AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE

10.1002/2014WR016520

Key Points:
- Multiobjective model calibration using many hydrological signatures is developed
- Proposed calibration formulations are compared with approaches from literature
- Proposed calibration formulations yield higher hydrological consistency

Correspondence to:
M. Shafii,
mshafiih@uwaterloo.ca

Citation:
Shafii, M., and B. A. Tolson (2015), Optimizing hydrological consistency by incorporating hydrological signatures into model calibration objectives, *Water Resour. Res.*, *51*, 3796–3814, doi:10.1002/2014WR016520.

Received 7 OCT 2014
Accepted 8 MAY 2015
Accepted article online 12 MAY 2015
Published online 31 MAY 2015

Optimizing hydrological consistency by incorporating hydrological signatures into model calibration objectives

Mahyar Shafii[1] and Bryan A. Tolson[2]

[1]Department of Earth and Environmental Sciences, University of Waterloo, Waterloo, Ontario, Canada, [2]Department of Civil and Environmental Engineering, University of Waterloo, Waterloo, Ontario, Canada

Abstract The simulated outcome of a calibrated hydrologic model should be hydrologically consistent with the measured response data. Hydrologic modelers typically calibrate models to optimize residual-based goodness-of-fit measures, e.g., the Nash-Sutcliffe efficiency measure, and then evaluate the obtained results with respect to hydrological signatures, e.g., the flow duration curve indices. The literature indicates that the consideration of a large number of hydrologic signatures has not been addressed in a full multiobjective optimization context. This research develops a model calibration methodology to achieve hydrological consistency using goodness-of-fit measures, many hydrological signatures, as well as a level of acceptability for each signature. The proposed framework relies on a scoring method that transforms any hydrological signature to a calibration objective. These scores are used to develop the hydrological consistency metric, which is maximized to obtain hydrologically consistent parameter sets during calibration. This consistency metric is implemented in different signature-based calibration formulations that adapt the sampling according to hydrologic signature values. These formulations are compared with the traditional formulations found in the literature for seven case studies. The results reveal that Pareto dominance-based multiobjective optimization yields the highest level of consistency among all formulations. Furthermore, it is found that the choice of optimization algorithms does not affect the findings of this research.

1. Introduction

1.1. Background

The calibration problem in the context of hydrological modeling is to find a parameter set whose model output best resembles the observed system's behavior [*Andreassian et al.*, 2014; *Duan et al.*, 1992, 1993; *Kollat et al.*, 2012; *Pechlivanidis et al.*, 2014; *Yapo et al.*, 1998]. The classical statistical approach [*Bates and Watts*, 1988; *Seber and Wild*, 1989] minimizes the sum of squared residuals (i.e., the difference between the observed and simulated system's behavior) to find the most appropriate parameter set. Other statistical methods address the comparison of simulations and observations based on a Bayesian viewpoint [*Box and Tiao*, 1973] and specify a likelihood function, which is then utilized to find the posterior distribution of model parameters and outcomes [e.g., *Ajami et al.*, 2007; *Kavetski et al.*, 2006; *Kuczera*, 1983; *Thiemann et al.*, 2001; *Vrugt et al.*, 2009b]. Among other mathematical strategies for model calibration are generalized likelihood approaches [*Beven*, 2006], the application of fuzzy set theory [*Seibert and McDonnell*, 2002], multicriteria calibration [e.g., *Boyle et al.*, 2000; *Gupta et al.*, 1998; *Shafii and Smedt*, 2009; *Vrugt et al.*, 2003b], etc.

It has been argued by researchers [*Gupta et al.*, 1998, 2005; *Wagener and Gupta*, 2005] that the average or aggregate measures of model data similarity such as sum of squared errors or Nash-Sutcliffe Efficiency (NSE) measure [*Nash and Sutcliffe*, 1970] do not provide the modeler with enough power to make a meaningful comparative evaluation of the consistency in model form, the equations, and function, e.g., representations such as storage or discharge—also more detailed analyses on NSE are provided elsewhere [*Criss and Winston*, 2008; *Gupta et al.*, 2009; *Schaefli and Gupta*, 2007]. To effectively evaluate models, *Gupta et al.* [2008] advocate an approach to model evaluation based on model diagnostics with a clear diagnosing power to detect structural model deficiencies. The diagnostic problem is defined as the identification of modeling components which, when assumed to be functioning properly, explain the discrepancy between the simulated and observed behavior [*Gupta et al.*, 2008]. For instance, if there are any problems in the simulation of overland flow processes by models, the infiltration or saturation excess components of the model might be at fault. More recently, diagnostic model evaluation approach has become an attractive topic in the literature [e.g., *Herbst et al.*, 2009; *Hrachowitz et al.*, 2014; *Ley et al.*, 2011; *Oudin et al.*, 2010; *Viglione et al.*, 2013].

To develop rigorous diagnostic-based model evaluation methods, numerous studies employ hydrological signatures that reflect the functional behavior of the catchment [*Black*, 1997; *Wagener et al*., 2007] that a model should be able to reproduce [e.g., *Carrillo et al*., 2011; *Clark et al*., 2011; *Eder et al*., 2003; *Jothityangkoon et al*., 2001; *McMillan et al*., 2012; *Wagener and Montanari*, 2011]. The signature-based model evaluation approach first uses observed data to identify a number of indices or signatures that quantify relevant aspects of the system behavior or watershed functions. Then, the evaluation approach tests the ability of the watershed model to reproduce these signatures by analyzing the similarities and differences between observed and simulated signatures. The literature shows a variety of signatures proven to be useful for model evaluation including the flow duration curve (FDC) [*Westerberg et al*., 2011; *Yadav et al*., 2007; *Yilmaz et al*., 2008], the spectral density of runoff [*Montanari and Toth*, 2007; *Winsemius et al*., 2009], the rising limb density [*Shamir et al*., 2005; *Yadav et al*., 2007], the base flow index [*Arnold and Allen*, 1999; *Vrugt and Sadegh*, 2013], and the peak distribution [*Sawicz et al*., 2011], among others.

Studies applying hydrological signatures can be categorized as either signature-based model selection or signature-based model calibration. In signature-based model selection (elaborated on in section 1.2), signatures are employed to make a selection among candidate model structures for a given catchment. These model structures are calibrated, independent from the signatures analysis. Studies applying signatures to evaluate the suitability of a single model structure (model evaluation) would fall into this category as well. In contrast, signature-based model calibration (described in detail in section 1.3) involves the implementation of hydrological signatures in the calibration of a particular model structure. As such, the calibrated parameter values are dependent on the signatures considered.

1.2. Signature-Based Model Selection

Hydrological signatures serve as a link between process understanding and models and have been traditionally used to identify an acceptable model structure for a given catchment. For instance, *Jothityangkoon et al*. [2001] incorporated signatures such as the interannual variability of runoff, mean monthly variation of runoff, and FDC in a downward approach for identification of the model components required at different modeling time scales. More recently, *McMillan et al*. [2011] have used a number of diagnostic tests based on different field data, soil moisture, flow, precipitation, etc., to identify an appropriate structure for hydrological modeling at a given catchment. Note that the diagnostic tests developed in *McMillan et al*. [2011] are qualitative plots that are designed based on the information contained in field data.

Euser et al. [2013] demonstrate a new signature-based framework to determine the most appropriate model structure from the many structures available. This determination is accomplished by coupling the signatures with principal component analysis. *Martinez and Gupta* [2010] developed a diagnostic multicriteria model performance evaluation strategy applied to a set of 764 catchments across the United States, aiming to find the most parsimonious plausible model hypothesis. Their findings reveal that conventional regression-based measures (aggregate measures of model data similarity such as sum of squared errors) are not sufficient, and to be more confident about the consistency of model results, measures of water balance and hydrologic variability, among others, must be examined.

1.3. Signature-Based Model Calibration

Yadav et al. [2007] study the relationship between hydrological signatures and the physical characteristics of watersheds and consequently identify the hydrological indices important for constraining discharge such as the number of annual occurrences with considerably high flow, runoff ratio, and slope of FDC. For instance, *Yadav et al*. [2007] demonstrate that runoff ratio constrains the ensemble medium flows for all the studied catchments. This means that, for the simulation model to be working properly, the parameter set needs to be adjusted in such a way that the runoff ratio of simulations remains similar to that of observations. As a result, such hydrological signatures can be used as calibration objectives in the parameter adjustment process.

The literature shows that several approaches have been employed to incorporate hydrological signatures in model calibration and parameter estimation in hydrological modeling. For instance, researchers use Monte Carlo Simulation (MCS) to identify hydrologically consistent parameter sets (i.e., resulting in signature values contained in a predefined range that is thought to be acceptable) in the calibration process [*Hingray et al*., 2010; *Pfannerstill et al*., 2014; *Westerberg et al*., 2011; *Winsemius et al*., 2009]. This approach is inspired from the concept of level of acceptability in calibration and uncertainty estimation in hydrological modeling

[*Beven*, 2006; *Liu et al*., 2009]. Such an approach is also applied to flood frequency estimation in *Blazkova and Beven* [2009] where consistent parameter sets are identified based on limits of acceptability for hydrological summary information (e.g., rainfall characteristics, frequency characteristics of snow water equivalent, etc.). Consistent with the signature satisfaction concept, *Gharari et al*. [2014] provide an alternative approach for parameter identification in the absence of time series data focused on satisfying all parameter and process constraints derived from prior (or expert) knowledge. Their approach is based on stepwise Monte Carlo sampling, where the satisfaction of parameter and process constraints is used to restrict the solution space. *Vrugt and Sadegh* [2013] was the first to apply the approximate Bayesian computation (ABC) concept from the statistics literature [e.g., *Del Moral et al*., 2012; *Marjoram et al*., 2003; *Sisson et al*., 2007] to hydrology as a tool for using hydrological signatures in diagnostic model evaluation, aiming to pinpoint the malfunctioning parts of the hydrologic model.

Martinez and Gupta [2011] address the notion of hydrological consistency and suggest that the model structures and parameters obtained in the classical maximum likelihood estimation must be constrained to reproduce desired hydrological characteristics of the process under investigation. These characteristics can be quantified using multiple hydrological signatures. Thus, an ideal signature-based model calibration would use "numerous" hydrological signatures to guide the parameter search toward regions of hydrological consistency in the search space. The literature shows that a few studies have attempted to implement such an idea in model calibration. For instance, *Seibert and McDonnell* [2002] calibrate the rainfall-runoff model against both standard calibration measures and a number of criteria derived from the soft data (qualitative knowledge from the experimentalist that cannot be used directly as exact numbers). *Yilmaz et al*. [2008] utilize signature information to progressively constrain the ranges of parameters that are found to influence the signature indices, reducing the prior ranges to the intervals associated to the best signature values. *Vrugt and Sadegh* [2013] consider four signatures and build corresponding likelihood-type measures. They then use an iterative Monte Carlo sampling scheme to sample parameter sets that show hydrologic consistency according to these signature-based likelihood-type measures. *van Werkhoven et al*. [2009] used only two signature-based measures in their optimization-based model calibration scheme.

The aforementioned studies demonstrate that only a limited number of signature-based measures have been directly used in model calibration, which might be due to computational limitations in multiobjective optimization associated with many objective functions [*Pokhrel et al*., 2012]. Consequently, to implement many signatures, rather than a full multiobjective analysis, researchers conduct traditional calibration (optimization-based approaches) first, and then postprocess these results to filter out the inconsistent parameter sets. For example, in some of the experiments provided in *Pokhrel et al*. [2012] and *Martinez and Gupta* [2011], model calibration is performed using mean squared errors (MSE) to generate candidate-calibrated parameter sets, and, subsequently, hydrological signatures are employed to identify the most appropriate subset of these parameter sets. A promising and seemingly unexplored approach to signature-based model calibration would be to apply multiobjective optimization concepts with the explicit objective of optimizing the hydrological consistency with respect to many hydrological signatures.

1.4. Research Outline

This study develops novel calibration formulations that utilize both residual-based goodness-of-fit measures and many hydrological signatures as calibration objectives. Sampling in the proposed methodology is conducted using optimization algorithms, with objectives formulated to simultaneously optimize goodness-of-fit measures as well as hydrological consistency. The latter is equivalent to finding parameter sets that reproduce acceptable signature values considering a predefined acceptability level for each signature. If a parameter set yields acceptable values for all signatures, it is considered as fully hydrologically consistent.

The main objective of this research is to formulate and solve a calibration problem where signature quality (measured by a signature scoring function based on simulated versus measured signatures) is directly assessed in the process of parameter search performed by an optimization algorithm. As mentioned earlier, common practice in the literature is either based on signature-based Monte Carlo simulations or to postprocess the results of standard calibration routines with respect to many signatures to identify hydrologically consistent parameter sets. Therefore, the implementation of many signatures in a full optimization framework (i.e., this research) is novel.

The first contribution of this research is to quantify the hydrological consistency based on scoring signatures using their predefined acceptability thresholds. Scoring signatures based on deviations from the limit of acceptability has been previously reported in the literature [*Blazkova and Beven*, 2009]. However, this study develops a novel scoring approach that enables the optimization algorithm to employ the resulting consistency metric in the calibration process. The second contribution is to identify the most suitable optimization approach to obtain hydrologically consistent parameter sets. As such, multiple sampling formulations are introduced and also compared with the two most common signature-based formulations in the literature; this comparison is conducted on the basis of hydrological consistency. The formulations in this research are shown to handle as many as 15 objectives (including 13 signature-based criteria and 2 goodness-of-fit measures) simultaneously and can produce more hydrologically consistent parameter sets in comparison to traditional approaches in the literature [e.g., *Pokhrel et al*., 2012; *Winsemius et al*., 2009; *Yilmaz et al*., 2008] that do not explicitly use consistency to guide the sampling for candidate parameter sets.

2. Methodology

In signature-based model calibration, the modeler needs to identify a set of relevant hydrological signatures, and then, evaluate the hydrological consistency of the results. Consider a standard calibration that is conducted using a sampling procedure that continually adjusts the parameter set based on the optimization of the calibration objectives. In this way, each sampled parameter set is a potential solution to the calibration problem and (if evaluated against a number of hydrological signatures) satisfies a subset of all signatures. If all solutions evaluated in the calibration process are retained in an archive (called the sample set), the hydrological consistency of the results can be visualized using Figure 1—that is, a conceptual graph considering a hypothetical calibration problem. The horizontal and vertical axes in Figure 1 represent the proportion of satisfied signatures ($P_{signatures}$) and sampled solutions ($P_{solutions}$), respectively. For any point in Figure 1, the coordinates (i.e., proportions) are computed based on the following equations:

$$P_{signatures}=k/K \tag{1}$$

$$P_{solutions}=m/M \tag{2}$$

where $P_{signatures}$ and $P_{solutions}$ are the horizontal and vertical coordinates (proportions), k is the number of satisfied signatures, K is total number of signatures considered, m is the number of solutions satisfying at least k signatures, and M is total number of solutions in the archive. For instance, the coordinates of the point shown with a triangle in Figure 1 are 0.6 and 0.4, which means that 40% of the solutions in the sample set satisfy at least 60% of the signatures considered. Modelers are particularly interested in calibration approaches that maximize the number of signatures satisfied. As such, we view the number of signatures satisfied as being the primary criterion for judging calibration result quality while the proportion of satisfied signatures is a secondary criterion. Ideally, a significant proportion the M solutions would be fully hydrologically consistent and thus satisfy all K signatures. However, as long as models are imperfect, it will typically be impractical to expect calibration to lead to the satisfaction of all K signatures (e.g., assuming $K \geq 10$). The results obtained in the numerical experiments of this research will be illustrated in graphs based on Figure 1.

The methods used to compare simulated and observed hydrological signatures and then transform them into calibration criteria scores are described in section 2.1. Afterward, in section 2.2, we elaborate on the aggregation of these criteria scores into a single overall consistency metric to be implemented in different optimization-based calibration approaches. These calibration approaches (and all possible formulations) are subsequently detailed in section 2.3.

2.1. Hydrological Signature Quality and Scoring

The numerical experiments in this study consider 13 signatures that are described in the Appendix A. Each signature is computed from the observed data (S_i^{obs}) and the corresponding simulated model result (S_i^{sim}) and then the deviation between these two quantities must be translated into a normalized criterion value by way of some scoring function. The signature deviations (D_i) are computed on a relative basis, consistent with past studies such as *Yilmaz et al*. [2008] and *Blazkova and Beven* [2009], via equation (3) for all signatures:

$$D_i=\frac{S_i^{obs}-S_i^{sim}}{S_i^{obs}}\times 100 \tag{3}$$

Figure 2 shows the two different scoring functions utilized in this paper that return a signature score (C_i) in the 0–1 range as a function of the signature deviation. In both cases, the calibration objective for a given

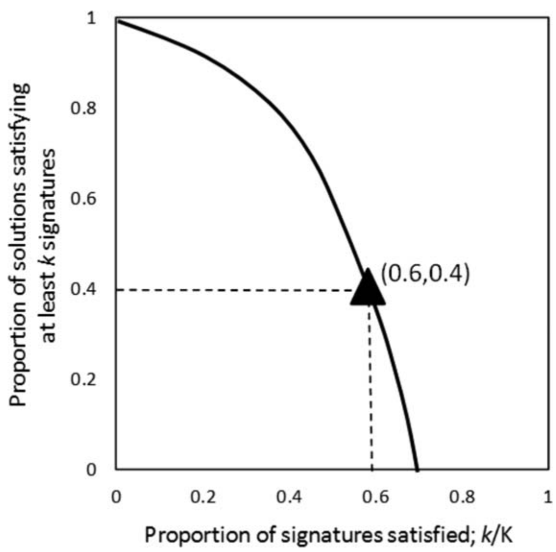

Figure 1. Conceptual graph reflecting the hydrological consistency in a hypothetical calibration problem.

signature is reduced to simply attaining a signature deviation that is within some modeler defined acceptable range based on the acceptability threshold, D_i^*. Here we define the scoring functions to be symmetric but in general, these functions need not be symmetric about 0, all that is required in our approach is to ensure that D_i^* is nonzero. The simplest approach is the binary scoring function, seen in Figure 2a, which defines a simulated signature result to be either satisfactory ($C_i=1$) or unsatisfactory ($C_i=0$). Although not common in hydrology, the binary scoring function has been used in *Komuro et al*. [2006]. Figure 2b shows a continuous scoring function (a piecewise linear function) that scores the distance a simulated signature is from being satisfactory and thus can take any fractional value between 0 and 1 when the signature is not satisfied. Continuous scoring functions are more common in hydrology [e.g., *Blazkova and Beven*, 2009] and strictly speaking need not be piecewise linear. The continuous scoring function, defined in equation (4), requires additional information from the modeler to define the function parameter $D_{i,\max}$ and here we select the range $[-100\%, 100\%]$ for all signatures.

$$C_{(i)}=\begin{cases} 1, & |D_i| \leq |D_i^*| \\ 0, & |D_i| > |D_{i,\max}| \\ \dfrac{|D_{i,\max}|-|D_i|}{|D_{i,\max}|-|D_i^*|}, & \text{Otherwise} \end{cases} \tag{4}$$

The proposed scoring methodology is independent from the type of signatures used in the calibration study, and it can utilize every signature for which an acceptability threshold can be defined. Furthermore, the proposed methodology can also implement standard goodness-of-fit measures such as NSE formulated as follows:

$$NSE=1-\sigma_\varepsilon^2/\sigma_o^2 \tag{5}$$

where σ_ε^2 is the variance of residuals (i.e., difference between simulations and observations), and σ_o^2 is the variance of observations; note that in this paper the measure calculated based on discharge is termed the NSE of high flows, whereas the measure calculated based on log-transformed discharge is termed the NSE of low flows. If the modeler considers an acceptability threshold for NSE, this measure can be directly implemented in the proposed framework. However, since it is not common to optimize NSE up to a threshold, we will also develop formulations (in section 2.3.2.3) that are capable of optimizing both NSE and the scoring functions in Figure 2.

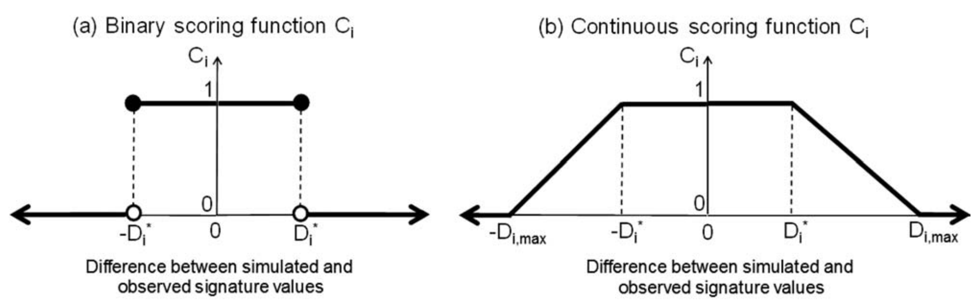

Figure 2. Scoring methods to transform hydrological signature deviation D_i to calibration criterion $C_{(i)}$ based on its acceptability threshold (D_i^*) and the maximum plausible value ($D_{i,max}$): (a) the binary method and (b) the linear method.

2.2. Overall Hydrological Consistency Metric

The overall hydrological consistency metric must be defined as a function that aggregates all of the signature scores defined in the previous section. This metric, designed to be maximized, is formulated based on the number of satisfied signatures (i.e., signature scores of 1.0) as well as some information about the amount of deviation from the acceptability threshold obtained from nonsatisfied scoring functions. Given a parameter set θ and I signatures, when the values of all I signature scores are computed, these values are sorted in a descending order where $C_{(1)}$ indicates the highest score value and $C_{(I)}$ indicates the lowest score value. Thus, the satisfied signatures (score of 1.0) are placed at the top after sorting, followed by nonsatisfied signatures (score less than 1.0). Next, the overall consistency metric ($C_{overall}$) is calculated in equation (6) as the sum of the highest n^*+1 score values, where n^* is the number of satisfied signatures:

$$C_{overall}(\theta)=\sum_{i=1}^{n^*+1} C_{(i)} \tag{6}$$

The signature located at rank (n^*+1) is the one that is the closest (among nonsatisfied signatures) to the acceptability level. The resulting sum is considered as the overall consistency value of parameter set θ. In other words, $C_{overall}$ is a number where the integer part equals the number of satisfied signatures and the fractional part is calculated based on the closest nonsatisfied signature to the acceptability threshold and is equal to C_{n*+1}. This intentionally ignores the simulated signatures that are the least consistent with the observed signatures based on our assumption that in typical model calibration problems, there will be signatures that cannot be satisfied and for these signatures there is little to no value in driving them, for example, from levels that are completely inconsistent to levels that are marginally better but still inconsistent. The overall consistency metric can be maximized to obtain the most hydrologically consistent parameter set. This consistency metric can be implemented in the calibration process using different approaches, which are described in the following section.

2.3. Signature-Based Calibration Approaches

As with all signature-based calibration literature noted in this study, it is assumed that the model calibration involves searching for parameter sets with good quality or optimal values of one or more goodness-of-fit measures (in addition to optimizing hydrologic consistency). Two fundamentally different approaches to signature-based calibration are to (1) implement a parameter sampling scheme that does not adapt or respond to the degree of hydrologic consistency (approach 1 described in section 2.3.1) and (2) utilize a parameter sampling scheme that adapts to the degree of hydrologic consistency of previously sampled parameter sets (approach 2 described in section 2.3.2).

2.3.1. Approach 1: Parameter Search Independent of Hydrologic Consistency

The first approach is consistent with most previous implementations of signature-based model calibration [*Hingray et al.*, 2010; *Martinez and Gupta*, 2011; *Pokhrel et al.*, 2012; *Winsemius et al.*, 2009; *Yilmaz et al.*, 2008]. Two sampling strategies are applied in the first approach including (i) Monte Carlo sampling using the prior parameters range, called A1-MC, and (ii) sampling using a biobjective optimization algorithm (AMALGAM in this study), called A1-BO. The latter considers two calibration objective functions, NSE of flow and NSE of log-transformed flow representing high and low-flow performance measures, respectively. In both strategies, once the sampling terminates, the entire sample set is postprocessed to evaluate the simulated and observed signature values for each parameter set and to calculate its consistency metric accordingly. Note that approach 1 is common practice in the signature-based calibration literature and is considered as the benchmark in this study.

2.3.2. Approach 2: Parameter Search Guided by Hydrologic Consistency

The second approach does not consider postprocessing and instead employs hydrological signatures throughout the calibration process, specifically during parameter optimization. This approach is in fact what is proposed in this research. As such, multiple formulations are considered in this approach based on either criteria aggregation or Pareto-based multiobjective optimization. This study aims at evaluating many different optimization formulations (including single, bi, and multiobjective) that can assess and respond to hydrological consistency throughout the parameter adjustment process.

2.3.2.1. Single-Objective Formulation: A2-SO

The first formulation, called A2-SO, aggregates the goodness-of-fit measures and hydrological signatures into a single objective, the consistency metric defined by equation (6), and applies a single-objective optimization algorithm to maximize the consistency metric. To implement goodness-of-fit measures (such as NSE) in this formulation, a suitable acceptability threshold must be defined. NSE is used in the numerical experiments in this study, and two acceptability thresholds are considered; 0.5 and 0.7. The continuous scoring functions defined for goodness-of-fit measures are conceptually similar to Figure 2b except that for NSE, the ideal NSE occurs at the end of the NSE range (not 0). As such, the NSE scoring functions are piecewise linear; with a value of 0 for any NSE equal or below 0, a value of 1.0 for any NSE equal or above 0.5 (or 0.7), and a linearly interpolated 0–1 score for NSE values between 0 and 0.5 (or 0.7). Considering two NSEs for low and high flows in addition to 13 hydrological signatures scores using the continuous scoring function (Figure 2a), the calibration problem aggregates 15 calibration objectives into the single consistency metric via equation (6). To sample parameter sets from the search space in A2-SO, the DDS optimization algorithm [*Tolson and Shoemaker*, 2007] is applied.

2.3.2.2. Biobjective Formulation: A2-BO

Another formulation is considered in the second approach, called A2-BO, where two objectives are defined and then a multiple objective optimization algorithm is applied to approximate the trade-off between these objectives. The first objective aggregates the goodness-of-fit measures into one value (i.e., the average NSE of low and high flows). Thus, this continuous objective can be defined based on all desired goodness-of-fit measures assuming they have the same scale. The second objective—also continuous—focuses only on hydrologic signatures and is the consistency metric defined by equation (6) that aggregates only the signatures (i.e., excluding NSEs). This formulation yields a standard biobjective optimization problem, which can be solved using a multiobjective optimization algorithm; AMALGAM [*Vrugt and Robinson*, 2007] is used in this study.

2.3.2.3. Multiobjective Formulation: A2-MO

The third formulation in the second approach, called A2-MO, conducts calibration through Pareto-based multiobjective optimization where a separate objective is defined for each signature and each goodness-of-fit measure. This formulation considers the same 15 goodness-of-fit measures and hydrological signatures used in formulation A2-SO. However, the continuous NSE goodness-of-fit measures are used directly as objectives (to be maximized), i.e., no need to any corresponding acceptability thresholds. A2-MO also differs from A2-SO in the sense that it does not aggregate the objectives into a single objective via the hydrological consistency metric. For the signature-based calibration problem addressed in this study, the most straightforward Pareto-based optimization is to directly use individual scores defined in equation (4) as the calibration objectives. However, as elaborated in the rest of this section, there are major issues associated with such an implementation.

The multiobjective optimization literature shows that as the number of objectives increases, when the objectives are not aggregated, the proportion of candidate solutions sampled during optimization that are nondominated (or Pareto solutions) tends to become large. This issue is termed dominance resistance and is known to slow the search [*Farina and Amato*, 2004; *Fleming et al*., 2005; *Purshouse and Fleming*, 2007]. Moreover, *Hadka and Reed* [2012] demonstrate the necessity of an epsilon dominance archive to maintain diversity when applying multiobjective evolutionary algorithms (MOEAs) to problems with more than three objectives. They also report that most MOEAs do not use such an approach to archiving. Nonetheless, solving such a many-objective problem in this way can require an incredibly large computational budget. As a result, the multiobjective benchmarking study by *Hadka and Reed* [2012] as well as most other similar studies consider optimization problems with eight or fewer objectives.

Because more than 10 signatures are implemented in this study, it is assumed that a multiobjective calibration problem formulated as a many-objective optimization problem with 10 or more continuous value objectives (e.g., to minimize $|D_i|$ for all $i = 1, \ldots, 13$ signature-based measures and to maximize two NSEs; 15 objectives in total) will cause the dominance resistance issue. Indeed, preliminary tests using these continuous objectives demonstrated very poor algorithm performance. Therefore, the optimization of many continuous objectives is not considered here. In other words, the scoring method shown in Figure 2b is not appropriate to define objectives in this multiobjective formulation.

Instead, the multiobjective calibration problem (given a large number of objective functions) is approached by converting the continuous objectives into binary objectives utilizing the level of acceptability for each signature-based measure. In other words, the simple binary scoring function demonstrated in Figure 2a is used where the satisfied signatures are given a value of 1 whereas nonsatisfied signatures are assigned a value of 0. This means that the other difference between A2-MO2 and A2-SO is the type of scoring method. The resulting scores can be used to evaluate the dominance among a set of parameter vectors, which means that the non-dominated solutions are assigned rank 1 (i.e., Pareto front) and other dominated solutions are assigned rank 2, 3, etc. Parameter sets can be sorted based on these ranks, and eventually, a Pareto dominance-based multiobjective optimization algorithm can be applied to find the optimal solutions. This is the basis of formulation A2-MO, and AMALGAM is employed to solve this formulation of the calibration problem.

The binary signature scoring functions is hypothesized to make the multiobjective calibration problem of A2-MO more tractable for algorithms like AMALGAM, because the optimization algorithm optimizes the objectives up to the acceptability level, and not unconditionally. This hypothesis was supported by a preliminary experiment where 15 continuous maximization objectives (that were assigned random numbers between 0 and 2) were considered and 10,000 realizations of these sets of objectives were generated. Upon nondominated sorting, almost 70% of the entire set of realizations showed to be Pareto front solutions. However, converting the 15 objectives to 15 binary objectives (i.e., 1 if the random number was less than 1, and 0 otherwise) reduced the number of Pareto front solutions to less than 1%. As a result, the Pareto resistance issues in many-objective optimization [*Farina and Amato*, 2004; *Fleming et al*., 2005] seem to be at least partially mitigated by converting to binary objectives. The application of such a multiobjective optimization formulation is reported in *Komuro et al*. [2006] who develop an ecological process model using a similar multiobjective assessment.

Almost in all multiobjective optimization schemes, the vector optimization problem (that is, having multiple objectives for optimization) is converted into a scalar optimization problem (i.e., a single optimization objective) via some form of scoring (e.g., in A2-SO, the hydrological consistency is the score assigned to each parameter set) or ranking (e.g., in A2-MO, parameter sets are sorted based on the nondomination concept and a rank is assigned to each solution). In this way, the solutions with better ranks or scores are preferred to other solutions. Moreover, in the Pareto-based schemes such as A2-MO, when it comes to the comparison among solutions with the same rank, a form of selection metric is used to select the most appropriate solution. When the binary scoring function was initially utilized to define objectives in the optimization process using AMALGAM, it was still observed that the standard selection metrics performed identically as random selection, and did not effectively guide the search toward regions of high hydrological consistency. This failure was predictable as these selection metrics have not been designed for multiobjective optimization with many binary objectives. To overcome this issue, the consistency metric described in equation (6) was defined as the selection metric to direct the search and thus replaced the standard selection crowding distance selection metric in AMALGAM. Thus, whenever AMALGAM needs to conduct a tournament selection between two solutions (with the same rank), hydrological consistency of each solution is calculated and the one with larger consistency is retained and the other one is discarded.

To use the hydrological consistency in the algorithm's selection procedure is considered a key difference between the proposed methodology to solve formulation A2-MO and a standard Pareto-based multiobjective optimization approach. Furthermore, it is worth noting that a key difference between the single-objective formulation A2-SO and the Pareto-based multiobjective formulation A2-MO is that the former explicitly optimizes the hydrological consistency, whereas the latter uses this metric in the selection phase.

In summary, A2-MO applies AMALGAM to a Pareto-based multiobjective calibration problem with 15 objectives in total, i.e., 2 NSEs (continuous) and 13 signature-based objectives (binary), and furthermore, AMALGAM uses the consistency metric as the selection metric to guide the search toward high-consistency regions in the search space. It is worth noting that since A2-MO does not aggregate the objectives, it is possible to use scoring functions with acceptability thresholds for the goodness-of-fit measures (i.e., NSEs here). Our preliminary results—not shown in the paper—demonstrated that when 13 binary scoring functions were used for signature objectives and the two NSEs were maximization objectives, the results were superior to the case where binary scoring functions were applied to all 15. In this way, it is not necessary to define an acceptability threshold for NSE, which might be more desirable for the modeler.

2.3.3. Summary of Calibration Formulations Tested

Table 1 lists all calibration formulations applied in this work and summarizes the approach/formulation acronym, the precise optimization objectives, as well as the sampling strategy or optimization algorithm

Table 1. Summary of Calibration Formulations Applied in This Study

Approach	Formulation	Acronym	(#) and Type of Objectives Considered in Calibration Experiment	Use Signatures in Parameter Search?	Optimization Algorithm
A1: traditional	Monte Carlo sampling	A1-MC	N/A	No	
	Optimization-based	A1-BO	(2) Maximize NSE of low and high flows	No	AMALGAM
A2: proposed in this study	Single-objective	A2-SO	(1) Maximize hydrological consistency via equation (6) aggregating 15 continuous scoring functions	Yes	DDS
	Biobjective	A2-BO	(2) Maximize avg. NSEs of low and high flows and maximize hydrological consistency via equation (6) aggregating 13 continuous scoring functions for hydrologic signatures	Yes	AMALGAM
	Multiobjective	A2-MO	(15) Maximize 13 binary scoring functions for hydrologic signatures and maximize NSE of low and high flows	Yes	AMALGAMModified to use equation (6) aggregating 15 continuous scoring functions for selection

applied to solve each formulation. As will be pointed out later, two optimization algorithms are used in these formulations; AMALGAM and DDS. To investigate the impact of the choice of optimization algorithm on the results obtained in our numerical experiments, we will also use two other algorithms (AMALGAM-SO and PA-DDS) in some of our case studies, and will present the results in section 4.1.

3. Application

3.1. Case Studies

Calibration of four rainfall-runoff models is considered in this study, ranging from lumped to semidistributed with low to high number of parameters. Table 2 elaborates on these models and the study areas; in total, seven case studies are considered in this research. The first model is HYMOD, a five parameter rainfall excess lumped model connected with a series of linear reservoirs. Details about HYMOD parameters as well as their prior range are provided in the literature [e.g., *Boyle*, 2000; *Vrugt et al*., 2003a]. In this study, HYMOD is applied to two watersheds in the United States from the MOPEX experiment [*Duan et al*., 2006], one in Florida with USGS station code of 02296750, and one in Georgia with the code 02202500. Two years of daily hydrologic data are used for model calibration, and one more year as the evaluation period in both watersheds.

The second model is HBV, originally developed by Swedish Meteorological and Hydrological Institute (SMHI) in the 1970s [*Bergström*, 1976]. Recently, *Aghakouchak and Habib* [2010] developed a lumped version of HBV based on the modified version of this model [e.g., *Lindstrom*, 1997] for teaching purposes, which is also used in this study. This model is run at a daily time step with 10 parameters to be adjusted during calibration (more information in *Aghakouchak and Habib* [2010]). HBV is applied to two more catchments in the MOPEX experiment, one in Illinois with USGS station code of 07196500 and one in Oregon with the code of 11501000. Similar to HYMOD, 2 years of data are used as the calibration period, and 1 year as the evaluation period.

The third rainfall-runoff model used in this study is a semidistributed model called WetSpa that is a hydrologic model simulating water and energy transfer between soil, plants, and the atmosphere. WetSpa is applied at an hourly time scale considering two catchments, the Hornad River catchment located in Slovakia with 5 years of data for calibration and 1 year for evaluation, and the Baron River watershed in the United States using 4 years of data for calibration and 1 year for evaluation. Application of WetSpa to these catchments is previously studied in the literature [*Bahremand et al*., 2007; *Liu et al*., 2003; *Safari et al*., 2012; *Shafii and Smedt*, 2009]. The first and second WetSpa case studies require eight and six parameters, respectively, to be estimated in the calibration stage. All of these parameters are spatially constant over the entire watershed. The time step used to evaluate the WetSpa model output is daily.

The fourth model is SWAT2000 simulation model [*Neitsch et al*., 2001], referred to as SWAT in this paper. SWAT is a spatially distributed model maintained by U.S. Department of Agriculture (USDA) and distributed by the U.S. Environmental Protection Agency (EPA). SWAT is used for predictions of discharge in the Cannonsville Reservoir catchment in upstate New York, which is previously studied by *Tolson* [2005]. This SWAT calibration problem (with the same set of parameters) is studied by *Tolson and Shoemaker* [2007]. There are 26 SWAT parameters (spatially constant over the watershed) to be estimated for this catchment [*Tolson and*

Table 2. Summary of Case Studies Included in This Research

Case Study Number	Sim. Model (Type)	Number of Para.	Catchment (Location)	Acronym	Calibration Length	Evaluation Length	Refs. for More Info
1	HYMOD (lumped)	5	Florida (USA)	HYMOD (FL)	2 years	1 year	*Boyle* [2000] and *Duan et al.* [2006]
2			Georgia (USA)	HYMOD (OG)			
3	HBV (lumped)	10	Illinois (USA)	HBV (IL)			*Aghakouchak and Habib* [2010] and *Duan et al.* [2006]
4			Oregon (USA)	HBV (OR)			
5	WetSpa (semidistributed)	8	Hornad (Slovakia)	HO	5 years	1 year	*Bahremand et al.* [2007] and *Shafii and Smedt* [2009]
6		6	Baron (USA)	BA	4 years	1 year	*Safari et al.* [2012]
7	SWAT (semidistributed)	26	Cannonsville (USA)	TO	2 years	274 days	*Tolson* [2005] and *Tolson and Shoemaker* [2007]

Shoemaker, 2007, Tables 2 and 3]. The simulation period consists of 92 days as the warm-up period, 2 years as the calibration period, followed by 274 more days as the evaluation period. The time step used to evaluate the SWAT model output is daily.

3.2. Experimental Setup

Optimization algorithms solving all formulations are run in five independent trials of 5000 simulations in all case studies. Signature-based model calibration studies report a wide range of computational budgets, from 1000 model simulations [*Yilmaz et al.*, 2008] to 3 million Monte Carlo simulations [*Winsemius et al.*, 2009]. The number of simulations per trial in this paper (5000) was selected based on computational considerations as well as some preliminary experiments to make sure that the calibrated models can satisfy at least some signatures. Note that an identical computational budget was considered for all formulations. In Monte Carlo experiments, a larger computational budget (i.e., increased by 1 order of magnitude to 50,000 simulations) is also considered in each trial.

Three different scenarios are considered in defining the acceptability levels. The first scenario uses the acceptability threshold,D_i^*, at 20% so that $[-20\%, 20\%]$ is the acceptability interval for all signature deviations. Moreover, in terms of the threshold for NSE (in cases where NSE needs to be translated into a continuous scoring function value, e.g., A2-SO), 0.5 is used in the conjunction with $D_i^*=20\%$ for signature deviations. Similarly, the second and third scenarios consider acceptability threshold combinations $\pm10\%$ and 0.7, and $\pm5\%$ and 0.7, respectively. The last scenario applies the strictest thresholds (the most difficult calibration problem) in the consistency-based calibration problem. When the sampling process terminates, the entire set of solutions evaluated in each trial is postprocessed to evaluate satisfaction rates of the signatures and the goodness-of-fit measures in that trial (how many solutions satisfy 15/15 continuous scoring functions, how many solutions satisfy 14/15, etc.).

4. Results

4.1. Comparison of Calibration Formulations

Figure 3 illustrates the average results of the analysis detailed in the experimental setup section for all seven case studies presented by the model name (HBV, HYMOD, WetSpa, and SWAT) and catchment abbreviations (IL, OR, FL, OG, HO, BA, and TO). This figure is consistent with the conceptual graph depicted in Figure 1. The vertical axis in Figure 3 shows the average proportion of solutions satisfying at least $k$of 15 scoring functions in five trials, and the horizontal axis is the proportion of satisfied scoring functions. The vertical axis is logarithmic to show the difference among small proportions more clearly. For example, in the top left plot of Figure 3 (HBV (IL) 20%), A1-BO formulation could only satisfy 87% (or 13) of the 15 scoring functions (13 signatures and 2 NSEs) at best and only 0.7% (or 35) of the 5000 sampled solutions achieved this maximum hydrologic consistency. Note that not all solutions would satisfy the same 13 scoring functions. The left, middle, and right columns of graphs show the results for the first ($\pm20\%$), second ($\pm10\%$), and third ($\pm5\%$) set of acceptability thresholds.

Figure 3 shows that A2-MO, which maximizes the NSE objectives and maximizes the binary scoring functions for each signature, yields the largest proportion of satisfied scoring functions by a higher number of solutions

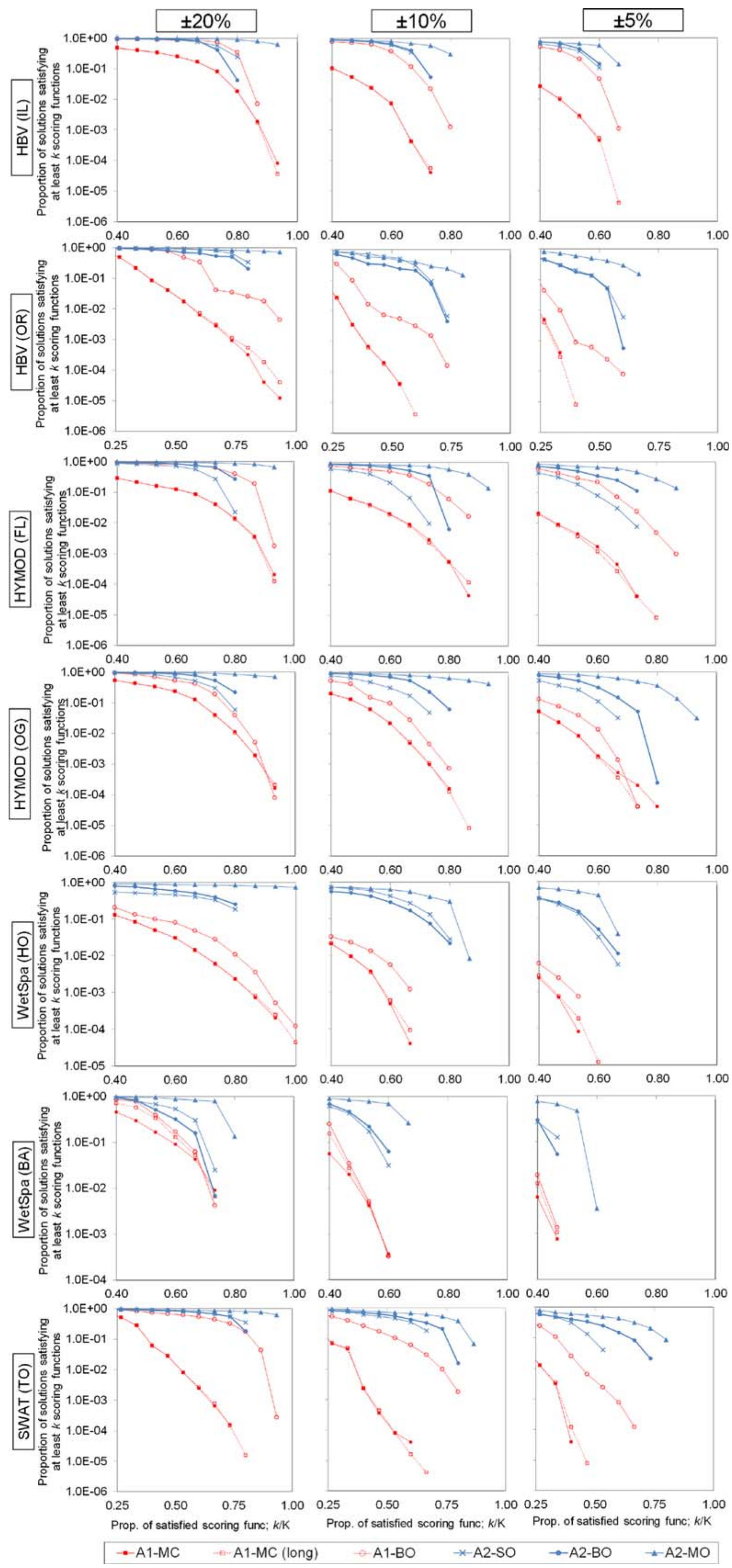

Figure 3. Proportion of the entire set of solutions versus proportion of satisfied scoring functions obtained in all case studies (listed in Table 2) in different formulations (listed in Table 1); A1-MC (long) is Monte Carlo simulations considering a large number of simulations (50,000).

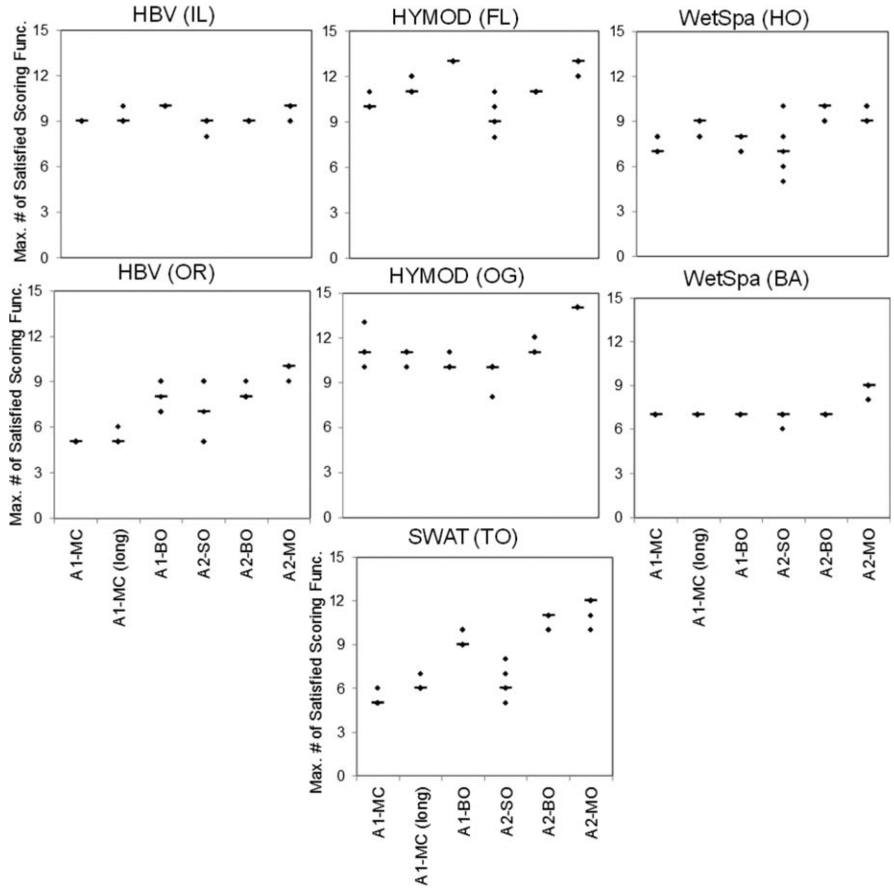

Figure 4. Maximum number of satisfied signatures obtained in five independent trials of different formulations (listed in Table 1) in all case studies (listed in Table 2) given the acceptability threshold of 5%.

in all case studies. The results of the approach A1 using both MC simulations (with low and high computational budget) and AMALGAM optimization algorithm are inferior to the results of the formulation A2-MO. Moreover, it is observed in Figure 3 that both A2-SO and A2-BO perform more promising than the approach A1 in 16 of 21 cases. Overall, this finding indicates that, in comparison to the traditional signature-based calibration approach, implementation of hydrological signatures during the calibration process results in larger number of consistent parameter sets, and a higher overall level of hydrological consistency.

Figure 4 shows the maximum number of satisfied scoring functions in five independent trials, as well as their median, obtained in different formulations for the strictest acceptability level of 5% (note that there are 15 scoring functions in total, 13 signature, and 2 NSEs). Each plot in Figure 4 is associated to one case study. Figure 4 shows that A2-MO generally yields higher consistency than other formulations (highest consistency in five of seven cases). In cases where the results in Figure 4 indicate another formulation is equal or better than A2-MO (A1-BO in HYMOD (FL) and A2-BO in WetSpa (HO)), the proportion of consistent parameter sets given by A2-MO is higher (see Figure 3). These results indicate that when satisfaction of signatures is considered explicitly in the search for highly consistent parameter sets, the high-consistency regions in the search space are explored more effectively. Figure 4 also demonstrates that the single-objective approach (A2-SO) shows substantially higher variability of results in some cases (HYMOD-FL, WetSpa-HO, HBV-OR, and SWAT-TO), which might be due to aggregation of all calibration objectives into a single overall objective. This indicates that while a single A2-SO result may closely approach or equal A2-MO results, A2-SO may be quite poor in another optimization trial. Nevertheless, as the results are based on only five independent trials, such performance variability should be interpreted with caution.

Figure 5 provides the results of evaluating the impact of tying NSE and signature-based scoring functions on parameters consistency for two example case studies, SWAT-TO 5% (left) and HYMOD-FL 5% (right). The horizontal axis in Figure 5 shows different number of hydrological signatures from 1 to 13 meaning that the

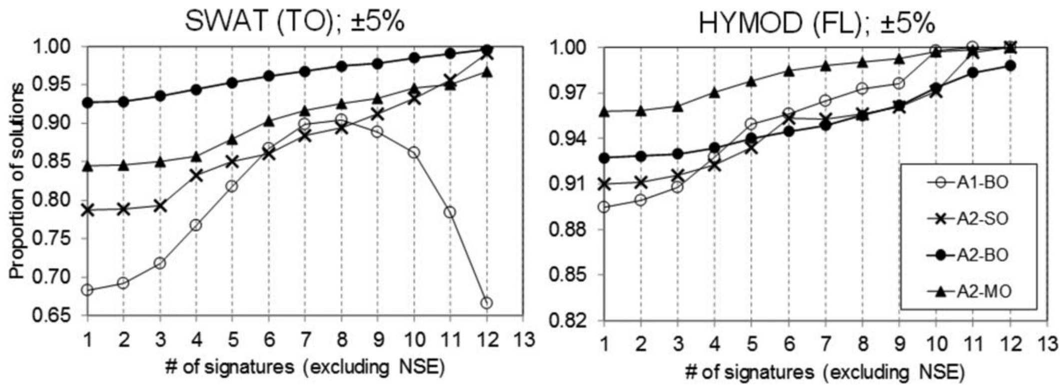

Figure 5. Proportion of solutions satisfying NSE-based measures versus number of satisfied signature-based scoring functions from 1 to 13 (indicating increase of hydrological consistency) in four formulations (listed in Table 1). (left) SWAT-TO 5% case study and (right) HYMOD-FL 5% case study.

signature scores are separated from two NSE measures. The vertical axis shows the proportion of solutions that satisfy both NSE measures; the vertical coordinates are dependent on horizontal coordinates and how they are calculated is explained as follows. For each number of hydrological signatures on the horizontal axis, say *K*, we look at the entire sample and identify the number of parameter sets that can satisfy at least *K* hydrological signatures—say n_1. Then we focus on these n_1 solutions and find those parameter sets that also satisfy both NSE scoring functions—say n_2. The vertical coordinate of each point is calculated as n_2/n_1, i.e., a proportion in the range (0,1). Note that the satisfaction of NSE scoring functions is based on the predefined NSE threshold, e.g., 0.7 in case of 5% acceptability threshold. For instance, in the left plot and looking at the results of A1-BO, the first point on the line indicates that 68% of the solutions that satisfy at least one signature will satisfy both NSE scoring functions as well.

Figure 5 (left) shows a case study where increasing the number of hydrological signatures in approach A1 does not correspond to an increase in the proportion of solutions with good NSE values. Figure 5 (right) demonstrates a case study where approach A1 yields increasing proportions. It was observed that approach A1 resulted in variations similar to the left plot in eight examples out of 21, i.e., seven case studies and three acceptability thresholds. In contrast, the formulations A2-SO, A2-BO, and A2-MO demonstrate near-monotonically increasing proportions in all case studies. The A2 formulations thus effectively tie NSE and hydrological signatures during calibration so that both improve together. These results indicate that the traditional approach of optimizing multiple goodness-of-fit measures with a multiobjective algorithm followed by postprocessing can result in a proportion of parameter sets that may appear consistent based on hydrologic signatures but show relatively poor performance with respect to goodness-of-fit measures.

Figure 6 illustrates the number of satisfied scoring functions on the vertical axis versus average NSE of low and high flows on the horizontal axis obtained in only one trial of the HBV case study in the Illinois catchment (IL). The plots correspond to $\pm$20% (left), $\pm$10% (middle), and $\pm$5% (right) acceptability thresholds. For all formulations, first the quantities plotted in Figure 6 are calculated for all 5000 samples, and then, the solutions are sorted based on the nondomination sorting scheme with respect to these two quantities. To make the figure visually presentable, Figure 6 shows only the Pareto front solutions among 5000 simulations, rather than the entire set of solutions. To show the improvements in the results of A2-MO considering a large computational budget, Figure 6 also demonstrates the Pareto front solutions obtained after 50,000 simulations in A2-MO (marked as A2-MO (long) in Figure 6).

Figure 6 demonstrates that NSE values obtained in formulation A1-BO (which is solely focused on optimizing the NSE values) are just slightly improved over the maximum NSE values of other formulations. In other words, the inclusion of signatures into the calibration objectives does not preclude the search algorithm from returning nearly optimal goodness-of-fit measures. In contrast, A1-BO cannot find overall consistency values as high as that of A2-MO. Figure 6 also shows that increasing the computational budget in A2-MO makes it possible to get more hydrologically consistent results with relatively higher NSE values. Overall, Figure 6 demonstrates that the application of formulation A2-MO generates a new part of the trade-off between goodness-of-fit measures and the consistency metric that the traditional formulation A1-BO does not generate.

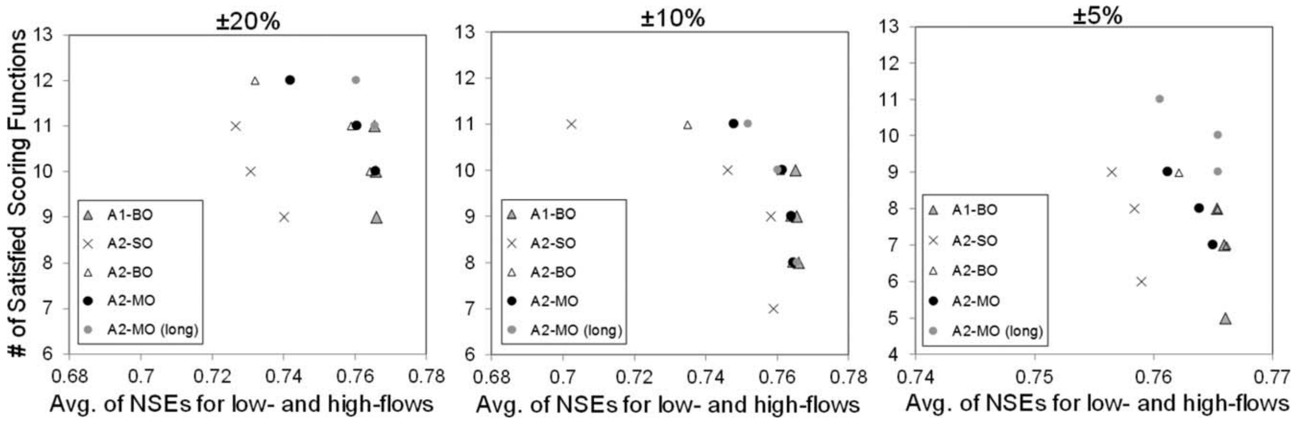

Figure 6. Number of satisfied scoring functions versus average NSE for low and high flows obtained from different formulations (listed in Table 1) in the HBV case study applied to the catchment in Illinois (IL) considering three acceptability thresholds. A2-MO (long) is the same as A2-MO, but with the computational budget of 50,000 simulations.

4.2. Effect of the Choice of Optimization Algorithms

The results illustrated previously are based on application of one set of optimization algorithms in different formulations. To investigate the impact of the choice of optimization algorithms on our findings, other optimization algorithms are also applied to a few of case studies. Among the new experiments is the application of AMALGAM-SO [*Vrugt et al*., 2009a] to calibration of WetSpa in Hornad catchment (HO) for the A2-SO formulation, and PADDS [*Asadzadeh and Tolson*, 2013] to calibration of HBV in the Illinois catchment (IL) for the A2-BO and A2-MO formulations. These algorithms are both among recently developed efficient algorithms, and we chose them because we were comfortable working with them. Any other single, bi, and multiobjective optimization algorithms could be used in this part. Figure 7 illustrates the proportion of satisfied signatures versus the average proportion of samples in five trials. Figure 7 (top) shows the results of the WetSpa case study in A2-SO. Figure 7 (bottom) demonstrates the results of the HBV case study in A2-BO and A2-MO.

Figure 7 indicates that, in the WetSpa case study, the application of AMALGAM-SO does not improve over DDS results and the A2-MO formulation is still clearly preferred. Moreover, in the HBV case study, in spite of some differences between the performance of PADDS and AMALGAM both in A2-BO and A2-MO, A2-MO still shows to be the most promising approach among the formulations considered in this study using both algorithms. For example, in four of six cases using PADDS, A2-MO yields better performance for any number of signatures satisfied. In the other two cases when using PADDS (HBV 10% and HBV 5%), A2-MO yields the best performance at the highest proportion of satisfied signatures. Hence, since the choice of the algorithm does not affect the resulting ranking of the formulations, the overall conclusions do not change, and the full multiobjective formulation (A2-MO) remains as the most promising formulation with respect to hydrological consistency. It should be pointed out that a full analysis of the sensitivity of the results to the type of optimization algorithm was not the goal of these experiments. Rather, we intended to investigate whether the best formulation would be impacted by a change in the type of optimization algorithm, which in fact was not the case.

5. Concluding Remarks

This study develops a multiobjective calibration methodology where both goodness-of-fit measures such as NSE and hydrological signature-based measures are implemented as explicit objectives in the optimization formulation. First, a scoring method is developed to transform any hydrological signature to a calibration objective on the basis of an acceptability threshold for the signature. These scores are aggregated to an overall metric that quantifies the hydrological consistency of each parameter set. The proposed scores and the consistency metric are subsequently implemented in various signature-based calibration frameworks where different formulations are provided and multiple optimization algorithms are applied to take samples from parameter space and adapt the sampling according to hydrologic signature values. The formulations include (i) single-objective optimization that maximizes the hydrological consistency metric; (ii) biobjective

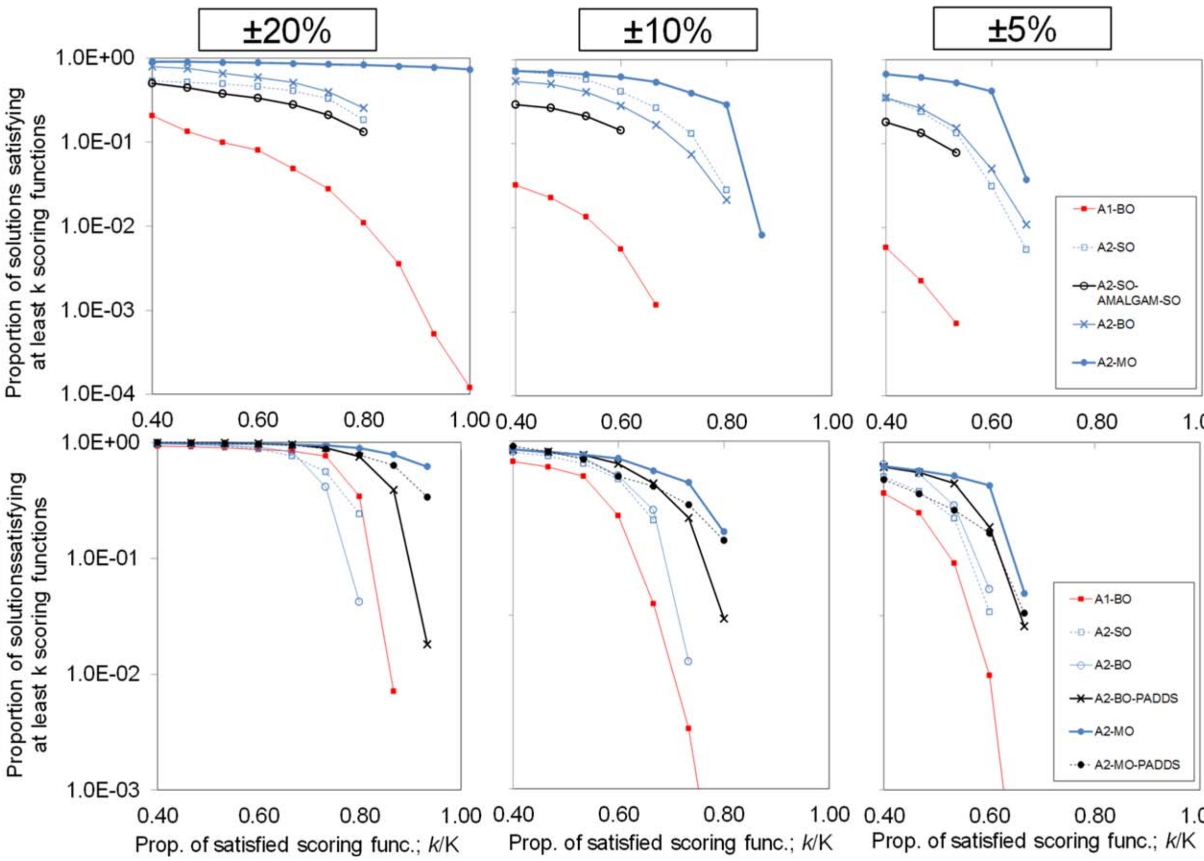

Figure 7. Proportion of the entire set of solutions versus proportion of satisfied scoring functions obtained in the (top) HBV and (bottom) WetSpa case studies in the formulations listed in Table 1. Results of A1-BO, A2-SO, A2-BO, and A2-MO are the same as Figure 3 while new optimization algorithms used in A2-SO-AMALGAM-SO and A2-BO-PADDS.

optimization where the first and second objectives are aggregated goodness-of-fit measures and the hydrological consistency metric, respectively; and (iii) Pareto-based multiobjective optimization implementing individual objectives for each hydrologic signature and each goodness-of-fit measure. In the last formulation, due to particular issues with many-objective optimization, the signatures are scored 1/0 (satisfied/non-satisfied) and the hydrological consistency is used as the multiobjective algorithm selection metric.

These proposed formulations are compared with the traditional signature-based model evaluation approaches that either are based on Monte Carlo simulations or postprocess the calibration results using hydrological signatures. The comparison among calibration formulations is on the basis of the hydrological consistency of the parameter sets quantified in terms of our proposed consistency metric. Proposed calibration formulations as well as the traditional approach are tested in seven case studies using a set of optimization algorithms.

The results show that the full multiobjective Pareto-based calibration approach yields the highest level of hydrological consistency in all case studies in comparison to other optimization schemes, Monte Carlo sampling, and the traditional calibration approaches that only optimize residual-based measures. In other words, consideration of individual hydrological signatures throughout the calibration process enhances the chance of finding more hydrologically consistent parameter sets. Moreover, this formulation results in a larger number of consistent parameter sets returned by the optimization algorithm. Superiority of this formulation in comparison to traditional formulations is more significant in calibration problems with stricter levels of acceptability (i.e., 5%). This finding highlights the advantage of multiobjective optimization concepts, especially Pareto dominance, in signature-based calibration of hydrologic models where algorithms respond directly to the quality of hydrologic signatures.

This study also evaluates the effect of the choice of optimization algorithm on the findings by using different single and multiobjective algorithms to some of the case studies. The results demonstrate that, despite slight changes in the outcome of applying different algorithms, the final ranking of the formulations against achieved hydrological consistency level remains the same when an alternative algorithm is selected. Therefore, the findings depend more on the formulation as opposed to the optimization algorithm.

Recently, *Vrugt and Sadegh* [2013] developed a signature-based model calibration using a sequential Monte Carlo sampling based on ABC, applied to a linear regression and a lumped conceptual modeling. In the present paper, the ABC approach is not implemented, because the ABC concept is very recent to hydrology and is only demonstrated on a few lumped modeling (models with very small computational burden) case studies. The traditional methods implemented in this research are more established, as they have been applied in multiple research studies. Given the difference between the proposed optimization-based methodology and the Monte Carlo-based ABC in *Vrugt and Sadegh* [2013], future work should compare these two methods with respect to the hydrological consistency as well as the computational expense.

Since signature-based model calibration constrains the model output from the hydrological perspective, it provides the modeler with insight into the performance of different modules in the hydrologic simulation model. Therefore, the components of the model that are at fault may be identified, and steps can be taken to improve these components accordingly. When many signatures are considered (similar to this study), the modeler will have more appropriate control on different elements of the model structure. A possible research avenue for future studies can be to perform further analysis of the calibration results to identify model structural adequacy.

Appendix A: Description of Hydrological Signatures

Table A1 in this appendix presents the mathematical formulation of 13 hydrological signatures used in this study. These signatures can be calculated for the observed or simulated time series of discharge.

Table A1. Hydrological Signatures and the Corresponding Measures Used in This Study; These Signature Measures Are Calculated Using N Time Steps in the Time Series of Flows Q

Hydrological Signature	Signature Description	Signature Calculation Based on Time Series of Flow Data (Q) (Q Can Either be Simulated of Observed)
Water balance	(1) Overall runoff ratio considering Q	$\sum_{t=1}^{N} Q_t / \sum_{t=1}^{N} P_t$, where P_t is precipitation in time step t
	(2) Overall runoff ratio considering $\ln(Q)$	$\sum_{t=1}^{N} \ln(Q_t) / \sum_{t=1}^{N} P_t$, where P_t is precipitation in time step t
Flow duration curve (FDC)	(3) FDC midsegment slope	$\log(Q_{m1}) - \log(Q_{m2})$, where $m1$ and $m2$ are the lowest and highest flow exceedance probabilities within the midsegment of FDC (0.2 and 0.7 in this research)
	(4) FDC high-segment volume	$\sum_{h=1}^{H} Q_h$, where $h = 1,2,\ldots,H$ are flow indices located within the high-flow segment (exceedance probabilities lower than 0.02); H is the index of the maximum flow
	(5) FDC low-segment volume	$-1.\sum_{l=1}^{L} (\log(Q_l) - \log(Q_L))$, where $l=1,2,\ldots,L$ are flow indices located within the low-flow segment (0.7–1.0 flow exceedance probabilities); L is the index of the minimum flow
Discharge statistics	(6) Mean discharge	$\sum_{t=1}^{N} Q_t / N$
	(7) Discharge variance	$\sqrt{\sum_{t=1}^{N} (Q_t - \overline{Q})^2 / (N-1)}$, where $\overline{Q}$ is the average of the entire flow data Q
	(8) Median discharge	$M(Q)$, where $M(Q)$is the median of the entire flow data Q
	(9) Peak discharge	$P(Q)$, where $P(Q)$is the peak of the entire flow data Q
	(10) Lag-1 autocorr. Coefficient	$\sum_{t=1}^{N-1} (Q_t - \overline{Q})(Q_{t+1} - \overline{Q}) / \sum_{t=1}^{N} (Q_t - \overline{Q})^2$
	(11) Mean log-transformed discharge	$\sum_{t=1}^{N} \ln(Q_t) / N$
	(12) Variance of log-transformed Q	$\sqrt{\sum_{t=1}^{N} (\ln(Q_t) - \overline{q})^2 / (N-1)}$, where $\overline{q}$ is the average of $\ln(Q)$
Monthly flow	(13) Maximum of monthly flow	$Max\{F(i)\}, i=1,2,\ldots,12$, where $F(i)$is the average flow in month i

Acknowledgments
The input data used in the modeling experiments are available and can be provided upon request from the first author. This research was supported with funding from Bryan Tolson's NSERC Discovery grant (50%), and also with funding from the Department of Civil and Environmental Engineering at University of Waterloo.

References

Aghakouchak, A., and E. Habib (2010), Application of a conceptual hydrologic model in teaching hydrologic processes, *Int. J. Eng. Educ.*, *26*(4), 963–973.
Ajami, N. K., Q. Duan, and S. Sorooshian (2007), An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction, *Water Resour. Res.*, *43*, W01403, doi:10.1029/2005WR004745.
Andreassian, V., F. Bourgin, L. Oudin, T. Mathevet, C. Perrin, J. Lerat, L. Coron, and L. Berthet (2014), Seeking genericity in the selection of parameter sets: Impact on hydrological model efficiency, *Water Resour. Res.*, *50*, 8356–8366, doi:10.1002/2013WR014761.
Arnold, J. G., and P. M. Allen (1999), Automated methods for estimating baseflow and ground water recharge from streamflow records, *J. Am. Water Resour. Assoc.*, *35*(2), 411–424.
Asadzadeh, M., and B. Tolson (2013), Pareto archived dynamically dimensioned search with hypervolume-based selection for multi-objective optimization, *Eng. Opt.*, *45*(12), 1489–1509.
Bahremand, A., F. De Smedt, J. Corluy, Y. Liu, J. Poorova, L. Velcicka, and E. Kunikova (2007), WetSpa model application for assessing reforestation impacts on floods in Margecany–Hornad Watershed, Slovakia, *Water Resour. Manage.*, *21*(8), 1373–1391.
Bates, D. M., and D. G. Watts (1988), *Nonlinear Regression Analysis and Its Applications*, John Wiley, N. Y.
Bergström, S. (1976), *Development and Application of a Conceptual Runoff Model for Scandinavian Catchments*, Lund Institute of Technology, University of Lund, Norrköping, Sweden.
Beven, K. (2006), A manifesto for the equifinality thesis, *J. Hydrol.*, *320*(1–2), 18–36.
Black, P. E. (1997), Watershed functions, *J. Am. Water Resour. Assoc.*, *33*(1), 1–11.
Blazkova, S., and K. Beven (2009), A limits of acceptability approach to model evaluation and uncertainty estimation in flood frequency estimation by continuous simulation: Skalka catchment, Czech Republic, *Water Resour. Res.*, *45*, W00B16, doi:10.1029/2007WR006726.
Box, G. E. P., and G. C. Tiao (1973), *Bayesian Inference in Statistical Analysis*, Addison-Wesley, Boston, Mass.
Boyle, D. P. (2000), *Multicriteria calibration of hydrological models, PhD dissertation*, Univ. of Ariz, Tucson.
Boyle, D. P., H. V. Gupta, and S. Sorooshian (2000), Toward improved calibration of hydrological models: Combining the strengths of manual and automatic methods, *Water Resour. Res.*, *36*(12), 3663–3674.
Carrillo, G., P. A. Troch, M. Sivapalan, T. Wagener, C. Harman, and K. Sawicz (2011), Catchment classification: Hydrological analysis of catchment behavior through process-based modeling along a climate gradient, *Hydrol. Earth Syst. Sci.*, *15*(11), 3411–3430.
Clark, M. P., D. Kavetski, and F. Fenicia (2011), Pursuing the method of multiple working hypotheses for hydrological modeling, *Water Resour. Res.*, *47*, W09301, doi:10.1029/2010WR009827.
Criss, R. E., and W. E. Winston (2008), Do Nash values have value? Discussion and alternate proposals, *Hydrol. Processes*, *22*, 2723–2725.
Del Moral, P., A. Doucet, and A. Jasra (2012), An adaptive sequential Monte Carlo method for approximate Bayesian computation, *Stat. Comput.*, *22*(5), 1009–1020.
Duan, Q., S. Sorooshian, and V. K. Gupta (1992), Effective and efficient global optimization for conceptual rainfall-runoff models, *Water Resour. Res.*, *28*(4), 1015–1031.
Duan, Q., V. K. Gupta, and S. Sorooshian (1993), A shuffled complex evolution approach for effective and efficient global minimization, *J. Opt. Theory Appl.*, *76*(3), 501–521.
Duan, Q., et al. (2006), Model parameter estimation experiment (MOPEX): An overview of science strategy and major results from the second and third workshops, *J. Hydrol.*, *320*(1–2), 3–17.
Eder, G., M. Sivapalan, and H. P. Nachtnebel (2003), Modelling water balances in an Alpine catchment through exploitation of emergent properties over changing time scales, *Hydrol. Processes*, *17*(11), 2125–2149.
Euser, T., H. C. Winsemius, M. Hrachowitz, F. Fenicia, S. Uhlenbrook, and H. H. G. Savenije (2013), A framework to assess the realism of model structures using hydrological signatures, *Hydrol. Earth Syst. Sci.*, *17*, 1893–1912.
Farina, M., and P. Amato (2004), A fuzzy definition of "optimality" for many-criteria optimization problems, *IEEE Trans. Syst. Man Cybern. Part A, Syst. Humans*, *34*(3), 315–326.
Fleming, P. J., R. C. Purshouse, and R. J. Lygoe (2005), Many-objective optimization: An engineering design perspective, in *Evolutionary Multi-Criterion Optimization*, edited by C. A. Coello Coello et al., LNCS 3410, pp. 14–32, Springer-Verlag, Berlin, Heidelberg, Germany.
Gharari, S., M. Shafiei, M. Hrachowitz, R. Kumar, F. Fenicia, H. V. Gupta, and H. H. G. Savenije (2014), A constraint-based search algorithm for parameter identification of environmental models, *Hydrol. Earth Syst. Sci.*, *18*(12), 4861–4870.
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), 751–764.
Gupta, H. V., K. J. Beven and T. Wagener (2005), Model calibration and uncertainty estimation, in *Encyclopedia of Hydrological Sciences*, edited by N. J. Hoboken, pp. 1–17, John Wiley, Hoboken, N. J., doi:10.1002/0470848944.hsa138.
Gupta, H. V., T. Wagener, and Y. Liu (2008), Reconciling theory with observations: Elements of a diagnostic approach to model evaluation, *Hydrol. Processes*, *22*(18), 3802–3813.
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.
Hadka, D., and P. Reed (2012), Diagnostic assessment of search controls and failure modes in many-objective evolutionary optimization, *Evol. Comput.*, *20*(3), 423–452.
Herbst, M., H. V. Gupta, and M. C. Casper (2009), Mapping model behaviour using self-organizing maps, *Hydrol. Earth Syst. Sci.*, *13*(3), 395–409.
Hingray, B., B. Schaefli, A. Mezghani, and Y. Hamdi (2010), Signature-based model calibration for hydrological prediction in mesoscale Alpine catchments, *Hydrol. Sci. J.*, *55*(6), 1002–1016.
Hrachowitz, M., O. Fovet, L. Ruiz, T. Euser, S. Gharari, R. Nijzink, J. Freer, H. H. G. Savenije, and C. Gascuel-Odoux (2014), Process consistency in models: The importance of system signatures, expert knowledge, and process complexity, *Water Resour. Res.*, *50*, 7445–7469, doi: 10.1002/2014WR015484.
Jothityangkoon, C., M. Sivapalan, and D. L. Farmer (2001), Process controls of water balance variability in a large semi-arid catchment: Downward approach to hydrological model development, *J. Hydrol.*, *254*(1–4), 174–198.
Kavetski, D., G. Kuczera and S. W. Franks (2006), Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory, *Water Resour. Res.*, *42*, W03407, doi:10.1029/2005WR004368.
Kollat, J. B., P. M. Reed, and T. Wagener (2012), When are multiobjective calibration trade-offs in hydrologic models meaningful?, *Water Resour. Res.*, *48*, W03520, doi:10.1029/2011WR011534.

Komuro, R., E. D. Ford, and J. H. Reynolds (2006), The use of multi-criteria assessment in developing a process model, *Ecol. Modell.*, *197*(3–4), 320–330.

Kuczera, G. (1983), Improved parameter inference in catchment models: 1. Evaluating parameter uncertainty, *Water Resour. Res.*, *19*(5), 1151–1162.

Ley, R., M. C. Casper, H. Hellebrand, and R. Merz (2011), Catchment classification by runoff behaviour with self-organizing maps (SOM), *Hydrol. Earth Syst. Sci.*, *15*(9), 2947–2962.

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

Liu, Y., J. Freer, K. Beven and P. Matgen (2009), Towards a limits of acceptability approach to the calibration of hydrological models: Extending observation error, *J. Hydrol.*, *367*, 93–103.

Liu, Y. B., S. Gebremeskel, F. De Smedt, L. Hoffmann, and L. Pfister (2003), A diffusive transport approach for flow routing in GIS-based flood modeling, *J. Hydrol.*, *283*(1–4), 91–106.

Marjoram, P., J. Molitor, V. Plagnol, and S. Tavare (2003), Markov chain Monte Carlo without likelihoods, *Proc. Natl. Acad. Sci. U. S. A.*, *100*(26), 15,324–15,328.

Martinez, G. F., and H. V. Gupta (2010), Toward improved identification of hydrological models: A diagnostic evaluation of the ''abcd'' monthly water balance model for the conterminous United States, *Water Resour. Res.*, *46*, W08507, doi:10.1029/2009WR008294.

Martinez, G. F., and H. V. Gupta (2011), Hydrologic consistency as a basis for assessing complexity of monthly water balance models for the continental United States, *Water Resour. Res.*, *47*, W12540, doi:10.1029/2011WR011229.

McMillan, H., T. Krueger, and J. Freer (2012), Benchmarking observational uncertainties for hydrology: Rainfall, river discharge and water quality, *Hydrol. Processes*, *26*(26), 4078–4111.

McMillan, H. K., M. P. Clark, W. B. Bowden, M. Duncan, and R. A. Woods (2011), Hydrological field data from a modeller's perspective: Part 1. Diagnostic tests for model structure, *Hydrol. Processes*, *25*, 511–522.

Montanari, A., and E. Toth (2007), Calibration of hydrological models in the spectral domain: An opportunity for scarcely gauged basins?, *Water Resour. Res.*, *43*, W05434, doi:10.1029/2006WR005184.

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

Neitsch, S. L., J. G. Arnold, J. R. Kiniry, and J. R. Williams (2001), Soil and water assessment tool user's manual version 2000, report 02-02, U.S. Dep. of Agric., Agric. Res. Serv., Temple, Tex.

Oudin, L., A. Kay, V. Andreassian, and C. Perrin (2010), Are seemingly physically similar catchments truly hydrologically similar?, *Water Resour. Res.*, *46*, W11558, doi:10.1029/2009WR008887.

Pechlivanidis, I. G., B. Jackson, H. McMillan, and H. Gupta (2014), Use of an entropy-based metric in multiobjective calibration to improve model performance, *Water Resour. Res.*, *50*, 8066–8083, doi:10.1002/2013WR014537.

Pfannerstill, M., B. Guse, and N. Fohrer (2014), Smart low flow signature metrics for an improved overall performance evaluation of hydrological models, *J. Hydrol.*, *510*, 447–458.

Pokhrel, P., K. K. Yilmaz, and H. V. Gupta (2012), Multiple-criteria calibration of a distributed watershed model using spatial regularization and response signatures, *J. Hydrol.*, *418*, 49–60.

Purshouse, R. C., and P. J. Fleming (2007), On the evolutionary optimization of many conflicting objectives, *IEEE Trans. Evol. Comput.*, *11*(6), 770–784.

Safari, A., F. De Smedt, and F. Moreda (2012), WetSpa model application in the Distributed Model Intercomparison Project (DMIP2), *J. Hydrol.*, *418–419*, 78–89.

Sawicz, K., T. Wagener, M. Sivapalan, P. A. Troch, and G. Carrillo (2011), Catchment classification: Empirical analysis of hydrologic similarity based on catchment function in the eastern USA, *Hydrol. Earth Syst. Sci.*, *15*(9), 2895–2911.

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

Seber, G. A., and C. J. Wild (1989), *Nonlinear Regression*, John Wiley, N. Y.

Seibert, J., and J. J. McDonnell (2002), On the dialog between experimentalist and modeler in catchment hydrology: Use of soft data for multicriteria model calibration, *Water Resour. Res.*, *38*(11), 1241–1254.

Shafii, M., and F. D. Smedt (2009), Multi-objective calibration of a distributed hydrological model (WetSpa) using a genetic algorithm, *Hydrol. Earth Syst. Sci.*, *13*, 2137–2149.

Shamir, E., B. Imam, H. V. Gupta, and S. Sorooshian (2005), Application of temporal streamflow descriptors in hydrologic model parameter estimation, *Water Resour. Res.*, *41*, W06021, doi:10.1029/2004WR003409.

Sisson, S. A., Y. Fan, and M. M. Tanaka (2007), Sequential Monte Carlo without likelihoods, *Proc. Natl. Acad. Sci. U. S. A.*, *104*(6), 1760–1765.

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

Tolson, B. A. (2005), *Automatic calibration, management and uncertainty analysis: Phosphorus transport in the Cannonsville watershed, PhD thesis*, Cornell Univ., Ithaca, N. Y.

Tolson, B. A., and C. A. Shoemaker (2007), Dynamically dimensioned search algorithm for computationally efficient watershed model calibration, *Water Resour. Res.*, *43*, W01413, doi:10.1029/2005WR004723.

van Werkhoven, K., T. Wagener, P. Reed, and Y. Tang (2009), Sensitivity-guided reduction of parametric dimensionality for multi-objective calibration of watershed models, *Adv. Water Resour.*, *32*(8), 1154–1169.

Viglione, A., J. Parajka, M. Rogger, J. L. Salinas, G. Laaha, M. Sivapalan, and G. Bloschl (2013), Comparative assessment of predictions in ungauged basins—Part 3: Runoff signatures in Austria, *Hydrol. Earth Syst. Sci.*, *17*(6), 2263–2279.

Vrugt, J., and B. Robinson (2007), Improved evolutionary optimization from genetically adaptive multimethod search, *Proc. Natl. Acad. Sci. U. S. A.*, *104*(3), 708–711.

Vrugt, J. A., and M. Sadegh (2013), Toward diagnostic model calibration and evaluation: Approximate Bayesian computation, *Water Resour. Res.*, *49*, 4335–4345, doi:10.1002/wrcr.20354.

Vrugt, J. A., H. V. Gupta, W. Bouten, and S. Sorooshian (2003a), A shuffled complex evolution metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters, *Water Resour. Res.*, *39*(8), 1201, doi:10.1029/2002WR001642.

Vrugt, J. A., H. V. Gupta, L. A. Bastidas, W. Bouten, and S. Sorooshian (2003b), Effective and efficient algorithm for multiobjective optimization of hydrologic models, *Water Resour. Res.*, *39*(8), 1214, doi:10.1029/2002WR001746.

Vrugt, J. A., B. A. Robinson, and J. M. Hyman (2009a), Self-adaptive multimethod search for global optimization in real-parameter spaces, *IEEE Trans. Evol. Comput.*, *13*(2), 243–259.

Vrugt, J. A., C. J. F. ter Braak, C. G. H. Diks, D. Higdon, B. A. Robinson, and J. A. Hyman (2009b), Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling, *Int. J. Nonlinear Sci. Numer. Simul.*, *10*(3), 273–290.

Wagener, T., and H. V. Gupta (2005), Model identification for hydrological forecasting under uncertainty, *Stochastic Environ. Res. Risk Assess.*, *19*, 378–387.
Wagener, T., and A. Montanari (2011), Convergence of approaches toward reducing uncertainty in predictions in ungauged basins, *Water Resour. Res.*, *47*, W06301, doi:10.1029/2010WR009469.
Wagener, T., M. Sivapalan, P. Troch, and R. Woods (2007), Catchment classification and hydrologic similarity, *Geogr. Compass*, *1*(4), 901–931.
Westerberg, I. K., J.-L. Guerrero, P. M. Younger, K. J. Beven, J. Seibert, S. Halldin, E. Freer and C.-Y. Xu (2011), Calibration of hydrological models using flow-duration curves, *Hydrol. Earth Syst. Sci.*, *15*, 2205–2227.
Winsemius, H. C., B. Schaefli, A. Montanari, and H. H. G. Savenije (2009), On the calibration of hydrological models in ungauged basins: A framework for integrating hard and soft hydrological information, *Water Resour. Res.*, *45*, W12422, doi:10.1029/2009WR007706.
Yadav, M., T. Wagener, and H. Gupta (2007), Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins, *Adv. Water Resour.*, *30*(8), 1756–1774.
Yapo, P. O., H. V. Gupta, and S. Sorooshian (1998), Multi-objective global optimization for hydrologic models, *J. Hydrol.*, *204*(1–4), 83–97.
Yilmaz, K. K., H. V. Gupta, and T. Wagener (2008), A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, *Water Resour. Res.*, *44*, W09417, doi:10.1029/2007WR006716.