic calibration for hydrologic models: comparison with multilevel expert calibration. J. Hydrol. Eng. 4 (2), 135–143.

Hall, R.O., Baker, M.A., Arp, C.D., Koch, B.J., 2009. Hydrologic control of nitrogen removal, storage, and export in a mountain stream. Limnol. Oceanogr. 54 (6), 2128–2142.

Han, W., Yang, Z., Di, L., Mueller, R., 2012. CropScape: a Web service based application for exploring and disseminating US conterminous geospatial cropland data products for decision support. Comput. Electron. Agr. 84, 111–123.

Hawley, N., Johengen, T.H., Rao, Y.R., Ruberg, S.A., Beletsky, D., Ludsin, S.A., Eadie, B.J., Schwab, D.J., Croley, T.E., Brandt, S.B., 2006. Lake Erie hypoxia prompts Canada-US study. Eos, Trans. Am. Geophys. Union 87 (32), 313–324.

He, C., Zhang, L., Fu, L., Luo, Y., Li, L., DeMarchi, C., 2013. Simulation of river flow for downstream water allocation in the Heihe River Watershed, Northwest China. In: Fu, B., Bruce, J.K. (Eds.), Landscape Ecology for Sustainable Environment and Culture. Springer, pp. 173–186.

Im, S., Brannan, K.M., Mostaghimi, S., Kim, S.M., 2007. Comparison of HSPF and SWAT models performance for runoff and sediment yield prediction. J. Environ. Sci. Health. A Toxicol. Hazard. Subst. Environ. Eng. 42 (11), 1561–1570.

Jakeman, A.J., Letcher, R.A., Norton, J.P., 2006. Ten iterative steps in development and evaluation of environmental models. Environ. Model. Softw. 21 (5), 602–614.

Krause, P., Boyle, D.P., Base, F., 2005. Comparison of different efficiency criteria for hydrological model. Assessment. Adv. Geosci. 5, 89–97.

Lake, J., 1978. In: Morrison, J. (Ed.), Environmental Impact of Land Use on Water Quality: Final Report on the Black Creek Project. US Environmental Protection Agency, Chicago IL. EPA-905/9-77-007B.

Lian, Y., Demissie, M., Xie, H., Singh, J., Knapp, H., 2007. Comparison of flow and sediment modeling using SWAT and HSPF for Watersheds in the Illinois River Basin. In: World Environmental and Water Resources Congress 2007, pp. 1–9.

Michalak, A.M., Anderson, E., Beletsky, D., Boland, S., Bosch, N.S., Bridgeman, T.B., Chaffin, J.D., Cho, K.H., Confesor, R., Daloğlu, I., DePinto, J., Evans, M.A., Fahnenstiel, G.L., He, L., Ho, J.C., Jenkins, L., Johengen, T., Kuo, K.C., Laporte, E., Liu, X., McWilliams, M., Moore, M.R., Posselt, D.J., Richards, R.P., Scavia, D., Steiner, A.L., Verhamme, E., Wright, D.M., Zagorski, M.A., 2013. Record-setting algal bloom in Lake Erie caused by agricultural and meteorologycal trends consistent with expected future conditions. Proc. Natl. Acad. Sci. U.S.A. 110 (16), 6243–6244.

Mion, J.B., Stein, R.A., Marschall, E.A., 1998. River discharge drives survival of larval walleye. Ecol. Appl. 8 (1), 88–103.

Moriasi, D.N., Arnold, J.G., Van Liew, M.W., Bingner, R.L., Harmel, R.D., Veith, T.L., 2007. MODEL evaluation guidelines for systematic quantification of accuracy in watershed simulations. Trans. ASABE 50 (3), 885–900.

Morris, M.D., 1991. Factorial sampling plans for preliminary computational experiments. Technometrics 33 (2), 161–174.

Nature Conservancy, 2009. Indicators of Hydrologic Alteration Version 7.1 User's Manual.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models: Part 1. A discussion of principles. J. Hydrol. 10 (3), 282–290.

Nasr, A., Bruen, M., Jordan, P., Moles, R., Kiely, G., Byrne, P., 2007. A comparison of SWAT, HSPF and SHETRAN/GOPC for modeling phosphorus export from three catchments in Ireland. Water Res. 41, 1065–1073.

NRCS, 2005. Western Lake Erie Basin Water Resources Protection Plan Ohio, Indiana and Michigan. United States Department of Agriculture Natural Resources Conservation Service. Available at: ftp://ftp-fc.sc.egov.usda.gov/OH/pub/Programs/Lake_Erie/erie_basin_plan.pdf (online accessed 23.07.13.).

Ohio EPA, 2010. Ohio Lake Erie Phosphorus Task Force. Final report. Ohio Environmental Protection Agency, p. 90.

O'Neil, J.M., Davis, T.W., Burford, M.A., Gobler, C.J., 2012. The rise of harmful cyanobacteria blooms: the potential roles of eutrophication and climate change. Harmful Algae 14 (2012), 313–334.

Pai, N., Saraswat, D., 2011. SWAT2009_LUC: a tool to activate land use change module in SWAT 2009. Trans. ASABE 54 (5), 1649–1658.
Poff, N.L., Ward, J.V., 1989. Implications of streamflow variability and predictability for lotic community structure: a regional analysis of streamflow patterns. Can. J. Fish. Aquat. Sci. 46, 1805–1818.
Poff, N.L., Allan, J.D., Bain, M.B., Karr, J.R., Prestegaard, K.L., Richter, B.D., Sparks, R.E., Stromberg, J.C., 1997. The natural flow regime. BioScience 47 (11), 769–784.
Pushpalatha, R., Perrin, C., Moine, N.L., Andréassian, V., 2012. A review of efficiency criteria suitable for evaluating low-flow simulations. J. Hydrol. 420–42, 171–182.
Rabalais, N.N., Turner, R.E., Díaz, R.J., Justić, D., 2009. Global change and eutrophication of coastal waters. ICES J. Mar. Sci. 66, 1528–1537.
Radcliffe, D.E., Lin, Z., 2007. Modeling phosphorus with hydrologic Simulation Program-Fortran. In: Radcliffe, D.E., Cabrera, M.L. (Eds.), Modeling Phosphorus in the Environment. CRC Press, Boca Raton, Florida, pp. 189–214.
Reed, S., Koren, V., Smith, M., Zhang, Z., Moreda, F., Seo, D.-J., DMIP Participants, 2004. Overall distributed model intercomparison project results. J. Hydrol. 298 (1–4), 27–60.
Refsgaard, J.C., Storm, B., 1995. MIKE SHE. In: Singh, V.P. (Ed.), Computer Models of Watershed Hydrology. Water Resources Publications, Highlands Ranch, Colo., pp. 809–846 (chapter 23).
Reichert, J.M., Fryer, B.J., Pangle, K.L., Johnson, T.B., Tyson, J.T., Drelich, A.B., Ludsin, S.A., 2010. River-plume use during the pelagic larval stage benefits recruitment of a lentic fish. Can. J. Fish. Aquat. Sci. 67, 987–1004.
Richards, R.P., Baker, D.B., Eckert, D.J., 2002. Trends in agriculture in the LEASEQ watersheds, 1975–1995. J. Environ. Qual. 31, 17–24.
Richards, R.P., Baker, D.B., Crumrine, J.P., Stearns, A.M., 2010. Unusually large loads in 2007 from the Maumee and Sandusky Rivers, tributaries to Lake Erie. J. Soil Water Conserv. 65, 450–462.
Richter, B.D., Baumgartner, J.V., Powell, J., Braun, D.P., 1996. A method for assessing hydrologic alteration within ecosystems. Conserv. Biol. 10 (4), 1163–1174.
Royer, T.V., David, M.B., Gentry, L.E., 2006. Timing of riverine Export of nitrate and phosphorus from agricultural watersheds in Illinois: implications for reducing nutrient loading to the Mississippi river. Environ. Sci. Technol. 40 (13), 4126–4131.
Rucinski, D.K., Beletsky, D., DePinto, J.V., Schwab, D.J., Scavia, D., 2010. A simple 1-dimensional, climate based dissolved oxygen model for the central basin of Lake Erie. J. Great Lakes Res. 36, 465–476.
Saleh, A., Du, B., 2004. Evaluation of SWAT and HSPF within BASINS program for the Upper North Bosque River watershed in Central Texas. Trans. ASAE 47 (4), 1039–1049.
Scavia, D., Allan, J.D., Arend, K.A.K., Bartell, S., Beletsky, D., Bosch, N.S., Brandt, S.B., Briland, R.D., Daloğlu, I., DePinto, J.V., Dolan, D.M., Evans, M.A., Farmer, T.M., Goto, D., Han, H., Höök, T.O., Knight, R., Ludsin, S.A., Mason, D.M., Michalak, A.M., Richards, R.P., Roberts, J.J., Rucinski, D.K., Rutherford, E., Schwab, D.J., Sesterhenn, T., Zhang, H., Zhou, Y., 2014. Assessing and addressing the re-eutrophication of Lake Erie: central basin hypoxia. J. Great Lakes Res. 40 (2), 226–246.
Smith, M.B., Seo, D.-J., Koren, V.I., Reed, S., Zhang, Z., Duan, Q.-Y., Moreda, F., Cong, S., 2004. The distributed model intercomparison project (DMIP): motivation and experiment design. J. Hydrol. 298 (1–4), 4–26.
Smith, M., Koren, V., Zhang, Z., Zhang, Y., Reed, S., Cui, Z., Moreda, F., Cosgrove, B.A., Mizukami, N., Anderson, E.A., DMIP 2 Participants, 2012. Results of the DMIP 2 Oklahoma experiments. J. Hydrol. 418–419, 17–48.
Srinivasan, R.S., Gerard-Marchant, P., Veith, T.L., Gburek, W.J., Steenhuis, T.S., 2005. Watershed scale modeling of critical source areas of runoff generation and phosphorus transport. J. Am. Water Resour. Assoc. 41 (2), 361–375.
Steinheimer, T.R., Scoggin, K.D., Kramer, L.A., 1998. Agricultural chemical movement through a field-size watershed in Iowa: surface hydrology and nitrate losses in discharge. Environ. Sci. Technol. 32 (8), 1048–1052.
Stumpf, R.P., Wynne, T.T., Baker, D.B., Fahnenstiel, G.L., 2012. Interannual variability of cyanobacterial blooms in Lake Erie. PloS One 7 (8), e42444.
U.S. EPA (Environmental Protection Agency), 2013. Watershed Modeling to Assess the Sensitivity of Streamflow, Nutrient, and Sediment Loads to Potential Climate Change and Urban Development in 20 U.S. Watersheds. EPA/600/R-12/058F. National Center for Environmental Assessment, Washington, DC. Available from the National Technical Information Service, Alexandria, VA, and online at: http://www.epa.gov/ncea.
USDA, 2010. Field Crops Usual Planting and Harvesting Dates. United States Department of Agriculture, National Agricultural Statistics Service. Agricultural Handbook Number 628.
VanDeGriend, A.A., Engman, E.T., 1985. Partial area hydrology and remote sensing. J. Hydrol. 81, 211–251.
Vanni, M.J., Renwick, W.H., Headworth, J.L., Auch, J.D., Schaus, M.H., 2001. Dissolved and particulate nutrient flux from three adjacent agricultural watersheds: a five-year study. Biogeochemistry 54 (1), 85–114.
Vidon, P., Hubbard, L.E., Soyeux, E., 2009. Seasonal solute dynamics across land uses during storms in glaciated landscape of the US Midwest. J. Hydrol. 376 (1–2), 34–47.
White, M.J., Storm, D.E., Busteed, P.R., Smolen, M.D., Zhang, H., Fox, G.A., 2010. A quantitative phosphorus loss assessment tool for agricultural fields. Environ. Model. Softw. 25, 1121–1129.
Willmott, C.J., 1981. On the validation of models. Phys. Geogr. 2, 184–194.
Willmott, C.J., Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim. Res. 30, 79–82.
Zambrano-Bigiarini, M., 2012. hydroGOF: goodness-of-fit functions for comparison of simulated and observed hydrological time series. R package version 0.3-4. http://CRAN.R-project.org/package=hydroGOF.