Model Performance Sensitivity to Objective Function during Automated Calibrations

Misgana K. Muleta

Abstract: Previous studies have reported limitations of the efficiency criteria commonly used in hydrology to describe goodness of model simulations. This study examined sensitivity of model performance to the objective function used during automated calibrations. Nine widely used efficiency criteria were evaluated for their effectiveness as objective function, and goodness of the model predictions were examined using 13 criteria. Two cases (Case I: Using observed streamflow data and Case II: Using simulated streamflow) were considered to accomplish objectives of the study using a widely used watershed model (SWAT) and good-quality field data from a well-monitored experimental watershed. Major findings of the study include (1) automated calibration results are sensitive to the objective function group—group that work based on minimization of the absolute deviations (Group I), group that work based on minimization of square of the residuals (Group II), and groups that use log of the observed and simulated streamflow values (Group III)—but not to objective functions within the group; (2) efficiency criteria that belong to Group I were the most effective when used as objective function for accurate simulation of both low flows and high flows; (3) Group I and Group II objective functions complement each other's performance; (4) with regard to the capability to describe goodness of model simulations, efficiency criteria that belong to Group I showed superior robustness; (5) for the study watershed, use of the long-term interannual calendar day mean as baseline model did not improve capability of an efficiency criterion to describe model performance; and (6) even for ideal conditions where uncertainty in input data and model structure are fully accounted for, identifying the so-called global parameters values through calibration could be daunting as parameter values that were significantly divergent from predetermined values produced model simulations that can be considered near perfect even when judged using multiple efficiency criteria.

Introduction

Computer models are routinely used for planning and management of water resources. Because models are simplifications of the real world, accuracy of their predictions cannot be taken for granted. Consequently, models must be calibrated to ensure that simulation results are sound and defensible (U.S. EPA 2002). Calibration refers to the process of identifying model parameters (i.e., nonmeasurable model inputs) that produce model outputs that closely match the observed watershed characteristics. Calibration can be performed manually with trial-and-error procedure, automatically with the help of optimization methods or using their combination. Automated calibration is favored as manual calibration is often tedious, time-consuming, and requires experienced personnel (Muleta and Nicklow 2005). Irrespective of the method pursued, the ultimate goal of calibration is to develop a model that simulates the watershed characteristics as accurate as possible.

How well a model simulation fits the observed data is evaluated graphically (i.e., visual comparison of observed and simulated values using various kinds of plots) as well as by using one or more statistical measures commonly referred to as efficiency criteria, goodness-of-fit criteria, or efficiency measures. Efficiency criteria are derived from the residual (error) between the simulated and observed values. Many such measures have been used in hydrologic modeling (Nash and Sutcliffe 1970; Legates and McCabe 1999; Krause et al. 2005; Moriasi et al. 2007; Gupta et al. 2009). The Nash-Sutcliffe efficiency (NSE) criterion originally proposed by Nash and Sutcliffe (1970) and the root of mean of square of errors (RMS) are the most commonly used criteria (Gupta et al. 2009). In spite of its popularity, NSE has many well-documented limitations (ASCE 1993; Legates and McCabe 1999; Seibert 2001; Krause et al. 2005; McCuen et al. 2006; Moriasi et al. 2007; Schaefli and Gupta 2007; Criss and Winston, 2008; Jain and Sudheer 2008; Gupta et al. 2009).

The first concern is tha.t the use of mean of the observations (O_{mean}) as a benchmark model would lead to overestimation of the hydrologic model performance for highly seasonal watershed variables (Gupta et al. 2009; Schaefli and Gupta 2007; Sharad and Sudheer 2008). This means that interpretation of model performance is inconsistent as the reference model has different meaning for different watersheds depending on seasonality of the watershed variable. Large value of NSE can be obtained with a poor model if the data have high variability. As an example, Schaefli et al. (2005) showed that a trivial model that assigns mean observed discharge for each calendar day resulted in NSE of 0.85 for mountainous watersheds that have strong annual discharge cycle. In the contrary, if observations exhibit less variability and the values are close to O_{mean}, NSE can approach negative infinity even if the model is good predictor of the observations. This makes model performance communication very misleading is NSE alone is used. The second limitation is oversensitivity of NSE to peak flows as it uses the

squared deviations making it less than adequate to measure ability of the hydrologic model to simulate low flows (Legates and McCabe 1999; Krause et al. 2005; Criss and Winston 2008).

Several solutions have been proposed to address shortcomings of the NSE. The major recommendations include (1) using more meaningful baseline model that a hydrologist is likely to adopt in the absence of a model for the watershed. Suggested baseline models include interannual mean for each calendar day (Schaefli and Gupta 2007), simple models such as the storage and lag (Schaefli and Gupta 2007) or rational formula (Seibert 2001); (2) using other efficiency criteria that are modified to lessen oversensitivity of NSE to peak flows (Legates and McCabe 1999; Krause et al. 2005; Criss and Winston 2008); and (3) evaluating and reporting model performance using multiple efficiency criteria including the NSE (Legates and McCabe 1999; Krause et al. 2005).

Besides describing how well model simulations fit observed data, one or more efficiency criteria are also used as objective function(s) during automated calibrations to help identify optimal parameter sets. Many studies have demonstrated sensitivity of calibration results to the objective function used as a calibration criterion (Sorooshian et al. 1983; Yan and Haan 1991; Gupta et al. 1998). Recommendations to address the sensitivity include using the objective function(s) consistent with the anticipated application of the model and/or implementing multiobjective calibration approaches (Gupta et al. 1998). While application of multiobjective calibration has been steadily rising (Yapo et al. 1998; Madsen 2000; Tang et al. 2006; Bekele and Nicklow 2007; Efstratiadis 2010), single objective calibration is still more widely used in many practical applications. Using an objective function that best represents the expected application of the model may work well if the anticipated application of the model is distinct (e.g., accurate simulation of peak flows for flood control application). For general (broad) purpose models that are expected to accurately simulate all parts of a hydrograph including low flows and peak flows; however, identification of an objective function that meets all requirements may be daunting.

The objectives of this study are to (1) examine sensitivity of model performance to the efficiency criteria used as objective function during automated calibration; (2) identify objective functions that are reasonably sensitive to both low flows and high flows and should be used for single objective automated calibration attempts; (3) identify objective functions that complement each other and should be used for multiobjective calibration applications; and (4) identify efficiency criteria that are robust in describing model performance and should be used to report model results. While most of these objectives have been examined by several authors in the past using lumped conceptual models (Sorooshian et al. 1983; Yan and Haan 1991; Gupta et al. 1998), this study revisits the issues using a widely used, spatially distributed watershed model on a data rich experimental watershed that has dense precipitation and streamflow gauges and good-quality geospatial data including topography, land cover, and soil. The multiple streamflow gauges available in the study watershed helped to examine robustness of the results to reasonably simulate internal responses of the watershed.

Model Efficiency Criteria: Overview

The efficiency criteria described below have been used in this study either as objective function during optimization and/or to test how well the calibrated model fits the observed data. The efficiency criteria have been selected based on recommendations in the literature.

Nash-Sutcliffe Efficiency

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{N}(S_i - O_i)^2}{\sum_{i=1}^{N}(O_i - O_{\mathrm{mean}})^2} \tag{1}$$

where S = model simulated output; O = observed hydrologic variable; O_{mean} = mean of the observations that the NSE uses as a benchmark against which performance of the hydrologic model is compared; and N = total number of observations. NSE values range from negative infinity to 1, where 1 shows a perfect model. NSE is zero, implies the observed mean is as good a predictor as the model, and if NSE is less than zero, then the model is worse predictor than Q_{mean}.

Root Mean Square Error

$$\mathrm{RMS} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(S_i - O_i)^2} \tag{2}$$

RMS ranges from zero (for the ideal model) to positive infinity (worst model). RMS is biased toward peak flows.

Nash-Sutcliffe Efficiency with Calendar Day Mean as Reference Model

One of the major limitations of NSE is the use of mean of all observations as the baseline model. Various studies, including Schaefli and Gupta (2007), have suggested alternative benchmark models such as the interannual calendar day mean. Accordingly, NSE formulation has been modified as follows to test the effect of using the interannual calendar day average ($\bar{O}_D$) as a reference model

$$\mathrm{NSD} = 1 - \frac{\sum_{i=1}^{N}(S_i - O_i)^2}{\sum_{i=1}^{N}(O_i - \bar{O}_D)^2} \tag{3}$$

For this study, $\bar{O}_D$ has been calculated for each calendar day of a year at five streamflow gauging sites in the study watershed from 39 years (i.e., 1967–2006) of observed data. The premise of using $\bar{O}_D$ as a baseline model is that in the absence of any model $\bar{O}_D$ would be better predictor of streamflow at the site for that specific calendar day of the year than the mean of observations (O_{mean}) used in the original NSE formulation.

Modified Forms of Nash-Sutcliffe Efficiency and Nash-Sutcliffe Efficiency with Calendar Day Mean

Oversensitivity of NSE and Nash-Sutcliffe Efficiency with Calendar Day Mean (NSD) to peak flows can be minimized by modifying formulation of the measures as follows (Krause et al. 2005):

$$\mathrm{MNS} = 1 - \frac{\sum_{i=1}^{N}|S_i - O_i|}{\sum_{i=1}^{N}|O_i - O_{\mathrm{mean}}|} \tag{4}$$

$$\mathrm{MNSD} = 1 - \frac{\sum_{i=1}^{N}|S_i - O_i|}{\sum_{i=1}^{N}|O_i - \bar{O}_D|} \tag{5}$$

where MNS and MNSD = modified forms of NSE and NSD, respectively. These modified forms are expected to better describe model performance as they are more evenly sensitive to low flows as well as high flows.

Nash-Sutcliffe Efficiency and Nash-Sutcliffe Efficiency with Calendar Day Mean Calculated from Logarithmic Values

Another proposal to ease oversensitivity of NSE to high flows was to use logarithmic of the observed and predicted values. The approach was tested for both NSE as well as NSD

$$\mathrm{LNS} = 1 \quad \frac{\sum_{i=1}^{N}(\ln(S_i + 0.001) - \ln(O_i + 0.001))^2}{\sum_{i=1}^{N}(\ln(O_i + 0.001) - O_{\ln,\mathrm{mean}})^2} \tag{6}$$

$$\mathrm{LNSD} = 1 \quad \frac{\sum_{i=1}^{N}(\ln(S_i + 0.001) - \ln(O_i + 0.001))^2}{\sum_{i=1}^{N}(\ln(O_i + 0.001) - \bar{O}_{\ln,D})^2} \tag{7}$$

where LNS and LNSD = the NSE and NSD calculated using log of the observed and simulated values, respectively. $O_{\ln,\mathrm{mean}}$ and $\bar{O}_{\ln,D}$ = mean of the log of all observations and mean of the log of each calendar day observations, respectively.

Mean Absolute Error (MAE)

MAE is expected to be less sensitive to high flows and more sensitive to low flows than NSE and RMS and is expected to describe model performance more evenly

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|S_i \quad O_i| \tag{8}$$

Volumetric Efficiency

Criss and Winston (2008) proposed the volumetric efficiency (VE) criterion to address the limitations of NSE and its modifications

$$\mathrm{VE} = 1 \quad \frac{\sum_{i=1}^{N}|S_i \quad O_i|}{\sum_{i=1}^{N} O_i} \tag{9}$$

Percent Bias

Moriasi et al. (2007) recommended percent bias (PBIAS) as one of the measures that should be included in model performance reports. Percent bias describes whether the model simulations overestimate or underestimate the observations

$$\mathrm{PBIAS} = 100 * \frac{\sum_{i=1}^{N}(O_i \quad S_i)}{\sum_{i=1}^{N} O_i} \tag{10}$$

Ratio of Standard Deviation of Observations to RMS

Ratio of standard deviation of observations to RMS (RSR) standardizes RMS using standard deviation of the observations and can be used to compare performances across watersheds or various constituents (Moriasi et al. 2007)

$$\mathrm{RSR} = \frac{\sqrt{\sum_{i=1}^{N}(S_i \quad O_i)^2}}{\sqrt{\sum_{i=1}^{N}(O_i \quad O_{\mathrm{mean}})^2}} \tag{11}$$

Coefficient of Determination

Coefficient of determination (R^2) is an indicator of the extent to which the model explains the total variance in the observed data. A major limitation of R^2 is that it describes the linear relationship between the two data sets, and one may obtain large R^2 value with a poor model that consistently overestimates or underestimates the observations.

$$R^2 = \left\{\frac{\sum_{i=1}^{N}(O_i \quad O_{\mathrm{mean}})(S_i \quad S_{\mathrm{mean}})}{[\sum_{i=1}^{N}(O_i \quad O_{\mathrm{mean}})^2]^{0.5}[\sum_{i=1}^{N}(S_i \quad S_{\mathrm{mean}})^2]^{0.5}}\right\}^2 \tag{12}$$

where S_{mean} = mean of the model simulations.

Index of Agreement

Willmott (1981) proposed the index of agreement (D) to overcome the limitation of R^2 described previously. D suffers from oversensitivity to extreme flows (Legates and McCabe 1999)

$$D = 1 \quad \left\{\frac{\sum_{i=1}^{N}(O_i \quad S_i)^2}{\sum_{i=1}^{N}(|S_i \quad O_{\mathrm{mean}}| + |O_i \quad O_{\mathrm{mean}}|)^2}\right\} \tag{13}$$

PBIAS, RSR, R^2, and D were used in the study only to describe goodness of model results, but not as objective function during optimization. For easy reference, all the efficiency criteria considered in the study are summarized in Table 1 with regard to the acronym used, range of values they assume, the formulation group they belong to, and whether they are used as objective function.

Methods and Materials

Study Watershed and Simulation Model

As shown in Fig. 1 headwaters of the Little River Experimental Watershed (LREW), one of the USDA-ARS's (AU: Please spell out acronym.) experimental watersheds, located in Georgia, was used to demonstrate the research objectives. The LREW was selected because it is heavily gauged for rainfall as well as streamflow (Bosch et al. 2007), and because data are readily accessible online (ftp://www.tiftonars.org/) from the Southeast Watershed Research Laboratory (SEWRL) (SEWRL 2010). The watershed consists primarily of low-gradient streams and is located mainly on sandy soils underlain by limestones that form locally confined aquifers. Land use within the watershed consists of about 31% row crop agriculture, 10% pasture, 50% forest, and 7% urban area (Bosch et al. 2006). Only the upper 116 km^2 of the LREW was used for this study to minimize computational demand of the model and also because the headwater subwatersheds have denser streamflow and rainfall gauges.

Table 1. Summary of Efficiency Criteria Considered for Study

Criterion	Range of values	Group	Application: Eff. criterion	Application: Obj. function
NSE	∞ to 1.0	II	x	x
RMS	0.0 to ∞	II	x	x
NSD	∞ to 1.0	II	x	x
MNS	∞ to 1.0	I	x	x
MNSD	∞ to 1.0	I	x	x
LNS	∞ to 1.0	III	x	x
LNSD	∞ to 1.0	III	x	x
MAE	0.0 to ∞	I	x	x
VE	∞ to 1.0	I	x	x
PBIAS	∞ to ∞	IV	x	No
RSR	0.0 to ∞	II	x	No
R^2	0 to 1.0	II	x	No
D	0 to 1.0	II	x	No

Note: Group I represents efficiency criteria that work based on minimization of absolute deviations; Group II represents efficiency criteria that work based on minimization of square of deviations; Group III represents efficiency criteria that use log of observed and simulated values; Group IV measure deviations between observed and simulated values; x = yes.

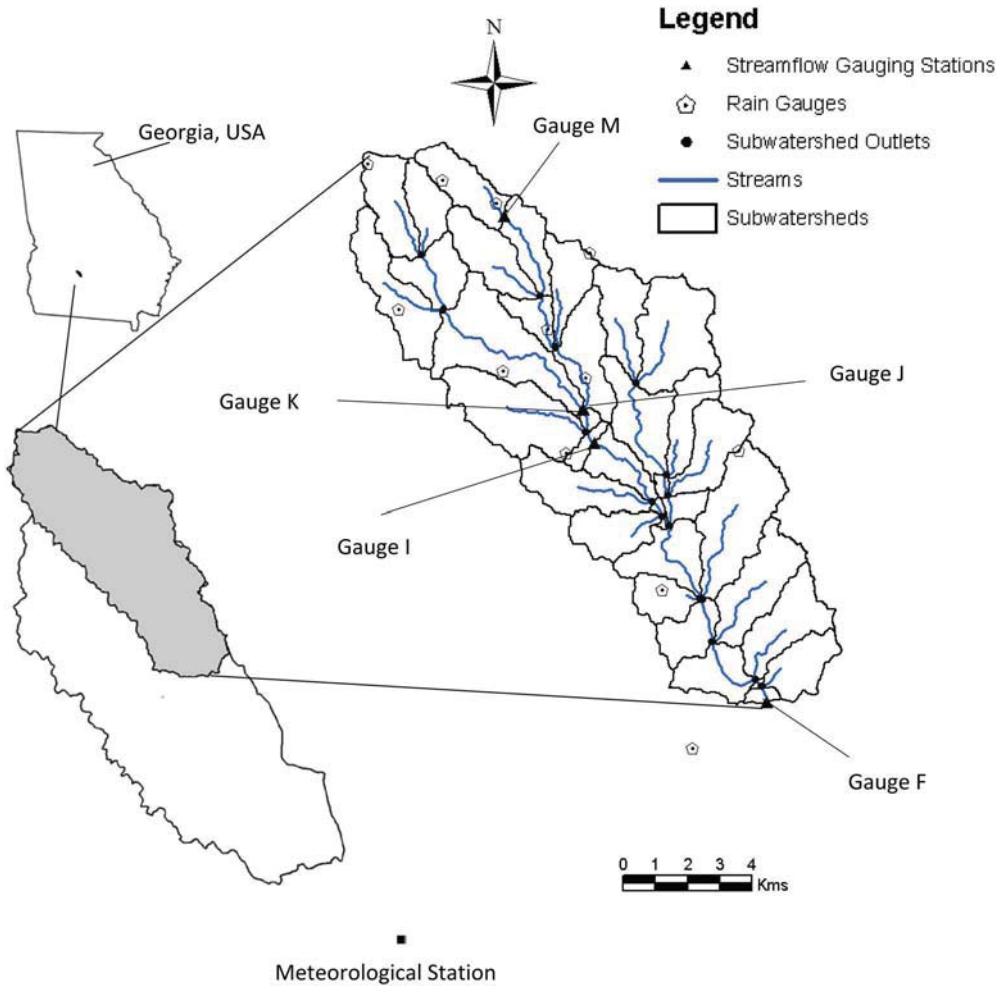

Fig. 1. Location map of study area and gaging stations

Twelve precipitation gauges and five streamflow gauges (see Fig. 1) with long-term daily data (i.e., 1967–2006) are available for the headwaters from the SEWRL. Daily minimum and daily maximum temperature data for a station near the watershed was obtained from the U.S. Historical Climatology Network (http://cdiac.ornl.gov/epubs/ndp/ushcn/ushcn.html) as the air temperature data available from SEWRL starts only from 2004. The geographic data used to set up SWAT model including topography, land use, stream networks, and rainfall and streamflow gauging locations were obtained from the SEWRL. Soil survey geographic (SSURGO) soil map was obtained from the U.S. Department of Agriculture, Natural Resources Conservation Service (NRCS) soil data mart (http://soildatamart.nrcs.usda.gov/). SWATioTools (Sheshukov et al. 2009), an ArcMap geographic information system (GIS) extension tool that converts SSURGO soils into the format readable by ArcSWAT (Winchell et al. 2008) was used to preprocess the SSURGO soil map. The land cover image used for the study was for 2003. After the climate, streamflow, land use, and soil data were preprocessed, the 116 km^2 study watershed was delineated and subdivided into 37 subwatersheds and 71 hydrologic response units (HRUs) using ArcSWAT as shown in Fig. 1.

Soil and water assessment tool (SWAT) (Arnold et al. 1999), the simulation model used for this study, is one of the most widely used watershed simulation models in use today (Gassman et al. 2007). SWAT is a physically based, spatially distributed model that uses information regarding climate, topography, soil properties, land cover, and human activities such as land management practices to simulate numerous physical processes including surface runoff, groundwater flow, streamflow, and many water quality fluxes. Spatially, the model subdivides a watershed in to subwatersheds and, potentially, further partitions them into hydrologic response units based on land cover, soil, and the overland slope diversity in the subwatershed. The reader is referred to Neitsch et al. (2005) for details on SWAT.

Automated Calibration Method

Automated model calibration was performed to identify optimal values of 12 most sensitive streamflow parameters of SWAT that were previously identified by the author (Muleta 2012) using a global sensitivity analysis model known as Sobol (Sobol' 1993). All 12 parameters were assumed to follow uniform distribution as done in Muleta and Nicklow (2005), and the bounds recommended in Neitsch et al. (2005) were used for majority of the parameters. List of the parameters and their bounds are provided in Table 2. Some of these model parameters (e.g., NRCS curve number, CN2) can vary spatially depending on soil, land cover, slope, and/or other watershed characteristics. During calibration, the baseline values originally assigned to the spatially varying parameters were altered by adding or multiplying the baselines by a sampled value as described in Table 2. This way, the parameters were adjusted while preserving their spatial variability. The dynamically dimension search (DDS) algorithm described in Tolson and Shoemaker (2007) was obtained from the first writer and integrated with SWAT for automated calibration.

Table 2. SWAT's Most Sensitive Streamflow Parameters and Ranges Used for Analysis

Name	Description	Bounds of values	
		Min.	Max.
Alpha_Bf	Base flow alpha factor (days)	0.1	1
Canmx	Maximum canopy storage index (unit less)	0	10
Ch_K2	Effective hydraulic conductivity in main channel alluvium (mm/hr)	0	150
Ch_N2	Manning's n for the main channels (unit less)	0.0	0.1
Cn2[a]	SCS runoff curve number for moisture condition II	25%	25%
Esco	Soil evaporation compensation factor (unit less)	0	1
Gwqmn[b]	Threshold depth of water in the shallow aquifer required for return flow to occur (mm)	5,000	5,000
Slope[a]	Average slope steepness (m/m)	50%	50%
Sol_Awc[a]	Available water capacity of the soil layer (mm/mm soil)	50%	50%
Sol_K[a]	Soil hydraulic conductivity (mm/hr)	50%	50%
Sol_Z[a]	Soil depth (mm)	50%	50%
Surlag	Surface runoff lag time (days)	0	10

[a]Indicates spatially varying parameters whose baseline values are adjusted during calibration by multiplier sampled from bound.
[b]Indicates spatially varying parameters whose baseline values are adjusted during calibration by adding it to value sampled from bound.

Methodology

Season-Based Calibration

Season-based calibration was used in this study based on the findings of Muleta (2012). In season-based calibration, model parameters that are physically expected to change from dry season to wet season are allowed to vary seasonally. In the conventional calibration technique, parameters are assumed constant during low flow seasons as well as high flow seasons. The study by Muleta (2012) demonstrated significant improvement in model performance when season-based calibration is pursued. Various diagnostic analysis studies have also shown sensitivity of dominant model characteristics to various seasons of a year (Li et al. 2012; Tian et al. 2012). According to Tang et al. (2007) and van Werkhoven et al. (2008), forcings, mainly rainfall, are responsible for temporal sensitivity for the model watershed they used. For the headwaters of LREW, however, Muleta (2012) showed that the observed rainfall and runoff exhibited relationships that cannot be described using rainfall alone. Monthly runoff coefficients determined from 39 years (i.e., 1968–2006) of rainfall and runoff data for the watershed, however, proved effective in defining dry and wet season for the watershed (Muleta 2012). Months with runoff coefficient greater than 0.1 (i.e., January to April) were considered wet season, and months with runoff coefficient less than 0.1 (i.e., June to November) were considered dry season. December and May were transition months where parameters values linearly vary from the dry season values to the wet season values and vice versa, respectively.

From the 12 sensitive parameters listed in Table 2, Slope, Sol_AWC, Sol_Z, and Sol_K were assumed season insensitive and the other eight parameters were allowed to vary seasonally. Single objective automated calibration was performed using nine different efficiency criteria as objective function, one objective function per calibration run. Performance of each calibration attempt was then tested using 13 different efficiency criteria. Streamflow data from gauge F (outlet of the study watershed as shown in Fig. 1) were used for the calibration. One-year data (i.e., 1999 data) were used as a spin-up period to diffuse the effect of antecedent conditions, and four-year data (i.e., 2000–2003) were used for calibration. Performance of the calibration result was verified using the split-sampling approach (i.e., 2004–2006 data at the calibration site were used for verification) as well as by examining capability of the calibrated model to simulate reasonable streamflow at internal gauges not used for calibration (i.e., gauges I, J, K, and M) using seven-year data (i.e., 2000–2006).

The DDS algorithm guarantees uniqueness of calibration results for a given objective function as long as the random seed generator used by the algorithm is not altered. Conducting multiple calibration runs for the same objective function and random seed does not affect calibration result. This was confirmed in the study by conducting 10 separate calibration attempts using NSE as objective function and same random seed for all 10 calibration runs. This implies that the difference exhibited between the results of two calibration attempts that use two different objective functions is attributed purely to the objective functions. As such, only one calibration run per objective function was performed to accomplish goals of the study.

Cases Considered

Two cases were considered for the research. In the first case, referred to as Case I, the actual data (i.e., soil, land use, topography, climate, and streamflow) available for the watershed were used to build the SWAT model. Then, DDS was used to calibrate the model using nine different objective functions using the season-based calibration method. For each calibration run, 3,000 SWAT simulations were used for DDS. In the second case, referred to as Case II, instead of the actual streamflow available for the watershed, the streamflow simulated by the calibrated SWAT in Case I using one of the nine objective functions was used as observed streamflow and calibrations were repeated for each of the nine objective functions.

In Case II, uncertainties from model structure, input data, and observed streamflow were eliminated from the modeling process. This is because (1) the streamflow simulated by the calibrated model is considered as observed streamflow implying that the output data (i.e., streamflow) used for calibration are error free; (2) forcings (inputs) have no error as the inputs that produced the observed streamflow are used to recalibrate the model; and (3) structure of the simulation model is perfect for the watershed characteristics being modeled as the streamflow simulated by the model is used to recalibrate the same model. Additionally, because optimal values of the 12 parameters that produced the observed streamflow are known, Case II helped to identify capability of the objective functions to reproduce these known optimal parameters values. Therefore, Case II was designed to further elucidate relative effectiveness of the nine objective functions. Total of nine calibration attempts, each with 3,000 SWAT simulations, were performed for Case II as well.

Results and Discussion

Case I: Using Observed Data

Sample results for Case I are given in Fig. 2 and Tables 3–5. Fig. 2 compares observed streamflow with the streamflow simulated using the optimal parameters values identified when NSE and MNSD

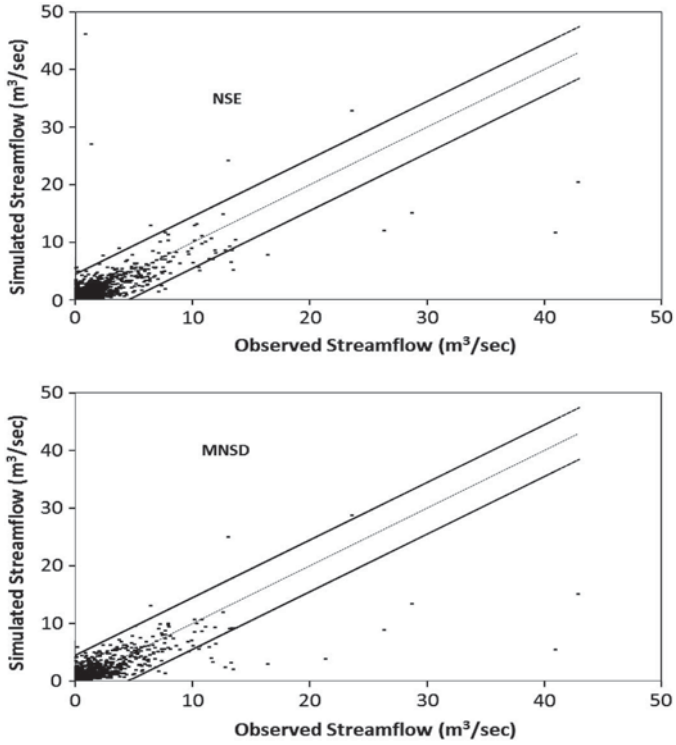

Fig. 2. Comparison of observed and simulated streamflow at gauge *F* for case I when NSE and MNSD are used as objective function; the light dash line shows perfect fit line and darker lines show bounds of 95% confidence interval

were used as objective functions. For easy observation, the perfect fit line (i.e., 1:1 line) and lower and upper bounds of the 95% confidence interval are also given in Fig. 2. The tables show efficiency criteria values calculated from the streamflow simulated using optimal solutions of each of the nine objective functions considered. The efficiency criteria given in the rows represent the criteria used as objective function for the attempt. The best efficiency criteria values are given in bold for easy comparison. Results are given for the calibration site (i.e., gauge *F*), and for an internal site (i.e., gauge *J*) for both calibration period (i.e., 2000–2003) and verification period (i.e., 2004–2006). Similar results are available for other internal gauges (i.e., gauges *I*, *K*, and *M*), but are not shown here for brevity. Gauge *J* is representative of the results obtained at the other internal gauges.

The efficiency criteria values determined considering all months of a year (i.e., dry months, wet months, and transition months combined) are given in Table 3. To examine sensitivity of model performance to season, the results given in Table 4 were calculated using streamflow from the wet months (i.e., January to April) and those in Table 5 were determined using streamflow simulated for the low flow months (i.e., June to November). Streamflow simulations are generally considered satisfactory if NSE > 0.5, PBIAS is within $\pm 25\%$, and RSR is ≤ 0.70 (Moriasi et al. 2007). According to these criteria, Tables 3–5 indicate, for the most part, satisfactory results for both calibration and verification periods for the year-round, wet season and dry season cases at the calibration site as well as at the internal gauges. Fig. 2 also shows satisfactory results as the model simulations, except for the high flows, align well with the perfect fit line and are within the 95% confidence interval. These results confirm robustness of the season-based calibration approach pursued in the study.

Performance Sensitivity to Objective Function

A closer look at Eqs. (1)–(9) reveals that the nine objective functions can be grouped (see Table 1) as (1) Group I: those that minimize absolute deviations of the simulated and observed flows (i.e., MAE, MNS, MNSD, and VE); (2) Group II: those that minimize square of the residuals (i.e., RMS, NSE, and NSD); and (3) Group III: those that use log of the observed and simulated values (i.e., LNS and LNSD). Interestingly, the results given in Table 3 exhibited sensitivity only to the three objective function groups, but not to the objective functions within the group. In other words, MAE, MNS, MNSD, and VE produced identical results when used as objective function except for the trivial differences observed in the PBIAS values. Likewise, RMS, NSE, and NSD generated identical solutions, and LNS and LNSD produced results that are alike. Similar performances were exhibited for the wet and the dry season results. This shows that model performance is not sensitive to the baseline model used for the objective function (e.g., NSD and NSE produced identical optimal solutions). Based on this finding and to avoid redundancy, Tables 4–6 show the results obtained using one objective function per group (i.e., NSE, MNSD, and LNSD). However, RMS results are also reported in Table 7 to demonstrate nonuniqueness of the optimal parameters values.

Sensitivity of the optimal solutions to the three objective function groups can be examined from Tables 3–5. As an example, Table 3 shows that the optimal solutions obtained when LNSD and NSE were used as objective functions produced NSE of 0.30 and 0.73, and PBIAS of 21% and 1.89%, respectively, at gauge *F* for the calibration period. None of the nine objective functions produced solutions that consistently outperformed the other objective functions when evaluated using all 13 efficiency criteria. In other words, using an NSE as an objective function may produce solution that improves one or more efficiency criteria, but would also result in inferior values of the other efficiency criteria. A typical example is the NSE result shown in Table 3 for gauge *J* for the calibration period, where it produced the best NSE of 0.71, but the worst LNSD (i.e., 0.35) and PBIAS (i.e., 16.90). Generally speaking, however, it can be concluded that performance of Group I objective functions (those minimizing the absolute deviations) was more robust. Group I objective functions either outperformed or produced fairly comparable results to that of Groups II and III for all year-round Table 3, wet season Table 4, and low flow season Table 5.

Another important finding is that Group I and Group II objective functions seem to complement each other's performance. In Tables 3–5, for the most part, the best performing objective function belongs to either of the two groups. This emphasizes the benefit of multiobjective calibrations (Gupta et al. 1998; Yapo et al. 1998) that use an objective function from each group (e.g., RMS and MNSD) to improve accuracy of simulating both low flows as well as peak flows (Madsen 2000; Efstratiadis and Koutsoyiannis 2010). One surprising observation is that no seasonal sensitivity has been exhibited by the performance of the objective functions. Objective functions in Group II did not show significant improvement in simulating wet season flows compared to how they simulated the dry season flows. This is counterintuitive given the notion that objective functions that minimize square of the residuals would be biased toward high flows and would more accurately simulate wet season flows.

Robustness in Describing Model Performance

With regard to explaining the goodness of model performance, with the exception of LNS and LNSD, all the other efficiency criteria seem consistent. From Tables 3–5, it can be observed that, for the most part, a calibration result seems to either improve or deteriorate all efficiency criteria (except for LNS and LNSD) consistently.

Table 3. Efficiency Criteria Values Obtained for Case I for All Months

Criteria	MAE	RMS	NSE	NSD	MNS	MNSD	LNS	LNSD	VE	PBIAS	RSR	R^2	*D*
						Calibration period							
Gauge *F*													
MAE	**0.49**	1.06	0.68	0.64	**0.55**	**0.57**	0.20	0.30	**0.46**	8.71	0.01	0.69	0.89
RMS	0.51	**0.97**	**0.73**	**0.69**	0.53	0.55	0.24	0.23	0.44	1.95	0.01	**0.73**	**0.92**
NSE	0.51	**0.97**	**0.73**	**0.69**	0.53	0.55	0.24	0.24	0.44	**−1.89**	0.01	**0.73**	**0.92**
NSD	0.51	**0.97**	**0.73**	**0.69**	0.53	0.55	0.24	0.24	0.44	2.16	0.01	**0.73**	**0.92**
MNS	**0.49**	1.06	0.68	0.64	**0.55**	**0.57**	0.20	0.30	**0.46**	8.77	0.01	0.69	0.89
MNSD	**0.49**	1.06	0.68	0.64	**0.55**	**0.57**	0.20	0.30	**0.46**	8.71	0.01	0.69	0.89
LNS	0.66	1.58	0.30	0.20	0.40	0.42	**0.42**	**0.06**	0.28	21.01	0.02	0.36	0.75
LNSD	0.66	1.58	0.30	0.19	0.40	0.42	**0.42**	**0.06**	0.28	19.90	0.02	0.36	0.75
VE	**0.49**	1.06	0.68	0.64	**0.55**	**0.57**	0.20	0.30	**0.46**	8.71	0.01	0.69	0.89
Gauge *J*													
MAE	**0.09**	**0.20**	**0.71**	**0.72**	**0.51**	**0.62**	**0.27**	**−0.28**	**0.36**	9.43	0.01	0.71	0.91
RMS	0.10	0.21	**0.71**	0.71	0.48	0.60	0.23	0.35	0.31	16.76	0.01	**0.72**	**0.92**
NSE	0.10	0.21	**0.71**	0.71	0.48	0.60	0.23	0.35	0.31	16.90	0.01	**0.72**	**0.92**
NSD	0.10	**0.20**	**0.71**	**0.72**	0.48	0.60	0.23	0.35	0.31	16.50	0.01	**0.72**	**0.92**
MNS	**0.09**	**0.20**	**0.71**	**0.72**	**0.51**	0.63	**0.27**	**−0.28**	**0.36**	**9.35**	0.01	0.71	0.91
MNSD	**0.09**	**0.20**	**0.71**	**0.72**	**0.51**	**0.62**	**0.27**	**−0.28**	**0.36**	9.43	0.01	0.71	0.91
LNS	0.10	0.25	0.56	0.57	0.45	0.58	0.23	0.34	0.28	10.18	0.02	0.57	0.86
LNSD	0.10	0.25	0.58	0.59	0.46	0.58	0.23	0.34	0.29	8.91	0.02	0.59	0.87
VE	**0.09**	**0.20**	**0.71**	**0.72**	**0.51**	**0.62**	**0.27**	**−0.28**	**0.36**	9.43	0.01	0.71	0.91
						Verification period							
Gauge *F*													
MAE	**0.68**	**1.92**	**0.50**	**0.49**	**0.43**	**0.47**	0.23	0.15	**0.35**	7.41	0.02	**0.50**	**0.80**
RMS	0.71	2.57	0.12	0.09	0.41	0.44	0.24	0.13	0.31	0.28	0.03	0.40	0.77
NSE	0.71	2.54	0.13	0.11	0.41	0.44	0.24	0.13	0.32	0.15	0.03	0.41	0.78
NSD	0.71	2.56	0.12	0.10	0.41	0.44	0.24	0.13	0.31	**0.06**	0.03	0.40	0.77
MNS	**0.68**	**1.92**	**0.50**	**0.49**	**0.43**	**0.47**	0.23	0.15	**0.35**	7.50	0.02	**0.50**	**0.80**
MNSD	**0.68**	**1.92**	**0.50**	**0.49**	**0.43**	**0.47**	0.23	0.15	**0.35**	7.41	0.02	**0.50**	**0.80**
LNS	0.90	3.68	0.82	0.87	0.25	0.29	0.42	0.14	0.13	23.08	0.04	0.20	0.60
LNSD	0.90	3.70	0.84	0.89	0.24	0.29	**0.43**	**0.15**	0.13	22.62	0.04	0.20	0.59
VE	**0.68**	**1.92**	**0.50**	**0.49**	**0.43**	**0.47**	0.23	0.15	**0.35**	7.41	0.02	**0.50**	**0.80**
Gauge *J*													
MAE	**0.17**	0.48	0.55	0.54	**0.49**	**0.48**	**0.42**	**0.05**	**0.35**	33.06	0.02	**0.64**	0.80
RMS	**0.17**	0.46	**0.60**	**0.59**	**0.49**	**0.48**	0.34	0.07	**0.35**	**−26.93**	0.02	0.61	**0.87**
NSE	**0.17**	**0.45**	**0.60**	**0.59**	**0.49**	**0.48**	0.34	0.07	**0.35**	27.06	0.02	0.61	**0.87**
NSD	**0.17**	**0.45**	**0.60**	**0.59**	**0.49**	**0.48**	0.35	0.07	**0.35**	27.14	0.02	0.61	**0.87**
MNS	**0.17**	0.48	0.55	0.54	**0.49**	**0.48**	**0.42**	**0.05**	**0.35**	33.09	0.02	**0.64**	0.80
MNSD	**0.17**	0.48	0.55	0.54	**0.49**	**0.48**	**0.42**	**0.05**	**0.35**	33.06	0.02	**0.64**	0.80
LNS	0.21	0.64	0.21	0.19	0.37	0.36	0.28	0.18	0.20	45.29	0.03	0.36	0.75
LNSD	0.21	0.64	0.21	0.20	0.37	0.36	0.28	0.18	0.20	45.20	0.03	0.36	0.75
VE	0.17	0.48	0.55	0.54	**0.49**	**0.48**	**0.42**	0.05	0.35	33.06	0.02	0.64	0.80

Note: Table shows efficiency criteria values calculated using optimal solution identified when criterion shown in rows is used as objective function during optimization; MAE is in m^3/s; RMS is in m^3/s; PBIAS is in %; other efficiency criteria are unitless.

Solutions that improve LNS and LNSD, however, seem to result in inferior performance when judged by the other 11 efficiency criteria. This issue may be attributed to the fact that Little River is an intermittent river and that both LNS and LNSD are highly sensitive to low flows. Trivial inaccuracy in estimating low flows may significantly deteriorate LNS and LNSD. Application of similar analysis to a perennial river could further elucidate this observation.

Overall, the criteria in Group I (i.e., MNS, MNSD, VE, and MAE) seem more robust in describing model performance. Notice from Tables 3–5 that it is easier to obtain values that suggest better agreement between observed and simulated flows using the criteria in Group II. The same optimal solution would likely produce NSE value closer to 1.0 than it would for MNSD. In this regard, VE appears very conservative criterion. A solution that leads to higher value of VE is likely to produce more desirable values of the other efficiency criteria as well. However, VE seems stringent criterion as solutions that can easily be considered satisfactory using, for example, NSE produce poor values of VE. On the contrary, almost all solutions produced very high values of *D* indicating that *D* is less suitable to describe model performance. These findings lead to the conclusion that model performance should be described using multiple efficiency criteria, preferably

Table 4. Efficiency Criteria Values Obtained for Case I for Wet Season

Criteria	MAE	RMS	NSE	NSD	MNS	MNSD	LNS	LNSD	VE	PBIAS	RSR	R^2	*D*
						Calibration period							
Gauge *F*													
NSE	0.80	**1.29**	**0.78**	**0.74**	0.43	0.59	0.80	0.23	0.13	**−16.52**	**0.03**	**0.78**	**0.93**
MNSD	**0.72**	1.32	0.77	0.73	**0.49**	**0.63**	**0.90**	**0.37**	**0.21**	22.13	**0.03**	0.77	**0.93**
LNSD	1.05	2.07	0.43	0.34	0.25	0.46	0.84	0.02	0.15	47.01	0.05	0.46	0.79
Gauge *J*													
NSE	0.16	**0.26**	**0.82**	**0.82**	0.43	0.66	0.63	0.10	0.10	**−5.30**	**0.03**	**0.81**	**0.95**
MNSD	**0.15**	0.30	0.77	0.77	**0.46**	**0.68**	**0.78**	**0.34**	**−0.04**	7.41	0.04	0.77	0.92
LNSD	0.17	0.32	0.72	0.73	0.39	0.64	0.45	0.66	0.19	50.40	0.04	0.73	0.91
						Verification period							
Gauge *F*													
NSE	**0.84**	**1.90**	**0.72**	**0.72**	**0.41**	**0.61**	0.45	0.34	**0.19**	32.22	**0.04**	**0.72**	**0.91**
MNSD	0.86	2.11	0.65	0.65	0.39	0.60	**0.75**	**0.38**	0.17	34.42	**0.04**	0.66	0.89
LNSD	1.07	2.75	0.41	0.41	0.24	0.50	0.47	0.30	0.03	68.31	0.05	0.50	0.82
Gauge *J*													
NSE	**0.20**	**0.40**	**0.80**	**0.80**	0.44	0.61	0.03	0.71	0.21	60.74	**0.03**	**0.88**	**0.93**
MNSD	**0.20**	0.44	0.77	0.76	**0.45**	**0.62**	**0.51**	**0.19**	**0.24**	**−59.73**	0.04	0.86	0.91
LNSD	0.24	0.47	0.73	0.72	0.33	0.54	0.56	1.59	0.07	85.57	0.04	0.80	0.90

Note: Table shows efficiency criteria values calculated using optimal solution identified when criterion shown in the rows is used as objective function during optimization; MAE is in m^3/s; RMS is in m^3/s; PBIAS is in %; other efficiency criteria are unitless.

Table 5. Efficiency Criteria Values Obtained for Case I for Dry Season

Criteria	MAE	RMS	NSE	NSD	MNS	MNSD	LNS	LNSD	VE	PBIAS	RSR	R^2	*D*
						Calibration period							
Gauge *F*													
NSE	**0.36**	**0.76**	**0.66**	**0.60**	**0.62**	**0.48**	0.08	0.33	**0.60**	11.67	**0.01**	**0.66**	**0.88**
MNSD	0.37	0.91	0.50	0.42	0.61	0.47	0.19	0.47	**0.60**	2.25	0.02	0.53	0.78
LNSD	0.46	1.31	0.03	0.19	0.52	0.34	**0.19**	**0.00**	0.50	**0.77**	0.02	0.18	0.66
Gauge *J*													
NSE	0.07	0.17	0.00	0.07	0.52	0.46	0.14	0.58	0.52	31.82	**0.01**	0.46	0.78
MNSD	**0.06**	**0.13**	**0.40**	**0.44**	**0.57**	**0.51**	0.19	0.66	**0.57**	20.71	**0.01**	**0.47**	**0.81**
LNSD	0.07	0.20	0.47	0.37	0.52	0.46	**0.01**	**−0.38**	0.52	**15.79**	0.02	0.18	0.65
						Verification period							
Gauge *F*													
NSE	0.60	2.94	0.60	0.70	0.47	0.28	0.22	0.11	0.42	14.75	0.04	0.27	**0.66**
MNSD	**0.55**	**1.90**	**0.34**	**0.30**	**0.52**	**0.34**	0.16	0.21	**0.47**	2.39	**0.03**	**0.35**	0.63
LNSD	0.82	4.35	2.50	2.71	0.28	0.03	**0.45**	**0.21**	0.21	**0.69**	0.06	0.10	0.43
Gauge *J*													
NSE	**0.15**	**0.50**	**0.41**	**0.39**	**0.54**	**0.35**	**0.44**	**0.06**	**0.42**	**−13.04**	**0.03**	0.27	**0.66**
MNSD	**0.15**	0.53	0.34	0.31	0.53	0.34	**0.44**	**0.06**	0.41	24.27	**0.03**	**0.35**	0.63
LNSD	0.19	0.75	0.34	0.38	0.40	0.16	**0.44**	**0.06**	0.25	27.83	0.04	0.10	0.43

Note: Table shows efficiency criteria values calculated using optimal solution identified when criterion shown in rows is used as objective function during optimization; MAE is in m^3/s; RMS is in m^3/s; PBIAS is in %; other efficiency criteria are unitless.

one criterion from each of the three efficiency groups (e.g., NSE, MAE, LNS) and PBIAS.

Case II: Using Simulated Data

Case II results are summarized in Figs. 3 and 4, and Tables 6–8. Fig. 3 uses scatter plot analysis to compare model simulated and observed streamflow at the calibration site (gauge *F*) for the calibration and the verification durations (i.e., 2000–2006) when NSE and MNSD were used as objective functions. Fig. 4 shows similar information for one internal site (gauge *K*). The perfect fit line and the lower and upper bounds of the 95% confidence interval are also given in both Figs. Table 6 shows values of the 13 efficiency criteria for the calibration period and the verification period at gauges *F* and *J* using solutions obtained when NSE, MNSD, and LNSD were used as objective functions. Table 7 compares percent deviation of the known parameters values from the optimal parameters values identified using RMS, NSE, MNSD, and LNSD as objective functions. The known parameters values are also given in Table 7.

Table 6. Efficiency Criteria Values Obtained for Case II for All Months

Criteria	MAE	RMSE	NSE	MNSD	LNSD	VE	PBIAS	RSR	R^2	*D*
					Calibration period					
Gauge *F*										
NSE	0.05	0.11	**1.00**	0.95	**1.00**	0.94	1.41	0.00	0.99	1.00
MNSD	**0.03**	**0.09**	**1.00**	**0.97**	**1.00**	**0.96**	**−0.37**	0.00	**1.00**	1.00
LNSD	0.04	0.13	0.99	0.96	**1.00**	0.95	0.63	0.00	0.99	1.00
Gauge *J*										
NSE	0.02	0.04	0.99	0.93	0.98	0.88	0.63	0.00	0.98	1.00
MNSD	**0.01**	0.04	0.99	0.94	**0.99**	0.90	**0.16**	0.00	0.99	1.00
LNSD	**0.01**	**0.03**	**1.00**	**0.96**	**0.99**	**0.92**	3.72	0.00	0.99	1.00
					Verification period					
Gauge *F*										
NSE	0.07	0.29	0.99	0.94	0.99	0.93	0.32	0.00	0.99	1.00
MNSD	**0.05**	**0.18**	**1.00**	**0.97**	**1.00**	**0.96**	1.25	0.00	**1.00**	1.00
LNSD	0.06	0.24	0.99	0.95	**1.00**	0.94	**0.24**	0.00	**1.00**	1.00
Gauge *J*										
NSE	0.02	0.05	0.99	0.94	0.98	0.93	**−0.68**	0.00	0.98	1.00
MNSD	**0.01**	**0.04**	**1.00**	**0.96**	0.99	**0.95**	0.87	0.00	0.98	1.00
LNSD	**0.01**	**0.04**	**1.00**	**0.96**	**1.00**	**0.95**	1.18	0.00	0.98	1.00

Note: Table shows efficiency criteria values calculated using optimal solution identified when criterion shown in rows is used as objective function during optimization; MAE is in m^3/s; RMS is in m^3/s; PBIAS is in %; other efficiency criteria are unitless.

Effectiveness as Objective Function

Figs. 3 and 4 and Table 6 show that all objective functions performed excellent. The efficiency criteria values given in Table 6 are very close to their respective ideal values and the figures show almost perfect fit between the simulated and observed streamflow. Objective functions in Group I (e.g., MNSD) seem to have slightly outperformed objective functions in Group II (e.g., NSE). Unlike in Case I, objective functions in Group III (e.g., LNSD)performed very well in Case II. Overall, however, MNSD (a representative of Group I) either outperformed or produced very comparable results to the other objective functions.

Uniqueness of Optimal Parameters

Table 7 reveals several interesting findings. From observation of Figs. 3 and 4 and Table 6 it can be easily concluded, as done in the preceding paragraph, that the model performance is excellent. As such, it is expected that the parameters values identified by the calibration process be very close to the known values as well. However, Table 7 shows that the percent deviation observed between the identified optimal values and the known values is significantly high for many of the calibrated parameters. The smallest percent deviation, and hence the best result, observed for each parameter has been given in bold. Interestingly, no clear superiority was observed among the objective functions with regard to producing parameters values that are the closest to their respective known values. The "Within ±10%" column shows the number of parameters (out of the 12 calibrated parameters) whose optimal value is within ±10% of the respective known value. In that regard, MNSD and LNSD performed better than the other objective functions. Another interesting observation from Table 7 is that RMS and NSE, both of Group II, produced very different optimal parameters values.

Comparison of Tables 6 and 7 leads to the following important conclusions. First, even for ideal modeling conditions, such as Case II, where uncertainties from input data, observed data and

Table 7. Known Values and Percent Deviation of Optimal Parameter Values from Known Values for Case II

Criteria	α_Bf	Canmx	Ch_K2	Ch_N2	Cn2	Esco	Gwqmn	Slope	Sol_Awc	Sol_K	Sol_Z	Surlag	Within ±10%
						Wet season							
Known values	0.10	0.05	53.07	0.08	8.18	0.94	23.88	19.4	22.86	9.65	49.85	0.59	
RMS	**−0.2**	9879.0	24.4	38.3	57.9	6.1	9,016.6	**−4.1**	0.6	−157.4	**0.2**	**−28.5**	5
NSE	**−0.2**	5117.0	22.5	40.0	54.1	3.7	**−339.1**	108.4	0.6	**−143.9**	6.9	**−28.5**	4
MNSD	5.4	1378.4	15.0	5.6	**0.6**	**1.9**	7,699.4	125.7	1.2	409.9	1.1	42.5	6
LNSD	−5.1	**−92.7**	**9.5**	**−5.1**	37.0	8.2	8,992.7	143.3	**0.3**	306.8	5.8	32.4	6
						Dry season							
Known values	0.98	9.89	84.71	0.09	17.3	0.02	4995.70	19.4	22.86	9.65	49.85	4.94	
RMS	86.4	5,250.4	39.6	0.4	31.8	**0.7**	590.1	**−4.1**	0.6	157.4	**0.2**	438.3	5
NSE	83.8	6.2	12.4	4.4	1.0	22.3	43.8	108.4	0.6	**−143.9**	6.9	35.1	5
MNSD	28.1	3.4	**1.8**	1.1	0.3	97.2	**−0.2**	125.7	1.2	409.9	1.1	**30.1**	7
LNSD	**−15.5**	**0.7**	3.2	**0.3**	**−0.1**	225.1	0.4	143.3	**0.3**	306.8	5.8	44.9	7

Note: "Within ±10%" column indicates number of parameters whose optimal values are within ±10 percent of respective known values when corresponding efficiency measure is used as objective function; see Table 2 for unit of known values.

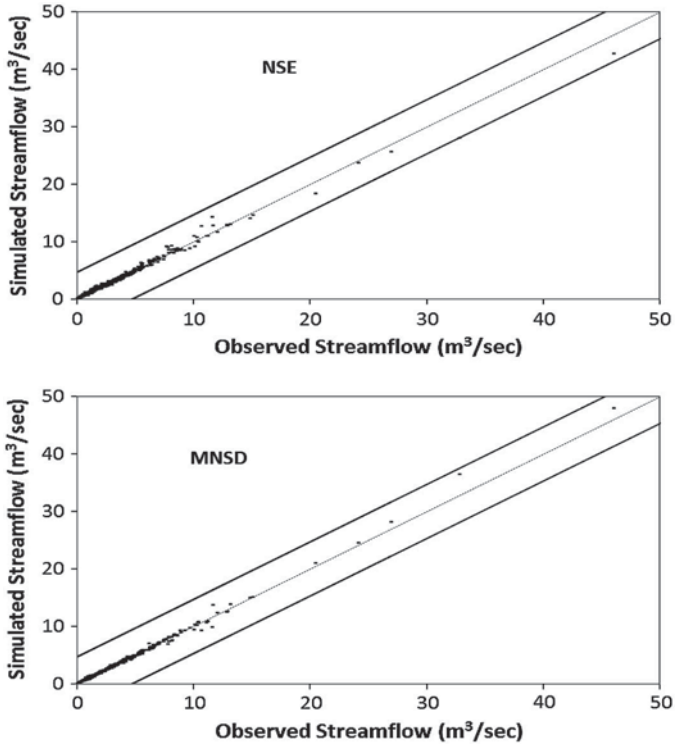

Fig. 3. Comparison of observed and simulated streamflow obtained at gauge *F* for case II when NSE and MNSD are used as objective function; light dash line shows perfect fit line and darker lines show bounds of 95% confidence interval

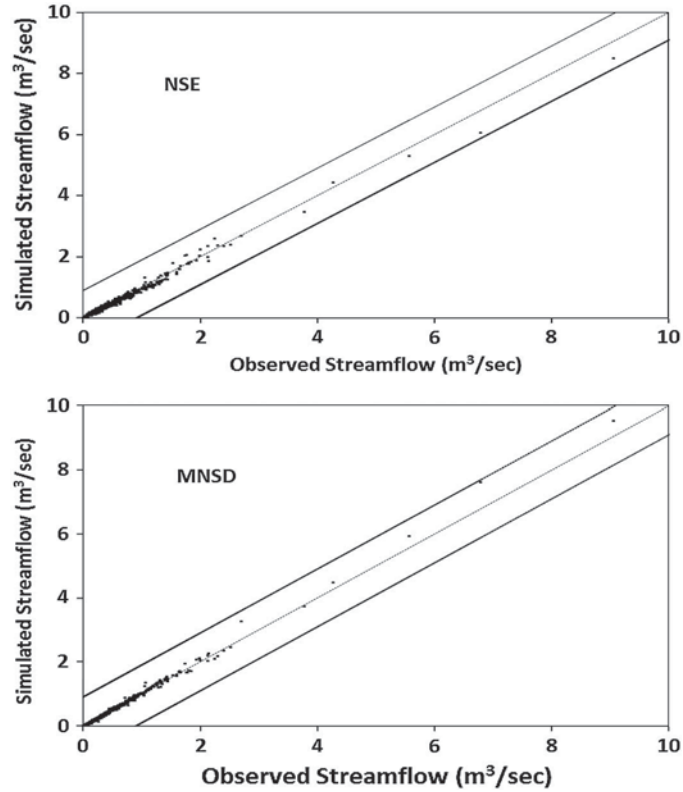

Fig. 4. Comparison of observed and simulated streamflow obtained at gauge *K* for case II when NSE and MNSD are used as objective function; light dash line shows perfect fit line and darker lines show bounds of 95% confidence interval

model structure are fully accounted for (which is very difficult, if not impossible, to achieve for real world applications), obtaining efficiency criteria values that are close to their respective ideal values does not guarantee optimal parameters values that are close to the known (if any) parameter values. Here it should be emphasized that there is no such thing as known parameter value in real world applications. Second, optimal parameters values are not unique as values that are very different from the known values produced streamflow simulations that could be considered excellent when judged by 13 different efficiency criteria. This conclusion is consistent with the equifinality concept of Beven and Freer (2001), which argues that multiple sets of parameters can simulate watershed characteristics reasonably well. Third, because parameters values that are significantly divergent from the known values produced very accurate streamflow simulations, for practical applications, the minor limitations exhibited in calibration methods (e.g., calibration algorithm used, objective function selection) in identifying the global optimal parameters values may contribute only minimally to the overall hydrologic modeling uncertainty. Most of the prediction uncertainty experienced in hydrologic modeling can be attributed to the uncertainty from input and output data (i.e., poor quality and/or insufficient quantity of input and output data that do not accurately represent spatial and temporal heterogeneity of the watershed characteristics) and uncertainty from model structure (i.e., incapability of the hydrologic model to accurately characterize the watershed). The third conclusion is consistent with Kuczera et al. (2006) who pointed out that "…poorly determined parameters do not necessarily lead to high predictive uncertainty."

Table 8. Intercorrelation Coefficient among Various Efficiency Criteria

Criteria	NSE	NSD	MNS	MNSD	LNS	LNSD	VE	AbsPBIAS	RSR	R^2	*D*
NSE	**1.00**	**1.00**	**0.83**	**0.83**	0.25	0.60	**0.79**	**−0.77**	**−0.92**	0.43	**0.82**
NSD		**1.00**	**0.83**	**0.83**	0.24	0.60	**0.79**	**−0.76**	**−0.90**	0.45	**0.82**
MNS			**1.00**	**1.00**	0.63	**0.90**	**0.99**	0.67	**−0.88**	**0.74**	**0.92**
MNSD				**1.00**	0.63	**0.89**	**0.99**	0.66	**−0.88**	**0.75**	**0.93**
LNS					**1.00**	**0.88**	0.66	0.00	0.40	**0.77**	0.58
LNSD						**1.00**	**0.92**	0.37	0.67	**0.84**	**0.82**
VE							**1.00**	0.61	**−0.83**	**0.80**	**0.93**
AbsPBIAS								**1.00**	**0.76**	0.07	0.50
RSR									**1.00**	0.48	**−0.82**
R^2										**1.00**	**0.84**
D											**1.00**

Robustness in Describing Model Performance

Relative robustness of the efficiency criteria in describing goodness of model performance was further examined using the intercorrelation coefficients given in Table 8. The table shows intercorrelation coefficients among the efficiency criteria determined from the results of Case I and Case II for the calibration period and the verification period. Total of 180 sets of efficiency criteria values (i.e., 90 from Case I and 90 from Case II) were used to determine the coefficients. For each case, values of the efficiency criteria were calculated at five sites (i.e., gauges F, I, J, K, and M) for each of the nine objective functions. This leads to 45 sets of efficiency criteria values (i.e., results of nine objective functions at five sites) for the calibration period and another 45 sets for the verification period producing 90 data sets per case.

Clearly, the efficiency criteria highly correlated to most of the other efficiency criteria is deemed the most robust in describing model performance. Intercorrelation coefficients with magnitude ≥ 0.70 are considered good in this study and are shown in bold in Table 8. MAE and RMS have been excluded from the intercorrelation analysis because of their scale dependence (i.e., MAE and RMS values depend on how big or small a river is at the gauge site) as results from five different gauges (i.e., F, I, J, K, and M) were used to calculate the correlation coefficients. Table 8 indicates that except for LNS and AbsPBIAS (i.e., absolute value of PBIAS), all the other efficiency criteria are well correlated. MNS, MNSD, and D are best correlated to the other efficiency criteria. MNSD was least correlated to LNS (correlation coefficient $= 0.63$) and to AbsPBIAS (correlation coefficient $= 0.66$), which still show reasonably good correlation. As described in Case I, D alone is not a suitable efficiency criterion to report model results as poor models could yield attractive D values. Therefore, MNS and MNSD are the most robust efficiency criteria to report model results among the efficiency criteria considered for the study. MNS and MNSD also outperformed the other efficiency criteria as objective functions for both Case I and Case II.

Conclusions

The study investigated effectiveness of the most commonly used efficiency criteria for use as objective function during automated calibration and examined their robustness in describing the goodness of model performance. Two cases were considered to accomplish the research objectives. In Case I, actual data were used to build SWAT model for the watershed and calibrations were performed using nine different objective functions. In Case II, the optimal parameters values identified during the calibration performed in Case I using NSE as objective function were used to simulate streamflow at gauges F, I, J, K, and M using SWAT. The simulated streamflow was then considered as observed data and calibration attempts were repeated using nine objective functions. Case II was designed to eliminate uncertainties from input data, observed data used for calibration, and model structure and to further investigate relative effectiveness of the objective functions under ideal condition.

Major conclusions of the study are (1) automated calibration results are sensitive to the objective function group; group that work based on minimization of the absolute deviations (Group I); group that work based on minimization of square of the residuals (Group II); and groups that use log of the observed and simulated streamflow values (Group III), but not to objective functions within the group; (2) efficiency criteria that belong to Group I (i.e., MAE, MNS, MNSD, and VE) were the most effective when used as objective function for both low flows and high flows. Based on these two conclusions, either of MAE, MNS, MNSD, and VE is recommended as objective function for single objective calibration applications; (3) objective functions in Group I and those in Group II complement each other's performance quite well. This implies that accuracy of multiobjective calibrations can be improved by using one objective function from each group; (4) with regard to the capability to describe the goodness of model simulations, except for LNS and LNSD, the other 11 efficiency criteria behaved consistently (i.e., a calibration result would either improve or deteriorate all the 11 measures consistently). The intercorrelation analysis, however, showed relative robustness of MNS and MNSD in describing model performance because the two were fairly well correlated (correlation coefficient ≥ 0.6) to all other efficiency criteria considered for the correlation analysis; (5) the study also showed that for the Little River watershed, use of long-term interannual daily mean as baseline model for NSE and MNS did not improve capability of the measures to describe model performance. However, this finding is likely to be watershed specific; (6) even for ideal conditions where uncertainty in input data and model structure are fully accounted for, as in Case II of this study, identifying the so-called *global* parameter values through calibration exercise could be daunting as the parameters values exhibited nonuniqueness. Parameter values that were significantly divergent from the known (global) values produced model performance that may be considered near perfect even when judged using multiple efficiency criteria; (7) the performance gap exhibited between the calibration attempts made for Case I (satisfactory but not very impressive performance) and Case II (excellent performance) can be attributed to uncertainty arising from errors in input data, the observations used for calibration, and the model structure. In other words, more accurate input data and streamflow observations and model that more accurately describes characteristics of the watershed would reduce predictive uncertainty. As observed from Case II, parameter uncertainty (i.e., inability to identify the global (known) parameters values) contributed minimally toward the predictive uncertainty. Therefore, most of the predictive uncertainty experienced in hydrologic modeling may be attributed to inputs, outputs, and model structures. This implies that attention of hydrologists needs to shift toward building models that are better representative of the watershed characteristics and collecting and utilizing more reliable data. If properly applied, the automated calibration approaches practiced in hydrology today are good enough to reduce parameter uncertainty if more accurate data and models are available.

Acknowledgments

Source code for DDS (Tolson and Shoemaker 2007) was obtained from the first writer of the program.

References

Arnold, J. G., Williams, J. R., Srinivasan, R., and King, K. W. (1999). *SWAT: Soil and water assessment tool*, U.S. Dept. of Agriculture, Agricultural Research Service, Temple, TX.

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

Bekele, E. G., and Nicklow, J. W. (2007). "Multi-objective automated calibration of SWAT using NSGA-II." *J. Hydrol. (Amsterdam)*, 341(3–4), 165–176.

Beven, K. J., and Freer, J. (2001). "Equifinality, data assimilation, and uncertainty estimation in mechanistic modeling of complex environmental systems using the GLUE methodology." *J. Hydrol. (Amsterdam)*, 249(1–4), 11–29.

Bosch, D. D., et al. (2007). "Little river experimental watershed database." *Water Resour. Res.* 43(9), W09470.

Bosch, D. D., Sullivan, D. G., and Sheridan, J. M. (2006). "Hydrologic impacts of land-use changes in coastal plain watersheds." *Trans. ASABE*, 49(2), 423–432.

Criss, R. E., Winston, W. E. (2008). "Do Nash values have value? Discussion and alternate proposals." *Hydrol. Processes*, 22(14), 2723–2725.

Efstratiadis, A., and Koutsoyiannis, D. (2010). "One decade of multiobjective calibration approaches in hydrological modelling: A review." *Hydrol. Sci. J.*, 55(1), 58–78.

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

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martineza, G. F. (2009). "Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modeling." *J. Hydrol. (Amsterdam)*, 377(1–2), 80–91.

Gupta, H. V., Sorooshian, S., and Yapo, P. O. (1998). "Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information." *Water Resour. Res.*, 34(4), 751–763.

Jain, S. K., and Sudheer, K. P. (2008). "Fitting of hydrologic models: A close look at the Nash-Sutcliffe index." *J. Hydrologic Eng.*, 13(10), 981–986.

Krause, P., Boyle, D. P., and Bäse, F. (2005). "Comparison of different efficiency criteria for hydrological model assessment." *Adv. Geosci.*, 5, 89–97.

Kuczera, G., Kavetski, D., Franks, S., and Thyer, M. (2006). "Towards a bayesian total error analysis of conceptual rainfall-runoff models: Characterizing model error using storm-dependent parameters." *J. Hydrol. (Amsterdam)*, 331(1–2), 161–177.

Legates, D. R., and McCabe, G. J. (1999). "Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model evaluation." *Water Resour. Res.*, 35(1), 233–241.

Li, H., Sivapalan, M., and Tian, F., (2012). "Comparative diagnostic analysis of runoff generation processes in Oklahoma DMIP2 basins: The Blue River and the Illinois River." *J. Hydrol. (Amsterdam)*, 418–419, 90–109.

Madsen, H. (2000). "Automated calibration of a conceptual rainfall-runoff model using multiple objectives." *J. Hydrol. (Amsterdam)*, 235(3–4), 276–288.

McCuen, R. H., Knight, Z., and Gillian, C. A. (2006). "Evaluation of the Nash-Sutcliffe efficiency index." *J. Hydrol. Eng.*, 11(6), 597–602.

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

Muleta, M. K. (2012). "Improving model performance using season based evaluation." *J. Hydrol. Eng.*, 17(1), 191–200.

Muleta, M. K., and Nicklow, J. W. (2005). "Sensitivity and uncertainty analysis coupled with automated calibration for a distributed watershed model." *J. Hydrol. (Amsterdam)*, 306(1–4), 127–145.

Nash, J. E., and Sutcliffe, J. V. (1970). "River flow forecasting through conceptual models. Part I—A discussion of principles." *J. Hydrol. (Amsterdam)*, 10(3), 282–290.

Neitsch, S. L., Arnold, J. G., and Williams, J. R. (2005). *Soil and water assessment tool theoretical documentation—Version 2005*, Grassland, Soil and Water Research Laboratory, Temple, TX.

Schaefli, B., and Gupta, H. V. (2007). "Do Nash values have value?." *Hydrol. Processes*, 21(15), 2075–2080.

Schaefli, B., Hingray, B., Musy, A. (2005). "A conceptual glaciohydrological model for high mountainous catchments." *Hydrol. Earth Syst. Sci.*, 9(1–2), 95–109.

Seibert, J. (2001). "On the need for benchmarks in hydrological modeling." *Hydrol. Processes*, 15(6), 1063–1064.

Sheshukov, A., Daggupati, P., Lee, M. C., Douglas-Mankin, K. (2009). "ArcMap tool for pre-processing SSURGO soil database for ArcSWAT." *Proc., 5th Int. SWAT Conf.*, Texas A&M University, College Station, TX.

Sobol', I. M. (1993). "Sensitivity estimates for non-linear mathematical models." *Math. Model. Comput. Exp.*, 1(4), 407–414.

Sorooshian, S., Gupta, V. K., and Fulton, J. L. (1983). "Evaluation of maximum likelihood parameter estimation techniques for conceptual rainfall-runoff models: Influence of calibration data variability and length on model credibility." *Water Resour. Res.*, 19(1), 251–259.

Southeast Watershed Research Laboratory (SEWRL). (2010). "Little River experimental watershed data." ⟨ftp://www.tiftonars.org/⟩ (July 2010).

Tang, Y., Reed, P. M., van Werkhoven, K., and Wagener, T. (2007). "Advancing the identification and evaluation of distributed rainfall-runoff models using global sensitivity analysis." *Water Resour. Res.*, 43(6), W06415.

Tang, Y., Reed, P. M., and Wagener, T. (2006). "How effective and efficient are multiobjective evolutionary algorithms at hydrologic model calibration?." *Hydrol. Earth Syst. Sci.*, 10(2), 289–307.

Tian, F., Li, H., and Sivapalan, M. (2012). "Model diagnostic analysis of seasonal switching of runoff generation mechanisms in the Blue River basin, Oklahoma." *J. Hydrol. (Amsterdam)*, 418–419, 136–149.

Tolson, B. A., and Shoemaker, C. A. (2007). "Dynamically dimensioned search algorithm for computationally efficient watershed model calibration." *Water Resour. Res.*, 43(1), W01413.

U.S. Environmental Protection Agency (U.S. EPA). (2002). "Guidance for quality assurance project plans for modeling." *EPA QA/G-5M. Rep. EPA/240/R-02/007*, Washington, DC.

van Werkhoven, K., Wagener, T., Reed, P., and Tang, Y. (2008). "Rainfall characteristics define the value of streamflow observations for distributed watershed model identification." *Geophys. Res. Lett.*, 35(11), L11403.

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

Winchell, M., Srinivasan, R., Di Luzio, M., and Arnold, J. (2008). *ArcSWAT 2.1 Interface for SWAT 2005*, User's Guide: Blackland Research Center, Temple, TX.

Yan, J., Haan, C. T. (1991). "Multiobjective parameter estimation for hydrologic models—Weighting of errors." *Trans. ASABE*, 34(1), 0135–0141.

Yapo, P. O., Gupta, H. V., and Sorooshian, S. (1998). "Multi-objective global optimization for hydrologic models." *J. Hydrol. (Amsterdam)*, 204(1–4), 83–97.