water

MDPI

Article

Evaluation of Four GLUE Likelihood Measures and Behavior of Large Parameter Samples in ISPSO-GLUE for TOPMODEL

Huidae Cho [1,*], Jeongha Park [2] and Dongkyun Kim [2]

[1] Institute for Environmental and Spatial Analysis, University of North Georgia, Oakwood, GA 30566, USA
[2] Department of Civil Engineering, Hongik University, Seoul 04066, Korea; jungha1122@mail.hongik.ac.kr (J.P.); kim.dongkyun@hongik.ac.kr (D.K.)
* Correspondence: hcho@isnew.info

Received: 14 October 2018; Accepted: 26 February 2019; Published: 3 March 2019

Abstract: We tested four likelihood measures including two limits of acceptability and two absolute model residual methods within the generalized likelihood uncertainty estimation (GLUE) framework using the topography model (TOPMODEL). All these methods take the worst performance of all time steps as the likelihood of a model and none of these methods were successful in finding any behavioral models. We believe that reporting this failure is important because it shifted our attention from which likelihood measure to choose to why these four methods failed and how to improve these methods. We also observed how large parameter samples impact the performance of a hybrid uncertainty estimation method, isolated-speciation-based particle swarm optimization (ISPSO)-GLUE using the Nash–Sutcliffe (NS) coefficient. Unlike GLUE with random sampling, ISPSO-GLUE provides traditional calibrated parameters as well as uncertainty analysis, so over-conditioning the model parameters on the calibration data can affect its uncertainty analysis results. ISPSO-GLUE showed similar performance to GLUE with a lot less model runs, but its uncertainty bounds enclosed less observed flows. However, both methods failed in validation. These findings suggest that ISPSO-GLUE can be affected by over-calibration after a long evolution of samples and imply that there is a need for a likelihood measure that can better explain uncertainties from different sources without making statistical assumptions.

Keywords: generalized likelihood uncertainty estimation; hydrologic modeling; uncertainty analysis

1. Introduction

It is important to assess how much uncertainties are involved in hydrologic modeling because there are many different sources of uncertainty including model structures, parameters, and input and output data [1–4]. Model structural uncertainty is due to the fact that we cannot perfectly represent the natural processes involved in hydrologic modeling [4]. Parameter uncertainty indicates that many model parameters are not directly measurable (such as conceptual parameters) or can only be obtained with unknown errors (such as physical parameters) [3]. Measurement uncertainty in input and output data can be caused by unknown measurement errors, incommensurability issues, etc., [2,4]. Uncertainty is either epistemic or aleatory [5]. Epistemic uncertainty results from a lack of our knowledge while aleatory uncertainty arises from random variability. The former can sometimes be reduced by gathering more data and improving our knowledge while the latter cannot. A typical modeling process involves comparing the observed data with unknown errors (aleatory or epistemic) and model output that is simulated by an imperfect model with its own structural (epistemic), parameter (epistemic, aleatory, or both), and measurement (aleatory or epistemic) uncertainties. Since this process brings in different

types of uncertainty in a non-linear and complex manner, it is very challenging, if possible at all, to disaggregate the observational error (difference between the observed and simulated variables) into different sources without making strong statistical assumptions about each source of uncertainty [4]. For example, the parameter and input measurement uncertainties are propagated through the model structural uncertainty resulting in predictions with both epistemic and aleatory uncertainties, which are then compared to the observed data with its own uncertainty. The major problem with error disaggregation is that the observational error is not completely aleatory in nature and may not be appropriately modelled by a statistical model, which is often used to quantify aleatory uncertainty in random variables [5].

The generalized likelihood uncertainty estimation (GLUE) method [6] takes a different approach and does not strongly requires statistical assumptions on errors. This method has been widely used for uncertainty analysis in hydrologic modeling because of its simplicity, ease of implementation, and less strict statistical assumptions about model errors [6–15]. As its name implies, GLUE provides a general framework for uncertainty estimation and, unlike formal Bayesian methods, it does not require statistical assumptions about the structure of model errors although those assumptions may be made within the framework to use a formal likelihood function for statistical analysis [16]. GLUE allows the use of an informal likelihood measure that evaluates the fitness of the model to the observed data. The likelihood measure is a function of the model parameters and observed data, whose value is rescaled to sum to 1 over the ensemble of behavioral models. The likelihood measures before and after incorporating information from the observed data are referred to as the prior and posterior likelihood, respectively. The prior likelihood is typically defined based on the modeler's judgment and knowledge about the model and study area before incorporating information from the observed data. The posterior likelihood is obtained by multiplying the prior likelihood and the likelihood measure given the observed data as follows:

$$L_{posterior}(\theta|\xi,\ y) = \frac{L(\theta|\xi,y)\cdot L_{prior}(\theta)}{C}, \tag{1}$$

where θ is the model parameter set, ξ and y are the observed input and output data, respectively, $L(\theta|\xi,y)$ is the likelihood of θ given ξ and y, $L_{prior}(\theta)$ and $L_{posterior}(\theta|\xi,y)$ are the prior and posterior likelihood of θ before and after observing ξ and y, respectively, and C is a normalizing constant such that the sum of $L_{posterior}(\theta|\xi,y)$ becomes 1. Parameter samples are classified as either "behavioral" or "non-behavioral" depending on their posterior likelihood and predefined threshold value. Behavioral models simulate output variables acceptably giving a likelihood value greater than a preset threshold value or within the limits of acceptability while non-behavioral models do not.

The prediction percentile of the behavioral models is defined as follows:

$$P\left(\hat{Z}_t < z_t\right) = \sum_{i=1}^{n}\left\{L_{posterior}(\theta_i|\xi,y)|\hat{Z}_{i,t} < z_t\right\}, \tag{2}$$

where Z is the variable that the model predicts, t is the time step, $\hat{Z}_t$ is the predicted variable Z at time step t, $P\left(\hat{Z}_t < z_t\right)$ is the prediction percentile of the behavioral models predicting the variable Z as less than z_t at time step t, and n is the number of behavioral models. For each time step, the prediction percentile is evaluated independently, and the lower and upper tails of the prediction percentile are abandoned to build uncertainty bounds with a certain confidence level.

Unlike probabilistic uncertainty estimation methods that try to fit statistical error models to input, output, or parameters in the hope of being able to justify those statistical characteristics of different errors, GLUE does not explicitly separate the observational error into different sources and, instead, focuses on evaluating the "effective observational error," which is the deviation of model predictions from observed variables [1]. Beven extended the concept of model evaluation to include set-theoretic approaches because traditional likelihood measures aggregate errors into a single performance measure

in which errors in certain time steps can easily be compensated for by errors in other time steps [1]. In set-theoretic model evaluation, at each time step during the simulation period, a fuzzy membership function is defined where the observed value takes a peak value of 1 and gets assigned the minimum and maximum acceptable limits of model predictions with a likelihood value of 0. A predicted value at each time step is assigned a membership value from this function. The shape of the membership function can vary depending on applications and available information. These limits of acceptable model predictions are referred to as the "limits of acceptability" in GLUE. Unlike aggregation-based approaches, set-theoretic approaches adopt fuzzy logic operators, especially fuzzy intersection or fuzzy AND, which is the minimum of all membership values. That is, the likelihood value of a model is the minimum membership value of all predicted values across the entire calibration period. Even if a predicted value at just one time step falls outside the limits of acceptability, this model is considered non-behavioral. Beven and Smith also discussed how to formulate a likelihood measure when there is disinformation in the data [17]. In this research, they did not construct the limits of acceptability as a fuzzy membership function, but instead they used the absolute model residual and informative data to express a likelihood measure and assigned a likelihood of 0 to models that fail to predict any rainfall events during simulation. In this study, we refer to this method as the "absolute model residual method."

There are different parameter sampling methods that have been used with GLUE including the nearest neighbor method [6], a hybrid sampling strategy of the genetic algorithm and artificial neural network (GAANN) [18], adaptive Markov Chain Monte Carlo (MCMC) sampling [19], and uniform random sampling that is generally used with GLUE [19], among others. One of recent sampling techniques used with GLUE includes the work of Cho and Olivera [20]. They combined GLUE with a multi-modal heuristic algorithm called isolated-speciation-based particle swarm optimization (ISPSO) [21] to introduce a hybrid uncertainty estimation method, ISPSO-GLUE. They applied the ISPSO-GLUE method to the Soil and Water Assessment Tool (SWAT) [22] and compared ISPSO-GLUE and a random sampling version of GLUE using the Nash–Sutcliffe (NS) coefficient [23] as the likelihood measure. They concluded that, with a faster convergence rate and a smaller number of parameter samples, ISPSO-GLUE only slightly underestimated the predictive uncertainty and likelihood-weighted ensemble predictions compared to GLUE using random sampling. They used the same likelihood measure for both methods, so the only differences between the two methods were the sampling method and how ISPSO-GLUE compensated for biased sampling. Since SWAT is a long-term hydrologic model that is computationally expensive, they conducted uncertainty analysis based on 46,000 parameter samples because of limited computational resources and time constraints. However, considering the 17 model parameters they changed during calibration, 46,000 samples might not have explored the parameter space comprehensively and so might not be enough to evaluate the performance of a large number of samples in the ISPSO-GLUE and GLUE methods. For example, a slight underestimate of predictive uncertainty in their study might have been due to a relatively small number of samples compared to the number of model parameters and ISPSO-GLUE may have not performed favorably if they took a lot more samples because of the effects of overly optimized (or overly conditioned on the calibration data) parameter values even with its compensation for bias. Cho et al. [24] investigated the efficiency of the ISPSO-GLUE method in hydrologic modeling using the topography model (TOPMODEL) [25] and the NS coefficient as the likelihood measure, and found that the cumulative model performance of ISPSO-GLUE converged much faster than GLUE with random sampling and its 95% uncertainty bounds contained 5.4 times more observed streamflows than those of GLUE. Their results showed that random sampling in GLUE initially performed better, but parameter sets from ISPSO-GLUE had improved at a much faster rate and reached an NS coefficient value of 0.82 compared to 0.47 by GLUE. We believe that this faster discovery of better behavioral models is one of the favorable characteristics of ISPSO-GLUE, but their simulation period was limited to only one year because of the lack of input and output data, and the number of parameter sets was 10,000. Also, they only assessed the NS coefficient as a likelihood measure.

Our first objective is to evaluate four likelihood measures including two limits of acceptability methods [1] and two absolute model residual methods [17] within the GLUE framework for TOPMODEL. Since the limits of acceptability and absolute model residual methods are more focused on rejecting models on a time-step-by-time-step (or event-by-event) basis rather than evaluating the models by aggregating individual observational errors from all time steps such as in the NS coefficient, it would be useful to see how these approaches perform in a long-term hydrologic simulation because we believe that the longer the simulation period is the harder it would be to find behavioral models that can make predictions within the limits of acceptability across the entire period. In fact, we reported total failures of these methods at least for our study area and discussed our findings.

The second objective is to compare ISPSO-GLUE and GLUE to observe how each method performs when the number of samples becomes a lot larger (half a million) than 46,000 with SWAT [20] and 10,000 with TOPMODEL [24]. We used random sampling with GLUE because, compared to most of other sampling techniques, this method has an advantage that the modeler does not need to modify the likelihood to reflect non-uniform sampling results [5] and unbiased uniform samples can serve as a reference population for performance comparisons with other uncertainty analysis methods. We compared the performance of ISPSO-GLUE and GLUE in terms of the wall-clock time [26] of simulations, convergence rate, and percentage of enclosed observed data. Finally, we examined samples in the parameter space from both methods to see how well the ISPSO-GLUE samples explored and exploited the problem space, and assessed the applicability of ISPSO-GLUE to the uncertainty analysis of TOPMODEL.

In this study, we did not try to construct unknown statistical models to consider different sources of uncertainty in GLUE explicitly because doing so could easily overestimate information content in the observed data. Instead, we made a subjective assumption that informal likelihood measures implicitly incorporate these uncertainty sources. We also examined the impact of widening the limits of acceptability (adding more uncertainty) on the result of uncertainty analysis.

2. Materials and Methods

2.1. Study Area and Data

The Village Creek watershed in Texas shown in Figure 1 was selected for this study. This watershed has a drainage area of 2228 km^2. There are no major waterbodies and man-made dam structures within the watershed except a couple of small lakes on tributaries. There is the U.S. Geological Survey (USGS) streamflow gage 08041500 at the outlet, which has daily streamflow data from 1 January 2002 to 31 December 2013 [27]. Daily precipitation (PRCP) and pan evaporation (EVAP) data for the same period were available from the Global Historical Climatology Network-Daily (GHCND) dataset [28]. We used this pan evaporation data, which represents potential evapotranspiration [29], as an input to the model. Daily weather data from multiple weather stations (seven rainfall gages and two evapotranspiration gages) were combined by taking the area-weighted average based on the Voronoi diagram [30]. We split the entire data from 2002 to 2013 into five sets of calibration and validation periods to evaluate the impact of the simulation period on the performance of ISPSO-GLUE and GLUE. The calibration and validation periods are 2–6 years starting from 2002 and 2008, respectively, including the first year as a warm-up period (i.e., 2002 and 2008). Actual model evaluation was conducted starting from 2003 and 2009 without the warm-up period, hence 1–5 years of calibration and validation. The mean annual averages of precipitation and pan evaporation during the 12-year period are 997 mm and 1308 mm, respectively. The mean annual runoff ratio is 0.31. Based on the National Land Cover Database (NLCD) [31], the watershed consists of 0.3% open water, 6.3% developed land, 0.1% barren land, 45.0% forest, 32.4% shrub/grass/pasture, and 15.9% wetlands.

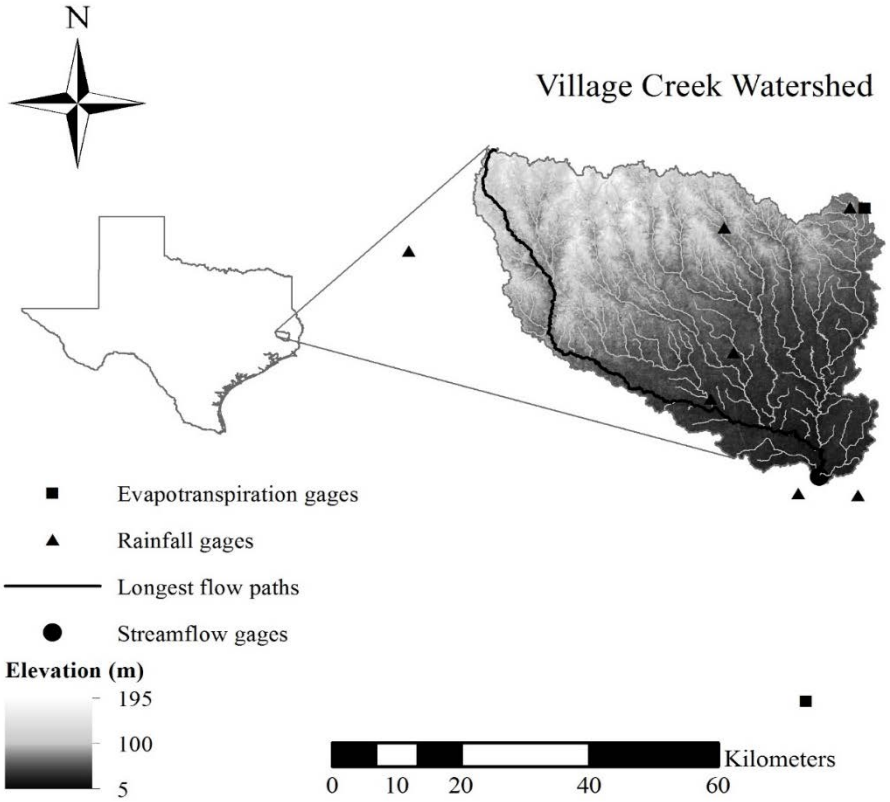


Figure 1. The Village Creek watershed in Texas.

We used the 1 arc-second National Elevation Dataset (NED) [32], which is approximately equivalent to a 30 m resolution digital elevation model (DEM), for delineating catchments, calculating the longest flow path, and computing the topographic index [25] map.

In this study, we used a daily time step because of the lack of observed streamflow data at a finer temporal resolution. We estimated the time of concentration using the TR-55 method [33] and it was greater than one day. Since the time of concentration is longer than the modeling time step, we assumed that the impact of daily time steps on estimated contributing areas would not be significant [34]. A comparison between the TR-55 method and TOPMODEL result will be discussed later.

2.2. Isolated-Speciation-Based Particle Swarm Optimization (ISPSO)

Cho et al. [21] introduced a new variant of particle swarm optimization (PSO) [35,36] called ISPSO. They enhanced the behavior of particles in species-based particle swarm optimization (SPSO) [37] to search multi-modal search spaces more efficiently by incorporating "isolated-speciation" and low-discrepancy sampling. In hydrologic modeling, different sets of the model parameters may produce similar modeling results, which leads to "equifinality" [1]. The multi-modal nature of ISPSO helps particles find solutions from regions of the search space that are substantially different yet producing similar results. ISPSO was implemented in the R language [38] and has successfully been applied for floodway optimization [39], stochastic rainfall generation [40–43], storm tracking [44], and climate change studies [45,46].

In ISPSO, parameter samples are referred to as particles, which are collectively called a swarm. Individual particles share information about the parameter space with their local neighbors and the entire swarm. By sharing the local and global experiences, particles update their positions so that they approach favorable regions of the search space in general. When particles in one neighbor fly around a very small region of the search space and do not converge significantly anymore, they claim to have found a solution referred to as a "nest" whose fitness is the greatest within a predefined radius of the particle. Particles are not allowed within a specified radius from existing "nests" to prevent

further findings at the same locations. This radius is referred to as the "nesting radius." Neighbors with this solution-finding feature are created based on spatial proximity within the search space, and they explore and exploit the search space independently while sharing the global information. For this reason, particles in ISPSO can search the parameter space more efficiently for solutions possibly spread throughout the search space.

2.3. ISPSO-Generalized Likelihood Uncertainty Estimation (GLUE)

Cho and Olivera [20] introduced the ISPSO-GLUE uncertainty analysis method, which incorporates parameter samples from ISPSO into the GLUE method for efficient uncertainty analysis of computationally expensive hydrologic models. The ISPSO-GLUE method is a hybrid uncertainty analysis method that calibrates the model parameters and estimates uncertainty in the model outputs at the same time. This method has been applied to SWAT by Cho and Olivera [20] and to TOPMODEL [25] by Cho et al. [24].

The main difference between ISPSO-GLUE and GLUE with random sampling is that random parameter samples in GLUE are replaced by particles from an ISPSO optimization run and the likelihood measures of those non-random particles from ISPSO are weighted to compensate for an irregular density of particles. Since ISPSO is a multi-modal optimization algorithm rather than a uniform or random sampling technique, those particles sampled from an optimization run are not uniformly nor randomly distributed and need to be treated properly so that the likelihood is not over-predicted where particles cluster together. However, as mentioned earlier, particles from ISPSO are not completely based on the evolution of the particle swarm and many of them are deterministically sampled from the search space based on low-discrepancy sampling. Low-discrepancy sampling ensures that samples are distributed "uniformly" rather than "randomly" in the search space by locating new samples deterministically such that they fill empty regions of the search space that are not occupied by existing low-discrepancy samples. This sampling strategy can help reduce sampling bias caused by particle evolution. Meanwhile, parameter samples in GLUE are randomly distributed across the search space and do not have this irregular distribution bias. Following van Griensven and Meixner [47], once an ISPSO run is completed, the weighted likelihood is calculated using the likelihood weight as follows:

$$\hat{L}(\theta|\xi,\ y) = \omega(\theta) \cdot L(\theta|\xi, y) \tag{3}$$

where $\hat{L}(\theta|\xi,\ y)$ is the weighted likelihood measure, which replaces $L(\theta|\xi, y)$ in Equation (1), and $\omega(\theta)$ is the likelihood weight of the parameter set θ. The likelihood weight $\omega(\theta)$ is calculated by dividing each axis of the search space into equally-spaced multiple intervals and taking the inverse geometric mean of the numbers of samples falling in the same intervals as each particle.

Van Griensven and Meixner [47] used this weighting method to address an over-sampling bias for uncertainty analysis. Their uncertainty analysis method is similar to ISPSO-GLUE in that their method combines an optimization algorithm with an uncertainty analysis framework. However, their method uses a probabilistic likelihood function within a Bayesian framework combined with a global search algorithm rather than a subjective likelihood measure in GLUE with a multi-modal search algorithm, which ISPSO-GLUE employs.

2.4. Topography Model (TOPMODEL)

TOPMODEL is a physically-based distributed hydrologic model and uses the topographic index represented by $\ln \frac{a_i}{\tan \beta_i}$ where a_i is the area of the hillslope per unit contour length draining into point i and β_i is the local slope at this point [25]. The model assumes that areas with similar topographic index values respond hydrologically in a similar manner. The total flow per unit area q_t is expressed as:

$$q_t = q_d + q_r + q_s \tag{4}$$

where q_d, q_r, and q_s are the direct precipitation on saturated areas, return flow, and subsurface flow, respectively. The direct precipitation on saturated areas is calculated by:

$$q_d = \frac{1}{A}\int_{A\in A_s} PdA \tag{5}$$

where A is the total watershed area, A_s represents saturated areas where the storage deficit at point i or S_i is less than or equal to 0, and P is the precipitation intensity. The return flow is written as:

$$q_r = \frac{1}{A}\int_{A\in A_s} |S_i|dA \tag{6}$$

and the subsurface flow can be derived from five assumptions made in TOPMODEL: (1) the recharge rate into the subsurface water table is homogeneous, (2) the hydraulic gradient in the saturated zone is approximated by the local slope, (3) the flow dynamics in the saturated zone is steady state, (4) the downslope transmissivity is an exponential function of the storage deficit, and (5) the lateral transmissivity is also homogeneous. Under assumption (1), the inflow into the saturated zone q_i is described by:

$$q_i = ra_i \tag{7}$$

where r is the recharge rate and a_i is the hillslope area per unit contour length draining into point i. Under assumptions (2) and (3), the outflow from the saturated zone can be expressed as $T_0 \tan\beta_i e^{-fz_i}$ where f is the scaling parameter and z_i is the subsurface water table depth, and set equal to the inflow as in:

$$ra_i = T_0 \tan\beta_i e^{-fz_i}. \tag{8}$$

By solving Equation (8) for z_i, taking its integration, and rearranging, we can obtain the following equation:

$$f(\overline{z} - z_i) = \frac{\overline{S} - S_i}{m} = \left(\ln\frac{a_i}{\tan\beta_i} - \lambda\right) - (\ln T_0 - \ln T_e) \tag{9}$$

where $\overline{z}$ is the average water table depth, $\overline{S}$ is the watershed average storage deficit, m is the soil transmissivity parameter, λ is the areal average of the topographic index or $\lambda = \frac{1}{A}\int \ln\frac{a_i}{\tan\beta_i}dA$, T_0 is the lateral transmissivity at the soil surface, and $\ln T_e$ is the areal average of the soil surface transmissivity or $\ln T_e = \frac{1}{A}\int \ln T_0 dA$. By assumption (4), the downslope transmissivity can be expressed as:

$$T = T_0 e^{-fz_i} = T_0 e^{-\frac{S_i}{m}}. \tag{10}$$

Now, by integrating the right-hand side of Equation (8) with respect to the channel length L and dividing it by the watershed area, we can obtain the areal average subsurface flow q_s as follows:

$$q_s = e^{-\lambda}e^{-\frac{\overline{S}}{m}}\frac{1}{A}\int T_0 a_i e^{\ln T_e - \ln T_0}dL. \tag{11}$$

Under assumption (5), $\ln T_e - \ln T_0 = 0$ and Equation (11) can be simplified as:

$$q_s = T_0 e^{-\lambda}e^{-\frac{\overline{S}}{m}} = e^{-\Lambda}e^{-\frac{\overline{S}}{m}} \tag{12}$$

where $\Lambda = \frac{1}{A}\int \ln\frac{a_i}{T_0\tan\beta_i}dA$. $\ln\frac{a_i}{T_0\tan\beta_i}$ in Λ is referred to as the soil-topographic index. The streamflow is routed using the longest stream distance upstream from the outlet and the channel routing velocity.

The vertical flux from the unsaturated zone q_v is expressed by:

$$q_v = \frac{S_{uz}}{S_i t_d} \tag{13}$$

where S_{uz} is the storage in the unsaturated zone and t_d is the unsaturated zone time delay per unit storage deficit. Actual evapotranspiration E_a is estimated by

$$E_a = E_p\left(1 - \frac{S_{rz}}{S_{r,max}}\right) \tag{14}$$

where E_p is potential evaporation, S_{rz} is the root zone storage deficit, and $S_{r,max}$ is the maximum allowable storage deficit. For infiltration, the infiltration model in [48] that is based on the Green–Ampt assumptions [49] is used.

2.5. r.topmodel

We used the TOPMODEL module called r.topmodel [50] in the Geographic Resources Analysis Support System (GRASS) geographic information system (GIS) [51]. GRASS GIS is a free and open source GIS package with more than 350 geospatial analysis and scientific modules [51]. Cho [50] has reimplemented the FORTRAN 77 code for TOPMODEL (TMOD9502.FOR) in the C language and incorporated the model into GRASS GIS for efficient pre- and post-processing of the model inputs and outputs. Buytaert [52] and Conrad [53] implemented the TOPMODEL program for the R language [38] and the System for Automated Geoscientific Analyses (SAGA) GIS [54], respectively, based on the source code of r.topmodel.

Table 1 shows the list of the r.topmodel parameters that were calibrated in this study. There are 11 model parameters and, since there are no spatially distributed parameter values, the ISPSO algorithm has to solve an 11-dimensional problem. The prior distribution of the parameters was assumed to be uniform because of the lack of a priori information about the model parameters for the watershed. We enabled infiltration excess calculation for all simulations.

Table 1. r.topmodel parameters and their ranges for uncertainty analysis.

Name	Description	Min	Max
qs0	Initial subsurface flow per unit area in m/h	0	0.0001
lnTe	Areal average of the soil surface transmissivity in ln(m^2/h)	−7	10
m	Scaling parameter describing the soil transmissivity in m	0.001	0.25
Sr0	Initial root zone storage deficit in m	0	0.01
Srmax	Maximum root zone storage deficit in m	0.005	0.08
Td	Unsaturated zone time delay per unit storage deficit in h	0.001	40
vch	Main channel routing velocity in m/h	50	2000
vr	Internal subcatchment routing velocity in m/h	50	2000
K0	Surface hydraulic conductivity in m/h	0.0001	0.2
psi	Wetting front suction in m	0.01	0.5
dtheta	Water content change across the wetting front	0.01	0.6

2.6. Four Likelihood Measures in GLUE

2.6.1. Limits of Acceptability

We first investigated the use of the limits of acceptability based on the estimated observational error [1] as a likelihood measure. Given the lack of any information about observation errors provided by the data provider, we estimated the observational error at any time step using two error deviation estimators introduced in [55]. The constant error deviation estimator $\hat{\sigma}$ is defined as follows:

$$\hat{\sigma} = \sqrt{\frac{1}{2(n-1)}\sum_{t=2}^{n}(y_t - y_{t-1})^2} \tag{15}$$

where t is the time step, n is the number of observed values, and y_t is the observed data at time step t. The non-parametric error deviation estimator $\hat{\sigma}_t$ is defined as follows:

$$\hat{\sigma}_t = \sqrt{\begin{pmatrix} 2u \\ u \end{pmatrix}^{-1} (\Delta^u y_t)^2} \tag{16}$$

where Δ^u is the u-order difference operator. We chose $u = 3$ based on [55]. The constant error estimator assumes a constant error deviation across the entire observed time series while the non-parametric error estimator assumes local error deviations in the time series. After estimating observational errors, we decided to use the triangular relative weighting scheme [1], which is one of the simplest likelihood weighting schemes. These error estimators were used to calculate the acceptable range in the triangular relative weighting scheme at each time step by subtracting and adding them to the observed data, and the likelihood measure was defined as the minimum likelihood measure of all time steps (i.e., the poorest-performing time step). Because these error deviation estimators are solely based on the observed data, but there are other sources of uncertainty as well, implicitly we took into account other uncertainty sources in these fuzzy-set-based likelihood measures.

The first likelihood measure using the triangular relative weighting scheme based on $\hat{\sigma}$ can be written as:

$$L_i \propto \min_{1 \le t \le n} \left\{ \max\left(1 - \frac{\varepsilon_{s,i,t}}{\hat{\sigma}},\ 0\right) \right\} \tag{17}$$

where L_i is the likelihood measure of model i, t is the time step, n is the number of time steps, and $\varepsilon_{s,i,t}$ is the absolute model residual at time t. The second likelihood measure based on $\hat{\sigma}_t$ is similarly defined by:

$$L_i \propto \min_{1 \le t \le n} \left\{ \max\left(1 - \frac{\varepsilon_{s,i,t}}{\hat{\sigma}_t},\ 0\right) \right\}. \tag{18}$$

We examined the impact of extending the limits of acceptability on the selection of behavioral models to see how much more implicit uncertainty from other sources would result in better predictions.

2.6.2. Absolute Mode Residual Methods

Next, we evaluated two absolute model residual methods [17] as a likelihood measure. We divided the flow time series into informative and disinformative groups using the runoff coefficient, and discarded disinformative data when calculating the likelihood measure. Since we have daily rainfall and flow observed data, it was not appropriate to define storm events most of the time. To determine disinformative data, we first calculated the time of concentration using the TR-55 method and shifted rainfall time series by the time of concentration to match its overall peaks to those of flow time series. We then divided the flow time series (volume per day) by the drainage area and the rainfall time series (depth per day) to estimate the runoff coefficient. Any observed data with the runoff coefficient greater than 1 was considered disinformative because we assumed that the model is unlikely to be able to simulate an output flow volume greater than its corresponding input rainfall volume. Using informative data only, we considered a model to be non-behavioral if the model's absolute error is greater than a half of the absolute error of data from the mean observed flow at any time step. That is, if a model performs no better than 50% of the data-only-based model across the entire simulation period, the model is considered non-behavioral. The likelihood measure is defined as follows:

$$L_i \propto I_i \cdot \exp\left(-\frac{1}{n_K} \sum_{k \in K} \frac{\varepsilon_{s,i,k}}{\varepsilon_{d,k}} \right) \tag{19}$$

where I_i is the indicator function with 1 for behavioral models and 0 for non-behavioral models, K is a set of informative time steps, n_K is the number of flows in set K, $\varepsilon_{s,i,k}$ is the absolute model residual at

time step k, and $\varepsilon_{d,k}$ is the absolute data error from the mean observed flow at time step k. If $\varepsilon_{s,i,k}/\varepsilon_{d,k}$ is equal to or smaller than 0.5 for all time steps, the model is behavioral and I_i becomes 1. Otherwise, I_i is 0. We tried another denominator as in

$$L_i \propto I_i \cdot \exp\left\{-\frac{1}{n_K}\sum_{k\in K}\frac{\varepsilon_{s,i,k}}{\left(\frac{1}{n_K}\sum_{k\in K}\varepsilon_{d,k}\right)}\right\} \tag{20}$$

where I_i becomes 1 if $\varepsilon_{s,i,k}/(\sum_{k\in K}\varepsilon_{d,k}/n_K)$ is equal to or smaller than 0.5 for all time steps. Otherwise, I_i is 0. $(\sum_{k\in K}\varepsilon_{d,k})/n_K$ denotes the mean absolute data error from the mean observed flow for the informative data.

2.6.3. Random Sampling and Simulation Periods

We took an independent set of 500,000 random parameter samples for each calibration period (i.e., 1–5 years independently) for GLUE simulations, and used only 2003 (1 year) and 2003–2007 (5 years) for this analysis. All these five sets of parameter samples were used for GLUE in the comparison analysis in Section 2.7.

2.7. Comparison of ISPSO-GLUE and GLUE Using the Nash–Sutcliffe (NS) Coefficient

Finally, we used the NS coefficient as the likelihood measure to compare the performance of ISPSO-GLUE and a random sampling version of GLUE. The NS coefficient is calculated as:

$$NS(f(\theta_i,\xi),y) = 1 - \frac{\sum_{t=1}^{n}(y_t - f(\theta_i,\xi)_t)^2}{\sum_{t=1}^{n}(y_t-\overline{y})^2} \tag{21}$$

and the likelihood measure is defined as follows:

$$L_i \propto \max[NS(f(\theta_i,\xi),y),\ 0] \tag{22}$$

where the function $f(\cdot)$ represents the model structure, $f(\theta_i,\xi)$ is the model output, $NS(f(\theta_i,\xi),y)$ is the NS coefficient comparing the model output and observed data, y_t is the observed output at time t, and $f(\theta_i,\xi)_t$ is the simulated output at time t. Since the NS coefficient varies from the negative infinity up to 1, the likelihood measure ranges from 0 to 1. Models with an NS coefficient below 0 perform worse than the "no-model" [34], which is defined as the mean observed streamflow with no model structures at all. For this reason, these models were assigned a likelihood measure of 0. The threshold value for behavioral and non-behavioral model classification was set to 0.6.

For GLUE, each model run is independent from previous runs and no states are saved for further simulations. For the ISPSO-GLUE method, ISPSO stores the current states of particles and keeps track of their local and global optima. Storing this information in memory and running the algorithm to evolve particles based on the saved information can cause an overhead as compared to independent random sampling conducted by GLUE. To see the impact of this overhead on performance, we started the comparison of ISPSO-GLUE and GLUE (using the NS coefficient) by measuring computational time. No matter how good the performance of any uncertainty analysis method is, it is very important to consider how much time would be required to finish such an analysis in terms of wall-clock time [26]. Wall-clock time is the actual amount of time a worker spends to complete a task from start to end. In computing, it includes CPU time for the task itself, CPU time for other non-related tasks, and idle time. Since we performed a huge number of computer simulations, we have run most of those simulations in parallel to speed up the total simulation time. However, we performed both ISPSO-GLUE and GLUE for the 5-year period in a controlled computational environment to measure wall-clock times fairly. In this test, we used Linux kernel version 4.4.14 as the operating system on a desktop computer with the Intel Xeon E5620 2.40GHz CPU and 48GB system memory. R version

3.3.3 was used to run r.topmodel a half million times using each method. For the GLUE method, we used the runif function in R to take random samples while, for the ISPSO-GLUE method, we used the ISPSO R script developed by Cho et al. [21]. Before running each method, we rebooted the computer and did not run any other programs manually other than R to be fair. We used the modification time (last time when file contents were modified) of output files to measure wall-clock time.

We also compared the performance of ISPSO-GLUE and GLUE in terms of the convergence rate of the cumulative NS coefficient, and how much observed data each method predicts successfully within its uncertainty bounds, and discussed the effects of simulation periods on both methods. For each simulation period, a total of 500,000 parameter samples were taken randomly from the parameter space for the GLUE method while a swarm of 20 particles was allowed to evolve 25,000 times to find optimal solutions for the ISPSO-GLUE method, which results in another set of 500,000 parameter samples. The same two sets of random parameter samples (2003 and 2003–2007) were used for the likelihood measure analysis in Section 2.6.

3. Results and Discussion

3.1. Limits of Acceptability and Absolute Model Residual Methods

All four methods with GLUE (Equations (17)–(20)) found no behavioral models at all for the calibration period and, consequently, for the validation period as well. In other words, for all four approaches, not a single model out of a half million random samples was able to simulate flows within the acceptable error deviation for all time steps. We found that it would be very challenging for one model to be able to simulate flows for 5 years or 1826 days within certain limits of acceptability at all times. There are dry and wet seasons during the time of a year and a similar pattern repeats five times. A single set of the model parameters may perform well in either a dry or wet season, but that performance does not guarantee a similar efficiency in the other types of flow seasons.

Specifically, the constant error deviation estimator $\hat{\sigma}$ in Equation (15) was not able to explain error variations across different flow seasons. Figure 2 shows the simulated streamflows from the best NS and 100 other rejected models along with the limits of acceptability using the constant error deviation. As can be seen in this figure, TOPMODEL either highly overestimated or underestimated peak flows and simulated flows often fell outside the acceptable limits defined by the error estimator. The lack of any behavioral models may be attributed to the fact that we did not explicitly consider uncertainties in other sources than the observed data in the likelihood measure itself. We examined how much the acceptable range should be expanded to classify the model with the highest NS coefficient value (the "best" model) as behavioral. We divided the absolute error by the constant error deviation estimator to compute the error ratio relative to the error estimator (ER). About 28% of all time steps were found outside the acceptable range (ER > 1). Even after stretching the acceptable range two times wider, we found about 11% of all time steps outside the wider acceptable range (ER > 2). The largest deviation of the simulated flow from the observed flow was about 23 times larger than the error estimator (ER $\approx$ 23). However, this largest deviation relative to the error estimator occurs at the highest peak flow and does not necessarily mean that the model failed to simulate other time steps altogether. We found that we needed to widen the acceptable range 2.1 times the initial range to find about 90% of time steps within the widened acceptable range. However, even in this case, about 10% of the time, simulated flows fell outside the 2.1 times wider acceptable range. In other words, incorporating more uncertainties from other sources into the acceptable range implicitly by expanding it by 2.1 times would not result in any behavioral models either. Provided that the constant error deviation estimator is 65.6 times greater than the mean observed data and 1.3 times greater than the maximum observed data, more than doubling the observational error to implicitly consider other sources of uncertainty might overestimate the total uncertainty. That is, even explicitly incorporating other sources of uncertainty into the acceptable range might have not resulted in any behavioral models either even after ignoring the 10% worst simulated flows. We needed to expand the limits of acceptability by 23 times to find any behavioral models.

The non-parametric error deviation estimator $\hat{\sigma}_t$ in Equation (16) tries to estimate error deviations locally using local information at any time step in the observed time series. However, when there are very similar flows in consecutive time steps, the error estimator estimates a very small error deviation, which makes it even more challenging to accept any model as behavioral.

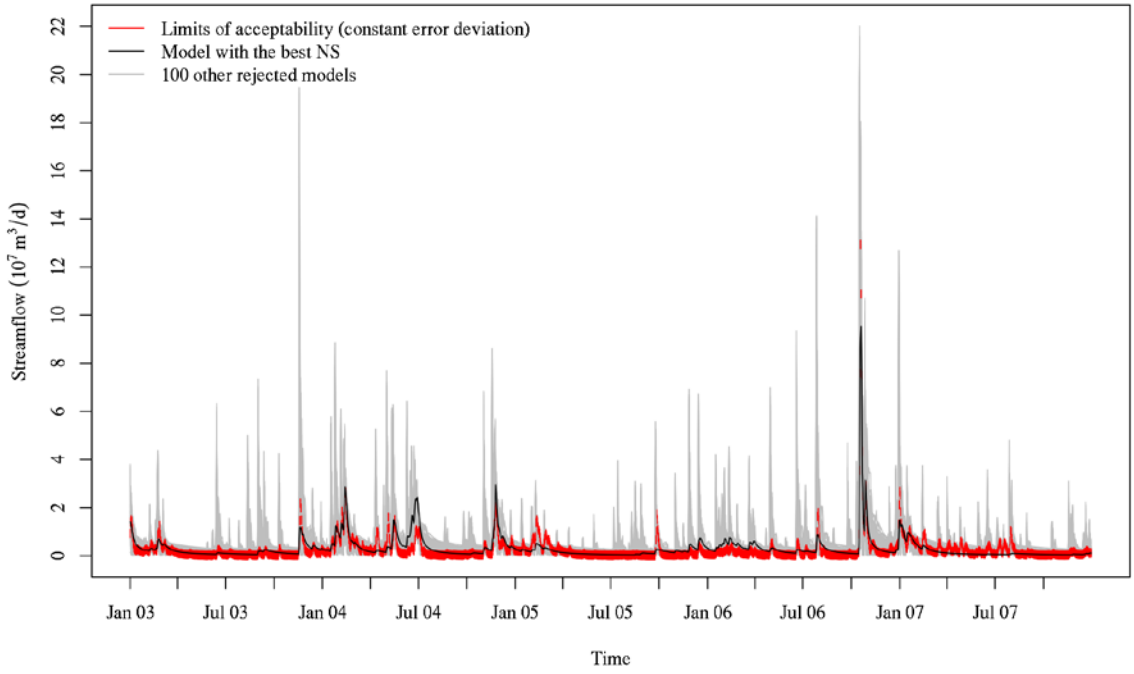


Figure 2. Constant error limits of acceptability, simulated flows with the best Nash–Sutcliffe (NS), and 100 other rejected models. Rejected models are plotted first to clearly show the limits of acceptability. The limits of acceptability using the non-parametric error deviation are not shown because the width of the limits is very narrow.

The two absolute model residual methods in Equations (19) and (20) were not very different in this regard. As mentioned above, the reason why all models were rejected as non-behavioral might be attributed to a long simulation period of 5 years, so we conducted another set of experiments using the two absolute model residual methods for two sets of 1-year calibration (2003 and 2004) and validation (2009 and 2010). However, the results for these simulations have not produced any behavioral models either. For these methods, we estimated the time of concentration as 112 hours or 4.7 days using the TR-55 method for identifying disinformative data. For comparison, we calculated this hydrologic parameter using the best behavioral model from ISPSO-GLUE (NS = 0.80). The internal subcatchment routing velocity vr was found to be 1038 m/h. By dividing the longest flow path 124 km by this routing velocity, we obtained the time of concentration of 119 hours or 5 days. Even using daily time steps, TOPMODEL agreed well with the TR-55 method within a 7% difference when estimating the time of concentration. We believe that this cross validation between TR-55 and TOPMODEL, and 5 days of the time of concentration, imply that the coarse daily temporal scale had minimal impacts on the estimated contributing areas.

This total failure of the four likelihood methods might have revealed the weakness and challenge of evaluating model predictions using the fuzzy AND operator (the minimum likelihood of all time steps), especially in long simulations in such a restrictive way. At the same time, it brings up a question about how we can effectively incorporate input uncertainty into the effective observational error without making unjustifiable statistical assumptions about the structure of the input error. All these questions are left for future research at this point.

We did not run ISPSO-GLUE using these four likelihood methods in this study. Unlike random sampling for GLUE, in case of ISPSO-GLUE, since parameter samples (particles) respond differently

to different objective function surfaces, it is necessary to conduct separate optimization runs using these likelihood methods for comparing ISPSO-GLUE and GLUE, but limited computational resources and time constraints prohibited us from performing this comparison using the limits of acceptability. However, it is not impossible to run ISPSO-GLUE using this fuzzy-set-based objective function with one modification. Unlike random samples that are independent on each other, particles from ISPSO-GLUE depend strongly on information from their own past experience and other particles when evolving, and identification of "better" models with a higher likelihood measure is a crucial requirement for the exploration and exploitation of the search space. With the current limits of acceptability, two non-behavioral models would get assigned a likelihood measure of 0 and an optimization run would not be able to tell in which direction particles should evolve. This information loss occurs because the limits of acceptability function is truncated at 0. The fuzzy membership function (limits of acceptability function) would need to be extended below 0 indefinitely to be able to "sort" the performance of non-behavioral models to produce the necessary information for swarm evolution. Investigating this simulation is left for future research.

3.2. Wall-Clock Time of Simulations

At the end of each method, the total wall-clock times were 725 min for ISPSO-GLUE and 730 min for GLUE. The mean run times per model run were 0.087 s for ISPSO-GLUE and 0.088 seconds for GLUE. GLUE was marginally faster than ISPSO-GLUE until about 364,000 model runs, but later ISPSO-GLUE became slightly faster than the GLUE method. On average, the overhead by particle evolution was almost negligible at the beginning and ISPSO-GLUE finished its simulation about 5 min earlier than GLUE. The wall-clock times of both methods were comparable and may not be one of major factors when choosing ISPSO-GLUE over GLUE.

3.3. Model Performance

Since a half million random samples failed to make predictions with the limits of acceptability and absolute model residual methods, at this point we asked ourselves if those samples would really be all useless in estimating predictive uncertainties in hydrologic modeling for the Village Creek watershed. We believe that it is a valid question because we obtained decent NS coefficient values for the same 5-year simulation (2003–2007 with GLUE). Table 2 summarizes the results of ISPSO-GLUE and GLUE. As can be seen in this table, the maximum NS coefficient was 0.80 for the 5-year GLUE calibration and the number of behavioral models was 28,289 out of a half million. The NS coefficient is an aggregated performance measure where it summarizes all simulated flows and observed flows in a single performance index rather than treating those simulated flows individually and taking the worst performing simulated flow as the representative performance measure just like in the limits of acceptability approach. For this reason, any periods where the model performs poorly can be compensated for by other periods where the model performs very well. This kind of performance compensation may not be acceptable for the limits of acceptability approach, but, sometimes, we may not be able to afford to reject a half million models as non-behavioral just because some portions of the entire simulation period were not simulated acceptably. In this study, we made that compromise and chose to use the NS coefficient as the likelihood measure to compare the two uncertainty analysis methods.

Table 2. Number of behavioral models, percentage of enclosed observed flows, maximum NS, and number of model runs for the maximum NS. The threshold value for behavioral models is a NS coefficient of 0.6. Dashes indicate that there were no behavioral models for the calibration period and no simulations were performed for the validation period. For comparison purposes, all models including behavioral and non-behavioral models were run to calculate the maximum NS coefficient in the validation periods, which are displayed within parentheses.

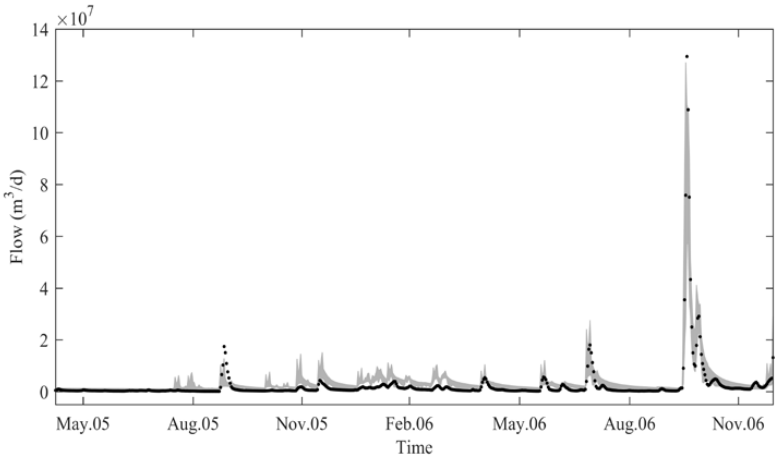
Method	Simulation Period		Number of Behavioral Models	Percentage of Enclosed Observed Flows (%)	Maximum NS	Number of Model Runs for Maximum NS
ISPSO-GLUE	Calibration	2003	497,474	6.6	0.84	478
		2003–2004	-	-	0.51	860
		2003–2005	-	-	0.50	257,475
		2003–2006	492,241	25.7	0.81	574
		2003–2007	491,483	17.4	0.80	871
	Validation	2009	0	0	0.18 (0.28)	290 (3676)
		2009–2010	-	-	- (0.37)	- (19,693)
		2009–2011	-	-	- (0.39)	- (121)
		2009–2012	0	0	0.40 (0.42)	183,567 (183,667)
		2009–2013	0	0	0.42 (0.42)	71,723 (71,723)
GLUE with random sampling	Calibration	2003	6708	22.2	0.84	198,826
		2003–2004	-	-	0.50	1550
		2003–2005	-	-	0.49	17,579
		2003–2006	28,716	30.9	0.81	1402
		2003–2007	28,289	37.6	0.80	19,239
	Validation	2009	0	0	0.22 (0.31)	320,279 (70,871)
		2009–2010	-	-	- (0.40)	- (346,578)
		2009–2011	-	-	- (0.40)	- (1248)
		2009–2012	0	0	0.42 (0.44)	153,849 (185,951)
		2009–2013	0	0	0.44 (0.45)	79,442 (48,571)

However, even the NS coefficient approach has produced a small number of behavioral models only for 3 out of 5 calibration periods for GLUE (1.3% to 5.7% of all model runs) and both methods failed to find any behavioral models for the 2003–2004 and 2003–2005 simulations. ISPSO-GLUE found a lot more behavioral models for the same calibration periods (98.3% to 99.5% of all model runs), but most of those models were found around a small region of the search space, especially on the axes of a few sensitive parameters such as lnTe, m, and vr, which will be discussed later in Section 3.4.

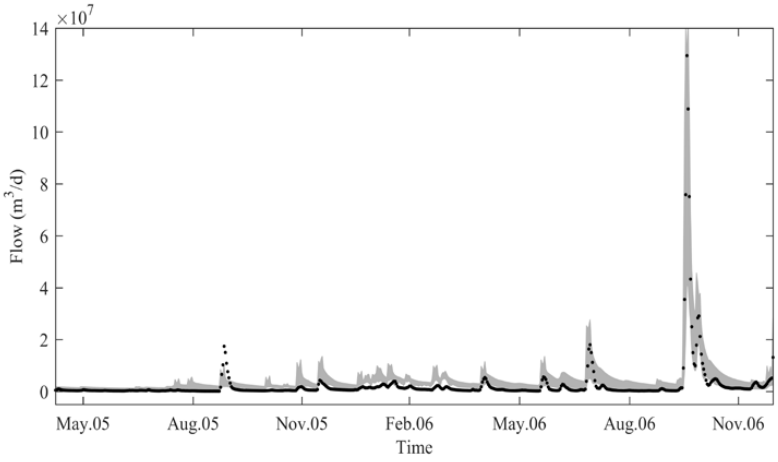
ISPSO-GLUE required up to two orders of magnitude fewer model runs to obtain a similar NS coefficient when both methods found any behavioral models. GLUE outperformed ISPSO-GLUE in terms of the percentage of enclosed observed flows at the expense of more model runs, but the maximum NS coefficients for both methods were comparable. The mean observed flow for the 5-year calibration period (2,598,131 m^3/d) is more than two times that for the 5-year validation period (1,124,422 m^3/d). Also, the variability of the observed flow in the calibration period is much higher than in the validation period. These two simulation periods exhibit very different characteristics of the observed time series and those models that performed well in the calibration period have failed to show similar performance in the validation period for this reason. However, running all half a million random models—not just those behavioral models from the calibration period—in the validation period has produced the maximum NS coefficient of 0.45 for the GLUE method, which is smaller than the threshold NS value of 0.6. In other words, for the GLUE method, if we used the validation period (2009–2013) for calibration, we would have obtained no behavioral models at all from half a million random samples and there would be no behavioral models left to simulate the validation period (2003–2007). We performed a separate set of optimization runs using ISPSO-GLUE for the validation period and obtained the maximum NS coefficient of 0.45. This result confirms that the performance of TOPMODEL in the validation period may not exceed the threshold NS coefficient of 0.6.

The major increase in model performance between the 2003–2005 and 2003–2006 calibration periods can be explained by the highest peak flow on 19 October 2006 and the higher mean observed

flow because of this peak flow. A higher NS coefficient indicates that a model performs relatively better than the mean observed flow. As shown in Figures 3 and 4, both methods predicted the peak flow really well while they did not perform well during the low flow seasons. Since the peak flow was well simulated, its squared residual did not contribute much to the sum of squared residuals in the numerator of Equation (21) while it increased the mean observation and, consequently, the denominator. As a result, the NS coefficient increased significantly, but it does not mean that both methods made good predictions even during the low flow seasons. We could have excluded these high peak flows from NS evaluation from the beginning and rejected all models as non-behavioral based on lower NS coefficient values. This strategy is well aligned with the concept of disinformative data in the absolute model residual methods. Using only informative data to evaluate the NS coefficient after filtering out extremely high peak flows may address this issue with the NS coefficient, but, in our study, we did not explore this NS coefficient approach with informative data only because of computational and time constraints. However, we believe that this approach is worth an investigation in the future.

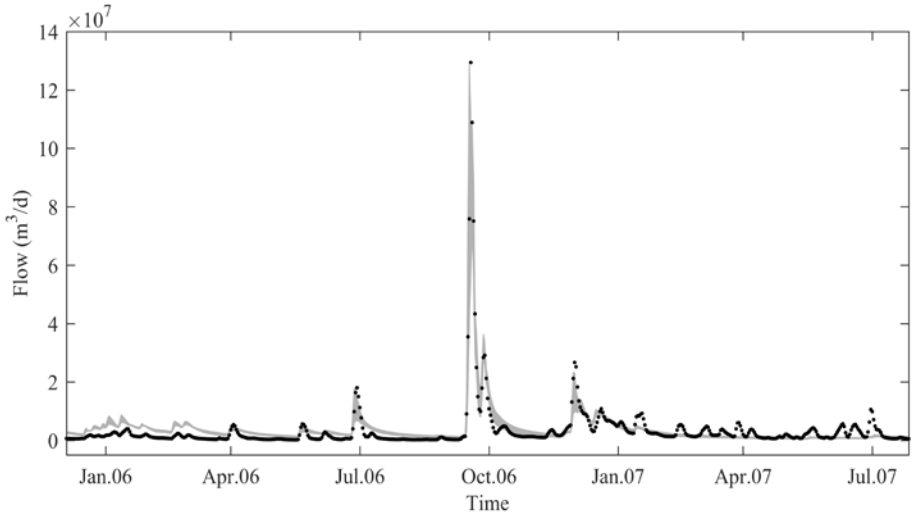


(a) ISPSO-GLUE for the 2003-2006 in calibration period

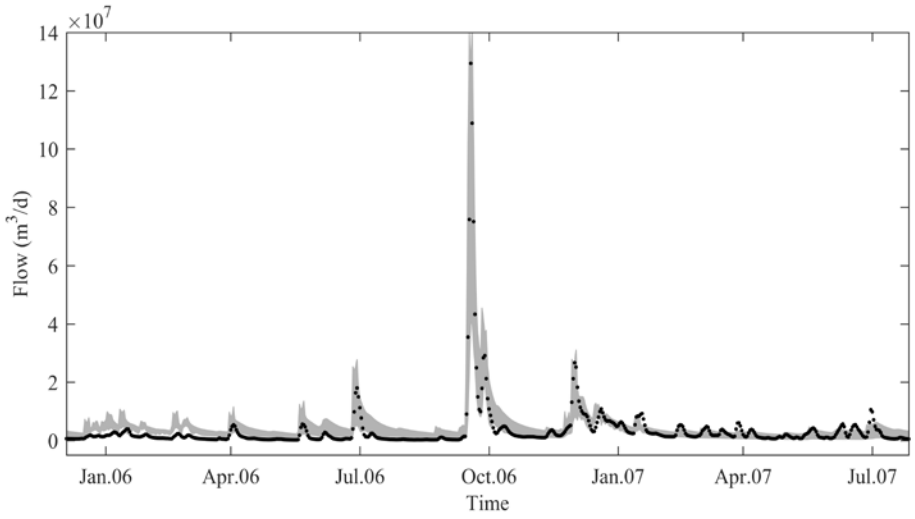


(b) GLUE for the 2003-2006 in calibration period

Figure 3. The 95% uncertainty bounds for the 2003–2006 calibration period from May 2005 to November 2006 (wettest period with the highest peak flow). The shaded area and dotted line represent the uncertainty bounds and observed flows, respectively.



(a) ISPSO-GLUE for the 2003-2007 in calibration period



(b) GLUE for the 2003-2007 in calibration period

Figure 4. The 95% uncertainty bounds for the 2003–2007 calibration period from January 2006 to July 2007 (wettest period with the highest peak flow). The shaded area and dotted line represent the uncertainty bounds and observed flows, respectively.

Overall, the performance of the behavioral models was not great because validation has failed for all five simulation periods. This result was unexpected in the beginning because we assumed that half a million samples should be enough to generate behavioral models for those five sets of calibration and validation periods at least for the GLUE method. For the ISPSO-GLUE method, half a million particles may have been over-conditioned to the calibration data, which can result in poor performance in the validation period. Comparing the maximum NS coefficients for the validation periods between the two methods, we found weak evidence of over-conditioning of particles to the calibration data because ISPSO-GLUE performed marginally worse than GLUE for all validation periods (difference in NS varies from -0.04 to -0.02). However, at the same time, ISPSO-GLUE enclosed fewer observed flows than GLUE did. This observation reveals that finding most models from favorable regions of the search space and weighting them based on the inverse geometric mean of the number of samples may not adequately address the over-sampling bias caused by selective sampling through optimization. The threshold value for the NS coefficient of 0.6 might have been too high when the highly wet period in 2006 inflated the NS coefficient for 2003–2006 and 2003–2007.

However, the threshold value represents the modeler's expectation before analysis and we believe that lowering the threshold value just to obtain more behavioral models is not the correct way to perform uncertainty analysis. The failure of validation for all simulation periods suggests that other likelihood measures might work better for this specific watershed or a rather coarse daily time step due to limited available data may have affected surface runoff produced by infiltration excess or saturation excess. This complete failure in validation can also mean that the model structure of TOPMODEL might not be good enough to describe this catchment. We observed one potential problem in how TOPMODEL routes the total flow to the outlet. If the main channel within a subcatchment is long enough that more than one time step is required to drain the flow, the contributing area of the subcatchment is divided proportionally to the time step. As a result, the flow generated within the subcatchment is also divided in the same way and routed to the outlet. This proportional contribution of the total flow may or may not work depending on the structure of the stream network within the subcatchment and increases the model structural uncertainty in the simulated output.

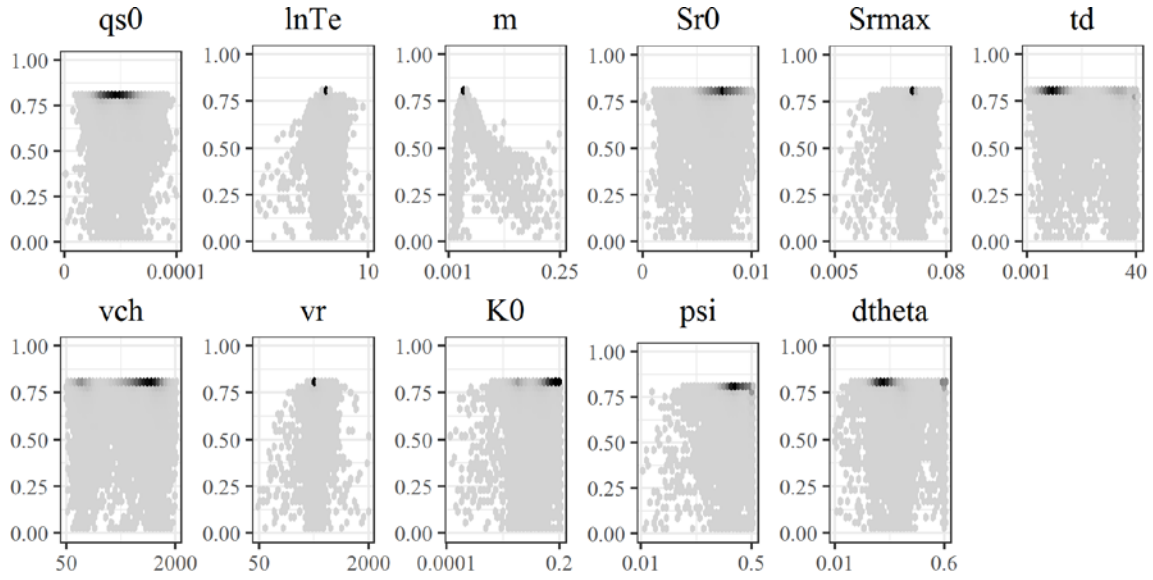
3.4. Behavior of Parameter Samples

The objective function surface needs to be examined to investigate how sensitive the model performance is to parameter values. We used the 5-year calibration and validation periods for this analysis because both methods performed well in the calibration period. Because the dimension of the problem is 11, two-dimensional projections of the objective function surface are presented as the hexagonal bin plots [56] of NS versus each parameter as shown in Figure 5. We chose to plot samples on a hexagonal bin plot because there are half a million points that overlap significantly in a small scatterplot and cannot be represented well in terms of densities if they were plotted individually. For our study, a hexagonal plot was created by first drawing imaginary hexagons in the 2-dimensional space of the NS coefficient versus each parameter, counting the number of samples falling in each hexagon, and color-coding those hexagons with some samples. The color of the hexagons represents the sample density. Figure 5b shows the parameter distributions of random samples from GLUE, which represent the overall shape of the parameter space better than those of selective particles from ISPSO-GLUE as shown in Figure 5a.

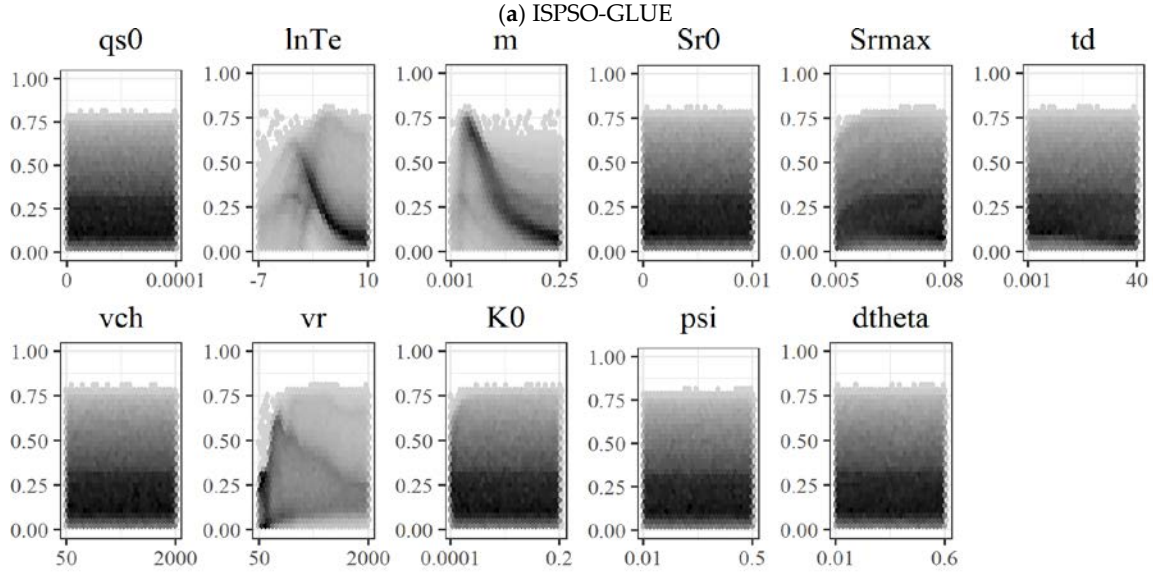
The parameters lnTe, m, and vr showed a similar tendency of random samples clustering around a certain value. The parameter Srmax performed worse closer to the lower limit, but overall, the model performance was not highly sensitive to this parameter. Comparing the two sets of hexagonal bin plots for ISPSO-GLUE and GLUE, we can clearly see that random samples in GLUE were more likely to cluster around regions of low likelihood, as the bottom portion of the plots in Figure 5b is much darker than in Figure 5a. At the same time, particles from ISPSO-GLUE show a tendency to move toward regions with higher likelihood as darker hexagons are more focused around the maximum NS coefficient. In other words, most random samples in GLUE were found in regions of low likelihood while particles in ISPSO-GLUE explored and exploited the search space in regions of high likelihood. Another important observation is that the lnTe, m, and vr parameters exhibit single modality in general, which causes particles from ISPSO-GLUE to cluster around a small region of single optimal parameter values in the search space. This particle behavior reduces the diversity of behavioral models significantly and limits the width of uncertainty bounds, which will be discussed later.

Figure 6 shows the convergence rate of the NS coefficient with respect to the number of model runs for different simulation periods. The validation plots were created using both behavioral and non-behavioral models to see how both methods perform as the number of model runs increases. The cumulative maximum NS coefficient is obtained from the best model found so far at each iteration and indicates how fast the performance of each method converges to the final state. Overall, the convergence rate of ISPSO-GLUE was faster than GLUE. Although ISPSO-GLUE found better samples than GLUE at the very beginning of the simulation in the 2003–2005 calibration as shown in Figure 5e, GLUE took a small number of model runs to move ahead of ISPSO-GLUE. After about 250,000 model runs, ISPSO-GLUE started outperforming GLUE. This observation suggests that taking random

samples from the parameter space can be more beneficial initially until the strategic evolution of particles reaches an optimal solution eventually. To investigate the effect of randomness in sampling on the performance, a bootstrapping analysis without replacement was conducted 1000 times with the GLUE samples for the 3-year calibration period. All re-sampled parameter sets exceeded the performance of ISPSO-GLUE before about 5000 model runs. However, after the initial discovery of a good solution, GLUE had never experienced a significant performance improvement afterward. This analysis shows that random sampling may find good solutions early by luck depending on the landscape of the objective function surface, but it takes more strategic efforts to explore the search space and find better optimal solutions. The number of model runs in ISPSO-GLUE to exceed the performance of GLUE depends on the study area and hydrologic data used. For example, for the Village Creek watershed, the number of required model runs in ISPSO-GLUE varied from 250,000 runs for the 3-year calibration period to less than 1000 runs for the other calibration periods.



(**a**) ISPSO-GLUE



(**b**) GLUE

Figure 5. Hexagonal bin plots of NS vs. model parameters for ISPSO-GLUE and GLUE for the 5-year calibration period. Darker hexagons represent a higher density of samples.

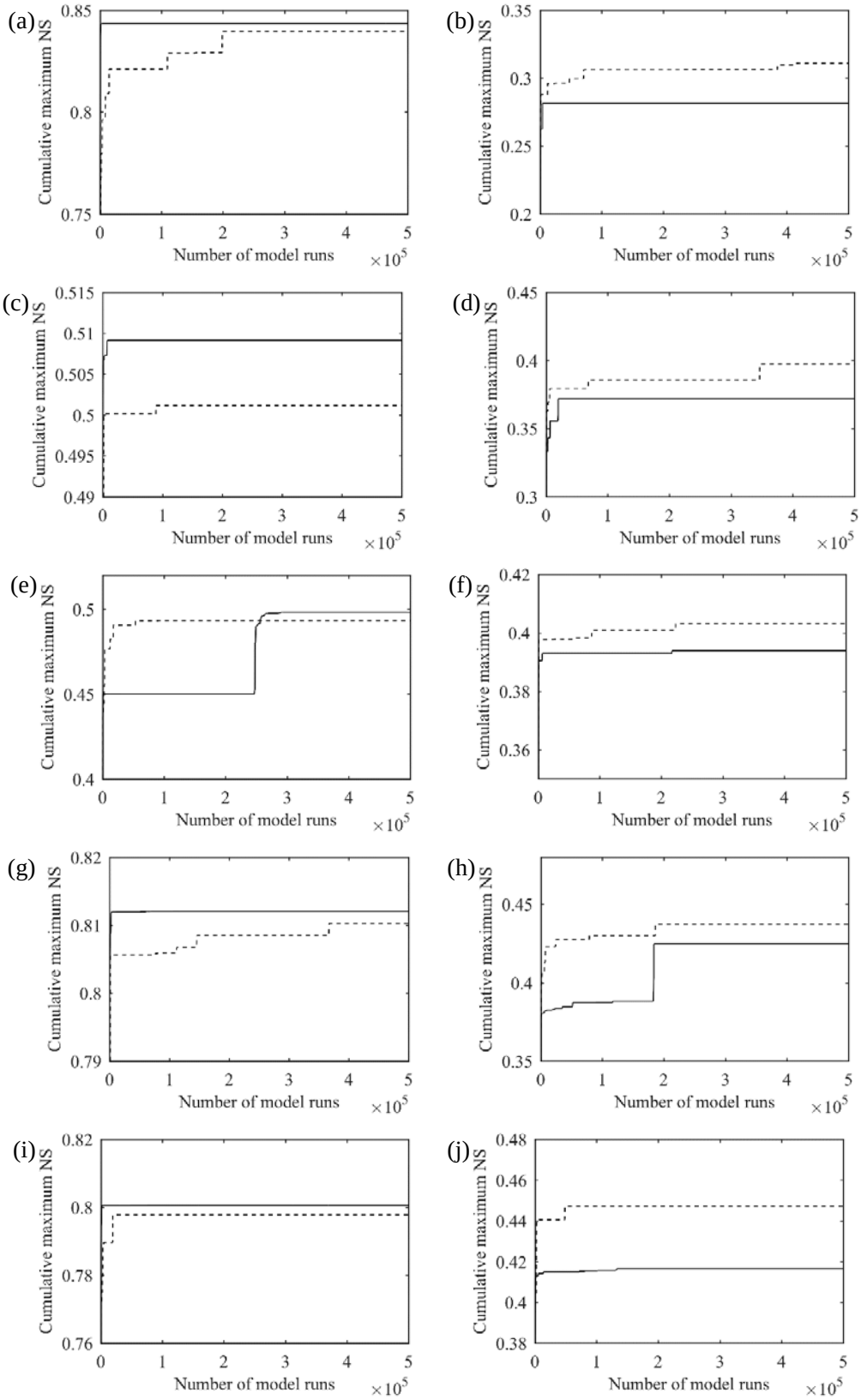


Figure 6. Convergence plot for the NS coefficient. The solid and dashed lines represent ISPSO-GLUE and GLUE, respectively. (**a**) 2003 calibration, (**b**) 2003 validation with all models, (**c**) 2003–2004 calibration, (**d**) 2003–2004 validation with all models, (**e**) 2003–2005 calibration, (**f**) 2003–2005 validation with all models, (**g**) 2003–2006 calibration, (**h**) 2003–2006 validation with all models, (**i**) 2003–2007 calibration, (**j**) 2003–2007 validation with all models.

3.5. Uncertainty Bounds

The number of behavioral models in Table 2 shows biased sampling of ISPSO-GLUE towards optimal solutions with more than 98% of 500,000 models being behavioral while GLUE only found 1%–6% in the calibration periods. Because of a sampling bias in ISPSO-GLUE, the higher number of behavioral samples does not necessarily mean that this method has found that many unique samples that are significantly different from each other. However, the "nesting radius" in ISPSO makes sure that no samples are duplicated within this distance. Also, the weighting method in Equation (3) penalizes highly dense regions with a lot of samples that are similar.

As can be seen in Table 2, even with more behavioral models, the uncertainty bounds for ISPSO-GLUE enclosed less observed flows as compared to those for GLUE. The average percentages of enclosed observed flows for ISPSO-GLUE and GLUE are 16.6% and 30.2%, respectively. Overall, the ISPSO-GLUE uncertainty bounds enclosed 13.6 pp (percentage points) less observed flows than the GLUE uncertainty bounds. However, Cho et al. [24] found contradictory results using TOPMODEL for another Texas watershed during a dry period where ISPSO-GLUE enclosed 79 pp more observed daily streamflows than GLUE. Behavioral models in their GLUE method using random sampling performed worse than those in the ISPSO-GLUE method because of low baseflows. In this study, this relative performance of the uncertainty bounds is reversed and GLUE performed better than ISPSO-GLUE. One of the major differences between this study and their study is the total number of samples. Compared to their study, the total number of samples is 50 times larger, and evolving the particle swarm 50 times more can result in extremely highly populated regions with good likelihood. These over-populated regions can give too much weight to good simulations, which closely simulate observed flows, but may not encompass a lot of them together as an ensemble. In other words, behavioral models from GLUE showed a higher variability than those from ISPSO-GLUE and produced highly variable model outputs with small to large errors surrounding the observed flows. This higher variability of the performance of GLUE was translated into wider uncertainty bounds, which enclose more observed flows.

Figures 3 and 4 show the 95% uncertainty bounds for the wettest period of the 4-year and 5-year calibration periods, respectively. The ISPSO-GLUE uncertainty bounds were generally narrower and missed extreme peak flows as compared to those of the GLUE method. These plots show that, even if ISPSO-GLUE produced uncertainty bounds with a narrow prediction band, and their predictive accuracy was not high enough to enclose a reasonable amount of the observed data when compared to GLUE. This result clearly reveals the weakness of a hybrid uncertainty analysis method like ISPSO-GLUE, where samples are taken to calibrate the model parameters as well as to perform uncertainty analysis. The result presented in this study is less desirable than the result of Cho and Olivera [20] using SWAT, which is highly multi-modal with many spatially distributed parameters. Figure 5 shows the single modality of TOPMODEL in the parameters lnTe, m, and vr, and the insensitivity of the other parameters to the NS coefficient. While single modality makes it easier to calibrate the model parameters, it reduces the diversity of particle sampling once the optimization algorithm found the global solution. Reduced diversity in parameter sampling can lead to low variability in simulated flows in ensemble predictions and cause the uncertainty bounds to miss more observed data.

3.6. Suggestions

Particles from the ISPSO-GLUE method successfully identified sensitive parameters and found regions with behavioral models. However, the strategic evolution of the particle swarm resulted in over-sampling of good behavioral models around preferable solutions even with low-discrepancy sampling, which increases the diversity of samples across the parameter space. Weighting the likelihood measure based on the population density of particles somehow alleviated the adverse effects of over-sampling bias, but the small variability of their high performance led to smaller intervals of the uncertainty bounds compared to those of the GLUE uncertainty bounds. On average, ISPSO-GLUE enclosed about 13.6 pp less observed data than GLUE did. It does not necessarily

mean that ISPSO-GLUE will always enclose a lot less observed data than GLUE as shown by Cho and Olivera [20] and Cho et al. [24], but it is advised that the objective function surface be closely inspected after an ISPSO-GLUE run to interpret and understand the uncertainty analysis results of the ISPSO-GLUE method. Since the single modality of the objective function surface adversely affects the performance of ISPSO-GLUE, the likelihood weighting method should be improved to gain better coverage of the observed data while particle samples are allowed to find optimal solutions. Lastly, it would be worth trying to evaluate an aggregated likelihood measure without disinformative data so that any extreme or abnormal hydrologic responses to forcing data do not highly affect the single index of the model performance in an adverse manner.

4. Conclusions

We investigated the use of four likelihood approaches with GLUE including two limits of acceptability and two absolute model residual methods. Half a million random samples were used to evaluate the four likelihood approaches. All these methods did not produce any behavioral models because it was very challenging for any models to make predictions within the acceptable effective observational error for all time steps. This failure highlighted the challenge of the limits of acceptability approach, especially in long simulations, and moved our attention to how we can better take into account different sources of uncertainty in model evaluation without strong statistical assumptions.

We also examined the applicability of ISPSO-GLUE to TOPMODEL by comparing its results to those of GLUE. A half million samples were taken randomly for the GLUE method to provide unbiased random reference samples for comparison and the same number of samples were generated by ISPSO. We used the NS coefficient for comparing the ISPSO-GLUE and GLUE methods. We split 10 years of data into 5 sets of calibration and validation periods, and performed ISPSO-GLUE and GLUE with random sampling for each calibration period. Behavioral models found during calibration were evaluated for the validation period. ISPSO-GLUE was able to identify sensitive parameters, but both methods failed to find any behavioral models for the 2–3 years' calibration periods. More importantly, both methods failed in all the validation periods given an NS coefficient value of 0.6 as the behavioral threshold, which suggests that other likelihood measures might work better for this watershed or TOPMODEL might not be good enough to describe our watershed. For calibration, it was shown that random sampling may perform better in the beginning because of uniformly distributed samples, but after a certain number of iterations, particles in ISPSO-GLUE started outperforming the random samples with the help of strategic evolution of the particle swarm. ISPSO-GLUE achieved similar performance to GLUE in terms of the maximum NS value with up to two orders of magnitude fewer model runs for the simulations with any behavioral models. However, the uncertainty bounds of ISPSO-GLUE were generally narrower than those of GLUE and, consequently, its uncertainty bounds enclosed 13.6 pp less observed flows on average as compared to the GLUE uncertainty bounds. These relative differences were mainly caused by different parameter sampling strategies employed by the two methods that are highly affected by the shape of the objective function surface, and the length and characteristics of observed data. While the GLUE method randomly takes parameter samples from the search space, the ISPSO-GLUE method takes more samples from regions with high likelihood in the parameter space, leading to a much smaller variability of their model outputs and, hence, narrower uncertainty bounds. The single modality of TOPMODEL for the study area exaggerated this effect as compared to a similar study using SWAT, which exhibits multi-modality.

These findings suggest that an uncertainty analysis method for hydrologic modeling should be chosen carefully by considering the hydrological properties of the study area and observed data, because all these factors eventually affect the landscape of the objective function surface and the uncertainty bounds. At the same time, since the choice of a likelihood measure also strongly affects the performance of calibration and validation, other likelihood measures will need to be evaluated including aggregated performance measures without disinformative data and different limits of acceptability approaches. Future work on the ISPSO-GLUE method will include addressing the small

variability of simulation outputs by further diversifying the search for single-modal problems and the investigation of the time complexity depending on its optimization parameters.

Author Contributions: Conceptualization, H.C.; Methodology, H.C.; Software, H.C.; Validation, H.C., J.P., and D.K.; Formal Analysis, H.C., J.P., and D.K.; Investigation, H.C., J.P., and D.K.; Resources, D.K.; Data Curation, J.P.; Writing—Original Draft Preparation, H.C.; Writing—Review and Editing, H.C., J.P., and D.K.; Visualization, H.C. and J.P.; Supervision, H.C.; Project Administration, D.K.; Funding Acquisition, D.K.

Funding: This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology [NRF-2015-041523 (50%), NRF-2017R1C1B2003927 (50%)].

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Beven, K. A Manifesto for the Equifinality Thesis. *J. Hydrol.* **2006**, *320*, 18–36. [CrossRef]
2. ASME PTC Committee. *Standard for Verification and Validation in Computational Fluid Dynamics and Heat Transfer*; American Society of Mechanical Engineers: New York, NY, USA, 2009.
3. Gorgoglione, A.; Bombardelli, F.A.; Pitton, B.J.; Oki, L.R.; Haver, D.L.; Young, T.M. Uncertainty in the Parameterization of Sediment Build-Up and Wash-Off Processes in the Simulation of Sediment Transport in Urban Areas. *Environ. Model. Softw.* **2019**, *111*, 170–181. [CrossRef]
4. Beven, K. On the Concept of Model Structural Error. *Water Sci. Technol.* **2005**, *52*, 167–175. [CrossRef] [PubMed]
5. Beven, K.; Binley, A. GLUE: 20 Years on. *Hydrol. Process.* **2014**, *28*, 5897–5918. [CrossRef]
6. Beven, K.; Binley, A. The Future of Distributed Models: Model Calibration and Uncertainty Prediction. *Hydrol. Process.* **1992**, *6*, 279–298. [CrossRef]
7. Freer, J.; Beven, K.; Ambroise, B. Bayesian Estimation of Uncertainty in Runoff Prediction and the Value of Data: An Application of the GLUE Approach. *Water Resour. Res.* **1996**, *32*, 2161–2173. [CrossRef]
8. Aronica, G.; Hankin, B.; Beven, K. Uncertainty and Equifinality in Calibrating Distributed Roughness Coefficients in a Flood Propagation Model with Limited Data. *Adv. Water Resour.* **1998**, *22*, 349–365. [CrossRef]
9. Beven, K.; Freer, J.; Hankin, B.; Schulz, K. The use of Generalised Likelihood Measures for Uncertainty Estimation in High Order Models of Environmental Systems. In *Nonlinear and Nonstationary Signal Processing*; Fitzgerald, W.J., Smith, R.L., Walden, A.T., Young, P.C., Eds.; Cambridge University Press: Cambridge, UK, 2000; pp. 115–151. ISBN 978-052-180-044-0.
10. Beven, K.; Freer, J. Equifinality, Data Assimilation, and Uncertainty Estimation in Mechanistic Modelling of Complex Environmental Systems using the GLUE Methodology. *J. Hydrol.* **2001**, *249*, 11–29. [CrossRef]
11. Christiaens, K.; Feyen, J. Constraining Soil Hydraulic Parameter and Output Uncertainty of the Distributed Hydrological MIKE SHE Model using the GLUE Framework. *Hydrol. Process.* **2002**, *16*, 373–391. [CrossRef]
12. Makowski, D.; Wallach, D.; Tremblay, M. Using a Bayesian Approach to Parameter Estimation; Comparison of the GLUE and MCMC Methods. *Agronomie* **2002**, *22*, 191–203. [CrossRef]
13. Freer, J.E.; McMillan, H.; McDonnell, J.; Beven, K. Constraining Dynamic TOPMODEL Responses for Imprecise Water Table Information using Fuzzy Rule Based Performance Measures. *J. Hydrol.* **2004**, *291*, 254–277. [CrossRef]
14. Muleta, M.K.; Nicklow, J.W. Sensitivity and Uncertainty Analysis Coupled with Automatic Calibration for a Distributed Watershed Model. *J. Hydrol.* **2005**, *306*, 127–145. [CrossRef]
15. Zheng, Y.; Keller, A.A. Uncertainty Assessment in watershed-scale Water Quality Modeling and Management: 1. Framework and Application of Generalized Likelihood Uncertainty Estimation (GLUE) Approach. *Water Resour. Res.* **2007**, *43*, W08407. [CrossRef]
16. Smith, P.; Beven, K.J.; Tawn, J.A. Informal Likelihood Measures in Model Assessment: Theoretic Development and Investigation. *Adv. Water Resour.* **2008**, *31*, 1087–1100. [CrossRef]
17. Beven, K.; Smith, P. Concepts of Information Content and Likelihood in Parameter Calibration for Hydrological Simulation Models. *J. Hydrol. Eng.* **2014**, *20*, A4014010. [CrossRef]

18. Khu, S.; Werner, M.G. Reduction of Monte-Carlo Simulation Runs for Uncertainty Estimation in Hydrological Modelling. *Hydrol. Earth Syst. Sci. Discuss.* **2003**, *7*, 680–692. [CrossRef]
19. Blasone, R.; Vrugt, J.A.; Madsen, H.; Rosbjerg, D.; Robinson, B.A.; Zyvoloski, G.A. Generalized Likelihood Uncertainty Estimation (GLUE) using Adaptive Markov Chain Monte Carlo Sampling. *Adv. Water Resour.* **2008**, *31*, 630–648. [CrossRef]
20. Cho, H.; Olivera, F. Application of Multimodal Optimization for Uncertainty Estimation of Computationally Expensive Hydrologic Models. *J. Water Resour. Plan. Man.* **2012**, *140*, 313–321. [CrossRef]
21. Cho, H.; Kim, D.; Olivera, F.; Guikema, S.D. Enhanced Speciation in Particle Swarm Optimization for Multi-Modal Problems. *Eur. J. Oper. Res.* **2011**, *213*, 15–23. [CrossRef]
22. Arnold, J.G.; Srinivasan, R.; Muttiah, R.S.; Williams, J.R. Large Area Hydrologic Modeling and Assessment Part I: Model Development 1. *J. Am. Water Resour. Assoc.* **1998**, *34*, 73–89. [CrossRef]
23. Nash, J.E.; Sutcliffe, J.V. River Flow Forecasting through Conceptual Models Part I—A Discussion of Principles. *J. Hydrol.* **1970**, *10*, 282–290. [CrossRef]
24. Cho, H.; Kim, D.; Lee, K. Efficient Uncertainty Analysis of TOPMODEL using Particle Swarm Optimization. *J. Korea Water Resour. Assoc.* **2014**, *47*, 285–295. [CrossRef]
25. Beven, K.J.; Kirkby, M.J. A Physically Based, Variable Contributing Area Model of Basin hydrology/Un Modèle à Base Physique De Zone d'Appel Variable De l'Hydrologie Du Bassin Versant. *Hydrol. Sci. J.* **1979**, *24*, 43–69. [CrossRef]
26. Institute of Electrical and Electronics Engineers (IEEE); The Open Group. The Open Group Base Specifications Issue 7, 2018 Edition, IEEE Std 1003.1-2017 (Revision of IEEE Std 1003.1-2008). 2018. Available online: https://pubs.opengroup.org/onlinepubs/9699919799/ (accessed on 5 December 2018).
27. U.S. Geological Survey (USGS). Surface-Water Daily Data for the Nation. 2016. Available online: http://waterdata.usgs.gov/nwis/sw (accessed on 10 March 2016).
28. National Oceanic & Atmospheric Administration-National Climatic Data Center (NOAA-NCDC). Global Historical Climatology Network-Daily. 2016. Available online: http://www.ncdc.noaa.gov/cdo-web/search?datasetid=GHCND (accessed on 10 March 2016).
29. Hobbins, M.T.; Ramírez, J.A. Trends in Pan Evaporation and Actual Evapotranspiration Across the Conterminous U.S.: Paradoxical or Complementary? *Geophys. Res. Lett.* **2004**, *31*, L13503. [CrossRef]
30. Voronoi, G. Nouvelles Applications Des Paramètres Continus à La Théorie Des Formes Quadratiques. Deuxième Mémoire. Recherches Sur Les Parallélloèdres Primitifs. *J. Reine Angew. Math.* **1908**, *134*, 198–287. [CrossRef]
31. U.S. Geological Survey (USGS). NLCD 2011 Land Cover (2011 Edition, amended 2014)—National Geospatial Data Asset (NGDA) Land Use Land Cover. 2014. Available online: https://www.mrlc.gov/data (accessed on 5 December 2018).
32. U.S. Geological Survey (USGS). National Elevation Dataset (NED). 2016. Available online: http://nationalmap.gov/elevation.html (accessed on 10 March 2016).
33. USDA-NRCS. Urban Hydrology for Small Watersheds, TR-55. 1986. Available online: https://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/stelprdb1044171.pdf (accessed on 16 March 2017).
34. Cho, H.; Olivera, F. Effect of the Spatial Variability of Land use, Soil Type, and Precipitation on Streamflows in Small Watersheds 1. *J. Am. Water Resour. Assoc.* **2009**, *45*, 673–686. [CrossRef]
35. Eberhart, R.; Kennedy, J. A New Optimizer using Particle Swarm Theory. In *Micro Machine and Human Science, MHS'95, Proceedings of the Sixth International Symposium, Nagoya, Japan, October 1995*; IEEE: New York, NY, USA, 1995; pp. 39–43. [CrossRef]
36. Kennedy, J.; Eberhart, R. Particle Swarm Optimization. In Proceedings of the Fourth IEEE International Conference on Neural Networks, Perth, WA, Australia, 27 November–1 December 1995; IEEE: New York, NY, USA, 1995; pp. 1942–1948.
37. Li, X. Adaptively Choosing Neighbourhood Bests using Species in a Particle Swarm Optimizer for Multimodal Function Optimization. In Proceedings of the Genetic and Evolutionary Computation Conference, Seattle, WA, USA, 26–30 June 2004; pp. 105–116. [CrossRef]
38. R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2006. Available online: http://www.r-project.org (accessed on 3 November 2015).

39. Cho, H.; Yee, T.; Heo, J. Automated Floodway Determination using Particle Swarm Optimization. *Water* **2018**, *10*, 1420. [CrossRef]
40. Kim, D.; Olivera, F.; Cho, H. Effect of the Inter-Annual Variability of Rainfall Statistics on Stochastically Generated Rainfall Time Series: Part 1. Impact on Peak and Extreme Rainfall Values. *Stoch. Environ. Res. Risk Assess.* **2013**, *27*, 1601–1610. [CrossRef]
41. Kim, D.; Olivera, F.; Cho, H.; Lee, S.O. Effect of the Inter-Annual Variability of Rainfall Statistics on Stochastically Generated Rainfall Time Series: Part 2. Impact on Watershed Response Variables. *Stoch. Environ. Res. Risk Assess.* **2013**, *27*, 1611–1619. [CrossRef]
42. Kim, D.; Olivera, F.; Cho, H.; Socolofsky, S.A. Regionalization of the Modified Bartlett-Lewis Rectangular Pulse Stochastic Rainfall Model. *Terr. Atmos. Ocean. Sci.* **2013**, *24*. [CrossRef]
43. Kim, D.; Cho, H.; Onof, C.; Choi, M. Let-it-Rain: A Web Application for Stochastic Point Rainfall Generation at Ungaged Basins and its Applicability in Runoff and Flood Modeling. *Stoch. Environ. Res. Risk Assess.* **2017**, *31*, 1023–1043. [CrossRef]
44. Cho, H.; Kim, D.; Lee, K.; Lee, J.; Lee, D. Development and Application of a Storm Identification Algorithm that Conceptualizes Storms by Elliptical Shape. *J. KOSHAM* **2013**, *13*, 325–335. [CrossRef]
45. Heo, J.; Yu, J.; Giardino, J.R.; Cho, H. Impacts of Climate and land-cover Changes on Water Resources in a Humid Subtropical Watershed: A Case Study from East Texas, USA. *Water Environ. J.* **2015**, *29*, 51–60. [CrossRef]
46. Heo, J.; Yu, J.; Giardino, J.R.; Cho, H. Water Resources Response to Climate and Land-Cover Changes in a Semi-Arid Watershed, New Mexico, USA. *Terr. Atmos. Ocean. Sci.* **2015**, *26*. [CrossRef]
47. Van Griensven, A.; Meixner, T. A Global and Efficient Multi-Objective Auto-Calibration and Uncertainty Estimation Method for Water Quality Catchment Models. *J. Hydroinform.* **2007**, *9*, 277–291. [CrossRef]
48. Beven, K. Infiltration into a Class of Vertically Non-uniform Soils. *Hydrol. Sci. J.* **1984**, *29*, 425–434. [CrossRef]
49. Green, W.H.; Ampt, G.A. Studies on Soil Physics 1. The Flow of Air and Water Through Soils. *J. Agric. Sci.* **1911**, *4*, 11–24. [CrossRef]
50. Cho, H. A GIS Hydrological Modeling System by Using the Programming Interface of GRASS GIS. Master's Thesis, Kyungpook National University, Daegu, Korea, 2000.
51. Neteler, M.; Bowman, M.H.; Landa, M.; Metz, M. GRASS GIS: A Multi-Purpose Open Source GIS. *Environ. Model. Softw.* **2012**, *31*, 124–130. [CrossRef]
52. Buytaert, W. TOPMODEL R Package. 2009. Available online: https://source.ggy.bris.ac.uk/wiki/Topmodel (accessed on 3 November 2015).
53. Conrad, O. SAGA-GIS Module Library Documentation (v2.1.3): Module TOPMODEL. 2003. Available online: http://www.saga-gis.org/saga_module_doc/2.1.3/sim_hydrology_2.html (accessed on 3 November 2015).
54. Olaya, V. *A Gentle Introduction to SAGA GIS*; The SAGA User Group eV: Gottingen, Germany, 2004.
55. Vrugt, J.A.; Diks, C.G.; Gupta, H.V.; Bouten, W.; Verstraten, J.M. Improved Treatment of Uncertainty in Hydrologic Modeling: Combining the Strengths of Global Optimization and Data Assimilation. *Water Resour. Res.* **2005**, *41*, W01017. [CrossRef]
56. Carr, D.B.; Littlefield, R.J.; Nicholson, W.; Littlefield, J. Scatterplot Matrix Techniques for Large N. *J. Am. Stat. Assoc.* **1987**, *82*, 424–436. [CrossRef]