Tradeoffs among watershed model calibration targets for parameter estimation

Katie Price,[1] S. Thomas Purucker,[1] Stephen R. Kraemer,[1] and Justin E. Babendreier[1]

Received 16 February 2012; revised 12 September 2012; accepted 18 September 2012; published 24 October 2012.

[1] Hydrologic models are commonly calibrated by optimizing a single objective function target to compare simulated and observed flows, although individual targets are influenced by specific flow modes. Nash-Sutcliffe efficiency (NSE) emphasizes flood peaks in evaluating simulation fit, while modified Nash-Sutcliffe efficiency (MNS) emphasizes lower flows, and the ratio of the simulated to observed standard deviations (RSD) prioritizes flow variability. We investigated tradeoffs of calibrating streamflow on three standard objective functions (NSE, MNS, and RSD), as well as a multiobjective function aggregating these three targets to simultaneously address a range of flow conditions, for calibration of the Soil and Water Assessment Tool (SWAT) daily streamflow simulations in two watersheds. A suite of objective functions was explored to select a minimally redundant set of metrics addressing a range of flow characteristics. After each pass of 2001 simulations, an iterative informal likelihood procedure was used to subset parameter ranges. The ranges from each best-fit simulation set were used for model validation. Values for optimized parameters vary among calibrations using different objective functions, which underscores the importance of linking modeling objectives to calibration target selection. The simulation set approach yielded validated models of similar quality as seen with a single best-fit parameter set, with the added benefit of uncertainty estimations. Our approach represents a novel compromise between equifinality-based approaches and Pareto optimization. Combining the simulation set approach with the multiobjective function was demonstrated to be a practicable and flexible approach for model calibration, which can be readily modified to suit modeling goals, and is not model or location specific.

Citation: Price, K., S. T. Purucker, S. R. Kraemer, and J. E. Babendreier (2012), Tradeoffs among watershed model calibration targets for parameter estimation, *Water Resour. Res.*, *48*, W10542, doi:10.1029/2012WR012005.

1. Introduction

[2] In simulating streamflow from watershed characteristics, physically based watershed hydrologic models benefit from some degree of calibration prior to use. Most modeling efforts involve analysis of a suite of parameters that are modified to affect various watershed processes. These parameters act as "tuning knobs" to adjust output, ideally such that the model closely simulates the watershed's observed hydrological behavior [*Madsen*, 2000]. Because the parameters generally have ostensible physical meaning (related to groundwater storage capacity, runoff potential, etc.), there are bounds within which parameter values must fall to be realistic for any given study area. There is ample discussion in the literature regarding the pitfalls of model calibration and the importance of minimizing subjectivity and overfitting during the calibration process [*Beven*, 2006; *Efstratiadis and Koutsoyiannis*, 2010]. Expert-guided, manual calibration processes are time consuming and require extensive familiarity with model operation, structure, and underlying assumptions [*Lindström*, 1997; *Madsen et al.*, 2002; *Ewen*, 2011]. Reproducible model evaluation is desirable for management and policy [*Matott et al.*, 2009], and manual calibration is potentially subjective. Given these complications, most hydrologists have come to favor automated calibration processes [*Madsen et al.*, 2002], which typically involve some type of Monte Carlo sampling to explore combinations of calibration parameters across defined ranges. While eliminating many of the problems associated with manual calibration, autocalibration is accompanied by other drawbacks. Most notably, it may lead to drastically increased computational demands (due to large numbers of simulation replicates), unrealistic parameter values or combinations resulting from inadequate parameter specification or model structure, and greater challenges to model output interpretation due to issues of equifinality [*Beven and Binley*, 1992; *Beven*, 2006].

[3] Methodological advancements for process-based Monte Carlo simulations have resulted in more efficient searches over a multidimensional parameter space to more

[1] Ecosystems Research Division, National Exposure Research Laboratory, Office of Research and Development, U.S. Environmental Protection Agency, Athens, Georgia, USA.

Corresponding author: K. Price, Ecosystems Research Division, National Exposure Research Laboratory, Office of Research and Development, U.S. Environmental Protection Agency, 960 College Station Rd., Athens, GA 30605, USA. (price.katie@epa.gov)

quickly find a "best" maximum likelihood parameter set. While a number of effective optimization functions are available for finding the parameter set with the highest likelihood, recent advances for efficiently exploring the posterior likelihood distribution have been developed for rejection sampling [*Robert and Casella*, 2004], Markov chain Monte Carlo (MCMC) methods [*Andrieu et al.*, 2003], and particle filtering [*Arulampalam et al.*, 2002]. These efficient search methods are often combined with a summary statistic-based goodness of fit function, rather than the joint likelihood product of all the observations [*Hartig et al.*, 2011], and a likelihood-based inference strategy suitable for stochastic simulations [e.g., *Beaumont*, 2010; *Beven*, 2006; *Grimm et al.*, 2005; *Wood*, 2010]. For non-MCMC applications, Latin hypercube sampling (LHS) has been shown to ameliorate the drawback of increased computational requirements. In LHS, the range for each parameter is subdivided and a sample value is taken from each division. This reduces the number of simulations required to explore a representation of the full parameter space, compared to truly random Monte Carlo approaches [*Uhlenbrook and Sieber*, 2005].

[4] Calibrations using a single optimized parameter set are susceptible to over-fitting, in that the single best fit may represent such a tailored function for the calibration data set that validation fits and prediction accuracy suffer. Despite hydrologists' long-standing awareness of equifinality and other weaknesses associated with single "best-fit" calibrated parameter sets [*Beven*, 1993; *Yapo et al.*, 1998, *Uhlenbrook et al.*, 1999; *Matott et al.*, 2009], the use of simulation sets (e.g., confidence bands, ensembles, likelihood distributions, etc.) remains under-explored. Several studies have demonstrated that simulation sets are an efficient and successful approach [e.g., *Thiemann et al.*, 2001; *Uhlenbrook and Sieber*, 2005]. In addition, Pareto efficient subsets can be drawn from simulation sets to find parameter combinations that are not "dominated" by other members of the simulation set [*Goldberg*, 1989]. Despite the availability of simulation set and Pareto approaches, single best-fit simulation approaches for model calibration continue to be more commonly used [*Beven*, 2006]. In this study we present a novel compromise between these two calibration strategies.

[5] While marked advancements have been made in model calibration methods, there remains a major limitation: no single calibration target meets the needs of all modeling applications [*Yapo et al.*, 1998; *Matott et al.*, 2009; *Efstratiadis and Koutsoyiannis*, 2010]. Calibration relies on calculating an objective function (typically a goodness-of-fit statistic), which relates simulated flows to observed flows from a gauged watershed outlet. Commonly used individual objective functions to compare simulated and observed flows are the R^2 coefficient of determination, the Nash-Sutcliffe Efficiency (NSE), and the index of agreement *d* [*Nash and Sutcliffe*, 1970; *Willmott*, 1981]. These metrics square the difference between simulated and observed flows, with the effect of emphasizing peak flows in model calibration [*Legates and McCabe*, 1999]. To attenuate or eliminate overemphasis of large errors on fit scores, modified forms of NSE and *d* have been devised to increase sensitivity to lower values [*Krause et al.*, 2005]. Given the much larger proportion of low flows, compared to high flows in natural flow regimes, modified forms effectively bias low flows.

[6] Another important calibration target relates to flow variability. Neither standard nor modified forms of R^2, NSE, *d*, or similar metrics, explicitly address the success of simulated flows in replicating dynamics of observed streamflow. Such variability is critically important to aquatic habitat, sustainable water use, and other ecosystem services considerations and is, thus, central to the concept of environmental flows [*O'Keeffe*, 2009; *Poff and Zimmerman*, 2010]. While many metrics quantify flow variability over long time scales [e.g., *Gao et al.*, 2009], long-term summary statistics cannot be used in place of a fit statistic for daily streamflow calibration. Several calibration targets have been suggested to meet this modeling need, generally incorporating the standard deviation of flow as part of the objective function [e.g., *Moriasi et al.*, 2007].

[7] The aforementioned individual objective functions may well serve highly specific modeling goals; however, many practitioners seek to simulate streamflows in a manner that reasonably represents multiple hydrograph response modes simultaneously [*Madsen et al.*, 2002; *Fenicia et al.*, 2007]. In such applications, modelers may be willing to accept suboptimal performance of one aspect of flow in order to improve accuracy in one or more other flow modalities [*Tekleab et al.*, 2011]. Calibration approaches relying on multiple simultaneous objective functions, hereafter referred to as multiobjective functions, have become increasingly common in recent years [*Gupta et al.*, 2009; *Matott et al.*, 2009; *Efstratiadis and Koutoyian*nis, 2010]. Thus, our objectives were to (1) compare model results using a suite of objective functions for calibration, (2) explore a calibration scheme not linked to any one model or study area that aggregates these into a multiobjective function, while allowing for user-defined fit criteria, and (3) develop a methodology that uses strengths from equifinality-based and Pareto optimization approaches. To achieve this, we applied the Soil and Water Assessment Tool (SWAT) watershed model to two adjacent watersheds, with separate calibrations performed for each calibration target in each watershed. We hypothesized that calibration targets known to emphasize certain aspects of flow (e.g., peak flows) would result in more accurate streamflow simulations for that particular response mode, while proving less accurate in others. Furthermore, we expected to see that calibrations based on calibrated simulation sets would result in superior validations and, by extension, better predictive models and/or spatial extrapolations, than calibrations using a single best-fit approach.

2. Methods

[8] For this comparative analysis of model calibration strategies, we simulated streamflows for two watersheds, Mountain Creek (21 km^2) and Little River (201 km^2). These are located in the state of North Carolina in the upper Neuse River system (Figure 1) within the Piedmont physiographic province, which is characterized by moderate relief, crystalline bedrock, and saprolite, with deep, weathered soils [*Mills et al.*, 1987]. Climate normals (1981–2010) recorded at the nearby Durham weather station are 1220 mm annual precipitation and average January and July temperatures of 3.2 and 25.3 °C, respectively [*National Climate Data Center* (*NCDC*), 2011].

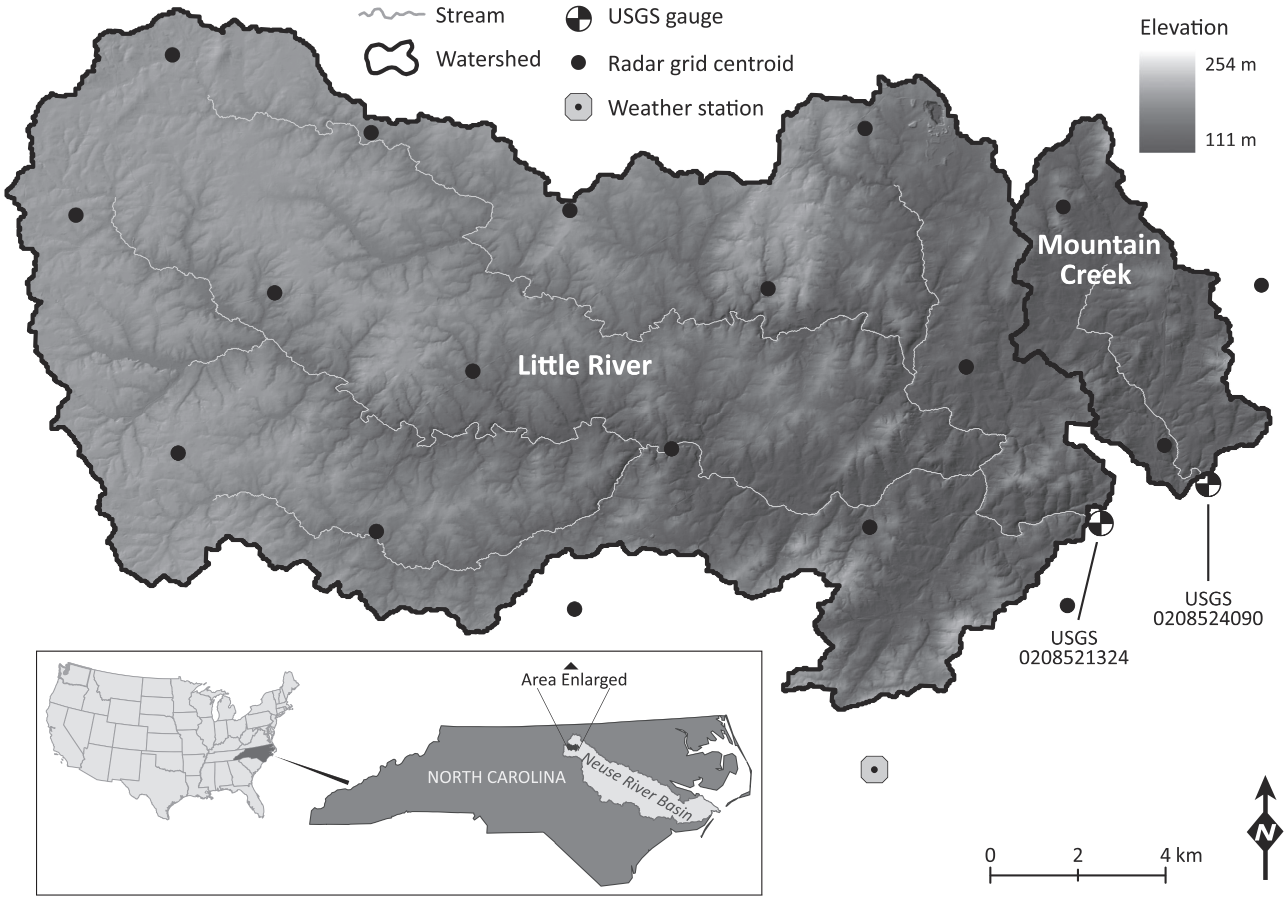

Figure 1. Study watersheds. Mountain Creek (21 km^2) and Little River (201 km^2), subwatersheds of the Neuse River system in the North Carolina Piedmont. Radar grid centroids correspond to 4 × 4 km multisensor precipitation estimator (MPE) pixels. The weather station is COOP 312515, Durham, NC, which was the source for maximum and minimum daily temperatures. Data were obtained from *NCAR Earth Observing Laboratory* [2011] and *NCDC* [2011].

The 2006 land use in Mountain Creek and Little River was predominantly forest (52% and 60%, respectively) and agriculture (40% and 34%), with smaller proportions of low to medium density developed (9% and 5%), and wetland (1% each) [*Fry et al.*, 2011].

[9] SWAT 2009 [*Neitsch et al.*, 2011] was used to simulate daily streamflows from 2001 to 2010. Streamflow was calibrated to observed flows from USGS 0208521324 (Little River at SR1461 near Orange Factory, NC) and USGS 0208524090 (Mountain Creek at SR1617 near Bahama, NC), located at the outlets of the two simulation watersheds (Figure 1). SWAT was chosen because it is a widely used tool within the U.S. Environmental Protection Agency (EPA) watershed modeling efforts, and is a very commonly used and well-supported model [*Easton et al.*, 2008]. Preprocessing and model setup were performed using the ArcSWAT extension for ArcGIS 9.x [*Winchell et al.*, 2007]. The 2001 simulation year output was discarded as model spin-up, and the last nine years (2002–2010) were retained for model evaluation. We used a split-sample approach for calibration and validation, which requires using sequential years within each period and accepting temporal autocorrelation [*Klemes*, 1986]. Our nine-year simulation period was split into a six-year calibration period (2002–2007) and a three-year validation period (2008–2010). Doppler radar-derived NEXRAD Stage IV Multisensor Precipitation Estimator (MPE) precipitation data were obtained from the *NCAR Earth Observing Laboratory* [2011], aggregated to a daily time step, and converted to time series for SWAT [*Price et al.*, 2011]. The centroid of each 4 × 4 km MPE pixel was treated as a virtual precipitation gauge in SWAT (Figure 1). Daily maximum and minimum temperature values were obtained from the National Climatic Data Center (NCDC), and SWAT's weather generator was used to estimate all other meteorological variables. Other input data requirements for SWAT included spatial coverages of soil, land use, and topography. Soil data from the U.S General Soil Map (STATSGO2) are available within ArcSWAT and were used in this analysis [*Soil Survey Staff*, 2011]. The 2006 National Land Cover Database was obtained from the Multi-Resolution Land Characteristics Consortium and combined with the USDA's 2009 Cropland Data Layer to represent land use [*Fry et al.*, 2011; *NASS*, 2011]. A 30 m DEM was obtained from the USGS seamless spatial data server. All default options for SWAT operation were used, with the exception that the temperature-based Hargreaves

option for evapotranspiration estimation was used in place of the Penman-Monteith equation. Hargreaves produced much better uncalibrated fits between simulated and observed flows, particularly at low flows, as has been observed in other SWAT applications [*Wang et al.*, 2006; *Setegn et al.*, 2008].

[10] Four SWAT calibrations were performed for each watershed, with identical data sets of land use, topography, soils, and temperature (i.e., only precipitation input varied within each watershed). We employed an importance sampling parameter estimation approach [*Matott et al.*, 2009], with separate parameter-updating calibration processes implemented to optimize three single-objective objective functions and one composite multiobjective function. A suite of 19 individual objective functions was originally considered, including the 18 metrics produced by the hydroGOF package for *R* statistical software [*Bigiarni*, 2010; *R Development Core Team*, 2011], along with the inverse error variance, as presented in *Beven and Binley*'s [1992] generalized likelihood uncertainty estimation (GLUE) methodology. To avoid selecting metrics that possess similar information, cross-correlation analysis was performed to avoid metric sets with high interdependence. Preference was given to nonredundant metrics that are known to hydrologists. The selected single objective functions (NSE, MNS, and RSD) had relatively low cross correlation and represent three distinct aspects of streamflow dynamics (high flows, low flows, and flow variability, respectively). Equations for NSE, MNS, and RSD for each simulation *i* are presented below, with *O* and *S* indicating observed and simulated flow values, respectively, time steps *t* ranging from 1 to *n*, and σ representing the standard deviation:

$$\mathrm{NSE}_i = \frac{\sum_{t=1}^{n} (O_t - S_t)^2}{\sum_{t=1}^{n} (O_t - \overline{O})^2}, \tag{1}$$

$$\mathrm{MNS}_i = \frac{\sum_{t=1}^{n} |O_t - S_t|}{\sum_{t=1}^{n} |O_t - \overline{O}|}, \tag{2}$$

$$\mathrm{RSD}_i = \frac{\sigma_s}{\sigma_o}. \tag{3}$$

In addition, an aggregated multiobjective function compositing the three selected metrics was created, in order to allow tradeoffs in fitting high flow, low flow, and flow variability components. Termed the composite likelihood index (hereafter referred to as CL), this index is an example of a "classical aggregation approach" to multiobjective calibration [*Efstratiadis and Koutsoyiannis*, 2010]. CL equally weights these three metrics and requires a similar scale for each so they can be coerced to a scalar for likelihood assignment. Therefore, CL components (NSE, MNS, and RSD) were transformed to a consistent range for aggregation:

$$\theta_{\mathrm{NSE}} = \frac{\max(0, \mathrm{NSE}_i)}{\sum_{i=1}^{n} \max(0, \mathrm{NSE}_i)}, \tag{4}$$

$$\theta_{\mathrm{MNS}} = \frac{\max(0, \mathrm{MNS}_i)}{\sum_{i=1}^{n} \max(0, \mathrm{MNS}_i)}, \tag{5}$$

$$\theta_{\mathrm{RSD}} = \frac{1 - \min(1, |1 - \mathrm{RSD}_i|)}{\sum_{i=1}^{n} [1 - \min(1, |1 - \mathrm{RSD}_i|)]}. \tag{6}$$

Where *i* represents each of 2001 SWAT runs in the simulation set. The resulting component scores potentially range from 0 (worst fit to observed flows) to 1 (perfect fit). A simple mean of the three scaled metrics is then used as the multiobjective CL of each simulated time series within the simulation set.

$$\mathrm{CL} = \mathrm{mean}(\theta_{\mathrm{NSE}}, \theta_{\mathrm{MNS}}, \theta_{\mathrm{RSD}}). \tag{7}$$

[11] Figure 2 presents our approach for model calibration, applied separately for each watershed using the four calibration targets. The process began with a preliminary SWAT run and the accompanying sensitivity analysis routine within ArcSWAT. In addition to parameters identified as sensitive by this routine, we also identified parameters that had been highly influential during our prior manual calibration efforts in these watersheds [*Price et al.*, 2011]. Realistic initial ranges for each parameter were defined based on existing literature and knowledge of the study area (Table 1) [*Wu and Xu*, 2006; *van Griensven et al.*, 2006; *Neitsch et al.*, 2011]. LHS was used to generate multidimensional parameter value combinations across a set of 2001 simulations with uniform distributions. The SUFI-2 routine of the SWAT Calibration and Uncertainty Program (SWAT-CUP) [*Abbaspour*, 2007] implemented LHS of the 12 sensitive parameters (Table 1). For each watershed, simulations with these parameter sets were executed in batches, or "passes," of 2001 runs, which is the maximum allowable within the current version of SUFI-2. A threshold of 2000 iterations has been previously shown to be useful in past multiobjective calibration studies [e.g., *Madsen*, 2000]. *R* statistical software was used to postprocess the SUFI-2 runs, in which fit scores were calculated for each simulation. A GLUE-like approach [*Beven and Binley*, 1992] assigned informal likelihoods to each parameter set. Our goal was to find a 12-dimensional hyper-rectangle with parameter ranges that yielded the best models for each objective function [*Yapo et al.*, 1998]. The 15.9 and 84.1 percentile (±1 standard deviation) values were extracted from the weighted likelihoods for each parameter (equations (4)–(7)). These in turn were used as input ranges for the next pass of LHS and 2001 simulations to home in on the optimal parameter ranges, block by multidimensional block, and iteratively reduce the feasible control space with each pass. This process represents a compromise between Pareto optimization and an approach based on equifinality and is illustrated in Figure 3 in a simplified two-dimensional schematic.

[12] We also imposed a stopping rule to avoid over-restricting the parameter space. Iteration was stopped when a pass failed to improve the average fit by 5% over the previous pass, with a maximum of five passes (10,005 total simulations). The parameter ranges associated with the best

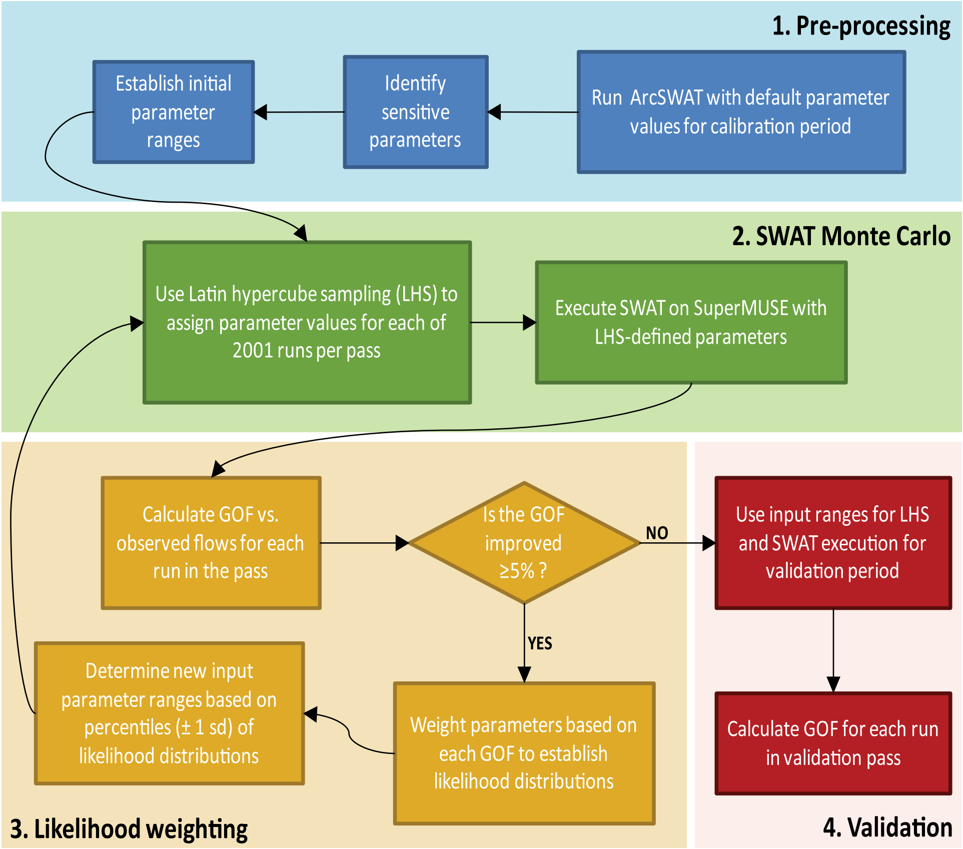
1. Pre-processing
Establish initial parameter ranges
Identify sensitive parameters
Run ArcSWAT with default parameter values for calibration period
2. SWAT Monte Carlo
Use Latin hypercube sampling (LHS) to assign parameter values for each of 2001 runs per pass
Execute SWAT on SuperMUSE with LHS-defined parameters
Calculate GOF vs. observed flows for each run in the pass
Is the GOF improved ≥5% ?
NO
YES
Use input ranges for LHS and SWAT execution for validation period
Determine new input parameter ranges based on percentiles (± 1 sd) of likelihood distributions
Weight parameters based on each GOF to establish likelihood distributions
Calculate GOF for each run in validation pass
3. Likelihood weighting
4. Validation

Figure 2. Conceptual diagram of model calibration approach. This process was applied separately for each watershed using the four calibration targets. First, sensitive parameters are identified and parameter ranges are established. The iterative part of the process is described in steps 2 and 3. In step 2, LHS is used to generate multidimensional parameter value combinations across a set of 2001 parallelized SWAT simulations. In step 3, goodness-of-fit (GOF) statistics are calculated for each simulation and informal likelihoods are assigned to each parameter set. The 15.9 and 84.1 percentile values are then extracted from the weighted likelihoods for each parameter and used as the input ranges for the next pass of LHS and 2001 simulations, beginning the next iteration of steps 2 and 3. Iteration is stopped when a pass fails to improve the average fit by at least 5%. The parameter ranges associated with the best calibration pass were used for the validation period (step 4), for which separate fit scores were calculated.

calibration pass were used to generate sets of 2001 simulations for the 2008–2010 validation period, for which separate fit scores were calculated. This process was repeated for each calibration, using all four objective functions in both watersheds (Figure 2). The SuperMUSE software system and distributed PC network in Athens, Georgia, was used to parallelize model simulations within each pass [*Babendreier and Castleton*, 2005]. One 2001-simulation pass for Mountain Creek (21 km^2 watershed area, 46 SWAT hydrologic response units, or HRUs) required 11 processor hours on 64-bit workstations with 8GB memory, and one pass for Little River (201 km^2 area, 241 HRUs) required 58 processor hours. Four separate calibrations involved five calibration passes and one validation pass, totaling 12,006 runs per watershed, for a total experiment time of 414 processor hours. All modeling routines herein (SWAT and SUFI-2) used 32-bit processes.

[13] In addition to fit statistics, multiple summary flow characteristics were calculated for the best pass of each calibration and validation set, as an a posteriori evaluation of the success of the various calibrations at simulating multiple streamflow regimes. As representations of low, medium, and high flows, the 5, 50, and 95 percentile flows were calculated from each simulation set and compared to corresponding percentiles of observed flows. To compare results of this simulation set approach, we also identified the single parameter set associated with the highest value of each objective function. Each of these sets was used to simulate a set of 2001 streamflow simulations for the 2008–2010 validation period (the "validation pass"), for which fit scores, as well as 5, 50, and 95 percentile flows, were calculated.

3. Results

[14] Results showed that different calibration targets (as individual goodness-of-fit metrics) produced substantially different ranges for calibration parameters, and, as a result, different simulated streamflow characteristics. This results section will detail differences among the calibration targets CL, NSE, MNS, and RSD in terms of the calibration iterations resultant parameter ranges, and simulated streamflow

Table 1. SWAT Calibration Parameter Definitions, Initial Ranges, and Calibrated Ranges

					Calibrated Ranges for Sensitive SWAT Parameters			
Parameter	Definition	Method[a]	Initial Range	Stream	CL	NSE	MNS	RSD
CN2	SCS Curve Number, Moisture Condition II	×	−1–1	MC	0.18–1	0.38–1	−0.06–0.68	−1–1
				LR	0.37–1	0.37–1	0.28–0.88	−1–1
Alpha_Bf	Baseflow Alpha Factor (days)	=	0–1	MC	0.25–1	0.64–1	0.98–1	0–1
				LR	0.53–1	0.53–1	0.92–1	0–1
CH_N2	Channel Manning's Coefficient	=	0–0.3	MC	0.06–0.26	0.11–0.23	0.12–0.18	0–0.3
				LR	0.05–0.25	0.05–0.25	0.12–0.18	0–0.3
CH_K2	Channel Hydraulic Conductivity (mm h^{-1})	=	0–150	MC	41.6–128	59.9–119	59.3–91.3	0–150
				LR	33.8–124	33.8–124	56.8–88.4	0–150
SOL_Z	Soil Depth (mm)	×	−0.25–0.25	MC	−0.16–0.17	−0.14–0.09	−0.05–0.05	−0.25–0.25
				LR	−0.18–0.16	−0.18–0.16	−0.06–0.05	−0.25–0.25
CANMX	Maximum Canopy Index	=	0–25	MC	4.92–21.06	5.81–16.86	9.97–15.27	0–25
				LR	4.62–21.32	4.62–21.32	9.83–15.24	0–25
ESCO	Soil Evaporation Compensation Factor	=	0.001–1	MC	0.13–0.81	0.32–0.84	0.40–0.62	0.001–1
				LR	0.15–0.84	0.15–0.84	0.41–0.62	0.001–1
GWQMN	Minimum Aquifer Depth for Flow (mm)	=	0–1000	MC	139–815	177–694	364–584	0–1000
				LR	151–827	151–827	371–591	0–1000
SOL_AWC	Available Soil Water Capacity	×	−0.2–0.6	MC	0.26–0.54	−0.01–0.34	0.36–0.45	−0.2–0.6
				LR	0.26–0.54	0.26–0.54	0.36–0.44	−0.2–0.6
SOL_K	Saturated Hydraulic Conductivity (mm h^{-1})	×	−0.5–0.5	MC	−0.35–0.36	−0.22–0.25	−0.10–0.11	−0.5–0.5
				LR	−0.34–0.33	−0.34–0.33	−0.11–0.11	−0.5–0.5
GW_REVAP	Capillary Fringe Coefficient	=	0.02–0.2	MC	0.05–0.17	0.06–0.14	0.09–0.13	0.02–0.2
				LR	0.05–0.17	0.05–0.17	0.09–0.13	0.02–0.2
SURLAG	Surface Runoff Lag Time (days)	=	0–24	MC	3.58–19.4	5.9–17.8	9.44–14.6	0–24
				LR	3.84–20.4	3.8–20.4	9.48–14.7	0–24

[a] ×: Original value is multiplied by the adjustment range, =: original value is replaced by the range.

properties at different levels of flow. In addition, we present the results from comparing our 2001-simulation set approach with the traditional model calibration method of using an individual winning set of parameters.

3.1. Iterative Calibration Results

[15] In accordance with the iterative calibration design, as passes progress from 1 to 5 for each calibration, the range for each parameter narrows. As a result, the distribution of simulated streamflows narrows, as shown in Figure 4a. This narrowing leads to lower variability within the simulation set, but in some cases, leads to more pronounced divergence from observed flow values. For each calibrated parameter, postprocessing of each pass increases the weight of the LHS-sampled parameter estimates if they result in better fit scores, and reduces the weight of less likely estimates of the same parameter. Figure 4b shows these iterations for CL calculations for the influential curve number parameter. As the range of the curve number is progressively winnowed, estimates resulting in poor simulations are dropped, and underlying metrics of the CL increase until a point of diminishing returns is reached. It is important to note that the set of best simulations within a pass is not constant across the different calibrated parameters—the weighted quantiles for updated parameter ranges are constructed separately for each parameter, and "good" individual parameter estimates can be associated with weak parameter estimates for other sensitive parameters via the LHS sampler, resulting in a poor fit score for that simulation. Figure 4b shows the progression of fit scores for a highly sensitive parameter (Curve number), whereas less sensitive parameters may show no identifiable improvement in fit scores as their range is narrowed.

[16] The fit distributions for CL (Figure 4c) demonstrate wider ranges in early passes. Although these may include the highest occurrences of fit scores, they are also accompanied by the lowest individual fit scores, the lowest average set scores, and a lack of clear central tendency. The winning pass was determined by the average fit score of each simulation set. To prevent over-constraining the parameter ranges, our stopping rule required that the sum of the fit scores increase by 5% for iterations to continue.

[17] The number of necessary pass iterations (or simulation sets) varied across the calibration targets. RSD calibrations did not retain passes beyond the initial, widest parameter ranges because the greater simulated flow variability with the widest parameter ranges most closely matched actual flow variability for both Mountain Creek and Little River. In contrast, MNS calibrations continued improving until the externally imposed limit of five passes; CL and NSE were intermediate, consistently satisfying the stopping rule within 2–3 passes.

3.2. Calibrated Parameter Ranges

[18] Like the pass iterations, calibrated parameter ranges also varied according to the calibration target (Table 1). Curve number and base flow recession coefficient showed the greatest influence over fit quality for both watersheds with respect to CL, NSE, and MNS. Curve number showed a positive skew in the updated ranges (within the initial proposed ranges) across all three calibrations. This was most pronounced with NSE, least pronounced with MNS, and intermediate with CL. Base flow recession also showed positive skew across the three calibrations, being most pronounced with MNS, least with CL, and intermediate with NSE. RSD retained the full initial range for both parameters and its influence on CL is apparent in base flow recession, given that the range for this parameter is greater in CL than in either MNS or NSE. The other sensitive parameters showed generally consistent behavior across the four calibrations, with the amount of narrowing corresponding

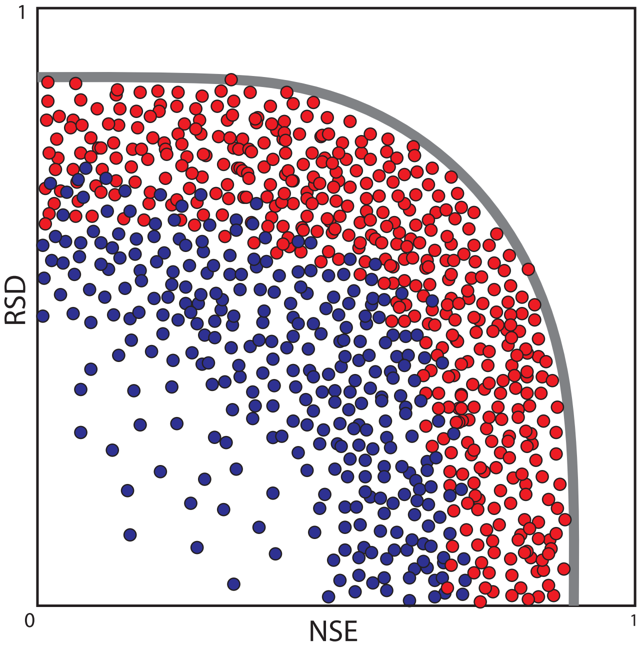

Figure 3. Relationship between the Pareto efficient frontier and the simulations selected by the composite likelihood (CL) approach. This schematic presents a simplified, two-dimensional representation of the aggregation function for a given calibration parameter. Each point represents a simulation within a set of 2001 model runs. After each pass, new ranges were constructed for each parameter by constructing weighted relative likelihood distributions (equations (4)–(7)). Red and blue points indicate simulations retained and rejected, respectively, as a result of this process. The gray line represents the theoretical Pareto efficient frontier, of which only a small number of simulations may actually be sampled within a given set of 2001 simulations used in our approach. Typically in a Pareto simulation study, the efficient frontier is estimated with a nondomination sort, which would be the subset of red points closest to the gray line not covered by other points. The approach here retains more simulations, consistent with the GLUE-like concepts of simulation sets used to address equifinality during optimization [*Beven*, 2006]. For example, two adjacent red points (one dominated and one nondominated) might result from two drastically different parameter values but be very close in performance space. This illustrates how our approach is a compromise between a true Pareto optimization and an approach based on equifinality, while accommodating computational limitations often associated with process-based hydrological models.

to the number of passes completed. One noteworthy exception is in the soil evaporation coefficient (ESCO), which is narrower in the MNS calibrations than the other three. For each objective function, parameter ranges were markedly consistent between the two watersheds (Table 1).

3.3. Calibrated Streamflows

[19] Figure 5 presents hydrographs for a representative six-month subset of the simulation period. Visual inspection of the simulated fan charts versus observed hydrographs for Mountain Creek (Figure 5a) and Little River (Figure 5b) reveals the wider variability within the winning pass for RSD, compared to the other three objective functions, which results from RSD calibration retaining the full initial parameter ranges. Among the four calibrations, MNS visually shows the best correspondence between simulated and observed low flows, while NSE produced lower uncertainty in high flow simulation than the other calibrations. As an aggregate measure, CL demonstrates the best compromise between low and high flows. These visual interpretations are corroborated by distributions of the fit scores (Figure 6) that, again, show the greatest uncertainty with RSD. Fit scores demonstrate improved validations compared to some of the calibrations, particularly CL and MNS, which are more strongly influenced than NSE and RSD by extreme low flow events during the calibration period. Figure 6 additionally shows that the calibration performance does not differ substantially between the two watersheds.

3.4. Low, Medium, and High Flows

[20] For a posteriori assessment of simulated low, medium, and high flows with observed values, we calculated 5, 50, and 95 percentile flows (Figure 7). Overall, the simulated flow distributions better match observed high flows than medium and low flows, regardless of the calibration target. MNS and RSD calibrations resulted in the highest uncertainty associated with high flows, while the central tendency of the NSE and CL high flow distributions matched well with observed high flows for both Mountain Creek and Little River. In general, simulated low and medium flows underestimated observed flows. Low and medium flows from the validation periods correspond better to observed flows than flows from the calibration period, which is likely due to the absence of extreme low flow events during the validation period. In the 5 and 50 percentile flows, influence of RSD on CL is evident, with the added variability due to incorporating RSD to help pull more simulated values closer to the observed values than with MNS and NSE. Simulated low flows show a tight distribution for all objective functions except RSD, with the lowest simulated low flow uncertainty associated with MNS.

3.5. Comparison of Simulation Set Approach to Single Best-Fit Simulation

[21] The individual best simulation (i.e., highest fit score from any pass) for each objective function was determined to evaluate how the simulation set approach used here compares with the traditional calibration approach of selecting parameter values associated with the single best-fit simulation. Relative to observed 5, 50, and 95 percentile flows, the single best-fit approach shows no consistent improvement or reduction in model performance when compared with the median of the simulation set distribution (Figure 7). While model performance may be effectively equivalent between these two approaches across calibration and validation periods, the simulation set approach provides bounds on the variability of model simulation results (as uncertainty estimates), which are important for a number of watershed applications and decision-making processes.

4. Discussion

[22] Our objectives were to (1) compare SWAT-simulated streamflows, using selected individual objective functions

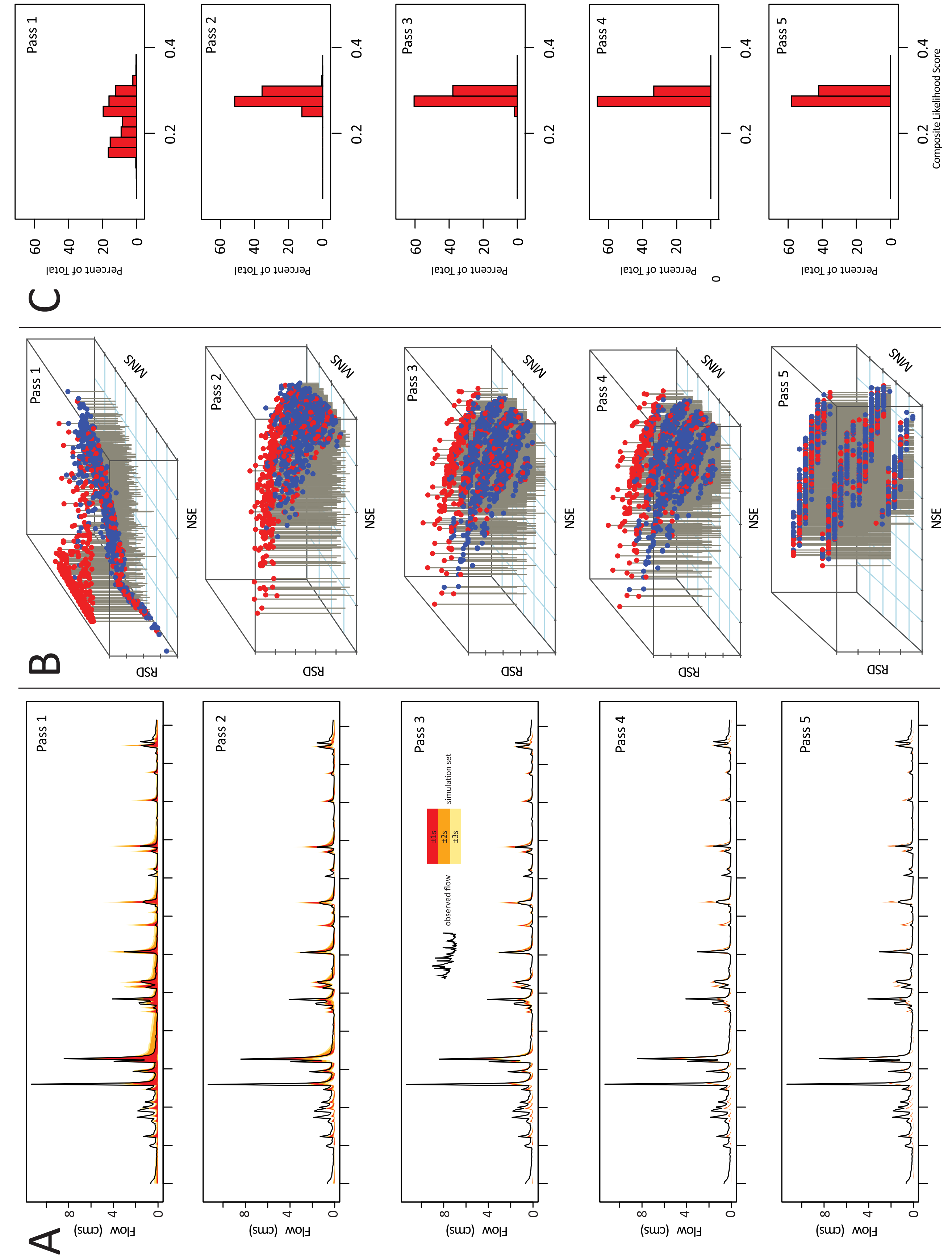
A
B
C
Pass 1
Pass 2
Pass 3
Pass 4
Pass 5
Flow (cms)
0
4
8
observed flow
simulation set
±1s
±2s
±3s
RSD
NSE
MNS
Percent of Total
0
20
40
60
0.2
0.4
Composite Likelihood Score

Figure 4

and a composite multiobjective function as calibration targets, (2) explore a calibration scheme that would provide uncertainty estimation and allow for user-defined fit criteria, and (3) develop a method borrowing from equifinality-based and true Pareto optimization approaches. In sum, our goal was to develop a robust calibration methodology that is computationally practicable at a management scale, not linked to any single watershed model, time step, study area, etc. We used LHS to reduce the runs required for Monte Carlo sampling, and an iterative likelihood-weighting scheme to narrow parameter ranges for simulation sets during calibration. The SUFI-2 environment for LHS and SWAT execution facilitated parallelization of simulation sets across multiple computers. Our methods allowed us to explore tradeoffs between calibrating on individual objective functions aimed at emphasizing low flows, high flows, and flow variability, as well as a composite multiobjective function to address all three flow response modes simultaneously. The approach used in this study satisfied the goals of the research and laid the foundation for a flexible method that can be broadly implemented, but our results also show that several parts of the process can be improved. Successes and weaknesses of our calibration approach were explored further, and suggestions for refinement are offered in this discussion.

[23] Overall, the optimized NSE values of ~0.6 observed in this study correspond well to other SWAT simulation studies in complex systems at daily time step [e.g., *Wang et al.*, 2006; *Santhi et al.*, 2008; *Setegn et al.*, 2008; *Almendinger and Ulrich*, 2010; *Looper et al.*, 2012]. NSE was expected to emphasize flood peaks during calibration, given that the difference between observed and simulated flows is squared in its calculation (equation (1)). Very large proportional differences in base flows may be of smaller absolute difference than even minor differences in flood peaks, thus resulting in little influence of low flows on the NSE score. It is evident from the fan charts that NSE calibrations result in better matches between simulated and observed high flows than low flows (Figure 4). The simulated 95 percentile flows from the NSE calibration are much closer to observed than median flows (Figure 7). It is interesting to note that the NSE calibration period for 5 percentile flows performed reasonably well, presumably due to multiple episodes of extreme low flows (zero values) during the 2002–2007 period. Thus, it appears that NSE performs well as a calibration target for extreme events of both high and low flows, with much larger proportional changes (i.e., lower return frequency) required at the low flow end than the high flow end to influence calibration. *Gupta et al.* [2009] show how this is possible, presenting a decomposition of NSE that functions as an aggregated multiobjective function—i.e., weighting components that correspond to bias, correlation, and flow variability. Termed the Kling-Gupta efficiency (KGE), this rearrangement of NSE shows that low flows can be well-represented in the NSE through embedded terms for bias and correlation, despite its reputation for emphasizing high flows. The formulation of KGE also allows for unequal weighting of the three components if one wishes to emphasize certain areas of the aggregate function tradeoff space; however, high variance simulations lead to low contributions of the bias component.

[24] Because MNS uses the first-order absolute difference between simulated and observed flows (equation (2)) instead of a second-order difference like NSE (equation (1)), it is generally considered superior for calibrating to low flows [*Bahremand et al.*, 2007; *Suleiman et al.*, 2007; *Van der Velde et al.*, 2009]. Indeed, our results show that MNS calibrations demonstrate the lowest uncertainty within the 5 percentile flow simulation sets and the central tendency of the MNS-calibrated 5 percentile flow simulation sets were closest to the observed values (Figure 7). However, it is evident from both the fan charts (Figure 4) and the 5 percentile flow simulation sets (Figure 7) that the NSE- and MNS-calibrated low flows are quite similar—much more than we anticipated. Other studies have used more aggressive low flow calibration targets, such as the log NSE, which is simply NSE computed using natural log-transformed flows:

$$\log \mathrm{NSE}_i = \frac{\sum_{t=1}^{n} [\ln(O_t) - \ln(S_t)]^2}{\sum_{t=1}^{n} [\ln(O_t) - \ln(\overline{O})]^2}. \qquad (8)$$

Tekleab et al. [2011] combined this log NSE with the standard NSE into a multiobjective function that captured both low and high flows satisfactorily for their objectives. MNS is included in many model evaluation software packages and maintains an established presence in the literature [e.g., *Bahremand et al.*, 2007; *Suleiman et al.*, 2007; *Van der Velde et al.*, 2008]. MNS being thus vetted as a low flow calibration target caused us to underestimate the utility

Figure 4. Streamflow simulations, simulation fit scores, and parameterization through the pass evolution process. (a) Example fan charts from the calibration period for Mountain Creek (CL calibration). With each pass iteration, input parameter ranges are narrowed, resulting in lower uncertainty within the simulation set. This is evident from narrowing of the first, second, and third standard deviation bands as passes progress from 1 to 5. (b) Shows a three-dimensional plot of each simulation within each pass for the constituent metrics of the CL calibration used to update the uniform distribution for the curve number. Red pinheads show Monte Carlo simulations within each pass that have relatively low likelihood; therefore, the curve number values that generated them are given low weight when constructing the updated parameter estimates for the next pass. Blue pinheads (roughly 2/3 of the 2001 simulations for each pass) signify simulations where model performance was better, and are included in the weighted likelihood distribution that updates the curve number range for the next pass. Significant movement occurred in the transition between the first and second passes, but the stopping rule indicates that subsequent passes do not sufficiently improve overall model fit. (c) Distribution of fit scores across calibration passes for Mountain Creek (corresponding to the fan charts in Figure 3a). Narrowing parameter ranges as passes progress from 1 to 5 results in truncation of both lowest and highest fit scores. The winning pass (determined by the average fit score) was pass 2.

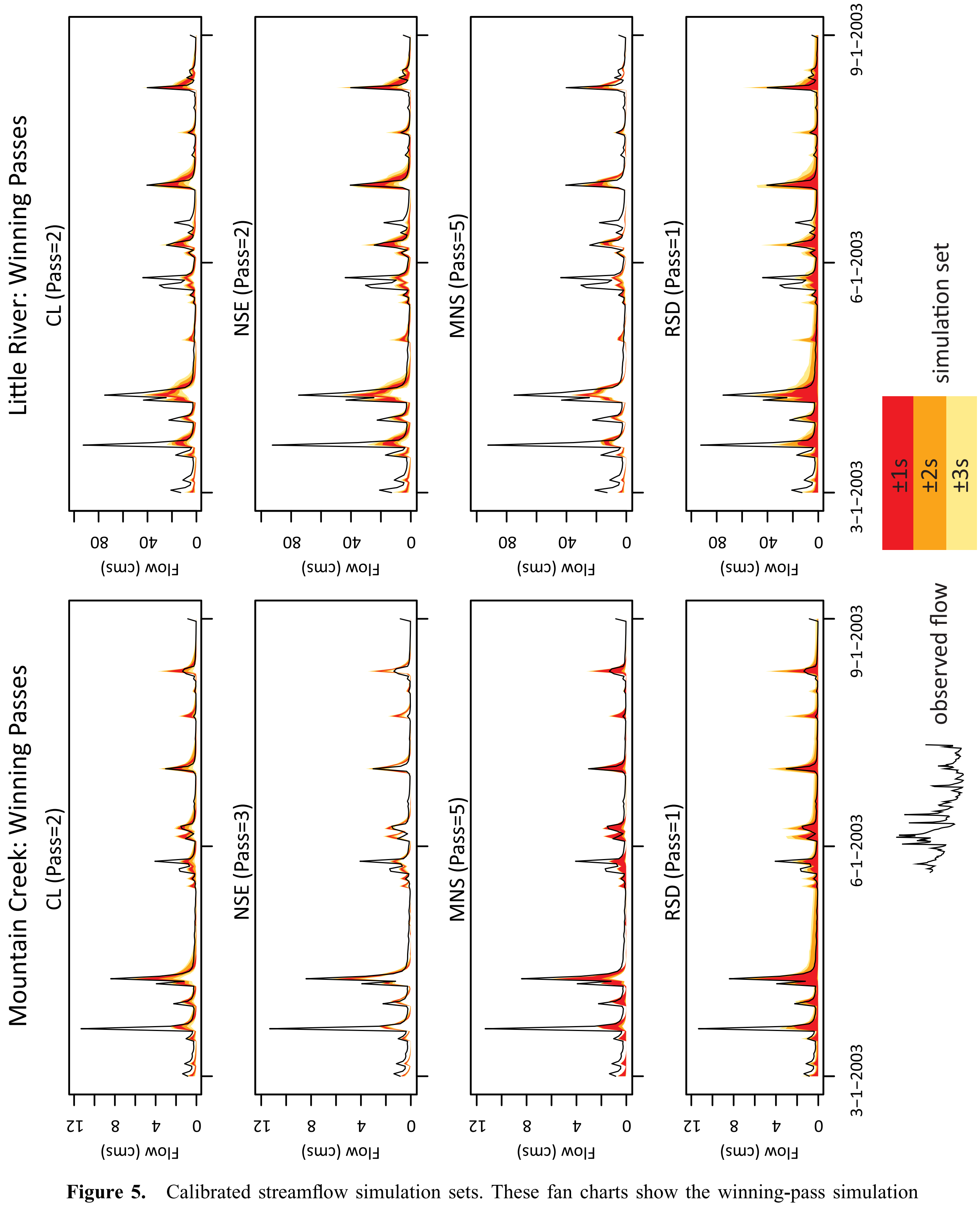

Figure 5. Calibrated streamflow simulation sets. These fan charts show the winning-pass simulation sets for a six-month, representative subset of the calibration "S" indicates sample standard deviation. The median simulated flow is the center of the ±1s band. CL = composite likelihood (multiobjective), NSE = Nash-Sutcliffe efficiency, MNS = modified Nash-Sutcliffe efficiency, and RSD = standard deviation ratio. These fan charts show that the NSE calibration produced the lowest simulation set uncertainty among high flows, but that CL corresponds better between simulated and observed peaks, and represents a compromise between NSE and MNS, which best matches observed low flows.

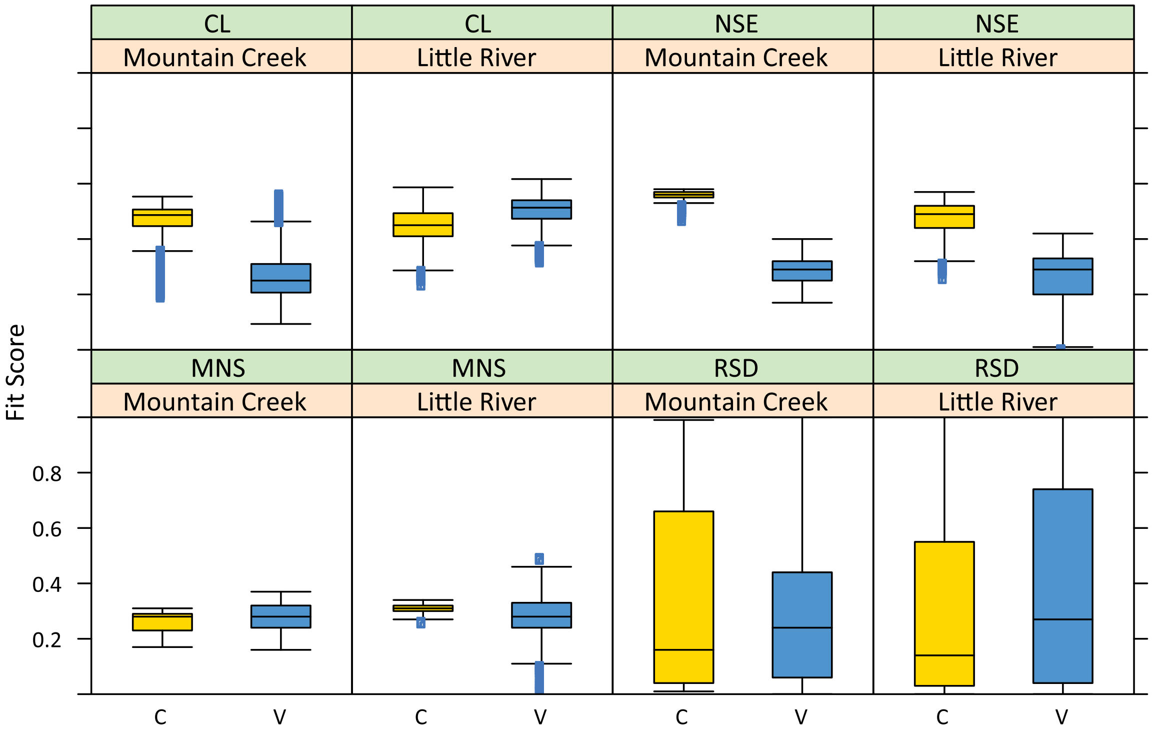

Figure 6. Fit score distributions for calibrated and validated simulation sets. These box plots show the distribution of fit scores for each objective function, with $n = 2001$ simulations per box. The yellow box (C) in each pair represents the calibration period (2002–2007), while the blue box (V) represents the validation period (2008–2010). CL = Composite Likelihood (multiobjective), NSE = Nash-Sutcliffe efficiency, MNS = modified Nash-Sutcliffe efficiency, and RSD = standard deviation ratio. Calibration and validation scores were consistent across the two watersheds. It is interesting to note that uppermost validation scores actually improved over calibration scores for CL and MNS, while validation scores are lower for NSE.

of log NSE in this role in our original research design, when in fact log NSE appears to be the superior low flow target between the two.

[25] We used RSD as a calibration target for flow variability. Its primary intended use was as part of the multiobjective function to avoid calibrating toward simulations that acted somewhat like a mean model—underpredicting high flows and overpredicting low flows, but minimizing error successfully. As an individual calibration target, RSD was optimized when flow ranges were greatest, which inevitably favored the widest parameter ranges corresponding to the first pass (Figure 5). As our initial parameter ranges were determined by physically realistic bounds, this favoring of RSD for the widest range should not serve as an implication that our initial ranges should have been wider. Instead, this indicates the smoothing effect that SWAT has on streamflow simulation, that even the widest bounds of physically realistic parameters fail to capture the full variability of observed streamflow. There exist other metrics of flow variability, such as base flow index [*Eckhardt*, 2008], flashiness index [*Baker et al.*, 2004] and the various pulse and timing summary statistics among the indicators of hydrologic alteration [*Gao et al.*, 2009], which could be used as individual flow variability calibration targets or in aggregated multiobjective functions. An alternative to RSD worth exploring is a distance measure integrating the discrepancy between simulated and observed flow duration curves across the simulation set, to supplement other flow duration curve analyses [e.g., *Mohamoud*, 2008; *Westerberg et al.*, 2011]. While RSD as used here may not be ideal as a freestanding calibration target, our results show it advantageously impacted our multiobjective calibrations. Figure 7 shows that CL-calibrated simulation sets for medium and low flows have much wider ranges than NSE or MNS high flow simulation sets, and this additional variability is due to the influence of RSD. While this greater simulation set variability is in some ways undesirable, wider ranges for medium and low flows result in higher likelihood of including the observed value, indicating that the high uncertainty introduced by RSD helps tie the calibration to reality and minimize chances of overfitting in the calibration process.

[26] Limitations of NSE, MNS, and RSD as individual objective functions and as components to CL are due not only to formulation of the objective functions, but also to model structure [*Matott et al.*, 2009]. Both SUFI-2 and SWAT impose limits on the parameter ranges that affect trajectories of the calibrations to an unknown degree. It should also be noted that these calibrations are for a predominantly process-based model, as opposed to an empirically conditioned model, so there are SWAT-specific limits on the amount of influence observed data and the calibration targets can have on how water is routed through the system. SWAT is a semidistributed model in which water is not spatially explicitly routed through the watershed but

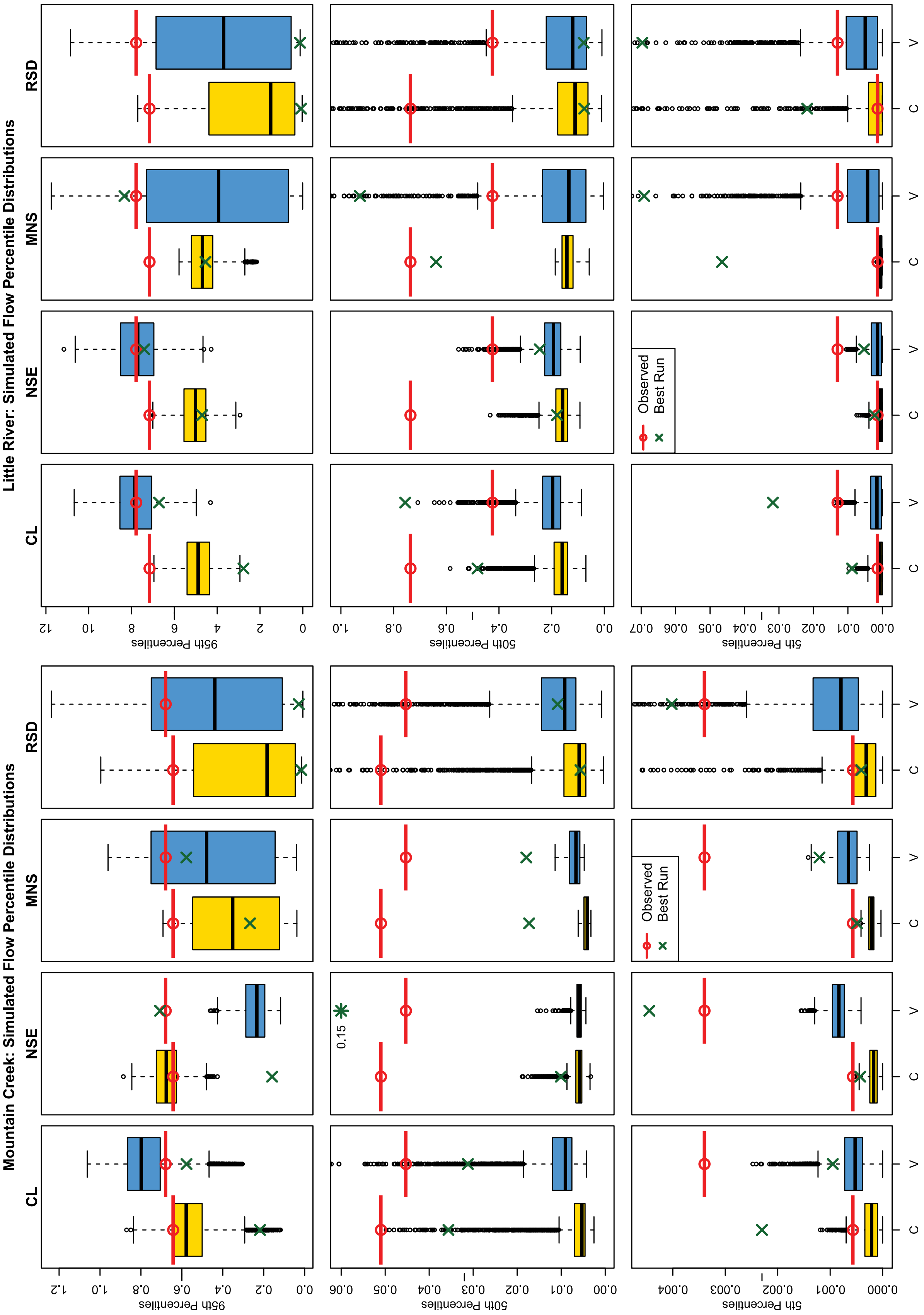
Little River: Simulated Flow Percentile Distributions
Mountain Creek: Simulated Flow Percentile Distributions
CL
NSE
MNS
RSD
95th Percentiles
50th Percentiles
5th Percentiles
C
V
Observed
Best Run
0.15
0 2 4 6 8 10 12
0.0 0.2 0.4 0.6 0.8 1.0
0.00 0.01 0.02 0.03 0.04 0.05 0.06 0.07
0.0 0.2 0.4 0.6 0.8 1.0 1.2
0.00 0.01 0.02 0.03 0.04 0.05 0.06
0.000 0.001 0.002 0.003 0.004

Figure 7

lumped at a series of linked spatial scales, as water budgets and transit times to the stream network are calculated for each HRU (finest scale), subbasin, and entire watershed (coarsest scale). Since most widely used watershed models are semidistributed, results from this study may inform calibration procedures for a range of popular models. In addition, many studies have shown difficulty in generating reliable low flow estimates with SWAT [e.g., *Peterson and Hamlett*, 1998; *Levesque et al.*, 2008; *Bosch et al.*, 2010], and this may be due to inadequacies in representation of subsurface processes or the very large influence of curve number in SWAT streamflow simulation.

[27] Our likelihood-weighted, aggregated multiobjective function approach (as a compromise between equifinality-based and Pareto optimization methods) worked well in this analysis. Figure 4b shows trajectories of the individual objective functions and the result of the iterative parameter range narrowing, illustrating that CL was a calibration compromise between NSE, MNS, and RSD. Pareto multiobjective optimization schemes [e.g., *Gupta et al.*, 1998; *Yapo et al.*, 1998] emphasize the "efficient frontier," a thin skin of parameter values associated with best-fit values of each component calibration target; our method is arguably more resistant to equifinality concerns and more practical to characterize in high-dimensional parameter spaces. A more Pareto-like implementation of our approach would swap out the composite metric (equation (7)) for a subset approach that would successively retain multiple Pareto layers, extracting and keeping subsets of nondominated solutions [*Ficici*, 2001; *Avigad et al.*, 2010] However, a potential weakness to the Pareto layer approach is that it will retain parameter sets that highly support one objective function, while producing poor results with respect to the other two, thereby failing to represent viable tradeoffs of the performance metrics. The likelihood-weighting aggregation approach used herein avoids this phenomenon.

[28] The multiobjective approach also shows strong potential for predictive and scenario-based modeling. While it is acknowledged that validation scores will be lower than calibration scores—even in process-based modeling [*Gupta et al.*, 2009]—our CL results show that the range of fit scores is retained from calibration to validation for both watersheds, with the larger watershed actually showing marked improvement over the period. Additionally, MNS shows good support from the validation period. NSE demonstrates the expected decline in fit score from calibration to validation, and the uncertainty of RSD is so high that it is difficult to draw comparisons between calibration and validation. Consistency between calibration and validation periods indicates the parameters are not overfit to the calibration period and boosts confidence in using the calibrated parameter set in predictive modeling [*Gupta et al.*, 2009].

[29] We compared our calibrated simulation sets with simulations from the highest-scoring individual parameter set for each objective function, which correspond to the single best-fit parameter set optimization approach traditionally used for watershed model calibration [*Beven*, 2006]. Focusing parameter ranges toward a compromise central tendency (Figure 4c) illustrates the purpose of simulation sets, as opposed to traditional choosing of parameter values associated with the highest individual objective function fit score. Overall, best-fit parameter set simulations had equivalent performance to medians of the simulation sets in matching observed low, medium, and high flows for the calibration and validation periods (Figure 7). The best-fit parameter set approach, however, does not provide uncertainty estimation. Uncertainty information is critical for policy-relevant modeling efforts and multimodel frameworks [*Matott et al.*, 2009]; the results of this study give confidence that using a simulation set approach to harness uncertainty information does not involve tradeoffs in model calibration or validation accuracy, when compared to traditional, single best-fit methods.

[30] An important goal of this research was to explore calibration methods that would generate uncertainty estimates, while remaining practical for use in management and policy-scale watersheds. This translates to minimizing the number of simulations required to home in on optimal parameter ranges and generate plausible confidence bands. Our LHS and informal likelihood-weighting approach was applied to simulation sets of 2001 runs, the limit imposed by SUFI-2. Our stopping rule required fit scores to improve by at least 5% to progress to additional passes, but we additionally imposed a limit of five total passes, equating to a maximum of 10,005 runs. This was accomplished in 66 total processor hours for our small watershed and in 348 total hours for the larger watershed, which at 201 km^2 is relatively small in terms of management relevance. We acknowledge two key modifications to our methods that could improve model calibration, but would be accompanied by increases in simulations required for optimization. First, the degree to which parameter ranges are narrowed between passes could be decreased. Here we used the 15.9 and 84.1 percentiles of the likelihood-weighted parameter distributions from the completed pass as the full range for the next pass, as ± one standard deviation. If we used

Figure 7. Comparison of simulated and observed low, median, and high flows. As an a posteriori assessment of the success of each calibration in simulation of low, medium, and high flows, three percentile values (5, 50, and 95 percentiles) were calculated for each simulation in the winning calibration and validation simulation sets. Each box represents 2001 simulations. CL = composite likelihood (multiobjective), NSE = Nash-Sutcliffe efficiency, MNS = modified Nash-Sutcliffe efficiency, and RSD = standard deviation ratio. The wide line with inset circle indicates the corresponding observed percentile flow value for the calibration period (2002–2007, left box) and validation period (2008–2010, right box). "X" indicates the percentile flow value calculated for the calibration and validation period, using the single best fit parameter set, for comparison of our simulation set approach to traditional optimization approaches. These results show that high flows are generally better predicted than median and low flows, and that NSE calibration performed best with high flows. While the MNS calibration produced the lowest uncertainty in simulated low flows, MNS, NSE, and CL simulated low flows equally well. The single best fit parameter set did not show any clear advantage or disadvantage over our simulation set approach in terms of accuracy, but it does not generate uncertainty estimates.

higher quantiles (e.g., 5 and 95 percentiles), we could home in on the optimal parameter ranges more precisely, but this would require more passes of simulation sets before the stopping rule was satisfied. Second, our LHS method in SUFI-2 allowed a maximum of 2001 simulations per pass, while more divisions of the hypercube would allow for greater exploration of parameter interactions. Clearly a greater number of simulations per pass would increase the total number of simulations required for calibration, which would lead to longer runtimes. While we utilized a parallel computing facility, other modelers may not have such access. The methods as used herein could be performed on an individual PC for small or very coarsely discretized watersheds, but high-resolution modeling of larger watersheds (100s to 1000s km^2) would benefit from parallelization. Thus, there is a balance to be struck between true optimization and practicability, and this method allows flexibility to define that tradeoff to meet case-specific objectives. Cloud-based computing will expand parallelization options for many researchers in years to come.

[31] SWAT is known to inaccurately represent overland flow processes in humid, heavily vegetated regions not dominated by infiltration-excess runoff processes [*Easton et al.*, 2008, 2010; *White et al.*, 2011]. This inaccurate process representation in large part explains the far-from-perfect simulations commonly seen in SWAT applications in humid regions [e.g., *Wang et al.*, 2006; *Santhi et al.*, 2008; *Setegn et al.*, 2008; *Rouhani et al.*, 2009; *Almendinger and Ulrich*, 2010], including the results we present in this paper. Nevertheless, SWAT remains one of the most widely used watershed models in all regions, due to its ease of use, strong user support, versatility of simulated water quality constituents, and feasible input data and computational requirements. Countless government agency and academic research projects investigating the effects of land use and climate change on water quantity and quality rely on SWAT model results, and the momentum behind this reliance on SWAT seems unlikely to shift in the near future. It can clearly be argued that the water resources community needs to develop better models instead of focusing continued energy on developing calibration strategies for models that do not fully represent hydrologic processes in all regions. However, the need for modeling results outpaces model development, and imperfect models continue to be used in all spheres of research. For this reason we need to understand how to best use these available tools until momentum shifts to improved models.

[32] One underexplored, yet highly practical and innovative approach is to modify existing tools to better represent regionally specific hydrology. For example, several recent studies have successfully used customized SWAT formulations to better represent variable source area hydrology, avoiding the limitations to infiltration-excess runoff seen in the standard formulation of SWAT [*Easton et al.*, 2008, 2010, *White et al.*, 2011]. While programming requirements for these types of customizations exceed the capabilities of most modeling practitioners, this is a fruitful area for development that needs wider recognition. It can be hoped that as development of such tools continues, general availability of improved options is ensured by the primary model providers. In addition, it is known that observed streamflow data and input meteorological data are imprecise, and more work needs to be done to quantify the uncertainty associated with such datasets when interpreting model performance (e.g., K. Price et al., Comparison of radar and gauge precipitation data in watershed models across spatial and temporal scales, submitted to Water Resources Research, 2012).

5. Conclusions

[33] We explored a suite of model evaluation criteria, or objective functions, to better understand how streamflow simulations are parameterized and interpreted under different calibration frameworks. We calibrated daily streamflow from two watersheds in the North Carolina Piedmont, using three individual objective functions: Nash-Sutcliffe efficiency (NSE), modified Nash-Sutcliffe efficiency (MNS), and the ratio of simulated and observed standard deviations (RSD). In addition, we implemented a fourth calibration using a multiobjective function combining NSE, MNS, and RSD, termed the composite likelihood (CL). We used an iterative, informal likelihood-weighting approach to generate simulation sets of 2001 model runs over a series of five evolutionary iterations, for a total of 10,005 runs. This approach represents an effort to develop flexible and computationally efficient methods that serve as a compromise between equifinality-based calibration approaches, such as the GLUE methodology developed by *Beven and Binley* [1992], and Pareto optimization approaches, such as those developed by *Yapo et al.* [1998] and *Gupta et al.* [1998].

[34] We expected that NSE and MNS calibrations would produce the best peak flow and general base flow calibrations, respectively, while combining these two functions with RSD in the multiobjective CL would lead to calibrations best matching a range of flow conditions. Results generally supported our expectations, but several key considerations for future multiobjective calibration attempts emerged. First, NSE was influenced by extreme high flows as expected, but also by extreme low flow events and, thus, may be a good individual calibration target in highly variable flow regimes. Second, MNS did not improve base flow calibrations over NSE as much as expected, so more aggressive base flow calibration targets should be explored. RSD failed to produce good calibrations as an individual calibration target, but was an important influence in the multiobjective CL. The CL calibration showed promising performance in model validation—greater than NSE—which encourages further use of this approach for scenario-based predictive modeling, while reformulations of NSE offer potential as well.

[35] In addition to comparing calibrations with selected objective functions, we also compared our simulation set approach with calibrations that used the single best-fit parameter set for each calibration target. The single best-fit approach represents traditional optimization methods, in which unique parameter values (as opposed to ranges) associated with the highest individual occurrence of the objective function score are considered optimal. Single best-fit calibration methods are associated with higher risks of overfitting challenges with model equifinality, and lack of critical uncertainty information used for management-scale modeling efforts. Our results showed that, compared to observed streamflows, the single best-fit approach did not lead to better validated models than our simulation set approach. This gives confidence that employing a simulation set approach to leverage accompanying uncertainty

information does not come with costs of reduced model accuracy or predictive modeling capabilities.

[36] **Acknowledgments.** This research was conducted in collaboration with Chris Knightes, John M. Johnston, Robin Dennis, and Ellen Cooter. Feedback on early drafts of this manuscript was provided by Mike Cyterski, Roy Sidle, and Rhett Jackson. Insightful and helpful peer reviews were graciously provided by Joshua Kollat, Tammo Steenhuis, and one additional anonymous reviewer. This paper has been reviewed in accordance with the U.S. Environmental Protection Agency's peer and administrative review policies and has been approved for publication. Approval does not signify that that the contents necessarily reflect the views and policies of the agency, nor does the mention of trade names or commercial products constitute endorsement or recommendation for use. MPE rainfall data were provided by NCAR/EOL under sponsorship of the National Science Foundation.

References

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

Almendinger, J. E., and J. M. Ulrich (2010), Constructing a SWAT model of the Sunrise River watershed, eastern Minnesota, St. Croix Watershed Research Station, Marine on St. Croix, MN. [Available at http://www.smm.org/static/scwrs/tapwaters_sunrise.pdf.]

Andrieu, C., N. de Freitas, A. Doucet, and M. I. Jordan (2003), An introduction to MCMC for machine learning, *Mach. Learning*, *50*, 5–43.

Arulampalam, M. S., S. Maskell, N. Gordon, and T. Clapp (2002), A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking, *IEEE Trans. Signal Process.*, *50*, 174–188.

Avigad, G., E. Eisenstadt, and A. Goldvard (2010), Pareto layer: Its formulation and search by way of evolutionary multi-objective optimization. *Eng. Opt.*, *42*(5), 453–470.

Babendreier, J. E., and K. J. Castleton (2005), Investigating uncertainty and sensitivity in integrated, multimedia environmental models: Tools for FRAMES-3MRA, *Environ. Model. Software*, *20*(8), 1043–1055.

Bahremand, A., F. DeSmedt, J. Corluy, Y. B. Liu, J. Poorova, L. Velcicka, and E. Kunikova (2007), WetSpa model application for assessing reforestation impacts on floods in Margecany-Hornad Watershed, *Slovakia, Water Res. Manage.*, *21*(8), 1373–1391.

Baker, D. B., R. P. Richards, T. T. Loftus, and J. W. Kramer (2004), A new flashiness index: Characteristics and applications to midwestern rivers and streams, *J. Am. Water Resour. Assoc.*, *40*(2), 503–522.

Beaumont, M. A. (2010), Approximate Bayesian computation in evolution and ecology, *Annu. Rev. Ecol. Syst.*, *41*, 379–406.

Beven, K. (1993), Prophecy, reality and uncertainty in distributed hydrologic modeling, *Adv. Water Resour.*, *16*, 41–51.

Beven, K. (2006), A manifesto for the equifinality thesis, *J. Hydrol.*, *320*(1–2), 18–36.

Beven, K., and A. Binley (1992), The future of distributed models—Model calibration and uncertainty prediction, *Hydrol. Processes*, *6*(3), 279–298.

Bigiarni, M. Z. (2010), R Package 'hydroGOF': Goodness-of-fit functions for comparison of simulated and observed hydrological time series. [Available at http://cran.r-project.org/web/packages/hydroGOF.]

Bosch, D. D., J. G. Arnold, M. Volk, and P. M. Allen (2010), Simulation of a low-gradient coastal plain watershed using the SWAT landscape model. *Trans. ASABE*, *53*(5), 1445–1456.

Easton, Z. M., D. R. Fuka, M. T. Walter, D. M. Cowan, E. M. Schneiderman, and T. S. Steenhuis (2008), Re-conceptualizing the soil and water assessment (SWAT) model to predict runoff from variable source areas, *J. Hydrol.*, *348*, 279–291.

Easton, Z. M., D. R. Fuka, E. D. White, A. S. Collick, B. B. Ashagre, M. McCartney, S. B. Awulachew, A. A. Ahmed, and T. S. Steenhuis (2010), A multi basin SWAT model analysis of runoff and sedimentation in the Blue Nile, Ethiopia, *Hydrol. Earth Syst. Sci.*, *14*, 1827–1841, doi:10.5194/hess-14-1827-2010.

Eckhardt, K. (2008), A comparison of baseflow indices, which were calculated with seven different baseflow separation methods, *J. Hydrol.*, *352*(1–2), 168–173.

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

Ewen, J. (2011), Hydrograph matching method for measuring model performance, *J. Hydrol.*, *408*(1–2), 178–187.

Fenicia, F., H. H. G. Savenije, P. Matgen, and L. Pfister (2007), A comparison of alternative multiobjective calibration strategies for hydrological modeling, *Water Resour. Res.*, *43*(3), W03434, doi:10.1029/2006WR005098.

Ficici, S. G. (2001), Pareto optimality in coevolutionary learning, *Proceedings of Advances in Artificial Life: 6th European Conference*, ECAL, Prague, Czech Republic, 10–14 September.

Fry, J., G. Xian, S. Jin, J. Dewitz, C. Homer, L. Yang, C. Barnes, N. Herold, and J. Wickham (2011), Completion of the 2006 National Land Cover Database for the conterminous United States, *Photogram. Eng. Remote Sens.*, *77*(9), 858–864.

Gao, Y. X., R. M. Vogel, C. N. Kroll, N. L. Poff, and J. D. Olden (2009), Development of representative indicators of hydrologic alteration, *J. Hydrol.*, *374*(1–2), 136–147.

Goldberg, D. E. (1989), *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, Reading, MA.

Grimm, V., E. Revilla, U. Berger, F. Jeltsch, W. M. Mooji, S. F. Railsback, H. H. Thulke, J. Weiner, T. Wiegand, and D. L. DeAngelis (2005), Pattern-oriented modeling of agent-based complex systems: Lessons from ecology, *Science*, *310*, 987–991.

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

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

Hartig, F., J. M. Calabrese, B. Reineking, T. Wiegand, and A. Huth (2011), Statistical inference for stochastic models—Theory and application, *Ecol. Lett.*, *14*, 816–827.

Klemes, V. (1986), Operational testing of hydrological simulation models, *Hydrol. Sci. J.*, *31*(1), 13–24.

Krause, P. (2005), Comparison of different efficiency criteria for hydrological model assessment, *Adv. Geosci.*, *5*(89), 89–97.

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

Levesque, E., F. Anctil, A. van Grievensen, and N. Beauchamp (2008), Evaluation of streamflow simulations by SWAT model for two small watersheds under snowmelt and rainfall, *Hydrol. Sci. J.*, *53*(5), 961–976.

Lindström, G. (1997), A simple automatic calibration routine for the HBV model, *Nord. Hydrol.*, *28*(3), 153–168.

Looper, J. P., B. E. Vieux, and M. A. Moreno (2012), Assessing the impacts of precipitation bias on distributed hydrologic model calibration and prediction accuracy, *J. Hydrol.*, *418–419*, 110–122.

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

Madsen, H., G. Wilson, and H. C. Ammentrop (2002), Comparison of different automated strategies for calibration of rainfall-runoff models, *J. Hydrol.*, *261*(1–4), 48–59.

Matott, L. S., J. E. Babendreier, and S. T. Purucker (2009), Evaluating uncertainty in integrated environmental models: A review of concepts and tools, *Water Resour. Res.*, *45*, W06421, doi:10.1029/2008WR007301.

Mills, H. H., G. R. Brakenridge, R. B. Jacobson, W. L. Newell, M. J. Pavich, and J. S. Pomeroy (1987), Appalachian mountains and plateaus, in *Geomorphic Systems of North America*, edited by W. L. Graf, pp. 5–50, Geological Society of America, Boulder, CO.

Mohamoud, Y. M. (2008), Prediction of daily flow duration curves and streamflow for ungauged catchments using regional flow duration curves, *Hydrol. Sci. J.*, *53*(4), 706–724.

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

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

National Agricultural Statistics Service (NASS) (2011), 2010 North Carolina Cropland Data Layer, [Available at http://datagateway.nrcs.usda.gov/], NASS, USDA, Washington, D. C.

National Climate Data Center (NCDC) (2011), Monthly climate station summaries, 1981–2010, [Available at www.ncdc.noaa.gov], U.S. Department of Commerce, Washington, D. C.

NCAR Earth Observing Laboratory (2011), GCIP/EOP Surface: Precipitation NCEP/EMC 4KM Gridded Data (GRIB) Stage IV Data, [Avialable at http://data.eol.ucar.edu/codiac/dss/id=21.093], National Center for Atmospheric Research, Boulder, Col.

Neitsch, S. L., J. G. Arnold, J. R. Kiniry, and J. R. Williams (2011), Soil and Water Assessment Tool Theoretical Documentation Version 2009, *Texas Water Resources Institute Technical Report* 406, Texas A&M University System, College Station, TX.

O'Keeffe, J. (2009), Sustaining river ecosystems: balancing use and protection, *Progress Phys. Geogr.*, *33*(3), 339–357.

Peterson, J. R., and J. M. Hamlett (1998), Hydrologic calibration of the SWAT model in a watershed containing fragipan soils, *J. Am. Water Resour. Assoc.*, *34*(3), 531–544.

Poff, N. L., and J. K. H. Zimmerman (2010), Ecological responses to altered flow regimes: A literature review to inform the science and management of environmental flows, *Freshwater Biol. 55*(1), 194–205.

Price, K., S. T. Purucker, and S. R. Kraemer (2011), Multi-scale comparison of stage IV NEXRAD (MPE) and gauge precipitation data for watershed modeling, *Proceedings of Georgia Water Resources Conference*, 11–13 April, Warnell School of Forestry and Natural Resources, University of Georgia, Athens, GA.

R Development Core Team (2011), R: A language and environment for statistical computing; visit R Foundation for Statistical Computing, visit www.R-project.org.

Robert, C. P., and G. Casella (2004), *Monte Carlo Statistical Methods*, Springer, New York.

Santhi, C., N. Kannan, J. G. Arnold, and M. Di Luzio (2008), Spatial calibration and temporal validation of flow for regional scale hydrologic modeling, *J. Am. Water Resour. Assoc.*, *44*(4), 829–846.

Setegn, S. G., R. Srinivasan, and B. Daraghi (2008), Hydrological modelling in the Lake Tana Basin, Ethiopia using SWAT model, *Open Hydrol. J.*, *2*, 49–62.

Soil Survey Staff (2011), U. S. General Soil Map (STATSGO2), [Available at http://soildatamart.nrcs.usda.gov], Natural Resources Conservation Service (NRCS), U.S. Dep. of Agriculture, Lincoln, Neb.

Suleiman, A. A., C. M. T. Soler, and G. Hoogenboom (2007), Evaluation of FAO-56 crop coefficient procedures for deficit irrigation management of cotton in a humid climate, *Agric. Water Manage.*, *91*(1–3), 33–42.

Tekleab, S., S. Uhlenbrook, Y. Mohamed, H. H. G. Savenije, M. Temesgen, and J. Wenninger (2011), Water balance modeling of Upper Blue Nile catchments using a top-down approach, *Hydrol. Earth Syst. Sci.*, *15*(7), 2179–2193.

Thiemann, M., M. Trosset, H. Gupta, and S. Sorooshian (2001), Bayesian recursive parameter estimation for hydrologic models, *Water Resour. Res.*, *37*(10), 2521–2535.

Uhlenbrook, S., and A. Sieber (2005), On the value of experimental data to reduce the prediction uncertainty of a process-oriented catchment model, *Environ. Modell. Software*, *20*(1), 19–32.

Uhlenbrook, S., J. Seibert, C. Leibundgut, and A. Rodhe (1999), Prediction uncertainty of conceptual rainfall-runoff models caused by problems in identifying model parameters and structure, *Hydrol. Sci. J.*, *44*(5), 779–797.

Van der Velde, M., F. Bouraoui, and A. Aloe (2009), Pan-European regional-scale modelling of water and *N* efficiencies of rapeseed cultivation for biodiesel production, *Global Change Biol.*, *15*(1), 24–37.

van Griensven, A., T. Meixner, S. Grunwald, T. Bishop, M. Di Luzio, and R. Srinivasan (2006), A global sensitivity analysis tool for the parameters of multi-variable catchment models, *J. Hydrol.*, *324*, 10–23.

Wang, X., A. M. Melesse, and W. Yang (2006), Influences of potential evapotranspiration estimation methods on SWAT's hydrologic simulation in a northwestern Minnesota watershed, *Trans. ASABE*, *49*(6), 1755–1771.

Westerberg, I. K., J. L. Guerrero, P. M. Younger, K. J. Beven, J. Seibert, S. Halldin, J. E. Freer, and C. Y. Xu (2011), Calibration of hydrological models using flow-duration curves, *Hydrol. Earth Syst. Sci.*, *15*(7), 2205–2227.

White, E. D., Z. M. Easton, D. R. Fuka, A. S. Collick, E. Adgo, M. McCartney, S. B. Awulachew, Y. G. Selassie, and T. S. Steenhuis (2011), Development and application of a physically based landscape water balance in the SWAT model, *Hydrol. Proc.*, *25*(6), 915–925, doi:10.1002/hyp.7876, 2010.

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

Winchell, M., R. Srinivasan, M. DiLuzio, and J. Arnold (2007), *ArcSWAT Interface for SWAT2005: User's Guide*, Texas Agricultural Experiment Station, Blackland Research Center, Temple, Tex.

Wood, S. N. (2010), Statistical inference for noisy nonlinear ecological dynamics, *Nature*, *466*, 1102–1104.

Wu, K., and Y. J. Xu (2006), Evaluation of the applicability of the SWAT model for coastal watersheds in southeastern Louisiana, *J. Am. Water Resour. Assoc.*, *42*(5), 1247–1260.

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