Journal of Hydrology 380 (2010) 277–288

Contents lists available at ScienceDirect

Journal of Hydrology

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

Watershed model calibration using multi-objective optimization and multi-site averaging

Xuyong Li [a,b,*], Donald E. Weller [b], Thomas E. Jordan [b]

[a] State Key Laboratory of Urban and Regional Ecology, Research Center for Eco-Environmental Sciences, Chinese Academy of Science, Shuangqing Road 18, Beijing 100085, China
[b] Smithsonian Environmental Research Center, 647 Contees Wharf Road, P.O. Box 28, Edgewater, Maryland 20137-0028, USA

ARTICLE INFO

Article history:
Received 17 July 2008
Received in revised form 9 October 2009
Accepted 10 November 2009

This manuscript was handled by L. Charlet, Editor-in-Chief, with the assistance of Associate Editor Jose D. Salas

Keywords:
Multi-site averaging
Multi-objective optimization
Watershed model calibration
Parameter estimation
GWLF watershed model
Chesapeake Bay

SUMMARY

Recent advances in optimizing watershed model calibration have focused mainly on incorporating multiple objective measures of model performance and improving optimization algorithms. However, some parameters vary widely among different calibration locations. We present a watershed model calibration method that combines multi-objective optimization with averaging across multiple calibration sites. Model parameters were first estimated by multi-objective optimization at each calibration site, and then finalized by weighted averaging the parameter values across sites. The weight for each site was calculated from the prediction error at that site. The calibration framework was applied to estimate 16 hydrological and nutrient parameters of the General Watershed Loading Function (GWLF) watershed model at the Rhode River basin, in Maryland, United States of America. When calibrated to a single watershed, GWLF gave reasonable predictions for monthly streamflow (r^2 = 0.71–0.78), monthly total nitrogen (TN) loads (r^2 = 0.55–0.65), annual streamflow (r^2 = 0.80–0.91), and annual TN loads (r^2 = 0.67–0.86); but success for total phosphorus (TP) loads varied among watersheds (r^2 = 0.41–0.68 for monthly TP loads and r^2 = 0.47–0.79 for annual TP loads). In comparison to the single-site calibrations, the multi-site weighted average approach combined with multi-objective optimization reduced the relative cumulative error of predictions in validation watersheds by 3.5–7.4% for monthly streamflow, 3.2–6.3% for monthly TN loads, and 4.3–5.9% for monthly TP loads, respectively.

Introduction

The problem of eutrophication has become a central theme of coastal research and management due to the ever-increasing proportion of the human population being concentrated in coastal watersheds (Rosenberg, 1985; Gray, 1992; Howarth et al., 2000; Kemp et al., 2005). Watershed models are increasingly used in management efforts to reduce nutrient and sediment loads. The prediction accuracy of watershed models depends on model structure, input data availability and data quality, and model parameter estimation. Most watershed models have many parameters that cannot be directly measured or cannot be determined at the required spatial scale; therefore some model parameters must be estimated through model calibration. Calibration is the process of adjusting parameters until model outputs are sufficiently similar to observed values (where sufficiency is specified by the modeler). A measure to be used for judging the similarity between predicted and observed outputs is called an objective function (van Griensven and Bauwens, 2003).

Management decisions to protect estuaries are being made in the context of unprecedented environmental changes. We developed a management-oriented modeling system to look at interacting effects of climate and land use change on shallow subestuaries, which are biologically critical to larger estuarine systems such as the Chesapeake Bay in the United States.

The modeling system has linked watershed and subestuary models. We are using relatively simple watershed and subestuary models to facilitate application to large numbers of subestuaries and to facilitate automated calibration and analysis of model uncertainties. The watershed is considered to be an upstream boundary where multiple stressors, such as nutrients and sediments, are discharged to the subestuary. We applied a widely-used watershed model, the Generalized Watershed Loading Functions (GWLF) model (Haith and Shoemaker, 1987), to provide predictions of local watershed loading for driving the subestuary model and to predict how watershed loadings will respond to changes in land use and climate. One significant challenge is the watershed model calibration.

* Corresponding author. Address: State Key Laboratory of Urban and Regional Ecology, Research Center for Eco-Environmental Sciences, Chinese Academy of Science, Shuangqing Road 18, Beijing 100085, China. Tel./Fax: +86 10 62849428.
E-mail address: lix@si.edu (X. Li).

0022-1694/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhydrol.2009.11.003

The laborious nature of manual trial and error model calibration has motivated the development of automatic calibration techniques, including gradient-based methods like the Gauss–Levenberg–Marquardt method (Doherty and Johnston, 2003), population-evolution-based algorithms like Shuffled Complex Evolution method (Duan et al., 1992), and regionalization or spatial generalization (Lamb and Kay, 2004). Gradient-based methods are computationally efficient, but may identify parameters associated with a local minimum of the objective function rather than the global best parameter values (Rode et al., 2007). Global search methods such as SCE-UA (a shuffled complex evolution algorithm developed at The University of Arizona, Duan et al., 1992) avoid this problem, but global algorithms are complex and computationally expensive. A third calibration approach, called regionalization or spatial generalization, attempts to relate calibrated model parameters from many gauged watersheds to measurable watershed characteristics using regression or statistical methods, and then the relationships can be used to estimate model parameters for ungauged basins (Wagener and Wheater, 2006). This approach requires a large number of gauged watersheds with reliable data describing watershed characteristics, and sometimes there may not be any significant statistical relationships between a model parameter and watershed characteristics.

Multi-objective global optimization (Yapo et al., 1998; Gupta et al., 1999; Boyle et al., 2000; Vrugt et al., 2003; Fenicia et al., 2007) has been developed because any single-objective function, no matter how carefully chosen, may not adequately measure the ways in which a model fails to match the important characteristics of the observed data (Yapo et al., 1998). In multi-objective optimization, it is not possible to minimize all criteria simultaneously. Instead, a set of solutions is commonly found, which is called the Pareto set (Gupta et al., 1998). The Pareto set represents the set of solutions that can objectively be considered to be better than all other possible solutions. For the Pareto optimum solution, multiple objective functions can be aggregated into one statistical criterion (Madsen, 2000, 2003; van Griensven and Bauwens, 2003, 2005). Madsen (2000) proposed a balanced aggregated objective function using a Euclidian distance function in which all the objective functions are transformed to have about the same distance to the origin near the optimum. Based on Bayesian theory, Van Griensven and Meixner (2007) proposed a simpler method to create a global optimization criterion (GOC) with the shuffled complex evolution (SCE) algorithm which reduces the computational burden.

Most applications of multi-objective optimization methods to watershed model calibration are implemented at a single site. No matter how well a model is calibrated at one site, extrapolating the parameter values to other watersheds may be problematic because of the uniqueness of model parameters for individual watersheds (Beven, 2000). Using calibration results from multiple sites (Vandewiele and Elias, 1995) can be an effective way to reduce the uncertainty of extrapolating parameter estimates from a single-site calibration. As more stream measurements become available, multi-site calibration should be an important step towards applying global multi-objective optimization methods to broader watershed modeling applications for management of eutrophication in coastal waters.

In this paper, we apply the global multi-objective optimization method to multi-site calibration of a watershed model. The paper has two objectives. First, we present a new method that integrates multi-objective global optimization with multi-site calibration to better estimate model parameters for unmonitored sites with no calibration data. Second, we test the new method with the widely-used GWLF watershed model. We hypothesized that the multi-site approach would yield better parameter choices and better model predictive performance than one would get from extrapolating parameters from a single site. To evaluate this hypothesis, we quantify the benefit of multi-site calibration in improving model performance over single-site calibration and compare the benefit for different model predictions of flow, nitrogen and phosphorus. We apply the GWLF model to subwatersheds of the Rhode River basin of the Chesapeake Bay drainage. We aggregated multiple objective functions into one single global optimization criterion, calibrated water yield and nutrient loads at single and multiple sites, and evaluated the single-site and multi-site calibrations.

Methods

Study site

The study watersheds are part of the 3332 ha Rhode River drainage basin (Fig. 1). The Rhode River is a tidal tributary to the Chesapeake Bay in Maryland, United States of America (38°52′N, 76°32′W), and its watershed is in the mid-Atlantic Coastal Plain physiographic province. Since 1974, the Smithsonian Environmental Research Center (SERC) has measured stream discharges from subwatersheds within the basin using automated stream sampling stations that record streamflow and collect water samples for chemical analysis (Correll, 1977; Jordan et al., 1997a, 2003; Correll et al., 1999a,b). The five study subwatersheds considered in this paper range in size from 0.17 to 2.47 km^2 and differed in land cover (Table 1).

The GWLF model

GWLF (Haith and Shoemaker, 1987) is a lumped parameter watershed model that predicts streamflow and the loads of sediment, nitrogen (N) and phosphorus (P). GWLF uses runoff predictions from the curve number (CN) method, the Universal Soil Loss Equation (USLE, Wischmeier and Smith, 1978), and with average nutrient concentrations based on land use (Fig. 2). Watershed attributes are assumed to be homogenous across each watershed. For subsurface loading, the model uses a water balance approach. Daily water balances are computed for an unsaturated zone as well as a saturated sub-surface zone, where infiltration is computed as the difference between precipitation and snowmelt minus surface runoff plus evapotranspiration.

The GWLF model calculates nutrient loads using a number of empirical equations. All nutrients in GWLF computations are either dissolved or solid-phase, and total N or P are simply the sums of solid and dissolved forms. Monthly N discharge is calculated as

$$N_d = N_{dp} + N_{dr} + N_{dg} + N_{du} \quad (1)$$

$$N_s = N_{sp} + N_{sr} + N_{su} \quad (2)$$

where N_d is total dissolved N; and N_{dp}, N_{dr}, $N_{dg,}$ and N_{du} are the dissolved N contributions from point sources, rural runoff, ground water, and urban runoff, respectively. N_s is total solid N; and N_{sp}, $N_{sr,}$ and N_{su} are solid N from point source, rural runoff and urban runoff N discharges, respectively. Monthly P discharge is calculated in the same way as N. The GWLF model assumes that all the nutrient loads from point source, groundwater and septic systems are dissolved, and from urban nutrient load is solid.

Here we briefly describe the calculations for nitrogen loads (phosphorus loads are calculated in the same way as nitrogen) from rural runoff, urban runoff, groundwater sources and septic systems. More details about the calculations can be found in the model manual (Haith et al., 1996).

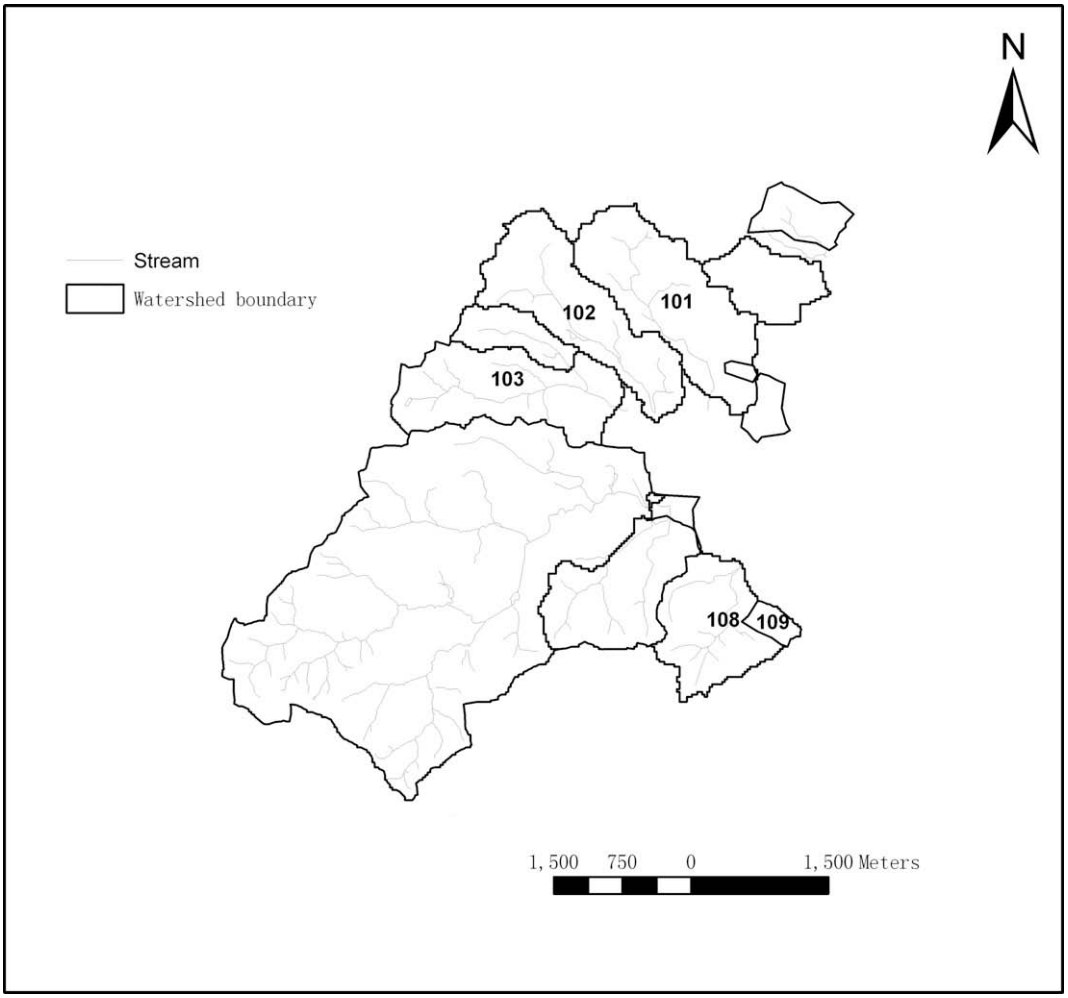

Fig. 1. Five study watersheds of Rhode River basin in Maryland, United States of America.

Table 1
Area and land cover (Homer et al., 2004) for study watersheds of Rhode River basin, Maryland, USA. Two minor land cover categories (wetlands and water) are not included.

Subwatershed	Land area (km^2)	Land cover percentage			
		Developed land	Cropland	Grassland	Forest
101	2.26	5.7	2.9	28.0	63.0
102	1.93	6.9	8.7	31.0	53.1
103	2.42	2.4	4.8	24.5	67.5
108	1.52	1.9	16.3	19.5	62.3
109	0.17	0.0	23.7	7.5	68.8

Rural runoff loads

The monthly dissolved nitrogen load from rural runoff is calculated as:

$$N_{dr} = 0.1 \cdot \sum_{k} \sum_{t=1}^{m} C_k \cdot Q_{kt} \cdot A_k \quad (3)$$

where C_k is nitrogen concentration in runoff from source area k (mg L^{-1}), Q_{kt} is runoff (cm) from source area k on day t, A_k is area (ha) of source area k, and m is number of days in the month.

The solid-phase nitrogen load from rural runoff N_{sr} is calculated from monthly sediment yield Y_m (Mg) and an average sediment nutrient concentration C_s (mg kg^{-1}):

$$N_{sr} = 0.001 \cdot C_s \cdot Y_m \quad (4)$$

Monthly sediment yield Y_m is estimated using the model developed by Haith (1985).

Urban runoff

The GWLF monthly nitrogen load from urban runoff is calculated as:

$$N_{su} = \sum_{k} \sum_{t=1}^{m} W_{kt} \cdot A_k \quad (5)$$

where m is number of days in the month and A_k is area (ha) of source area k. W_{kt} is a first-order wash off function:

$$W_{kt} = 1 - e^{-1.81 \cdot Q_{kt}} \quad (6)$$

where Q_{kt} is runoff (cm) from source area k on day t.

Groundwater sources

The monthly groundwater nitrogen discharge to the stream is calculated as:

$$N_{dg} = 0.1 \cdot C_g \cdot A \cdot \sum_{t=1}^{m} G_t \quad (7)$$

where m is the number of days in the month, C_g is the nitrogen concentration in groundwater (mg L^{-1}), A is area (ha) of source area k, and G_t is groundwater discharge (cm) to the stream on day t.

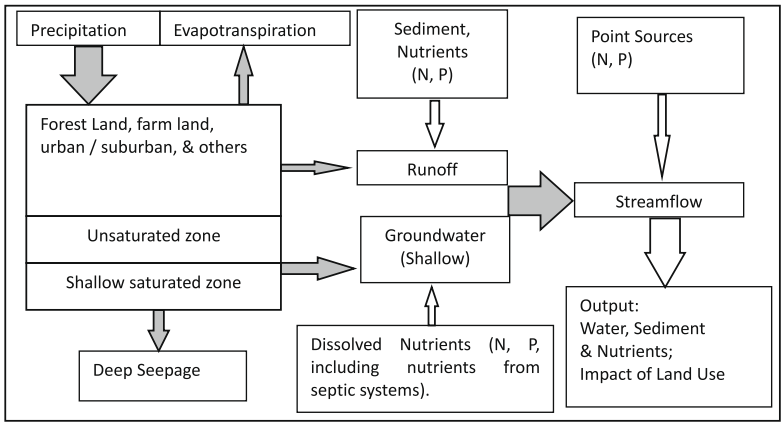

Fig. 2. Structure of the GWLF model. Shaded arrows are major fluxes of the hydrologic cycle.

Septic systems

The monthly nitrogen discharge from septic systems is the sum of discharges four septic system types:

$$N_{ds} = N_{ds1} + N_{ds2} + N_{ds3} + N_{ds4} \quad (8)$$

where N_{ds1}, N_{ds2}, N_{ds3} and N_{ds4} are the nitrogen discharges from normal, short-circuited, ponded, and direct discharge systems. These discharges are computed from per capita daily effluent loads and the populations served by each type of septic system.

Compared to other watershed-oriented water quality models, such as Soil and Water Assessment Tool (SWAT), Stormwater Management Model (SWMM), and Hydrological Simulation Program – Fortran (HSPF), GWLF has been classified by the US EPA as a model with "mid-level" complexity that includes most of the key mechanisms controlling nutrient fluxes within a watershed (US EPA, 1999). The data inputs for GWLF are generally easier to compile than many other models (Deliman et al., 1999). GWLF has been successfully applied to estimate water, total N, total P, and sediment discharges from ungauged watersheds (e.g., Lee et al., 2000, 2001; Schneiderman et al., 2002); to predict the effects of land use on downstream loads of nutrients and sediment (i.e., Howarth et al., 1991; Fisher et al., 2006); to forecast the effects of climate change on flow and nutrient discharges (i.e., Abler et al., 2002; Chang et al., 2001); and to calculate nutrient and sediment loads for TMDL (total maximum daily load) development and implementation (e.g., Borah et al., 2006).

Input data

We quantified watershed characteristics by using the ArcGIS (ESRI, Inc.) geographic information system (GIS) to intersect digital spatial data with watershed boundaries derived from digital topographic data and stream maps (Baker et al., 2006). Topographic variables, such as slope, were derived from the US National Elevation Data (1:24,000 DEM, Caruso, 1987; http://edc.usgs.gov/geodata/). Watershed land cover was calculated from the second generation of the US National Land Cover Dataset (30 m resolution; Homer et al., 2004; http://www.mrlc.gov), which was derived from Landsat 7 satellite imagery. We derived soil properties (permeability, soil erodibility and hydrologic group) from the SSURGO digital soil map for Anne Arundel County, Maryland (1:24,000, USDA-NRCS, 1995; http://www.ncgc.nrcs.usda.gov/products/datasets/ssurgo/).

We used daily meteorological data (precipitation and air temperature) from the SERC weather station (Correll et al., 1999a) to drive the model. Measurements of streamflow and nutrient loads, and sediment loads during 1980–2004 were from five monitoring stations operated by SERC (Correll et al., 1999a; Jordan et al., 2003). Human population data were from the US Census (www.census.gov). US Census data on households with and without public sewer service were used to estimate the number and density of septic systems.

Calibrated parameters

In applying GWLF in a Chesapeake coastal basin, Lee et al. (2000) grouped the parameters of GWLF into three classes based on their influences on model performance. Group 1 parameters must be adjusted to achieve adequate model performance. Group 2 parameters can improve the model performance, that is, they are useful but not essential. Group 3 parameters can be set to default values because they influence model output only for the first few months. In this paper, we calibrate only the hydrology and nutrient parameters in groups 1 and 2 (Table 2). The curve number and soil erosion parameters were not included in our calibrations. Curve number (CN) values were based on land use type (see CN values in Table 2). We derived soil erosion parameter values from mapped values of erosion parameters. These mapped values were derived from soil database, DEM and land cover maps (Boomer et al., 2008). The soil erosion parameters (soil erodibility K, the slope length-gradient LS, a crop/vegetation factor C, and a conservation practice factor P) are components of the Universal Soil Loss Equation (USLE, Wischmeier and Smith, 1978).

Model calibration procedure

We calibrated GWLF for monthly and annual water flow, TN load, and TP load. Three hydrological parameters were estimated by calibrating to monthly or annual flow: baseflow recession coefficient, deep seepage coefficient and unsaturated available water capacity (Table 2). Several other hydrological parameters, such as initial saturated and unsaturated water storage and initial snow, influence model output only for the first few months. We ran the model for 1 year prior to the time period of interest to avoid the effects from these initial conditions. Once the baseflow recession, deep seepage, and water capacity parameters were optimized, we estimated sediment and nutrient (TN and TP) parameters by comparing predicted monthly or annual sediment and nutrient loads with their measurements using multi-objective calibration at a single site or at multiple sites (see description below). We used a Latin-Hypercube sampling approach (McKay et al., 1979) to generate parameter sets for the optimization (below). The values for each parameter to be calibrated must be in a value range based on literature or site characteristics (Table 2). For a few hydrological parameters, we used empirical relationships to set the parameter

Table 2
GWLF parameters and their values or value ranges. Values ranges are from Haith et al. (1996) and Lee et al. (2000).

Parameter	Value or value range	Unit	Description
CN			
Forest	60	None	Curve number for computing surface runoff
Grassland	69		
Cropland	77		
Developed land	83		
Wetland	99		
Water	99		
Baseflow recession	0.01–0.2	cm day^{-1}	Groundwater recession coefficient
Seepage coefficient	0–0.08	cm day^{-1}	Deep seepage coefficient
UAWC	5–20	cm	Unsaturated available water capacity
Sediment delivery ratio	Site-dependent	None	The ratio of annual sediment yield to annual erosion
Sediment N	500–900	mg kg^{-1}	Solid-phase N from rural sources
Sediment P	120–393	mg kg^{-1}	Solid-phase P from rural sources
Groundwater N	0.1–19	mg L^{-1}	Dissolved N concentration in groundwater
Groundwater P	0.01–0.1	mg L^{-1}	Dissolved P concentration in groundwater
Agricultural runoff N	0–29	mg L^{-1}	Dissolved N concentration in runoff from agricultural land
Agricultural runoff P	0.1–5.1	mg L^{-1}	Dissolved P concentration in runoff from agricultural land
Grassland runoff N	0–29	mg L^{-1}	Dissolved N concentration in runoff from grassland
Grassland runoff P	0.1–5.1	mg L^{-1}	Dissolved P concentration in runoff from grassland
Forest runoff N	0.19–5	mg L^{-1}	Dissolved N concentration in runoff from forest
Forest runoff P	0.006–0.067	mg L^{-1}	Dissolved P concentration in runoff from forest
Urban build-up N	0.012–0.1	kg ha^{-1} d^{-1}	Solid N accumulation and wash off on urban surfaces
Urban build-up P	0.002–0.01	kg ha^{-1} d^{-1}	Solid P accumulation and wash off on urban surfaces

range. For example, we used a published relationship between the sediment delivery ratio and drainage area (see Haith and Shoemaker, 1987) to estimate the sediment delivery ratio, and then we assigned a range 50% below and 50% above that estimate. Similarly, Lee et al. (2000) observed a hyperbolic relationship between the groundwater recession coefficient and watershed drainage area, and we assigned the parameter value range to be 50% below and above the estimate from that relationship.

According to the principles of multi-objective calibration based on the concept of Pareto optimality (Gupta et al., 1998), the calibration problem can formally be stated as

$$\min F(\theta) = \{f_1(\theta), \ldots, f_m(\theta)\}\theta \subset \Theta \tag{9}$$

where the goal is to find values for θ (a set of model parameters) within the feasible parameter space Θ that minimize all of the objectives $f_i(\theta)$, $i = 1, 2, 3, \ldots, m$ simultaneously (where m is the number of objective functions). In this paper, the parameter space is defined by specifying lower and upper limits of each parameter. To solve the Pareto optimality, we used a global optimization criterion (GOC) approach that was proposed by Van Griensven and Meixner (2007). Several multi-objective functions are aggregated into one single objective, so that the multi-objective optimization problem becomes a single-objective problem:

$$\mathrm{GOC} = \sum_{i=1}^{m} \omega_i \cdot f_i(\theta) \tag{10}$$

where ω_i is the weight for objective function i ($i = 1, 2, 3, \ldots, m$) and $f_i(\theta)$ is the objective function. The weight is equal to the number of observations divided by the minimum of each objective function (van Griensven and Meixner, 2007), so the GOC can be calculated as

$$\mathrm{GOC} = \sum_{i=1}^{m} \frac{f_i(\theta) \cdot n}{f_i(\theta)_{\min}} \tag{11}$$

where $f_i(\theta)_{\min}$ are the minimum values to be found for each objective function, and n is the number of observations. Once we find

Table 3
Squared correlation coefficients (r^2) between GWLF model predictions and observations of monthly or annual streamflow, TN load, and TP load in Rhode River watersheds. The simulation time periods are October 1980–September 2004 for monthly streamflow, January 1985–December 2002 for monthly TN load and TP load. For annual results, water years numbered 1985–2001 extend from October 1 to September 30 of the following year. Statistical significance is evaluated with *t*-test and indicated as "not significant" ($p > 0.5$), "significant".

Variable	Watershed				
	101	102	103	108	109
Monthly streamflow	0.71**	0.78**	0.75**	0.73**	0.74**
Monthly TN	0.61**	0.61**	0.65**	0.55**	0.60**
Monthly TP	0.47*	0.42	0.68**	0.47*	0.41
Annual streamflow	0.88**	0.91**	0.84**	0.87**	0.80*
Annual TN	0.80*	0.69*	0.67*	0.86**	0.74*
Annual TP	0.70*	0.48	0.79*	0.56	0.47

* $0.01 < p < 0.05$ or "highly significant".
** $p < 0.01$.

Table 4
Values of objective functions calculated between GWLF model predictions and observations of monthly or annual streamflow, TN load, and TP load in Rhode River watersheds: cumulative error (CE), root mean squared error (RMSE), and the index of agreement (d). The simulation time periods were the same as in Table 3.

Watershed	101	102	103	108	109
Monthly streamflow					
CE (cm)	305	286	297	322	60
RMSE	1.56	1.39	1.44	1.62	1.20
d	0.86	0.88	0.86	0.87	0.85
Monthly TN					
CE (kg N)	3770	4928	6759	6848	1058
RMSE	33.06	36.02	51.02	60.48	11.95
d	0.82	0.84	0.84	0.81	0.82
Monthly TP					
CE (kg P)	1887	2057	22996	3360	691
RMSE	14.98	18.18	151.64	48.32	8.07
d	0.77	0.79	0.85	0.80	0.79
Annual streamflow					
CE (cm)	172	140	178	186	95
RMSE	6.97	6.06	7.78	8.07	4.33
d	0.9	0.92	0.88	0.9	0.89
Annual total N					
CE (kg N)	1517	2100	3092	3340	544
RMSE	140.24	167.21	216.69	235.76	44.27
d	0.86	0.85	0.85	0.88	0.86
Annual total P					
CE (kg P)	828	1196	4144	2138	353
RMSE	66.5	96.11	416.84	195.54	28.37
d	0.86	0.78	0.88	0.81	0.8

the minimum GOC in Eq. (11), we can identify the best parameter estimates using this multi-objective optimization.

In this study, we used three objective functions to calculate GOC in Eq. (11): CE, RMSE and d. These objective functions measure model error or residual. Their definitions are

$$\mathrm{RMSE} = \sqrt{\frac{1}{t}\sum_{i=1}^{t}(O_i - P_i)^2} \tag{12}$$

$$\mathrm{CE} = \sum_{i=1}^{t}|O_i - P_i| \text{ and} \tag{13}$$

$$d = 1 - \frac{\sum_{i=1}^{t}(O_i - P_i)^2}{\sum_{i=1}^{t}(|P_i - \overline{O}| + |O_i - \overline{O}|)^2} \tag{14}$$

Root mean squared error (RMSE) measures the generalized standard deviation between observed and model predicted values. Cumulative error (CE) is an important error measure for an entire prediction time period, and d is the index of agreement for the entire time period (Willmott, 1981). In Eqs. (12)–(14), t is the total number of time steps in the calibration period, O is the observed value, and P is the predicted value.

We normalized values of the three objective functions to a scale of 0–1 so that we could calculate the GOC values at the same magnitude with different objective functions. Each objective function was divided by the maximum value among objective functions of the five single-site calibrations to normalize its value. The three normalized functions thus received the values with same

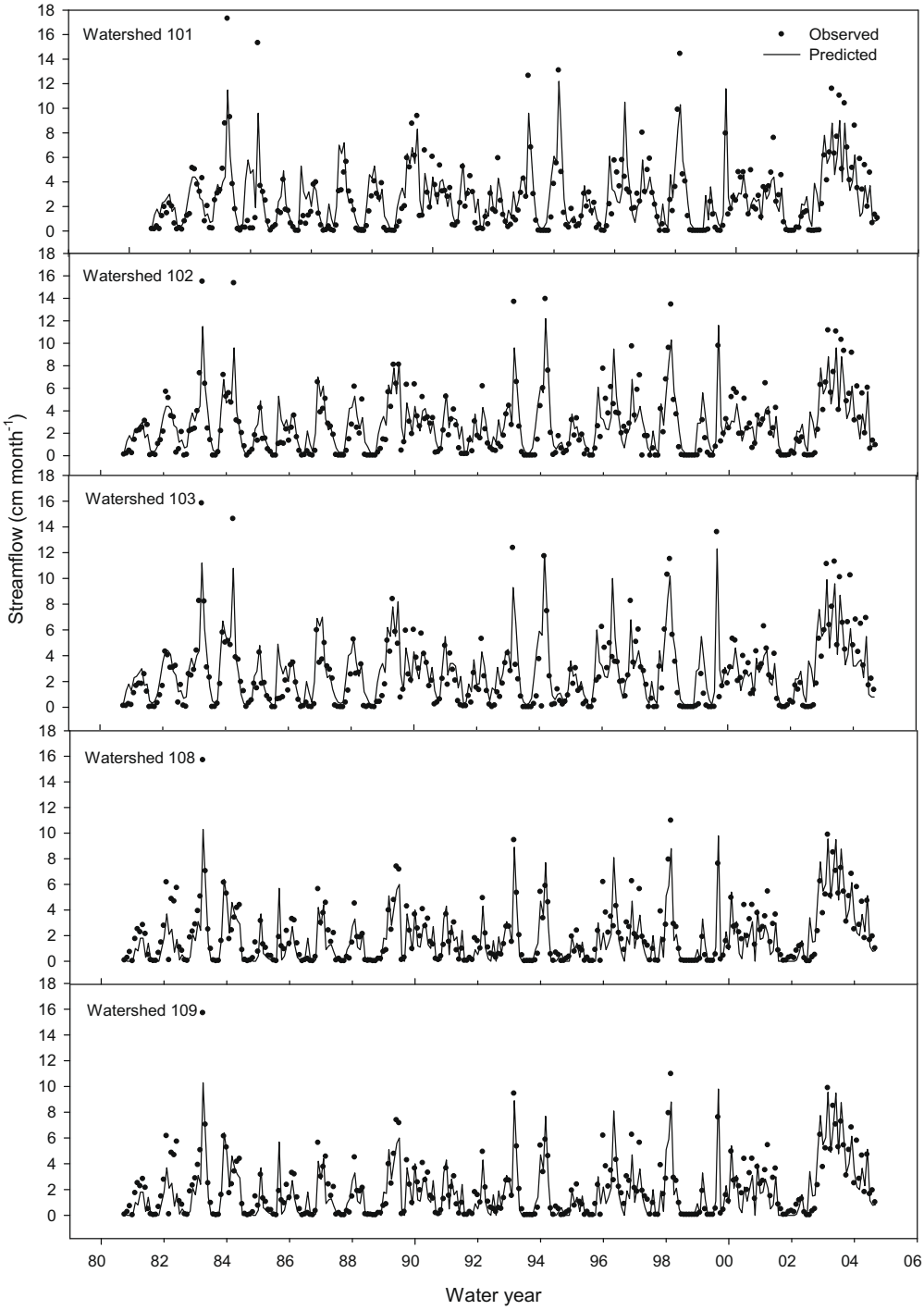

Fig. 3. Observed and predicted monthly streamflow (October-1980–September-2004) for five Rhode River watersheds. The predictions were based on single-site calibrations.

magnitude. We actually used $1 - d$ rather than d in the GOC because minimizing the GOC then tends to minimize $1 - d$ and thus maximize agreement (d). We also actually used relative CE (see Eq. (18) below) rather than absolute CE because relative CE remains comparable even when the duration of simulations is changed. For each calibration site, we estimated the parameters in Table 2 through minimizing GOC values in Eq. (11). We separately compared monthly or annual predictions of streamflow, TN loads, and TP loads with their observed values to identify the best parameter sets with minimum GOC values at both monthly and annual time scales. When we estimated nutrient concentration parameters, we required that agricultural runoff N (or P) > Grassland runoff N (or P) > Forest runoff N (or P). When the parameter values yielding a minimum GOC did not meet this constraint, we eliminated unreasonable parameter sets by choosing the parameter set with next lowest GOC that did satisfy the constraint.

For a single watershed, the best parameter set can be identified by optimizing GOC, but the parameter estimates will vary among watersheds. To reduce the possible bias in parameter estimates due to the arbitrary selection of a single calibration watershed, we used multi-site averaging to estimate the final parameter values. We tried averaging across calibration sites using two calculations. The equal weight average (hereafter called un-weighted average) is the simple average for each parameter from independent calibrations at multiple sites. For the multi-site weighted average (hereafter called weighted average), we assigned a higher weight to a site for which the difference between predicted and observed values is smaller. The weight w was the inverse of residual variance calculated as

$$w_j = \frac{1}{\sigma_j^2} \tag{15}$$

where σ_j is standard deviation of model residuals (predictions minus measured values) at site j ($j = 1, 2, 3, \ldots, n$). Then the weighted average parameter value is

$$\overline{\theta} = \sum_{j=1}^{n} \frac{w_j}{\sum_{j=1}^{n} w_j} \theta_j \tag{16}$$

where θ_j is the best estimate for parameter set based on Eq. (9) at site j ($j = 1, 2, 3, \ldots, n$). In Eq. (16), we normalized the weights.

Model validation and evaluation of model performance

Model validation using independent watersheds

We used data from an independent watershed to validate parameters estimated from various calibration approaches. We used any four of the five Rhode River watersheds for model

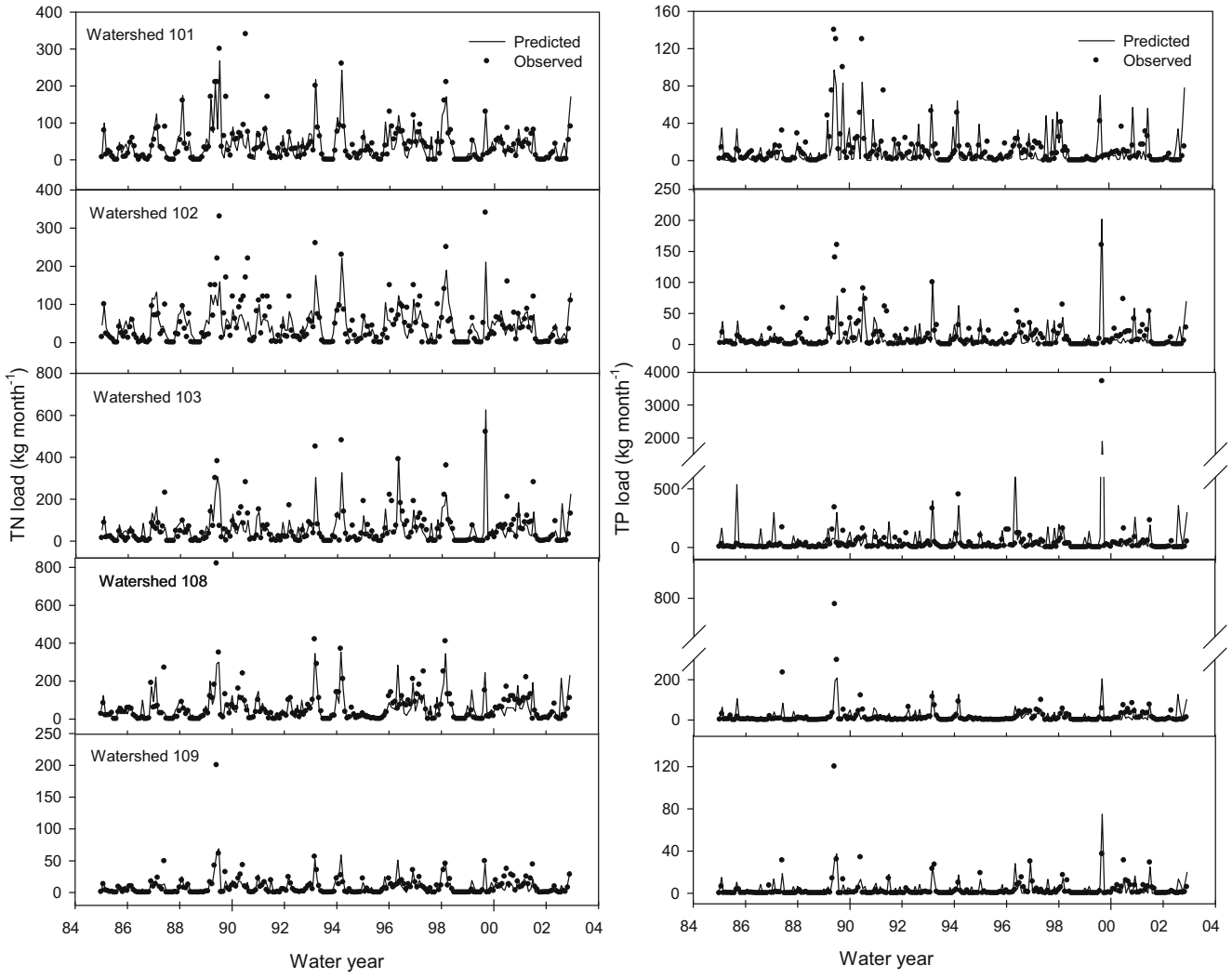

Fig. 4. Observed and predicted monthly discharges of TN and TP (January-1985–December-2002) for five Rhode River watersheds. The predictions were based on single-site calibrations.

calibrations and the remaining watershed for validation. We then changed the combination of the four calibration watersheds and the remaining one validation watershed, and finally implemented model calibrations for all five different combinations. The combinations for calibration and validation are: watersheds 101, 102, 103 and 108 for model calibration, and watershed 109 for validation; Watersheds 101, 102, 103 and 109 for calibration, and watershed 108 for validation; and so on.

Multi-site versus single-site calibration and validation

For each validation watershed, we compared model performance with different six parameter sets. There were four parameter sets calibrated from four single sites. The fifth parameter set was an un-weighted average of the four single-site sets, and the sixth parameter set was a weighted average of the four single-site sets (weighted by inverse residual variance, Eq. (15)). Thus, we had two multi-site calibration scenarios: one from un-weighted averaging and another from weighted averaging.

To quantify model performance, we calculated the squared correlation coefficient, r^2, between predicted and observed values (also called the coefficient of determination) and relative cumulative error (CE). r^2 is a widely-used and clearly interpretable goodness-of-fit indicator, while CE measures the error from an entire prediction time period. We used relative CE rather than absolute CE because relative CE remains comparable even when the duration of simulations is changed. The coefficient of determination r^2 is calculated as:

$$r^2 = \left(\frac{\sum_{i=1}^{t}(O_i - \overline{O}) \cdot (P_i - \overline{P})}{\sqrt{\sum_{i=1}^{t}(O_i - \overline{O})^2} \cdot \sqrt{\sum_{i=1}^{t}(P_i - \overline{P})^2}} \right)^2 \quad (17)$$

with O observed and P predicted values. CE is defined as the sum of absolute values of the difference between predictions and observations. The relative CE (%CE) is the ratio of absolute CE to the sum of

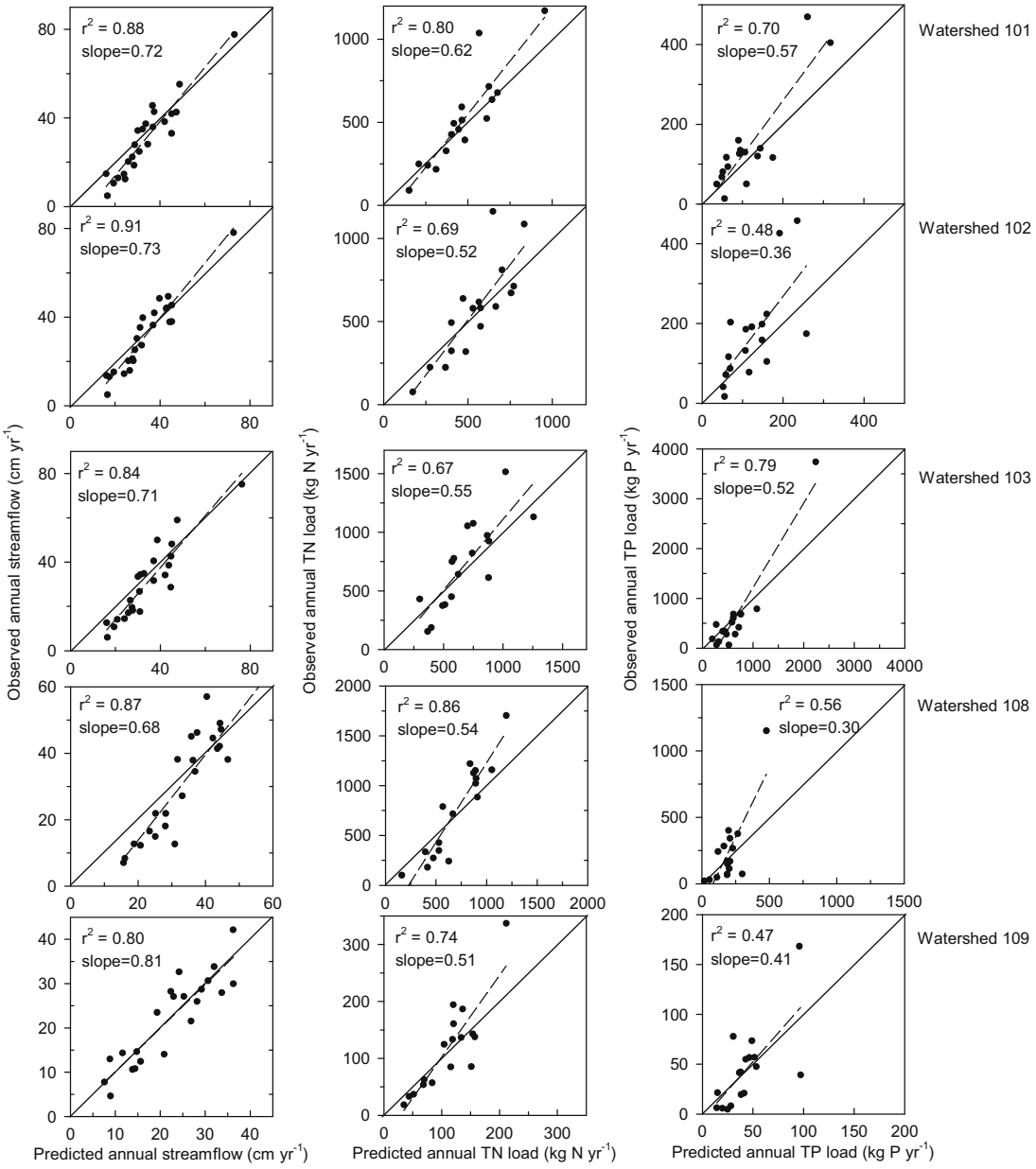

Fig. 5. Correlations between observed and GWLF predicted annual streamflow (1980–2003), TN load, and TP load (1985–2001) for five Rhode River watersheds. Water years numbered 1985–2001 extended from October 1 to September 30 of the following year. The predictions were based on single-site calibrations.

observed values over the entire study time period expressed as a percentage. The relative CE is calculated as:

$$\%CE = \frac{CE}{\sum_{i=1}^{t} O_i} \times 100 \quad (18)$$

where CE is calculated by Eq. (13).

Results

Single-site calibration

GWLF simulations calibrated with the multi-objective optimization approach at single sites showed various levels of success among responses and among the five calibration watersheds (Tables 3 and 4, Figs. 3–5). Overall, model predictions for monthly or annual streamflow were better than for TN load, and TN load predictions were better than TP load predictions. Simulations of monthly and annual streamflow and TN in all five watersheds were successful, with r^2 values of 0.71–0.78 for monthly streamflow, 0.55–0.65 for monthly TN, 0.80–0.91 for annual streamflow, 0.67–0.86 for annual TN. Model performance for monthly and annual TP loads was generally not as good as streamflow or TN loads, as shown by r^2 values of 0.41–0.68 for monthly TP loads, 0.47–0.79 for annual TP loads (Table 3).

Parameter estimates from multi-objective optimization at five single calibration sites varied among sites (Table 5). Some parameters varied widely, especially the site-dependent parameters identified by Lee et al. (2000). For example, the baseflow recession varied from 0.014 to 0.044 (a 67% difference between the highest and lowest values), and the sediment delivery ratio varied from 0.16 to 0.31 (a 48% difference between the highest and lowest values). Nutrient parameters varied less among watersheds than did the hydrological parameters. For example, the concentration of N in agricultural runoff N varied from 5.8 to 8.0 mg N L^{-1} (a 28% difference between the highest and lowest values).

Multi-site average parameters

The weights from different sites (Eq. (15)) used in estimating weighted average parameter values varied widely among watersheds and among model outputs in the same watershed (Table 6). The variation of weights for monthly streamflow was relatively lower than for monthly TN or TP load. For example, the weights for monthly streamflow ranged from 0.21 to 0.29 when watersheds 101, 102, 103 and 108 were selected as calibration sites and from 0.18 to 0.34 when watersheds 102, 103, 108 and 109 were selected as calibration sites. The weights for monthly TP or TP load were more variable. For example, the weights ranged from 0.12 to 0.39 for monthly TN loads and 0.01 to 0.56 for monthly TP loads when watersheds 101, 102, 103 and 108 were selected as calibration sites. The weight ranges were 0.09–0.53 for monthly TN loads and 0.01–0.51 when watershed 102, 103, 108 and 109 were selected as calibration sites.

The weighted average parameter estimates (Eq. (16)) improved model prediction in comparison to the average of single-site parameters for monthly and annual streamflow, TN and TP load in validation watersheds, but the un-weighted average parameters did not improve model performance (Table 7). In comparison to the average of r^2 values from single-site calibrations from four calibration watersheds, calibrations with weighted averaging increased squared correlation coefficients between predictions and observations in validation watersheds by 0.06–0.10 for monthly streamflow, 0.03–0.08 for monthly TN loads, and 0.02–0.05 for monthly TP loads. Similar improvements were found for predictions of annual streamflow, TN load, and TP load. Model prediction performance for each calibration scenario varied among validation watersheds. For example, squared correlation coefficients for monthly streamflow using the average of single-site parameters for calibration ranged from 0.60 to 0.68. The variations of model performance among validation watersheds for monthly and annual TN and TP loads were generally broader. For example, r^2 ranged 0.63–0.74 for annual TN loads with weighted average calibration.

The relative cumulative error (%CE) from validation simulations using weighted average parameters was lower than the %CE from simulations using un-weighted average parameters and also lower than the average of the four %CE values from single-site calibration (Table 8). For example, in validation watershed 101, the %CE for monthly streamflow was 33.8% with weighted average parameters and 41.6% with un-weighted average parameters. The average of the cumulative errors for the four single-site calibrations was 41.2%. However, the weighted calibration did not always produce a smaller %CE than every single-site calibration. For monthly TN load or TP load predictions in each validation watershed, there was at least one single-site calibration that produced a smaller %CE than did the weighted average. For example, at validation watershed 101, %CE for monthly TN loads was 40.8% for using calibration with weighted average, but 38.2% and 35.6% for the single-site

Table 5
GWLF parameter estimates from calibration of each single watershed of Rhode River basin using multi-o objective optimization approach.

Parameter	Calibration watershed				
	101	102	103	108	109
Baseflow recession	0.034	0.020	0.044	0.027	0.014
Seepage coefficient	0.016	0.016	0.024	0.020	0.015
UAWC	10.8	9.1	9.1	9.98	11.8
Sediment delivery ratio	0.31	0.30	0.31	0.27	0.16
Sediment N (mg kg^{-1})	631	688	601	703	719
Sediment P (mg kg^{-1})	172	172	187	248	253
Groundwater N (mg L^{-1})	1.5	3.7	1.3	3.3	3.7
Groundwater P (mg L^{-1})	0.082	0.082	0.073	0.085	0.087
Agricultural runoff N (mg L^{-1})	6.5	5.8	8.0	7.2	8.0
Agricultural runoff P (mg L^{-1})	1.3	1.0	1.6	1.8	2.0
Grassland runoff N (mg L^{-1})	4.3	2.5	5.8	5.0	4.3
Grassland run off P (mg L^{-1})	0.78	0.52	1.03	0.65	0.78
Forest runoff N (mg L^{-1})	1.3	1.6	2.3	1.9	2.0
Forest runoff P (mg L^{-1})	0.020	0.023	0.034	0.027	0.031
Urban build-up N (kg ha^{-1} day^{-1})	0.059	0.071	0.048	0.038	0.027
Urban build-up P (kg ha^{-1} day^{-1})	0.0064	0.0043	0.0045	0.0045	0.0041

Table 6
Weights for each watershed in different calibration combinations used for multi-site weighted average calibrations for monthly streamflow, TN and TP. The letter "V" represents the validation watershed.

Variable	Watershed				
	101	102	103	108	109
Streamflow	0.23	0.29	0.27	0.21	V
TN	0.39	0.33	0.16	0.12	V
TP	0.56	0.38	0.01	0.05	V
Streamflow	0.22	0.24	0.21	V	0.33
TN	0.19	0.16	0.08	V	0.57
TP	0.31	0.21	0.01	V	0.47
Streamflow	0.21	0.26	V	0.19	0.35
TN	0.20	0.16	V	0.06	0.58
TP	0.30	0.21	V	0.03	0.46
Streamflow	0.21	V	0.24	0.19	0.36
TN	0.28	V	0.12	0.08	0.52
TP	0.48	V	0.01	0.05	0.46
Streamflow	V	0.25	0.23	0.18	0.34
TN	V	0.25	0.13	0.09	0.53
TP	V	0.42	0.01	0.06	0.51

Table 7
Squared correlation coefficients (r^2) between GWLF model predictions and observations of monthly or annual streamflow, TN load, and TP load at validation watersheds. Four of five Rhode River watersheds were used for calibrations, and the remaining watershed for validation. The model predictions are based on parameter estimates from different calibration scenarios. The column labeled "single site mean" gives the mean value of r^2 for each validation site based on parameters from four single-site calibrations. The columns labeled "un-weighted average" and "multi-site weighted average" give r^2 values for each validation site based on parameters from multi-site un-weighted averaging and multi-site weighted averaging calibrations. Statistical significance is evaluated with t-test and indicated as "not significant" ($p < 0.5$), "significant".

Variable	Validation site	Calibration scenario					
		Single site Mean	Un-weighted Average	Weighted Average	Single site Mean	Un-weighted Average	Weighted Average
		Monthly			Annual		
Streamflow	101	0.60**	0.59**	0.70**	0.68*	0.67*	0.81**
TN	101	0.54*	0.54*	0.60**	0.65*	0.65*	0.74*
TP	101	0.47*	0.46*	0.49*	0.58	0.59	0.63*
Streamflow	102	0.68**	0.70**	0.77**	0.73*	0.75*	0.85**
TN	102	0.55**	0.54*	0.61**	0.57	0.56	0.64*
TP	102	0.37	0.39	0.42*	0.44	0.43	0.47
Streamflow	103	0.67**	0.68**	0.73**	0.69*	0.71*	0.77*
TN	103	0.60**	0.62**	0.64**	0.58	0.60	0.63*
TP	103	0.61**	0.61**	0.65**	0.69*	0.69*	0.72*
Streamflow	104	0.65**	0.67**	0.73**	0.75*	0.76*	0.82**
TN	104	0.52*	0.53*	0.55**	0.70*	0.71*	0.73*
TP	104	0.36	0.37	0.38	0.48	0.51	0.53
Streamflow	105	0.67**	0.69**	0.74**	0.73*	0.74*	0.80*
TN	105	0.52*	0.53*	0.60**	0.59	0.70*	0.72*
TP	105	0.45*	0.48*	0.47*	0.43	0.42	0.46

* $0.01 \leqslant p < 0.05$ or "highly significant".
** $p < 0.01$.

calibration from watersheds 102 and 109, respectively. With weighted average parameters for our five validation watersheds, the model prediction error was reduced by 3.5–7.4% for monthly streamflow, 3.2–6.3% for TN loads, and 4.3–5.9% for TP loads in comparison to average %CE from single-site calibrations (Table 9).

Weighted average parameters generally reduced prediction error compared to parameters from a single site (Table 8). For example, in predicting TN load, weighted average parameters produced lower %CE than two of four single-site calibrations at validation watershed 101 and three of four single-site calibrations at validation watershed 102, 103, 108, or 109. However, choosing the weighted average parameters instead of single-site parameters also decreased error more often and to a greater extent than it increased error (Table 9).

Table 8
Relative cumulative errors (%CE) for simulations using different parameters sets to predict monthly streamflow, TN load, and TP load. Columns giving results from single-site calibrations are labeled with the site number. The column labeled "Mean" gives the mean value of %CE for the four single-site calibrations. The last two columns give %CE values for parameters from multi-site un-weighted averaging and multi-site weighted averaging. In each row, four watersheds were used for calibration and the remaining one (marked with "V") for validation.

Watershed	Calibration scenario							
	101	102	103	108	109	Mean	Un-weighted average	Weighted average
Streamflow	V	39.3	42.0	44.0	39.5	41.2	41.6	33.8
TN	V	38.2	55.4	57.3	35.6	46.6	46.4	40.8
TP	V	81.2	87.7	98.9	78.0	86.5	85.9	81.3
Streamflow	35.6	V	41.6	42.4	40.4	40.0	38.7	34.1
TN	42.8	V	50.7	54.4	50.0	49.5	50.2	43.2
TP	76.3	V	96.7	93.3	78.1	86.1	85.6	80.3
Streamflow	40.7	42.8	V	43.5	37.0	41.0	40.2	37.0
TN	44.9	49.8	V	55.0	45.3	48.8	47.5	45.1
TP	76.3	92.2	V	96.9	85.3	87.7	88.0	81.8
Streamflow	43.4	42.9	45.1	V	36.9	42.1	40.9	36.8
TN	49.8	54.8	53.2	V	43.2	50.2	49.1	46.7
TP	96.1	93.5	83.7	V	78.6	88.0	85.7	82.5
Streamflow	41.7	39.2	40.1	42.5	V	40.9	42.5	37.4
TN	44.8	47.7	53.5	53.7	V	49.9	52.1	46.7
TP	76.4	78.9	95.8	93.3	V	86.1	87.8	81.8

Table 9
Difference (range and mean) of relative cumulative errors (%CE, see Table 8) for simulations between using multi-site weighted averaging (or un-weighted averaging) parameters and using a single-site calibration. A negative value indicates a decrease of %CE from using multi-site calibration compared to the single-site calibration and a positive value indicates an increase. A mean value indicates the mean of the differences in %CE.

Validation watershed	Un-weighted Average		Weighted Average	
	Range	Mean	Range	Mean
101				
Streamflow	(−2.7) ~ 2.4	0.4	(−10.5) ~ (−5.4)	−7.4
TN	(−12.1) ~ 12.0	−0.2	(−17.7) ~ 6.4	−5.8
TP	(−13.0) ~ 7.9	−0.5	(−17.6) ~ 3.3)	−5.2
102				
Streamflow	(−3.8) ~ 3.4	−1.3	(−8.5) ~ (−1.2)	−5.9
TN	(−4.7) ~ 8.1	0.8	(−11.7) ~ 1.1	−6.3
TP	(−11.1) ~ 9.3	−0.5	(−16.4) ~ 4.0	−5.8
103				
Streamflow	(−3.5) ~ 3.5	−0.8	(−6.7) ~ 0.3	−4
TN	(−8.2) ~ 3.0	−1.3	(−10.6)~0.6	−3.7
TP	(−8.9) ~ 11.7	0.3	(−15.1) ~ 5.5	−5.9
108				
Streamflow	(−4.4) ~ 4.3	−1.2	(−8.5) ~ 0.2	−5.3
TN	(−6.2) ~ 6.6	−1.2	(−8.6) ~ 4.2	−3.6
TP	(−10.4) ~ 7.1	−2.3	(−13.6) ~ 3.9	−5.5
109				
Streamflow	(−0.1) ~ 3.3	1.6	(−5.2) ~ (−1.8)	−3.5
TN	(−1.9) ~ 7.9	2.2	(−7.3) ~ 2.5	−3.2
TP	(−8.0) ~ 11.4	1.7	(−14.0) ~ 5.4	−4.3

Discussion

Parameter estimation with the multi-objective method can improve model prediction performance compared to a single objective calibration (e.g., Gupta et al., 1998; Yapo et al., 1998; Bekele and Nicklow, 2007). However, some hydrological and nutrient

parameters rely heavily on local watershed characteristics (such as soil physical and hydraulic features, nutrient concentrations in water flow, and nutrients in surface soil layers), which can vary widely among different locations. That variability complicates the task of extrapolating model parameters from calibration sites to other ungauged basins. We developed and tested a new method that integrates multi-objective optimization with multi-site calibration to better estimate model parameters for ungauged watersheds. We found that the new method provided a significant boost in model performance for validation watersheds (Tables 7 and 8).

Parameters from muli-site averaging always worked better when the individual site calibrations were weighted by their prediction errors using Eq. (9) (Tables 7 and 8). With such weighting, sites with better calibrations (lower prediction errors) contributed more to the average parameters than did sites poorer calibrations (higher prediction errors). Simulations with the weighted average parameters had lower cumulative errors than simulations using simple un-weighted average parameters (Table 8).

When selecting model parameters for unmonitored watersheds, parameters from the weighted averaging method are far more likely to minimize prediction bias than are parameters from single-site calibrations. This biases reduction occurs even though parameters from a single-site calibration might sometimes have lower prediction errors. Our study offers 15 test cases (5 validation watersheds $\times$ 3 materials) for comparing cumulative errors from weighted average parameters and single-site parameters (Table 8). Across these 15 cases, there are 60 possible choices of single-site parameter sets (15 cases $\times$ 4 possible calibration watersheds). Only 14 of these 60 choices of a single-site calibration (23%) would yield a cumulative error lower that the weighted average parameter set (Table 8). In other words, there is a better than 75% probability that the weighted average parameter set will have a lower cumulative error for an unmonitored watershed than one would get from choosing one of the single-site calibrations (Table 8).

It is not known whether or not model prediction performance will be further improved if the number of calibration watersheds is increased. Future studies should investigate the appropriate number of watersheds to use with the weighted average method. Future studies should also investigate the need for geographic proximity of the calibration sites (Vandewiele and Elias, 1995; Oudin et al., 2008). For example, in the Chesapeake region, Piedmont watersheds release more N than Coastal Plain watersheds with similar land cover proportions (Jordan et al., 1997b, 2003). Averaging watersheds from Piedmont and Coastal Plain could cause great variation in parameter estimates and lead to higher prediction error.

An improved calibration method does not help to reduce prediction errors that are caused by errors in model structure. For example, the poor model performance for TP loads may be caused by the inadequacy of model prediction in soil erosion with a curve number based method in Chesapeake region (Boomer et al., 2008). GWLF failed to predict unusually high discharges of monthly streamflow, TN and TP. For example, during April of 1983, GWLF underestimated monthly streamflow by 26–35% in our four calibration watersheds; during September of 1999, the remnants of hurricane Floyd passed through the mid-Atlantic area and deposited almost 23 cm of rain at Rhode River basin over an 18 h period, and GWLF underestimated monthly TN loads by 1–38%, and monthly TP loads by 26–102%. Poor model performance in extreme events could be due to poor ability to predict storm flow (Lee et al., 2000). The calibration for streamflow does not guarantee a good partitioning between base flow and stormflow. The uncertainty of flow partitioning affects the proportion of TN or TP discharges among surface runoff and groundwater recharge, as well as soil erosion process.

Conclusions

We developed a watershed calibration framework using a multi-objective optimization criterion and weighted average parameters from multiple neighboring watersheds. The new calibration method was tested using five watersheds at the Rhode River basin, in Maryland, United States of America. As we hypothesized, the multi-site calibration approach produced better parameter choices and better model predictive performance than did parameters extrapolated from a single site. The parameters derived from a multi-site weighted average approach reduced the relative cumulative error by 3.5–7.4% for monthly streamflow, 3.2–6.3% for monthly TN loads, and 4.3–5.9% for monthly TP loads in comparison to the average of single-site calibrations.

Acknowledgements

This work was initially supported by an award from the US EPA Science To Achieve Results (STAR) Program on Regional-Scale Modeling of Multiple Stressors to Aquatic Ecosystems (Grant No. RD83087801). Additional fund to support the first author was provided by State Key Development Program of Basic Research of China (Grant No. 2009CB421104) and State Key Laboratory Program (Grant No. SKLURE2008-1-05) of Urban and Regional Ecology Laboratory of Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences.

References

Abler, D., Shortle, J., Carmichael, J., Horan, R., 2002. Climatic change, agriculture, and water quality. Climatic Change 55, 339–359.
Baker, M.E., Weller, D.E., Jordan, T.E., 2006. Mapping watershed boundaries using digital elevation data: implications for landcover analysis of nutrient discharge. Photogramm. Eng. Rem. S. 72, 159–168.
Bekele, E.G., Nicklow, J.W., 2007. Multi-objective automatic calibration of SWAT using NSGA-II. J. Hydrol. 341, 165–176.
Beven, K.J., 2000. Uniqueness of place and process representations in hydrological modelling. Hydrol. Earth. Syst. Sci. 4 (2), 203–213.
Boomer, K.B., Weller, D.E., Jordan, T.E., 2008. Empirical models based on the universal soil loss equation fail to predict sediment discharges from Chesapeake Bay catchments. J. Environ. Qual. 37, 79–89.
Borah, D., Yagow, G., Saleh, A., Barnes, P.L., Rosenthal, W., Krug, E.C., Hauck, L.M., 2006. Sediment and nutrient modeling for TMDL development and implementation. T. ASAE 49, 967–986.
Boyle, D.P., Gupta, H.V., Sorooshian, S., 2000. Toward improved calibration of hydrologic models: combining the strength of manual and automatic methods. Water Resour. Res. 36, 3663–3674.
Caruso, V., 1987. Standards for digital elevation models. In: ASPRS-ACSM Annual Convention. American Society for Photogrammetry and Remote Sensing and American Congress on Surveying and Mapping, Falls Church, Virginia, vol. 4, pp. 159–166.
Chang, H., Evans, B., Easterling, D., 2001. Effects of climate change on streamflow and nutrient loading. J. Am. Water Resour. Assoc. 37, 973–986.
Correll, D.L., 1977. An overview of the Rhode River watershed research program. In: Correll, D.L. (Ed.), Watershed Research in Eastern North America – A Workshop to Compare Results. Smithsonian Institution, Edgewater, MD, pp. 105–123.
Correll, D.L., Jordan, T.E., Weller, D.E., 1999a. Effects of precipitation and air temperature on nitrogen discharges from Rhode River watersheds. Water Air Soil Poll. 115.
Correll, D.L., Jordan, T.E., Weller, D.E., 1999b. Effects of precipitation and air temperature on phosphorus fluxes from Rhode River watersheds. J. Environ. Qual. 28, 144–154.
Deliman, P.N., Glick, R.H., Ruiz, C.E., 1999. Review of Watershed Water Quality Models. US Army Corps of Engineers, Tech. Rep. W-99-1.
Doherty, J., Johnston, J.M., 2003. Methodologies for calibration and predictive analysis of a watershed model. J. Am. Water Resour. Assoc. 39 (2), 251–265.
Duan, Q., Sorooshian, S., Gupta, V.K., 1992. Effective and efficient global optimization for conceptual rainfall–runoff models. Water Resour. Res. 28, 1015–1031.
Fenicia, F., Savenije, H.H.G., Matgen, P., Pfister, L., 2007. A comparison of alternative multiobjective calibration strategies for hydrological modeling. Water Resour. Res. 43, W03434. doi:10.1029/2006WR005098.
Fisher, T.R., Hagy III, J.D., Boynton, W.R., Williams, M.R., 2006. Cultural eutrophication in the Choptank and Patuxent River estuaries of Chesapeake Bay. Limnol. Oceanogr. 51, 435–447.

Gray, J.S., 1992. Eutrophication in the sea. In: Colombo, G., Ferrari, I., Cecchereli, V.U., Rossi, R. (Eds.), Marine Eutrophication and Population Dynamics. Olsen & Olsen, Fredensborg, pp. 3–15.

Gupta, H.V., Sorooshian, S., Yapo, P.O., 1998. Toward improved calibration of hydrologic models: multiple and noncommensurable measures of information. Water Resour. Res. 34, 751–763.

Gupta, V.K., Bastidas, L.A., Sorooshian, S., Shuttleworth, W.J., Yang, Z.-L., 1999. Parameter estimation of a land surface scheme using multicriteria methods. J. Geophys. Res. 104, 19491–19503.

Haith, D.A., 1985. An event-based procedure for estimating monthly sediment yields. Trans. ASAE 28 (6), 1916–1920.

Haith, D.A., Shoemaker, L.L., 1987. Generalized watershed loading functions for streamflow nutrients. Water Resour. Bull. 23, 471–478.

Haith, D.A., Mandel, R., Wu, R.S., 1996. Generalized Watershed Loading Functions. Version 2.0 User's Manual.

Homer, C., Huang, C., Yang, L., Wylie, B., Coan, M., 2004. Development of a 2001 national land-cover database for the United States. Photogramm. Eng. Rem. S. 70, 829–840.

Howarth, R.W., Fruci, J.R., Sherman, D., 1991. Inputs of sediment and carbon to an estuarine ecosystem: influence of land use. Ecol. Appl. 1 (1), 27–39.

Howarth, R., Anderson, D., Cloern, J., Elfring, C., Hopkinson, C., Lapointe, B., Malone, T., Marcus, N., McGlathery, K., Sharpley, A., Walker, D., 2000. Nutrient pollution of coastal rivers, bays, and seas. Issues Ecol 7, 1–15.

Jordan, T.J., Correll, D.L., Weller, D.E., 1997a. Effects of agriculture on discharges of nutrients from coastal plain watersheds of Chesapeake Bay. J. Environ. Qual. 26, 836–848.

Jordan, T.J., Correll, D.L., Weller, D.E., 1997b. Relating nutrient discharges from watersheds to land use and streamflow variability. Water Resour. Res. 33, 2579–2590.

Jordan, T.E., Weller, D.E., Correll, D.L., 2003. Sources of nutrient inputs to the Patuxent River estuary. Estuaries 26, 226–243.

Kemp, W.M., Boynton, W.R., Adolf, J.E., Boesch, D.F., Boicourt, W.C., Brush, G., Cornwell, J.C., Fisher, T.R., Glibert, P.M., Hagy, J.D., Harding, L.W., Houde, E.D., Kimmel, D.G., Miller, W.D., Newell, R.I.E., Roman, M.R., Smith, E.M., Stevenson, J.C., 2005. Eutrophication of Chesapeake Bay: historical trends and ecological interactions. Mar. Ecol-Prog. Ser. 303, 1–29.

Lamb, R., Kay, A.L., 2004. Confidence intervals for a spatially generalized, continuous simulation flood frequency model for Great Britain. Water Resour. Res. 40, W07501. doi:10.1029/2003WR002428.

Lee, K.-Y., Fisher, T.R., Jordan, T.E., Correll, D.L., Weller, D.E., 2000. Modeling the hydrochemistry of the Choptank River basin using GWLF and Arc/Info: 1. Model calibration and validation. Biogeochemistry 49, 143–173.

Lee, K.-Y., Fisher, T.R., Rochelle-Newall, R., 2001. Modeling the hydrochemistry of the Choptank River basin using GWLF and Arc/Info: 2. Model application. Biogeochemistry 56, 311–348.

Madsen, H., 2000. Automatic calibration of a conceptual rainfall–runoff model using multiple objectives. J. Hydrol. 235, 276–288.

Madsen, H., 2003. Parameter estimation in distributed hydrological catchment modelling using automatic calibration with multiple objectives. Adv. Water Resour. 26, 205–216.

McKay, M.D., Beckman, R.J., Conover, W.J., 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21 (2), 239–245.

Oudin, Ludovic, Andréassian, V., Perrin, C., Michel, C., Le Moine, N., 2008. Spatial proximity, physical similarity, regression and ungaged catchments: a comparison of regionalization approaches based on 913 French catchments. Water Resour. Res. 44, W03413. doi:10.1029/2007WR006240.

Rode, M., Suhr, Wriedt, U.G., 2007. Multi-objective calibration of a river water quality model–information content of calibration data. Ecol. Model. 204, 129–142.

Rosenberg, R., 1985. Eutrophication – the future marine coastal nuisance? Mar. Pollut. Bull. 16, 227–231.

Schneiderman, E.M., Pierson, D.C., Lounsbury, D.G., Zion, M.S., 2002. Modeling the hydrochemistry of the Cannonsville watershed with Generalized Watershed Loading Functions (GWLF). J. Am. Water Resour. Assoc. 38, 1323–1347.

US EPA (US Environmental Protection Agency), 1999. Protocols for Developing Nutrient TMDLs. EPA 841-B-99-007. Office of Water (4503 F), Washington, DC.

USDA-NRCS (US Department of Agriculture-Natural Resources Conservation Service), 1995. Soil Survey Geographic (SSURGO) Data Base: Data Use Information. National Cartography and GIS Center, Fort Worth, Texas.

van Griensven, A., Bauwens, W., 2003. Multi-objective autocalibration for semi-distributed water quality models. Water Resour. Res. 39 (12), 1348. doi:10.1029/2003WR002284.

van Griensven, A., Bauwens, W., 2005. Application and evaluation of ESWAT on the Dender basin and the Wister lake basin. Hydrol. Process. 19 (3), 827–838.

van Griensven, A., Meixner, T., 2007. A global and efficient multi-objective auto-calibration and uncertainty estimation method for water quality catchment models. J. Hydroinform. 9 (4), 277–291.

Vandewiele, G.L., Elias, A., 1995. Monthly water balance of ungauged catchments obtained by geographical regionalization. J. Hydrol. 170, 277–291.

Vrugt, J.A., Gupta, H.V., Bastidas, L.A., Bouten, W., Sorooshian, S., 2003. Effective and efficient algorithm for multiobjective optimization of hydrologic models. Water Resour. Res. 39, 1–19.

Wagener, T., Wheater, H.S., 2006. Parameter estimation and regionalization for continuous rainfall–runoff models including uncertainty. J. Hydrol. 320, 132–154.

Willmott, C.J., 1981. On the validation of models. Phys. Geogr. 2, 184–194.

Wischmeier, W.H., Smith, D.D., 1978. Predicting Rainfall–Erosion Losses: A Guide to Conservation Planning. Agriculture Handbook #507. USDA, Washington, DC.

Yapo, P.O., Gupta, H.V., Sorooshian, S., 1998. Multi-objective global optimization for hydrologic models. J. Hydrol. 204, 83–97.