Developments in Informal Multi-Criteria Calibration and Uncertainty Estimation in Hydrological Modelling

by

Mahyar Shafii Hassanabadi

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Civil Engineering

Waterloo, Ontario, Canada, 2014

AUTHOR'S DECLARATION

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Hydrologic modelling has benefited from significant developments over the past two decades, which has led to the development of distributed hydrologic models. Parameter adjustment, or model calibration, is extremely important in the application of these hydrologic models. Multi-criteria calibration schemes and several formal and informal predictive uncertainty estimation methodologies are among the approaches to improve the results of model calibration. Moreover, literature indicates a general agreement between formal and informal approaches with respect to the predictive uncertainty estimation in single-criterion calibration cases. This research extends the comparison between these techniques to multi-criteria calibration cases, and furthermore, proposes new ideas to improve informal multi-criteria calibration and uncertainty estimation in hydrological modelling. GLUE is selected as a candidate informal methodology due to its extreme popularity among hydrological modellers, *i.e.,* based on the number of applications in the past two decades. However, it is hypothesized that improvements can be applied to other certain types of informal uncertainty estimation as well.

The first contribution of this research is an in-depth comparison between GLUE and Bayesian inference in the multi-criteria context. Such a comparison is novel because past literature has focused on comparisons for only single criterion calibration studies. Unlike the previous research, the results show that there can be considerable differences in hydrograph prediction intervals generated by traditional GLUE and Bayesian inference in multi-criteria cases. Bayesian inference performs more satisfactorily than GLUE along most of the comparative measures. However, results also reveal that a standard Bayesian formulation (*i.e.,* aggregating all uncertainties into a single additive error term) may not demonstrate perfect reliability in the prediction mode. Furthermore, in cases with a limited computational budget, non-converged MCMC sampling proves to be an appropriate alternative to GLUE since it is reasonably consistent with a fully-converged Bayesian approach, even though the fully-converged MCMC requires a substantially larger number of model evaluations.

Another contribution of this research is to improve the uncertainty bounds of the traditional GLUE approach by the exploration of alternative behavioural solution identification strategies. Multiple behavioural solution identification strategies from the literature are evaluated, new objective strategies are developed, and multi-criteria decision-making concepts are utilized to select the best strategy. The results indicate that the subjectivity involved in behavioural solution identification strategies impacts the uncertainty of model outcome. More importantly, a robust implementation of GLUE proves to require

comparing multiple behavioural solution identification strategies and choosing the best one based on the modeller's priorities. Moreover, it appears that the proposed objective strategies are among the best options in most of the case studies investigated in this research. Thus, it is recommended that these new strategies be considered among the set of behavioural solution identification strategies in future GLUE applications.

Lastly, this research also develops a full optimization-based calibration framework that is capable of utilizing both standard goodness-of-fit measures and many hydrological signatures simultaneously. These signatures can improve the calibration results by constraining the model outcome hydrologically. However, the literature shows that to simultaneously apply a large number of hydrological signatures in model calibration is challenging. Therefore, the proposed research adopts optimization concepts to accommodate many criteria (including 13 hydrologic signature-based objectives and two standard statistical goodness-of-fit measures). In the proposed framework, hydrological consistency is quantified (based on a set of signature-based measures and their desired level of acceptability) and utilized as a criterion in multiple calibration formulations. The results show that these formulations perform better than the traditional approaches to locate hydrologically consistent parameter sets in the search space.

Different hydrologic models, most of which are conceptual rainfall-runoff models, are used throughout the thesis to evaluate the performance of the developed strategies. However, the developments explored in this research are typically simulation-model-independent and can be applied to calibration and uncertainty estimation of any environmental model. However, further testing of these methods is warranted for more computationally intensive simulation models, such as fully distributed hydrologic models.

Acknowledgements

I would like to express my special gratitude to Dr. Bryan Tolson for his continuous guidance and support throughout my PhD studies. Bryan is a caring, intelligent, and responsible advisor, and I have learnt from him a lot ever since I got to know him. I would also like to thank my co-advisor, Dr. Loren S. Matott from the State University of New York at Buffalo for his constructive comments, which have helped me so much step forward in my research. I would also like to thank Dr. Donald Burn as another co-advisor during the first stage of my PhD for his helpful remarks on my comprehensive document.

It is my pleasure to thank my defense committee members Dr. Mazdak Arabi from the Colorado State University, Dr. Keith Hipel, Dr. John Sykes, and Dr. Scott Walbridge from the University of Waterloo.

I am also very thankful to my friends and colleagues, Masoud, Amin, and Saman for offering me their support as real friends.

Bahareh, my lovely caring wife, showed endless love and patience during my PhD studies, for which I am forever grateful. I would also like to express my sincere gratefulness to my Parents Ahmad and Roghayeh and my brother Mehran for providing me with everything they could so that I can continue my studies.

Dedication

I dedicate my dissertation work to my lovely wife, Bahareh, my parents, Ahmad and Roghayeh, and my brother, Mehran, whose words of encouragement ring in my ears. They have never left my side, and have been supporting me throughout the entire doctorate program.

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

1.1 Model Calibration and Uncertainty Estimation Overview

Hydrologic modelling has benefited from significant developments over the past two decades, including dramatic growth in computational power, ever increasing availability of distributed hydrologic observations, and improved understanding of the physics and dynamics of the hydrologic system. These factors have led to increasing complexity in hydrologic models and an advance from lumped conceptual models toward semi-distributed and distributed physics-based models. Hydrologic models include many parameters to be estimated via calibration, which is a parameter adjustment procedure using certain performance metrics and the historical observation data available. If the coherence of the underlying physical processes is not accounted for in calibration, the estimated parameters may not fall within acceptable physically meaningful ranges (*e.g.,* Burnash, 1995). Moreover, automatic calibration conducted without sufficient hydrological expertise may also yield improper parameter values, which can result in unreasonable regimes of model responses not controlled by measurements (Refsgaard, 1997; Wagener *et al.*, 2001). Furthermore, even well calibrated parameter values may result in poor performance with respect to an independent validation data set. As a result, improving the reliability of model calibration has been an attractive topic in the hydrological modelling literature.

Aggregation of all errors into a single criterion during calibration has been shown to cause issues in the calibration of hydrologic models because one criterion does not provide sufficient detail about model inadequacy (Gupta *et al.*, 1998), *e.g.,* single-criterion metrics do not distinguish between high-flow and low-flow model behaviour. This realization has motivated multi-criteria calibration approaches in which multiple sets of observations and/or multiple evaluation criteria are employed (Gupta *et al.*, 1998; Legates and McCabe, 1999; Madsen, 2000; Yapo *et al.*, 1998). Multi-criteria calibration uses more than one index to describe the characteristics of the error vector, *e.g.,* separate Nash-Sutcliffe values for high-flow and low-flow data, resulting in a criteria tradeoff curve and corresponding set of 'Pareto' optimal parameter values. Moreover, given the ever-increasing availability of multiple sets of data in catchment modelling, it seems necessary that modellers need to implement multi-criteria calibration concepts at some points in hydrological modelling experiments. Hence, this thesis focuses on model calibration in the multi-criteria context.

Uncertainty estimation is another crucially important factor in a reliable modelling effort in hydrology. In fact, for predictive hydrologic models to be useful and defensible, proper consideration of all uncertainty sources, *e.g.,* parameter estimation, meteorological input, upscaling, and other non-calibrated

forcing functions, is essential (Wagener and Gupta, 2005). As a result, research in the past decades has emphasized comprehensive statistical model assessment, often called model evaluation, procedures whereby parameter estimation is combined with formal predictive uncertainty estimation procedures, either using frequentist approaches (*e.g.,* Bates and Watts, 1988; Refsgaard *et al.*, 2005; Reichert, 1997; Seber and Wild, 1989), or Bayesian inference (*e.g.,* Bates and Campbell, 2001; Kavetski *et al.*, 2002; Kuczera, 1983; Thiemann *et al.*, 2001; Vrugt *et al.*, 2003b). These methods seek to elucidate appropriate probability density functions (PDF) for parameters and model outcome, which are then utilized for predictive uncertainty estimation.

Between the two classes of formal calibration and uncertainty estimation techniques, Bayesian inference has been increasingly applied in the hydrological modelling literature in the past decades, especially given recent advancements in computational power. At the heart of Bayesian inference approaches is the use of formal likelihood functions to analyse the predictive uncertainty. These likelihood functions must make explicit assumptions about the form of the model residuals, *i.e.,* deviations between simulations and observations (Box and Tiao, 1973; Gelman *et al.*, 2004; Kuczera, 1983; Thyer *et al.*, 2009). Because the appropriate statistical form for a given set of model residuals is not always clear in hydrological modelling, the Bayesian approaches are criticized for relying on potentially inappropriate likelihood functions (*e.g.,* Beven *et al.*, 2008). To address this issue, certain researchers emphasize the development of more appropriate likelihood functions (*e.g.,* Bates and Campbell, 2001; Schoups and Vrugt, 2010; Sorooshian and Dracup, 1980), and/or using hierarchical Bayesian structures that disaggregate different sources of uncertainties (*e.g.,* Huard, 2008; Kuczera *et al.*, 2006; Moradkhani *et al.*, 2005; Renard *et al.*, 2010; 2011; Wei *et al.*, 2010).

However, development and application of such formulations concerning complex non-linear hydrologic models is non-trivial and may be computationally intractable in certain case studies (Beven and Binley, 2013; Honti *et al.*, 2013; McMillan and Clark, 2009), especially in cases using existing state-of-the-art Markov Chain Monte Carlo (MCMC) samplers, which are a commonly used sampling-based strategy to conduct Bayesian inference in the non-analytical context. Furthermore, a proper application of these more recent uncertainty estimation approaches requires considerable understanding of mathematics, statistics, and details of implementations on a digital computer, especially in the absence of general-purpose software that enables the use of these state-of-the-art methods in a user-friendly environment (Vrugt *et al.*, 2008a). As a result, certain non-hierarchical Bayesian inference approaches involving standard likelihood functions based on sum of squared residuals have been applied in various research studies (*e.g.,* Dotto *et al.*, 2012; Laloy *et al.*, 2010; Li *et al.*, 2010b; Martin and Ayesa, 2010; Starrfelt and Kaste, 2014). Regardless of the type and complexity of likelihood functions applied in Bayesian

inference, benchmark studies point out that multiple post-diagnostic checks must be conducted after the parameter estimation stage to ensure that the results are statistically consistent (Kuczera, 1983; Thyer *et al.*, 2009). However, the hydrological literature (the previously cited papers that use standard Bayesian formulation) shows that insufficient attention is paid to these diagnostics, *i.e.,* no reports of diagnostic checks, or when these checks are reported, no efforts for changing the formulations towards appropriately meeting the assumptions (*e.g.,* Feyen *et al.*, 2007; Hutton *et al.*, 2014).

Given the associated issues with Bayesian inference, and despite robust theoretical underpinnings of these approaches, a variety of informal calibration and uncertainty estimation methods (categorized as uncertainty-based calibration techniques in this thesis) have been developed in the hydrological literature, among which simple importance sampling Monte Carlo Simulations (MCS) such as Generalized Likelihood Uncertainty Estimation (GLUE) (Beven and Binley, 1992) has been the most commonly used informal methodology in the hydrological literature. Informal methods are conceptually simpler than Bayesian approaches and are easier to implement and require less modifications to existing source codes of hydrologic models. Therefore, from the ease-of-use perspective, they have an important advantage over more sophisticated Bayesian approaches. As a result, the literature shows that many researchers and practitioners still prefer these techniques for model calibration and uncertainty estimation (Abebe *et al.*, 2010; Breinholt *et al.*, 2013; Liu *et al.*, 2012; Mannina, 2011; Setegn *et al.*, 2010; Tolson and Shoemaker, 2008; van Griensven and Meixner, 2007; Zhang *et al.*, 2012).

A general drawback of informal methods is that, because they do not accommodate a formal specification of error models, they might not be reliable for uncertainty estimation (Kavetski *et al.*, 2002). Moreover, some have argued that the combination of using pseudo-likelihood functions and the inherent subjectivity involved in applying GLUE yields incoherent results (*e.g.,* Mantovan and Todini, 2006; Stedinger *et al.*, 2008), which eventually makes the predictive uncertainty difficult to analyse. The debate on suitability of GLUE for uncertainty estimation in hydrological modelling continues in the current literature (Beven *et al.*, 2012a; Clark *et al.*, 2011; Clark *et al.*, 2012b). Other researchers have compared informal techniques with formal approaches in the hydrological modelling literature, and indicated relatively close agreement among alternative methods in terms of predictive capability (Beven *et al.*, 2008; Hutton *et al.*; Jeremiah *et al.*, 2011; Jin *et al.*, 2010; Li *et al.*, 2010a; Qian *et al.*, 2003; Vrugt *et al.*, 2008a; Yang *et al.*, 2008). The potential common ground between informal and formal uncertainty estimation approaches, and the fact that an alternative to a full Bayesian model evaluation may be needed in certain situations, pushes the focus of this thesis towards developments in informal calibration and uncertainty estimation. As mentioned earlier, this study focuses on the multi-criteria context, and thus, calibration and uncertainty estimation developments are explored in multi-criteria/multi-response cases.

The developments in this study are implemented in GLUE as a candidate informal methodology, and the reason why GLUE is selected is its popularity among hydrological modellers, *i.e.,* based on the number of applications in the past decades until present (Beven and Binley, 2013; Breinholt *et al.*, 2013; Liu *et al.*, 2012; Mannina, 2011). GLUE is based on the concept of 'equifinality' and classifies sampled parameter sets as having produced model output that is either 'behavioural' or 'non-behavioural'. This research hypothesizes that the developments can be applied to other informal methods that work on the basis of filtering behavioural solutions to characterize and propagate uncertainty such as behavioural identification techniques using optimization algorithms (*e.g.,* Tolson and Shoemaker, 2007; van Griensven and Meixner, 2007).

1.2 Research Outline

The developments explored in the present thesis include (i) in-depth comparison between informal (GLUE) and formal (Bayesian inference) multi-criteria calibration and uncertainty estimation methodologies, and (ii) the implementation of innovative ideas in GLUE to improve the reliability of uncertainty estimation and reduce the influence of GLUE subjectivity. These general research objectives are elaborated on as follows.

The comparison between GLUE and Bayesian inference aims to explore the common ground between these approaches with respect to the characterization of the predictive uncertainty in hydrological modelling. Previous comparative studies have considered only single-criterion cases. This thesis compares GLUE and Bayesian inference in the multi-criteria context against multiple comparative measures to analyse different aspects, *i.e.,* computational burden, complexity, and the predictive capacity. Note that informal methods can be easily extended to multi-criteria calibration problems (*e.g.,* Blasone *et al.*, 2008a; Blazkova and Beven, 2009; van Griensven and Meixner, 2007).

To improve the reliability of uncertainty bounds generated by GLUE, this thesis attempts to reduce the subjectivity inherent in GLUE. Previous studies have reported on the impact of varying the likelihood function definition on the model output uncertainty (Beven and Freer, 2001; Freer *et al.*, 1996) and on the impact of the behavioural threshold on prediction intervals (Montanari, 2005). This thesis evaluates the impact of varying the behavioural solution identification strategy (with a constant likelihood function definition) on the uncertainty bounds generated by GLUE, and furthermore, develops new objective identification strategies.

In the quest for improving the informal calibration and uncertainty estimation, this thesis also develops a framework whereby hydrological signatures are implemented in calibration and uncertainty estimation of hydrologic models. These signatures reflect the functional behaviour of the catchment, and have been

used in diagnostic-based models evaluation (*e.g.,* Carrillo *et al.*, 2011; Clark *et al.*, 2011; Eder *et al.*, 2003; Jothityangkoon *et al.*, 2001; McMillan *et al.*, 2012; Vrugt and Sadegh, 2013; Wagener and Montanari, 2011; Winsemius *et al.*, 2009; Yilmaz *et al.*, 2008). Hydrological signatures constrain model predictions with respect to hydrological consistency (Euser *et al.*, 2013; Martinez and Gupta, 2011), and are expected to improve the predictive capacity (Fenicia *et al.*, 2007; Kirchner, 2006). Robust diagnostic-based model evaluation requires multiple signatures to fully characterize the system's behaviour, which has not been previously implemented in a full multi-criteria context due to existing limitations in multi-criteria formulations with numerous criteria (Pokhrel *et al.*, 2012). This thesis utilizes optimization concepts to develop a novel multi-criterion informal uncertainty-based calibration framework that can accommodate the use of numerous signatures.

1.3 Thesis Structure

After the introduction and problems statement in Chapter 1 (this chapter), Chapter 2 provides an up-to-date review on the uncertainty estimation background in hydrological modelling, and elaborates on different formal and informal uncertainty estimation techniques applied in the hydrological literature. Chapter 3 presents the study of comparison between GLUE and the state-of-the-art MCMC-based Bayesian inference in multi-criteria context. Chapter 3 is previously published in the journal of Stochastic Environmental Research and Risk Assessment (Shafii *et al.*, 2014b). Chapter 4 provides the study on the reduction of subjectivity involved in GLUE by developing objective behavioural solution identification strategies. Chapter 4 is based on a manuscript that is currently in review (accepted with moderate revisions) in the *Journal of Hydrology*. Chapter 5 provides the details of the model calibration and uncertainty estimation framework utilizing signature-based measures. This chapter is prepared in the format of a manuscript to be submitted to Water Resources Research journal. Finally, extensive conclusions and recommendations for future work are detailed in Chapter 6.

1.4 Research Contributions

The specific contributions of the proposed research made to hydrological modelling uncertainty estimation are summarized as follows:

(1) An in-depth comparison between informal (GLUE) and formal (Bayesian inference) uncertainty estimation techniques in the multi-criteria context was performed. Literature indicates that such a comparison has not been previously done in the multi-criteria context. Results show that there can be considerable differences in hydrograph prediction intervals generated by GLUE and Bayesian inference in multi-criteria cases. The Bayesian inference methodology performs more satisfactorily than GLUE along most of the comparative measures. However, the results reveal that Bayesian

inference yields under-estimation of predictive uncertainty for low-flows in the future validation period. This finding indicates that a standard Bayesian formulation (*i.e.,* aggregating all uncertainties into a single additive error term) may not demonstrate perfect reliability in the prediction mode, which suggests using more complex hierarchical Bayesian structures. Moreover, results suggest that non-converged MCMC sampling can be considered as an alternate informal methodology for calibration and uncertainty estimation of rainfall-runoff models. Its performance is reasonably consistent with a fully-converged Bayesian approach, even though fully-converged MCMC requires a substantially larger number of model evaluations.

(2) Multiple behavioural solution identification strategies in GLUE were evaluated, motivating new ideas for objectively identifying behavioural solutions utilizing multi-criteria decision-making concepts. Results show that no single behavioural solution identification strategy yields the best performance in all case studies, and furthermore, a robust implementation of GLUE requires employing multiple strategies, comparing them during the calibration period against the comparative measures of interest, and selecting the best one for generating predictions in future. The objective strategies proposed in this research were among the best strategies selected in the majority of case studies.

(3) A full multi-criteria model calibration and uncertainty estimation framework was developed that can simultaneously accommodate both hydrological signatures and standard performance measures. According to the literature, to simultaneously consider a large number of hydrological signatures and multiple performance metrics in the process of model calibration is novel. A consistency measure is defined based on a set of signature-based measures and their desired level of acceptability. This measure is then used in a new consistency-based formulation for multi-criteria calibration and uncertainty estimation of hydrologic models. Compared to the traditional approach, the proposed formulation proves to locate more hydrologically consistent parameter sets in the search space. Moreover, the proposed approach properly facilitates the characterization of the model outcome uncertainty. Results also demonstrate that the findings are insensitive to the type of optimization algorithm.

Chapter 2
Uncertainty Estimation in Conceptual Rainfall-Runoff Modelling

Summary

Various hydrologic rainfall-runoff models have been used to represent the watershed physical processes which control the conversion of precipitation into streamflow and water storage changes. The performance of such hydrological models can be affected by many uncertainties, which makes uncertainty estimation one of the most popular topics in hydrology over the past two decades. This chapter provides a general overview of uncertainties in hydrological modelling and clarifies the terminology that is used in this thesis to describe the uncertainty bounds (Section 2.1). Moreover, multiple probabilistic (formal) and informal uncertainty estimation techniques applied in the field of hydrology are briefly described in Sections 2.2-2.4; these methods are categorized based on their applications to single- and multiple-criteria cases. Among all the methods introduced in this chapter, the techniques and formulations that are applied in the rest of this thesis are described in the most detail. Section 2.5 briefly point out advantages and disadvantages of different uncertainty estimation methods.

2.1 Background in Uncertainty Estimation in Hydrological Modelling

The predictions of hydrologic models are affected by uncertainties in the modelling process. Hydrologists deal with different sources of uncertainty, described as follows (*e.g.,* Gupta *et al.*, 2005b; Huard, 2008; Matott *et al.*, 2009; Refsgaard *et al.*, 2007):

- Parameter Estimation Uncertainty – Parameter values must be viewed probabilistically because alternative sets of parameter values can yield equally (or nearly equally) good fits to historical data.
- Data Uncertainty – data sets used such as forcing functions, initial conditions, and boundary conditions of a model, as well as response data used during calibration (*e.g.,* flow and water quality data), are uncertain (*i.e.,* measured imperfectly).
- Model Structure – models are inherent simplifications of reality and so-called structural uncertainty is related to the natural randomness in processes and spatio-temporal discretization chosen for a given modelling effort.

Because these sources of uncertainties in the inputs are the direct cause of model output uncertainty, it is always desirable to develop systematic, flexible, and comprehensive methodologies to estimate and quantify the modelling uncertainties. Considering a general hydrological modelling experiment formulated as,

$$\hat{Y}(t) = h(\boldsymbol{\theta}, \mathbf{X}, t) \qquad (1)$$

where, $Y(t)$ is the simulated outcome (*e.g.,* discharge) at time step t, and $h(\,)$ is the model output as a function of model parameters $\boldsymbol{\theta}$ and input data/boundary conditions $\mathbf{X}$, the goal of an uncertainty analysis is to determine the uncertainty in the elements of $\hat{Y}$ that results from uncertainty in the elements of $\mathbf{X}$ (Helton and Davis, 2003). When viewed in its most general form, uncertainty analysis simply involves the determination of the distribution for $\hat{Y}$ that results from the function $h(\,)$ and the distributions that characterize the uncertainty in the elements of $\mathbf{X}$. As a result, a systematic framework for uncertainty analysis consists of the following key elements:

- input uncertainty characterization—which involves fitting and/or assigning marginal and joint distributions to uncertain model inputs,
- uncertainty propagation—which involves translating the uncertainty in model inputs into the corresponding uncertainty in model outputs.

Note that the term 'uncertainty analysis' is sometimes referred to as an uncertainty importance assessment, which involves determining the key drivers of uncertainty of $\hat{Y}$ (Morgan and Small, 1992). Because of this delicate difference between the two interpretation of uncertainty analysis, this thesis uses the term 'uncertainty estimation' to refer to the combination of input uncertainty characterization and uncertainty propagation steps.

The early uncertainty estimation approaches, *e.g.,* Monte Carlo Simulations (MCS), assume that the probability distribution functions (PDF) of $\mathbf{X}$ are known, and they only deal with uncertainty propagation. Whereas, probabilistic calibration methods, *i.e.,* frequentist and Bayesian inference, are designed to simultaneously perform the determination of input PDFs and uncertainty propagation in the so-called 'calibration' period. In these methods, uncertainty propagation in model forecasts of the future (or the validation period, after the model is calibrated) simply involves sampling the PDFs of all input variables considered uncertain. The other main difference between MCS and probabilistic calibration approaches is that the former does not utilize the calibration data whereas the latter conditions the input uncertainty description to the system response, *e.g.,* observed discharge.

The probabilistic calibration methodologies work on the basis of analysing the statistical properties of the hydrologic model error series that occurred in reproducing observed historical river flow data (Montanari and Brath, 2004). The underlying mathematical formulation in probabilistic hydrological modelling is given in Equation (2), in which model parameters, model inputs, and modelling errors are

treated in a probabilistic context (Beven and Binley, 1992; Christensen and Cooley, 1999; Kavetski *et al.*, 2002; Troutman, 1983; Troutman, 1985; van Griensven and Meixner, 2007; Vrugt *et al.*, 2003b):

$$Y(t) = \hat{Y}(t) + \varepsilon_Y(t) \qquad (2)$$

where, $Y(t)$ is the observation and $\hat{Y}(t)$ is the simulation at time step t, and $\varepsilon_Y(t)$ is a normally distributed random variable (with mean zero and variance σ_ε^2) representing the total error at time step t. Equation (2) shows that the statistical methods combine the deterministic hydrologic model with the stochastic error term to generate the prediction bounds or predictive uncertainty. The total error aggregates all sources of uncertainty, which is shown that it can be projected into variation of $h(\,)$ with respect to different parameter values (Troutman, 1985). Such variation is often called the parameter uncertainty and is accounted for by seeking the optimal values of $\boldsymbol{\theta}$ (such that the deviations between simulations and observations becomes as low as possible) as well as a range, a distribution and/or a PDF of $\boldsymbol{\theta}$ and σ_ε^2. These PDFs are subsequently used to propagate the uncertainty through the model and derive the predictive uncertainty.

Various informal methods have also been developed where uncertainty estimation is performed using pseudo-likelihood functions, without relying on the estimation of the parameters of the error series. The uncertainty quantified by such methods is, in fact, the uncertainty of model output, rather than the parameter uncertainty or the predictive uncertainty defined in the formal literature (*e.g.,* Krzysztofowicz, 1999; Todini, 2008). Because of this discrepancy, Todini (2008) introduced the term "emulation uncertainty" to the informal type of uncertainty bounds. In this thesis, whenever informal uncertainty estimation methods are applied, the term ‘model output uncertainty’ is used to refer to the uncertainty that is propagated to the model output without probabilistically describing the error series.

The formal and informal methodologies described previously couple hydrologic models with an algorithm that repeatedly samples model parameter configurations to develop a sample of calibrated model parameters. Such approaches are categorized as uncertainty-based calibration methods in Razavi *et al.* (2010). Conceptually, uncertainty-based calibration is equivalent to uncertainty estimation (with its general definition described above). Hence, this thesis uses these two terms interchangeably. Note that uncertainty-based calibration may be employed in cases with single set of data, or multiple sets of observations, of which the latter is the main focus of this thesis.

In the rest of this chapter, different uncertainty estimation methods are described. Section 2.2 elaborates on the traditional methods for uncertainty estimation (analytic and Monte Carlo approaches),

while Sections 2.3 and 2.4 described calibration and uncertainty estimation methods in single- and multi-criteria context. The hydrological literature in uncertainty estimation is very extensive, and only a relevant subset of the entire literature is reviewed in this chapter.

2.2 Traditional Methods for Uncertainty Estimation

2.2.1 Monte Carlo Simulation (MCS)

Random sampling-based techniques are often employed in uncertainty propagation. The best known techniques in this class of methods is MCS (*e.g.,* Spear and Hornberger, 1980; van Straten and Keesman, 1991), which is a robust and generally applicable method to sample from, *i.e.,* approximate, the probability distribution of a model output (or prediction). MCS relies on the simulation of the model under a large number of parameter sets randomly drawn from prior known model parameter distributions. To estimate the model output uncertainty, the sampled parameter sets are used in the model, their corresponding model outputs are generated, and percentiles of the model simulations are derived, *e.g.,* 5^{th} and 95^{th} percentiles. If the errors associated with the modelling process are also included, the results will describe total or predictive uncertainty. Even though these methods are applicable to a wide range of linear and non-linear systems, their application is generally computationally expensive. As a result, other efficient analytical methods have been developed for uncertainty estimation, which are briefly described in the next section.

2.2.2 Analytical Methods

The analytical approximation approaches based on First-Order Second-Moment (FOSM) reliability methods (Wood, 1976) analyse model output reliability with respect to existing uncertainties. FOSM methods make some assumptions regarding the nature and distribution of the function, *e.g.,* hydrologic models, and derive the PDF of model outcome to reflect the uncertainty in underlying parameters. The basic assumption is that the imprecision of a model prediction can be described by an average value representing the central or expected tendency of prediction, and the variance describing the prediction variation around the mean. Garen and Burges (1981) employ a Mean-Value FOSM (MVFOSM) approach to derive model output bounds due to existing uncertainties. The MVFOSM takes a first-order Taylor series expansion of model performance at the mean values of the parameters. Melching *et al.* (1990) also applied FOSM and MVFOSM methods to evaluate the combined effect of uncertainties on the output hydrographs of a rainfall-runoff.

FOSM and MVFOSM are applicable for cases where the assumption of response's linearity is not greatly violated (Kuczera, 1988). These approximation techniques do not perform reliably for non-linear

systems such as hydrologic models, especially for the cases where the response of interest, *e.g.,* PDF of predicted discharge, deviates from multi-normal distribution, and/or has multiple local optima and discontinuous derivatives (Kuczera and Parent, 1998; Duan *et al.*, 1992). In these non-linear cases, other frequentist approaches are suggested whereby probability distributions are used to describe the stochastic behaviour of a system or at least of the measurement process. Model parameters are assumed to have true and fixed values, but if they are unknown, they are described by a probability distribution rather than a deterministic value. Finally, the uncertainty is propagated through the model. There is a huge literature on these analytical methods, *e.g.,* Ljung (1987) reviews on (mainly linear) techniques, and a description of existing methods for nonlinear systems is provided in Bates and Watts (1988) and Seber and Wild (1989). Since the focus of this thesis is not on these techniques, the formulations are not described in detail in this chapter.

2.3 Calibration and Uncertainty Estimation Techniques; Single-Criterion Cases

2.3.1 Bayesian Inference

In traditional approaches to uncertainty propagation, the determination of input PDFs must be specified. In contrast, in Bayesian inference, determination of input PDFs and propagation of uncertainties to model predictions are accounted for simultaneously. Bayes theorem allows updating the prior knowledge after observing present events. As such, the Bayesian approach treats the parameters as probabilistic variables having a joint PDF, which summarizes the probabilistic belief about the parameters. First, a prior PDF is considered for each parameter based on expert knowledge, which is subsequently combined with the likelihood derived based on confronting observed and simulated system's behaviours, *i.e.,* learning from data when they become available. This integration updates the joint PDF, which is then used to draw conclusions regarding parameter estimation and uncertainty propagation, *e.g.,* marginal distributions, most expected values, confidence intervals, etc.

Uncertainty estimation through Bayesian inference is conducted considering a regression model, whereby a deterministic component (the hydrologic model $h(\boldsymbol{\theta}, X, t)$ given the parameter vector $\boldsymbol{\theta}$, input data X, at time t) is combined with a random component which describes the remaining errors. Considering a time series of N streamflow observations, $Y_t \quad t = 1,..,N$, or $\mathbf{Y}$ in vector notation, Bayes' theorem uses the observed data to condition the posterior PDF of the parameter vector:

$$p(\boldsymbol{\theta} \mid \mathbf{Y}) = c \cdot p(\boldsymbol{\theta}) \cdot L(\boldsymbol{\theta} | \boldsymbol{\varepsilon}) \qquad (3)$$

where $p(\boldsymbol{\theta} \mid \mathbf{Y})$ is the posterior PDF of the model parameters conditioned on the calibration data set or observations $\mathbf{Y}$, c is a normalization constant, $p(\boldsymbol{\theta})$ is a prior PDF for the parameter vector $\boldsymbol{\theta}$, $\boldsymbol{\varepsilon}$ is a vector of model errors with individual elements computed as $\varepsilon_t = Y_t - \hat{Y}_t$, where $\hat{\mathbf{Y}} = h(\boldsymbol{\theta})$ is a vector of simulated streamflow, and $L(\boldsymbol{\theta}|\boldsymbol{\varepsilon})$ is the probability of observing $\mathbf{Y}$ given the set of model parameters $\boldsymbol{\theta}$. Assuming the model errors are mutually independent, uncorrelated, and normally distributed with mean of and variance of σ_ε^2, the posterior PDF conditioned on the observations has the following form:

$$p(\boldsymbol{\theta} \mid \mathbf{Y}) = c \cdot p(\boldsymbol{\theta}) \prod_{t=1}^{N} \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left(-\frac{\varepsilon_t^{\,2}}{2\sigma_\varepsilon^2} \right) \qquad (4)$$

It is often convenient to consider the natural logarithm of (4), which yields the log-posterior density (lp):

$$lp(\boldsymbol{\theta} \mid \mathbf{Y}) = \ln(c) + \ln(p(\boldsymbol{\theta})) - \frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma_\varepsilon^2) - \frac{1}{2\sigma_\varepsilon^2}\sum_{t=1}^{N} \varepsilon_t^{\,2} \qquad (5)$$

Inspection of Equations (4) and (5) reveals that the posterior PDF of the model parameters is dependent on both residuals, *i.e.,* deviation between simulations and observations, and corresponding error variance. In practice, the error variance is unknown and must be estimated in some fashion. Stedinger *et al.* (2008) suggest treating the error variance as a constant that is determined via maximum likelihood estimation or, equivalently, via standard least squares minimization. Alternatively, Box and Tiao (1973) suggest using a non-informative Jeffrey's prior distribution for the variance (*i.e.,* $p(\sigma_\varepsilon^2) \propto (\sigma_\varepsilon^2)^{-1}$), in which case the influence of the error variance can be integrated out, yielding a simplified form for the posterior PDF of the model parameters conditioned on the observations:

$$p(\boldsymbol{\theta} \mid \mathbf{Y}) \propto \left(\sum_{t=1}^{N} \varepsilon_t^{\,2} \right)^{-\frac{(N-1)}{2}} \qquad (6)$$

Another approach for dealing with the unknown variance term in Equation (4) is to assume that the error variance is conditional on the value of a given parameter vector. This makes it possible to determine a posterior distribution of the error variance jointly with the model parameters. Following Box and Tiao (1973), the resulting marginal PDF of σ_ε^2 follows an inverse chi-square distribution with N degrees of freedom with scale s, computed as follows:

$$s = \frac{1}{N}\sum_{t=1}^{N} \varepsilon_t^{\,2} \qquad (7)$$

Samples from the corresponding error variance distribution, conditional on a given parameter vector, is generated using the following formula:

$$\sigma_\varepsilon^2 = \frac{N}{z} \cdot s \qquad (8)$$

where z is a random variable drawn from a chi-squared distribution with N degrees of freedom and scale s. This formulation is suitable for the cases where errors are assumed to be normally distributed, independent, and homoscedastic. In order for the Bayesian inference to give meaningful and valid results, these assumptions must be met. The inconsistency between error model assumptions and observed residuals can introduce bias in estimated parameter values and affect parameter and predictive uncertainty (Thyer *et al.*, 2009).

In many real hydrological cases, errors are correlated, heteroscedastic (*i.e.,* in most cases, an increase in error variance with increase in streamflow discharge), and non-Gaussian (Kuczera, 1983). As a result, the standard Bayesian approach has been criticized in hydrology for relying too strongly on residual error assumptions that do not hold in many applications (Beven *et al.*, 2008). Several studies derive more complex formulations to extend the applicability of the Bayesian approach to the cases with autocorrelation, heteroscedasticity, and non-normality of model residuals. For instance, proper transformations, *e.g.,* Box-Cox transformation (Box and Cox, 1964), has shown to solve the heteroscedasticity issue (Sorooshian and Dracup, 1980). Alternatively, error models based on finite mixture of normal distributions for high-flows and low-flows show to be a promising solution to the heteroscedasticity problem (Schaefli *et al.*, 2007). Auto-correlation is also dealt with through development of autoregressive-based error models (Bates and Campbell, 2001; Kuczera, 1983; Laloy *et al.*, 2010; Yang *et al.*, 2007) – details are provided in Appendix A1.

More recently, Schoups and Vrugt (2010) have developed a new statistical error model to account for heteroscedasticity and non-normality. The methodology in Schoups and Vrugt (2010) model the error standard deviation as a linear function of simulated streamflow to induce homoscedasticity, and moreover, apply a parametric error distribution, *i.e.,* skew exponential power density, to account for non-normality. The formulation in Schoups and Vrugt (2010) is very complicated, but it is shown to significantly reduce the chance of obtaining inaccurate results due to dissatisfaction of statistical assumptions.

The analytical approach to derive the posterior joint PDF of model parameters is often challenging because it requires complicated integrals over probability distributions, which are often intractable. As a result, sampling-based methods have been applied to Bayesian parameter estimation as an alternative strategy, *e.g.,* Bayesian Monte Carlo (Dilks *et al.*, 1992), rejection sampling (Tarantola, 2005; von Neumann, 1951), Markov Chain Monte Carlo or MCMC sampling (Kavetski *et al.*, 2006; Kuczera and Parent, 1998; Thiemann *et al.*, 2001; Vrugt *et al.*, 2003b), etc. Among these methods, MCMC has been the most attractive sampling strategy applied to calibration and uncertainty estimation in hydrological modelling.

MCMC is based on the favorable mathematical properties of Markov chains, as they relate to Monte Carlo sampling and distribution estimation (Tierney, 1994). A Markov Chain is a sequence vectors $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots$ of stochastic variables where the value of the n^{th} vector depends only on the one-step past vector. In other words, $\tau(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_{t-2}, \ldots) = \tau(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1})$ where $\tau(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1})$ is called the transition kernel of the chain, and each new sample is taken from this kernel. It is assumed that the chain is time-homogenous, *i.e.,* τ does not depend on t. As time increases, the sample points increasingly resemble dependent samples from a unique stationary distribution, called 'target distribution'. In Bayesian inference, the application of MCMC is designed in such a way that the parameters posterior distribution becomes the target distribution. Due to ergodicity, reversibility, and convergence properties of MCMC algorithms (*e.g.,* Mengersen and Tweedie, 1996; Roberts and Tweedie, 1996), the successively generated values by the Markov Chain will be distributed with stable frequencies stemming from the posterior target distribution. Consequently, any statistical quantity of interest such as the PDF and its moments can be evaluated from the generated samples. Note that samples before MCMC's convergence are called 'burn-in' and discarded for the calculation of posterior statistics.

There are two distinguishable classes of algorithms commonly used for constructing Markov chains, those based on Metropolis-Hastings (MH) algorithm (Hastings, 1970; Metropolis *et al.*, 1953) and those related to Gibbs sampling (Geman and Geman, 1984). The Gibbs sample, brought into statistical application by Geman and Geman (1984), is a special case of MH for the cases where direct sampling is difficult, and the literature shows that most of MCMC applications in hydrology involves MH algorithms. MH algorithm uses an acceptance/rejection rule to converge to the specified target distribution. It starts from an initial value $\boldsymbol{\theta}_0$ from a starting distribution, draws a candidate $\boldsymbol{\theta}^*$ from a working conditional density or proposal distribution, computes the acceptance/rejection ratio α, and accepts $\boldsymbol{\theta}^*$ as a point in the chain with probability $\min(1, \alpha)$. If $\boldsymbol{\theta}^*$ is not accepted, the chain remains at the current point and

another candidate point is generated. MH algorithm converges to the posterior target distribution due to the ergodic properties (Gelman *et al.*, 2004).

Several efficient MCMC samplers have been developed based on MH algorithm. The numerical experiments in Chapter 3 utilize one of these samplers called DiffeRential Evolution Adaptive Metropolis (DREAM) (Vrugt *et al.*, 2009b). DREAM merges the strength of a previously developed algorithm called DE-MC (ter Braak and Vrugt, 2008) with delayed Rejection (DR) algorithm (Tierney and Mira, 1999) within a population-based evolutionary framework. DREAM is applied in numerous calibration and uncertainty estimation studies in hydrological modelling (*e.g.,* Laloy *et al.*, 2010; Vrugt *et al.*, 2008a; Vrugt *et al.*, 2008b). DREAM maintains ergodicity while showing excellent efficiency even if the target posterior distributions are complex, highly nonlinear, and/or multimodal. DREAM runs multiple Markov chains simultaneously to facilitate efficient global exploration of the parameter space. More details about this MCMC algorithm is provided in Vrugt *et al.* (2009b).

2.3.2 Informal Sampling-Based Approaches

The Bayesian inference approach is statistically rigorous, but at the same time, the utility of the approach depends upon validity of many underlying assumptions. Accordingly, many researchers have abandoned formal statistical inference in favour of more general, informal methods whereby certain informal performance metrics are employed in model calibration. Often these statistics are based on the familiar Pearson's correlation coefficient (r) or the coefficient of determination (R2), which describe the degree of collinearity between observations and model simulations (Legates and McCabe, 1999). Informal uncertainty estimation methods have been criticized that they do not require formal specification of an error model which might make them unreliable for uncertainty estimation (Kavetski *et al.*, 2002). However, comparison between informal and formal methods (in single-criterion cases) shows relatively close agreement among these methods from the predictive capability perspective (Beven *et al.*, 2008; Jeremiah *et al.*, 2011; Jin *et al.*, 2010; Li *et al.*, 2010a; Qian *et al.*, 2003; Vrugt *et al.*, 2008a; Yang *et al.*, 2008). A number of informal techniques for calibration and uncertainty estimation in hydrological modelling are described in the following sub-sections. Note that the informal techniques described below are not meant to exhaustively describe each type appearing in the literature. Instead, the methods described below are selected here because they are among the most commonly applied in the literature or have been selected for use in subsequent chapters of the thesis.

2.3.2.1 GLUE

The GLUE technique (Beven and Binley, 1992) is the most commonly applied method in the family of informal sampling-based methods. The merits of applying this methodology, especially in comparison to formal uncertainty estimation approaches, has been the subject of much debate (*e.g.,* Beven *et al.*, 2008; Mantovan and Todini, 2006; Montanari, 2005; Thiemann *et al.*, 2001). However, comparisons in hydrological literature have shown that formal Bayesian inference and informal GLUE have considerable common ground (Jin *et al.*, 2010; Li *et al.*, 2010a; Vrugt *et al.*, 2008a). More recently, Nott *et al.* (2012) showed that it is possible to give a formal Bayesian interpretation of GLUE even when the corresponding generalized likelihood is not a true likelihood. As a result, GLUE is still deemed to be a simple and practical approach uncertainty estimation, especially in rainfall-runoff modelling.

GLUE is an extension to MCS, and belongs to the family of importance sampling algorithms. GLUE is based on the concept of 'equifinality' which indicates that there are many different model structures and many different parameter sets within a chosen model structure that may be behavioural or acceptable in reproducing the observed behaviour of that system (Beven and Freer, 2001). As such, GLUE classifies Monte Carlo samples as having produced model output that is either 'behavioural' (*i.e.,* plausible, given the data and one's knowledge of the system) or 'non-behavioural'. The behavioural solutions are then used to derive the cumulative density function (CDF) of the model output of interest. Quantiles from these CDFs quantify the model output uncertainty, *e.g.,* the 5th and 95th percentile of the output. In GLUE, model output uncertainty accounts for all sources of uncertainty, because "the likelihood measure value associated with a parameter set reflects all these sources of error and any effects of the covariation of parameter values on model performance implicitly" (Beven and Freer, 2001). Other authors, however, challenge this assertion (*e.g.,* Clark *et al.*, 2011; Mantovan and Todini, 2006; Stedinger *et al.*, 2008). A GLUE analysis consists of the following four steps:

1. Defining the generalized informal likelihood measure $l(\boldsymbol{\theta})$. Generally, the measure $l(\boldsymbol{\theta})$ is a pseudo-likelihood function which demonstrates the model performance for a particular parameter set $\boldsymbol{\theta}$.

2. After defining $l(\boldsymbol{\theta})$, a large number of parameter sets are randomly sampled from the prior distribution and each parameter set is assessed as either ''behavioural'' or ''non-behavioural'' through a comparison of the likelihood measure with a subjectively selected threshold value.

3. Each behavioural parameter set is given a likelihood weight according to $\varpi_i = l(\boldsymbol{\theta}_i) \Big/ \sum_{k=1}^{N} l(\boldsymbol{\theta}_k)$, where N is the number of behavioural parameter sets.

4. Finally, the uncertainty of model outcome (*e.g.,* streamflow) is described by the quantiles of the outcome associated to the behavioural solutions, utilizing the weighted likelihood values at each time step, *e.g.,* 95% prediction intervals are constructed based on 2.5 and 97.5 percentiles.

The measure $l(\boldsymbol{\theta})$, which is used in GLUE, is generally a pseudo-likelihood function which demonstrates the model performance for a particular parameter set. The prior distributions reflect the uncertainty in parameters, which is propagated to predictions via sampling, followed by weighting the behavioural solutions. In spite of GLUE ease-of-use, researchers have criticised GLUE with respect to its statistical rigor (Mantovan and Todini, 2006; Montanari, 2005; Thiemann *et al.*, 2001). Beven *et al.* (2008) respond to these drawbacks, and mention "GLUE can indeed be coherent with the Bayesian approach in that identical formal likelihoods can be used where there is sufficiently strong prior information to justify formal likelihood assumptions". Beven *et al.* (2008) clarify that GLUE is intended to allow flexibility in dealing with the complexity of real applications, and furthermore, to make sure that there is a framework to allow the user to show explicitly how confidence in a model is being assessed and quantified, *i.e.,* in the likelihood measure used. Besides the question of statistical rigor, Kuczera and Parent (1998) and Tolson and Shoemaker (2008) show that the typical GLUE sampling approach of uniform random sampling can be very inefficient and this inefficiency will increase as the number of uncertain parameters increases.

2.3.2.2 Markov Chain Monte Carlo (MCMC) Implementations

The MCMC samplers, typically used in formal Bayesian inference, have also been applied as informal sampling-based approaches to address uncertainty in hydrology. The goal has been to use the functionality of MCMC samplers within informal formulations to take samples in the most likely region of parameters search space. Any non-statistically based likelihood function can be manipulated, so that its domain falls on the interval [0,1], and its values become monotonically increasing with improved model's performance. For instance, McMillan *et al.* (2009) use the Nash-Sutcliffe Efficiency (NSE) criterion as one of the most commonly used descriptors of rainfall-runoff model performance:

$$NSE = 1 - \frac{\sigma_\varepsilon^2}{\sigma_o^2} \qquad (9)$$

where σ_ε^2 is the error variance and σ_o^2 is the variance of the observed flow series. The NSE takes a value of 1 for a perfect model fit, a value of 0 for a model no better than the constant mean of the observed data. In order for this criterion to be used in MCMC, it is set to zero when negative values are returned. The

NSE is used via the acceptance/rejection ratio α of two samples, which can be expressed in the following form:

$$\alpha = \frac{1-\sigma_{\varepsilon 1}^2/\sigma_o^2}{1-\sigma_{\varepsilon 2}^2/\sigma_o^2} = \frac{\sigma_o^2-\sigma_{\varepsilon 1}^2}{\sigma_o^2-\sigma_{\varepsilon 2}^2} = \frac{K-SSE_1}{K-SSE_2} \qquad (10)$$

where SSE_1 and SSE_2 are the sums of squared errors for the two samples and K is a constant. The informal likelihood is rescaled to fall in the range [0,1] so that it could be a mirror of a statistical likelihood function. However, in practice, MCMC sampling can accommodate the measures that do not fall in [0,1]. Blasone *et al.* (2008a) also use the following informal likelihood function which is commonly used in the GLUE methodology:

$$l(\theta_i \mid \mathbf{Y}, \mathbf{X}) = \exp\left\{-N\sigma_{\varepsilon i}^2/\sigma_0^2\right\} \qquad (11)$$

where $l(\boldsymbol{\theta}_i \mid \mathbf{Y}, \mathbf{X})$ is the likelihood measure for the i^{th} model parameter set conditioned on the observations $\mathbf{Y}$ and input data $\mathbf{X}$. The exponent N is an adjustable parameter that sets the relative weightings of the better and worse solutions: higher N-values have the effect of giving more weight to the best simulations, thus increasing the difference between good and bad solutions (Freer *et al.*, 1996). The likelihood function $l(\,)$ can assume values between 0 and 1. The closer to 1 the likelihood is, the better the simulations are. Thus, this quantity has to be maximized within MCMC sampling. When a sufficient number of samples are taken by the MCMC sampler, the model output uncertainty is quantified using the quantiles (*e.g.,* 5th and 95th percentile) of the simulated ensemble of model outcome.

2.3.2.3 Optimization-based Methods

Other informal calibration and uncertainty estimation methods exist, in which optimization algorithms are employed to search the parameters space for regions of high goodness of fit between observations and simulations. The optimal and near-optimal parameter sets obtained by algorithms are used in the simulation model and the corresponding model outputs are used to provide the uncertainty bounds of model outcome. Among this class of uncertainty estimation methods are DDS-AU (Tolson and Shoemaker, 2008) and ParaSol (van Griensven and Meixner, 2007).

In DDS-AU, the DDS optimization algorithm is used to independently identify multiple acceptable or behavioural parameter sets. DDS-AU is an efficient uncertainty estimation approach that quantifies the model output uncertainty using prediction bounds rather than prediction limits. A typical optimization formulation in the application of DDS algorithm is as follows:

$$\max_{\theta \in \Theta} F(\boldsymbol{\theta}) \qquad (12)$$

where $F(\boldsymbol{\theta})$ is the vector of objective functions, $\boldsymbol{\theta}$ is an arbitrary parameter set, and Θ is the feasible parameters space. Considering this formulation, DDS-AU consists of the following steps: (1) maximum total number of model evaluations for the analysis (N_{Total}) is defined according to the available computational budget, and desired number (*i.e.,* maximum required) of behavioural solutions to identify (n_{beh}); number of model evaluations per DDS optimization trial (m_{DDS}) is then calculated as $m_{DDS} = N_{Total}/n_{beh}$ (rounded down to an integer); (2) n_{beh} DDS optimization trials are performed from n_{beh} random initial parameter sets, optimizing the informal likelihood function $F(\boldsymbol{\theta})$; (3) a threshold is considered for $F(\boldsymbol{\theta})$ and the behavioural solutions are classified among n_{beh} final best DDS, *i.e.,* the solutions resulting in likelihood values better than the threshold for $F(\boldsymbol{\theta})$; (4) the range of simulated hydrograph predictions across all behavioural solutions provides an approximate quantification of the predictive uncertainty. The simplest implementation of DDS-AU identifies only the final best parameter set from each trial as a possible behavioural solution (provided that the corresponding objective function value is better than the threshold).

The ParaSol method (van Griensven and Meixner, 2007) is a strategy to use the SCE-UA optimization algorithm (Duan *et al.*, 1992) to extract uncertainty information. ParaSol utilizes either a threshold for the objective function or Bayesian concepts to divide the simulations into acceptable and non-acceptable simulations. Then, the model output uncertainty is constructed by equally weighting all good simulations, and identifying the relevant percentiles of the generated hydrograph ensemble. It is worth noting that, in general, any type of optimization algorithms can be applied within the structure of aforementioned uncertainty estimation methodologies.

2.4 Calibration and Uncertainty Estimation Techniques; Multi-Criteria Cases

2.4.1 Formal Multi-Criteria Bayesian Inference

The single criterion Bayesian inference methods can be applied to multi-criterion calibration cases where multiple sets of observations are available and an explicit likelihood function can be defined for each set of observations. Unfortunately, this application presents added difficulties such as correlation among sets of observations, or more importantly, the need to a larger computational budget in cases where MCMC sampling is applied, *i.e.,* convergence might need larger number of simulation than single-criterion cases. The literature includes studies which present both analytical and sampling-based Bayesian approaches to deal with multi-criteria calibration (Balin-Talamba *et al.*, 2010; Hong *et al.*, 2005; Kuczera, 1983;

Kuczera and Mroczkowski, 1998; Mroczkowski *et al.*, 1997; Schaefli *et al.*, 2007). A typical multi-criteria Bayesian model calibration process can involve multiple likelihood functions used for different sets of measurements, *e.g.,* discharge, sediment, snow water equivalent, etc. As such, the literature also uses the term 'multi-response Bayesian calibration' in such cases.

The posterior distribution of model parameters conditioned on the observations in the multi-response Bayesian context, or $p(\boldsymbol{\theta} \mid \mathbf{Y})$, is identified as follows:

$$p(\boldsymbol{\theta} \mid \mathbf{Y}) = c \cdot p(\boldsymbol{\theta}) \prod_{k=1}^{K} \prod_{t=1}^{T_k} \frac{1}{\sqrt{2\pi\sigma_{\varepsilon,k}^2}} \exp\left(-\frac{\varepsilon_{t,k}^{\ 2}}{2\sigma_{\varepsilon,k}^2} \right) \qquad (13)$$

where $\mathbf{Y}$ is the vector of observations, k and t subscripts refer to response and time, respectively, $p(\boldsymbol{\theta})$ is the prior distribution of model parameters, c is the normalization constant, $\sigma_{\varepsilon,k}^2$ is the error variance of the response k. Basically, given a particular response k that is a set of T_k observation, the prior density is sequentially multiplied by the likelihood functions associated to T_k observations. The resulting distribution is subsequently updated utilizing the next set of observations, and this process is continued until all responses are used to update the parameters probability distributions.

There are also reports of multi-criteria Bayesian formulations using a single response, *e.g.,* Schaefli *et al.* (2007) consider multiple likelihood functions associated with high- and low-streamflows. Autoregressive-based likelihood functions are also developed in the multi-criteria context (Balin-Talamba *et al.*, 2010; Schaefli *et al.*, 2007). Details of multi-criteria autoregressive-based Bayesian formulations are provided in Appendix A2. Chapter 3 of the thesis applies such formulations in calibration and uncertainty estimation of multiple hydrologic models. Despite the differences between Bayesian formulations in single- and multi-criterion cases, parameter uncertainty is quantified the same in these two approaches, *i.e.,* deriving the PDF of model outcome and then describing the uncertainty using quantiles of model outcomes.

2.4.2 Informal Multi-Criteria Uncertainty Estimation Approaches

2.4.2.1 Pareto Optimality-Based Approaches

It is shown in the literature that the aggregation of all residuals into a single likelihood (objective) function during calibration does not provide detailed information about where the model is failing to perform well (Boyle *et al.*, 2000; Gupta *et al.*, 1998; Yapo *et al.*, 1998). As pointed out in the introduction section, Gupta *et al.* (1998) applied Pareto-based multi-criteria formulations for calibration of

hydrological models. This section elaborates on the Pareto optimization concepts, explains how model output uncertainty is estimated using the Pareto solutions, and briefly introduces the algorithmic aspects that are important in solving a Pareto-based calibration problem.

The Pareto optimality approach has the ability to simultaneously incorporate several objective functions. The multi-objective calibration problem can be stated as follows:

$$\min_{\theta \in \Theta} F(\theta) = \begin{bmatrix} F_1(\theta) \\ F_1(\theta) \\ \vdots \\ F_M(\theta) \end{bmatrix} \qquad (14)$$

where $F_i(\theta)$ is the i^{th} of M objective functions. The solution to this problem will in general be a set of parameter sets (the Pareto solution) corresponding to various trade-offs among the objectives, rather than a unique single parameter set. This way, feasible parameter space can be partitioned into non-inferior or Pareto solutions and inferior or dominated solutions. In the absence of additional information, it is impossible to distinguish any of the Pareto solutions as being objectively better than any of the other Pareto solutions. Furthermore, every member of the Pareto set will match some characteristic of the observed data better than any other member of the Pareto set, but the trade-off will be that some other characteristic of the observed data will not be as well-matched (Yapo *et al.*, 1998).

Gupta *et al.* (1998) also introduce the Pareto optimality as a simple method to deal with uncertainty. The Pareto optimality approach shares a similar notion with the "equifinality" concept of GLUE methodology in the sense that the number of the good or "behavioural" parameter sets, namely, Pareto solutions, defines model output uncertainty. The Pareto set of solutions defines the minimum uncertainty in the parameters that can be achieved without stating a subjective relative preference for minimizing one specific objective function at the expense of another (Vrugt *et al.*, 2003a). Thus, the Pareto optimality approach quantifies the total uncertainty simply in terms of the prediction range associated with Pareto parameter sets.

The literature shows that there are numerous studies that apply multi-criteria calibration to a wide range of hydrologic models, ranging from lumped rainfall–runoff models to semi-distributed and distributed hydrologic models. Recently, Efstratiadis and Koutsoyiannis (2010) have conducted a thorough review on multi-criteria calibration in hydrological modelling.

Pareto optimization is conducted using multi-objective evolutionary algorithms (MOEAs). Evolutionary algorithms (EAs) are well-established tools for handling nonlinear optimization problems,

and the key feature of these methods is the search of the parameters space, through a set (or population) of randomly generated points that evolves on the basis of stochastic transition schemes, *e.g.,* the genetic operators. MOEAs aim to spread the population along the Pareto front instead of converging around a single optimum. There are numerous MOEAs developed for solving multi-objective optimization problems, and multi-criteria calibration problems in the hydrological modelling. Recently, Reed *et al.* (2013) reviews state-of-the-art MOEAs and highlights key advances that the water resources field can exploit to better discover the critical trade-offs constraining the water resources systems.

In this thesis, most of the multi-criteria calibration experiments utilize A Multi-ALgorithm Genetically Adaptive Multi-objective (AMALGAM) method (Vrugt and Robinson, 2007) for locating the Pareto optimal solutions. AMALGAM uses a self-adaptive multi-method evolutionary search and simultaneously employs a diverse set of optimization algorithms. The algorithms considered to configure AMALGAM include Non-dominated Sorting Genetic Algorithm (NSGA-II) (Deb *et al.*, 2002), Particle Swarm Optimization (PSO) (Kennedy *et al.*, 2001), Adaptive Metropolis (AM) (Haario *et al.*, 2001), and Differential Evolution (DE) (Storn and Price, 1997). AMALGAM has a number of algorithmic parameters that are described in details in Vrugt and Robinson (2007). The only parameter that needs to be set in the numerical experiments of this thesis is the population size (*i.e.,* the number of individuals in each AMALGAM's generation).

Similar to almost all other MOEAs, AMALGAM uses the Pareto ranking (Goldberg, 1989) to sort the parameter sets, and moreover, maintains diversity to avoid convergence to single solutions (Coello Coello, 2005). The algorithms embedded in AMALGAM apply a mating selection process to select some solutions for generating new solutions (called offspring). Mating selection in most MOEAs is based on the dominance rank, *i.e.,* the lower the dominance rank the better. When two solutions have the same rank, they need to be compared with respect to another selection metric, which varies from algorithm to algorithm. Among these selection metrics are crowding distance (Deb *et al.*, 2002) or hypervolume contribution index (Emmerich *et al.*, 2005). Basically, the algorithm selects solutions with more desired values for these metrics with higher probabilities during optimization. Similar to other MOEAs, the final outcome of AMALGAM is a set of parameter sets that can be divided into Pareto optimal solutions and dominated solutions of which the former are used to provide the model output uncertainty bounds.

2.4.2.2 Criteria Aggregation-Based Approaches

The other set of methods to quantify the uncertainty in multi-criteria calibration of hydrologic models consider aggregation of multiple criteria into a single objective function. These methods then apply informal uncertainty estimation techniques to evaluate the uncertainty. As such, all the informal methods, which are used for single-criterion cases, are applicable to multi-criteria cases where the criteria are aggregated. Van Griensven *et al.* (2007) represented the aggregation of different criteria in terms of Global Optimization Criterion (GOC), and developed the ParaSol method to address the uncertainty in multi-criteria cases.

Other research papers dealing with criteria aggregation include those which applied GLUE for multi-criteria cases (Balin-Talamba, 2004; Blazkova *et al.*, 2002; Choi and Beven, 2007; Freer *et al.*, 2004; Lamb *et al.*, 1998). The aggregated generalized likelihood in multi-criteria implementations of GLUE can be calculated as follows:

$$l(\theta) = \prod_{i=1}^{M} \exp\left(- W_i \frac{\sigma_{\varepsilon,i}^2}{\sigma_{o,i}^2} \right) \qquad (15)$$

where W_i represents the weighting factor for criterion i, M is the number of criteria, $\sigma_{\varepsilon,i}^2$ and $\sigma_{o,i}^2$ are the variance of simulation errors and the variance of observed data, respectively, over the time window in which criterion i is calculated. The likelihood function $l(\theta)$ equals 1 if the observed and simulated data are the same for all criteria, and reduces towards zero as the similarity decreases.

To identify behavioural solutions in multi-criteria cases considering the formulation described in the previous paragraph, the standard behavioural solution identification strategy can be followed. As such, Lamb *et al.* (1998) and Balin-Talamba (2004) follow the original paper on GLUE by Beven and Binley (1992) and identify the behavioural solutions to be the top 10% of solutions set after sorting the parameter sets based on aggregated likelihood function. In their studies, once prior distributions are sampled from, the generalized likelihood function Equation (15) is calculated with equal weights as follows ($W_1 = 0.5$ and $W_2 = 0.5$ for cases with two responses):

$$l(\theta) = \exp\left(- W_1 \frac{\sigma_{\varepsilon,1}^2}{\sigma_{o,1}^2} \right) . \exp\left(- W_2 \frac{\sigma_{\varepsilon,2}^2}{\sigma_{o,2}^2} \right) = \exp\left[- \frac{1}{2} \left(\frac{\sigma_{\varepsilon,1}^2}{\sigma_{o,1}^2} + \frac{\sigma_{\varepsilon,2}^2}{\sigma_{o,2}^2} \right) \right] \qquad (16)$$

Lamb *et al.* (1998) show that relaxation of the rejection threshold to define a larger proportion of the total number of solutions as behavioural would cause only slight modifications of uncertainty bounds.

The reason for this insensitivity to the rejection threshold is that even after selecting a larger number of behavioural solutions, the majority of solutions would achieve only small likelihood values. Therefore, the predictions associated with these poor solutions would fall within the tails of the cumulative distributions of model outcome. Given the rescaling stage in GLUE (Step 3 in the framework described in Section 2.3.2.1), these predictions would have little effect on the location of uncertainty bounds (Lamb *et al.*, 1998).

2.4.2.3 Informal Multi-Criteria MCMC Sampling

MCMC samplers are designed to be employed in Bayesian inference structures, whereby formal statistical likelihood functions are combined with prior distributions to result in samples from the joint posterior distribution of the model parameters. These sampling techniques can be also applied to informal multi-criteria formulations to assess the uncertainty. Blasone *et al.* (2008a) develop an informal MCMC sampler to generate behavioural parameter sets in the GLUE framework. The likelihood functions are the weighted (aggregated) Root Mean Square Error (RMSE) of discharges and groundwater levels. In order to apply the aggregated objective function F in MCMC sampling, the likelihood function $l(\,)$ is calculated as the reciprocal of F, and then rescaled by a constant factor when all the behavioural solutions are identified:

$$l(\theta_i \mid \mathbf{Y}, \mathbf{X}) = \frac{1}{F_i} . \frac{1}{C} \qquad (17)$$

where C is a normalizing factor introduced to ensure that the sum of the likelihood functions of the behavioural solutions equals 1.

Vrugt *et al.* (2003a) also develop an informal MCMC sampler for multi-criteria calibration of a rainfall-runoff model based on the dominance concept. Their algorithm, called MOSCEM, calculates the likelihood value of each sample, sorts all samples based on the dominance concept, and identifies the set of non-dominated solutions (Pareto front solutions). Upon termination of the algorithm, the mode output uncertainty can be quantified using the Pareto front solutions in the exisiting set of samples. However, the main focus of the study by Vrugt *et al.* (2003a) is only on the calibrtion, rather than uncertainty estimation.

2.4.2.4 Multi-criteria GLUE Approach

This multi-criteria GLUE approach is distinguished from the GLUE approach that aggregates likelihood functions into one function in Section 2.4.2.2 in that the likelihood functions for each calibrated model

response are not aggregated. To apply GLUE in the multi-criteria context, certain changes must be applied to the main GLUE framework presented in Section 2.3.2.1. For instance, step 1 would require consideration of multiple likelihood values. The literature shows that there are at least two methods to do so. The first method is based on likelihood aggregation which was previously detailed in Section 2.4.2.2. The second approach, instead of aggregating likelihood values by multiplication, Equation (15) or Equation (16), separately accounts for individual likelihood values. Then, only the solutions which yield higher performance values than pre-defined thresholds for all the likelihood functions are retained in the behavioural set (*e.g.,* Blazkova *et al.*, 2002; Dean *et al.*, 2009; Freer *et al.*, 1996; Gallart *et al.*, 2007; Mitchell *et al.*, 2009; Muleta and Nicklow, 2005).

To define the behavioural threshold, instead of considering Equation (15), the Nash-Sutcliffe Efficiency (NSE) (Nash and Sutcliffe, 1970) can be employed for both responses as individual likelihood metrics. NSE is formulated in Equation (9). NSE takes value in the range of $(-\infty,1)$ where a perfect model fit yield 1. With NSE available for both responses, behavioural thresholds must be defined and many studies demonstrate this approach (*e.g.,* Arabi *et al.*, 2007; Blazkova *et al.*, 2002; Dean *et al.*, 2009; Freer *et al.*, 1996; Gallart *et al.*, 2007; Mitchell *et al.*, 2009; Muleta and Nicklow, 2005).

Note that when the behavioural solutions are identified in the aforementioned strategy, step 3 in GLUE framework (see Section 2.3.2.1) requires a single likelihood value for rescaling the parameter sets and deriving the prediction intervals. Literature shows that different approaches can be applied for this purpose, such as multiplication of performance criteria (Muleta and Nicklow, 2005), sum of likelihood values (Mitchell *et al.*, 2009), or sequentially updating the likelihood value of each parameter set (Freer *et al.*, 1996).

2.5 Discussion

Each of the uncertainty estimation approaches described in this chapter may have advantages and disadvantages that must be taken into consideration prior to applying them to uncertainty estimation. For instance, formal approaches rely on the satisfaction of a number of statistical assumptions and the ability of defining a proper likelihood function (Box and Tiao, 1973; Gelman *et al.*, 2004; Kuczera, 1983; Thyer *et al.*, 2009). These assumptions may not be satisfied in complex hydrological modelling experiments (*e.g.,* Beven *et al.*, 2008). Furthermore, a robust Bayesian inference requires certain post diagnostic checks in order to ensure that the results are statistically consistent (Kuczera, 1983; Thyer *et al.*, 2009). These tests are often neglected in calibration studies, *i.e.,* no reports of diagnostic checks, or when these checks are reported, no efforts for changing the formulations towards appropriately meeting the assumptions (*e.g.,* Feyen *et al.*, 2007; Hutton *et al.*, 2014). Furthermore, application of MCMC sampling

to the Bayesian inference in complex hydrological modelling experiments can become computationally expensive with some studies reporting the number of model runs required to be more than 100,000 (*e.g.,* Engeland *et al.*, 2005; Smith and Marshall, 2008; Vrugt *et al.*, 2003b).

On the other hand, informal methods such as GLUE are challenged by researchers because they do not accommodate an error series, and rely on a set of subjective decisions that impacts the model output uncertainty. For instance, the type of likelihood measure in GLUE impacts the uncertainty bounds. Also, in the Pareto optimality-based approaches, multiple optimization trials may result in different Pareto fronts, which means that each trial may yield a different uncertainty bound. Moreover, even when the criteria are aggregated in GLUE for example, depending on the criteria aggregation scheme, different sets of behavioural parameter sets may be identified, *i.e.,* different model output uncertainty bounds. Despite these challenges, informal methods are still employed by various hydrologists and practitioners because they are simple and easy-to-use. However, there is room to improve these informal methods, which is in fact the main objective of this research.

It should be pointed out that the original overall goal of the thesis was to propose improvements to uncertainty estimation methods for multi-criteria hydrologic model calibration. Therefore, the first step was to compare informal and formal methods (Chapter 3). Although they were found to be notably different, the formal method failed to satisfy some assumptions. Improving formal methods would require finding a way to make different assumptions and satisfy them all, and there is a substantial body of literature studying that issue currently. Therefore, the remainder of the thesis (Chapters 4 and 5) was focused on making improvements to informal methods in areas where there was either a small number of existing studies or no research.

Chapter 3
Uncertainty-Based Multi-criteria Calibration of Rainfall-Runoff Models: A Comparative Study

Summary

Chapter 1 and 2 of the thesis elaborated on the importance of uncertainty estimation in hydrological modelling, as well as existing formal and informal uncertainty estimation techniques applied to single- and multi-criteria cases. This chapter compares GLUE (as a commonly used informal approach) with formal Bayesian inference in the multi-criteria context. Bayesian inference is accomplished through Markov Chain Monte Carlo (MCMC) sampling based on an auto-regressive multi-criteria likelihood formulation. Non-converged MCMC sampling is also considered as an alternate informal methodology for calibration and uncertainty estimation. Methods are compared along multiple comparative measures calculated over the calibration and validation period in two case studies. Results demonstrate that there can be considerable differences in hydrograph prediction intervals generated by formal and informal uncertainty estimation strategies. Moreover, the performance of non-converged MCMC is reasonably consistent with those given by a formal and fully-converged Bayesian approach even though converged MCMC requires significantly larger number of model simulations. Section 3.1 provides a brief introduction, followed by elaboration on the material and methods in Section 3.2. Results and discussions are provided in Sections 3.3 and 3.4. Finally, Section 3.5 summarized the concluding remarks. This chapter is a mirror of the following published article with minor changes to increase its consistency with the body of the thesis and to avoid redundant material. Changes are only made in the Summary (abstract), Section 3.1, and Section 3.2.3. Moreover, references are unified at the end of the thesis.

Shafii, M., Tolson, B., Matott, L.S., (2014), Uncertainty-based multi-criteria calibration of rainfall-runoff models: a comparative study, Stoch. Env. Res. Risk., 28: 1493 – 1510.

3.1 Introduction

Calibration using a multi-criteria approach is suggested as a strategy to improve the reliability of models calibration (Gupta *et al.*, 1998). Moreover, proper uncertainty analysis proves to increase the usefulness of predictive hydrologic models. As such, recent calibration methodologies include an assessment of uncertainty in the calibration process, *i.e.,* traditional 'frequentist' approaches (*e.g.,* Bates and Watts, 1988; Reichert, 1997; Seber and Wild, 1989) and Bayesian inference approaches (Bates and Campbell, 2001; Box and Tiao, 1973; Gelman *et al.*, 2004; Kavetski *et al.*, 2002; Kuczera, 1983; Kuczera and

Parent, 1998; Thiemann *et al.*, 2001), or informal methods (*e.g.,* Beven and Binley, 1992; Blasone *et al.*, 2008a; Tolson and Shoemaker, 2008; van Griensven and Meixner, 2007).

To assess the strength of formal and informal uncertainty-based calibration techniques in uncertainty quantification, certain researchers have focused their efforts towards comparison between these methods in hydrological literature. Overall, these efforts generally indicate relatively close agreement among alternative methods, in terms of predictive capability (Beven *et al.*, 2008; Jeremiah *et al.*, 2011; Jin *et al.*, 2010; Li *et al.*, 2010a; Qian *et al.*, 2003; Vrugt *et al.*, 2008a; Yang *et al.*, 2008). Note that some studies have only considered informal methods in their comparisons (*e.g.,* Blasone *et al.*, 2008b). The observed similarity between the predictive capabilities of formal and informal approaches suggests that one might be able to gain insight into predictive uncertainty by means of informal approaches without getting involved in likelihood definition and corresponding assumptions.

Most of previous papers comparing formal and informal approaches have only considered single-criterion calibration scenarios. Balin-Talamba (2004) and Balin-Talamba *et al.* (2010) considered multi-criteria calibration of hydrologic models applying GLUE and MCMC sampling. These studies evaluated the impact of multi-response calibration on predictive uncertainty using GLUE and MCMC, in comparison with single-criterion calibration. However, the GLUE and MCMC techniques are only visually compared in Balin-Talamba (2004) and no comparative measures are reported. To the best of our knowledge, comparison among formal and informal techniques from a multi-criteria perspective using quantitative comparative measures has yet to be reported on in the literature.

The main objective of the research in this chapter is to evaluate the applicability of different uncertainty-based calibration approaches to multi-criteria calibration and uncertainty estimation of hydrologic models. The methodologies addressed in this study are statistically-based Bayesian inference using MCMC sampling (Bates and Campbell, 2001; Kuczera, 1983; Schaefli *et al.*, 2007; Vrugt *et al.*, 2009b), and sampling-based uncertainty estimation using GLUE (Beven and Binley, 1992; Blazkova and Beven, 2009). Bayesian inference is implemented using the DREAM MCMC sampler (Vrugt *et al.*, 2009b) through a robust multi-criteria formulation. Also, an alternative Bayesian method is considered based on the results of MCMC sampling up to a limited computational budget (*i.e.,* using the MCMC before convergence). Such a method cannot be viewed informal, as it uses formal likelihood function; however, it would not be formal either, as convergence has not occurred, meaning that the solutions in the chain could not be considered as samples from posterior distributions.

3.2 Methodology

The comparison framework of this study uses the posterior distribution of model parameters derived from MCMC sampling, as well as the behavioural or optimal parameter sets obtained from other methods. In order to be consistent in wording, the term "posterior" is applied to all of the considered techniques even though the results of non-converged MCMC sampling and GLUE are not a formal statistical posterior distribution. Results are then compared with respect to computational burden, complexity, and predictive capacity. Numerical experiments are aimed at exploring advantages and disadvantages of the uncertainty-based calibration techniques addressed in this study in multi-criteria calibration of rainfall-runoff models.

The GLUE approach of this study employs informal likelihood functions and results are compared with those obtained from formal Bayesian inference as well as non-converged MCMC sampling. The use of GLUE without a formal likelihood function has been the subject of much debate (*e.g.,* Beven *et al.*, 2008; Mantovan and Todini, 2006; Montanari, 2005; Thiemann *et al.*, 2001). Nevertheless, GLUE is used with an informal generalized likelihood function in this study because the objective of the study is to assess the performance of informal methods. Much of the reason informal methods like GLUE are so well utilized in practice is because they can use informal likelihood functions based on long utilized deterministic calibration objective functions like sum of squared errors or the Nash Sutcliffe coefficient. It is also worth noting that GLUE could also be applied using formal likelihood functions (Freni and Mannina, 2009; Romanowicz *et al.*, 1994), but this is not addressed in the present thesis.

The comparison approach (informal to formal methods) of this study is exactly consistent with previous comparative studies of uncertainty-based calibration in hydrological modelling (*e.g.,* Vrugt *et al.*, 2008a; Yang *et al.*, 2008). Beven (2009) noticed that in Vrugt *et al.* (2008a) the formal Bayes estimates are based on an autoregressive error model, while such information is not supplied to the GLUE simulations. Despite the difference between the formulations of the Bayesian approach and GLUE in Vrugt *et al.* (2008a), it is shown in that paper that formal and informal uncertainty estimation methods have some common ground with respect to the total predictive uncertainty in single-criterion calibration cases. In this study, multiple quantitative comparative measures are applied and the similarity in behaviour of MCMC and GLUE is evaluated in the multi-criteria context. As such, this study considers the same implementations of MCMC sampling and GLUE as used in Vrugt *et al.* (2008a).

3.2.1 Formal Multi-Criteria Bayesian Inference

Bayesian statistics have been shown to be a robust methodology for formal multi-criteria uncertainty-based calibration of hydrologic models (Balin-Talamba *et al.*, 2010; Hong *et al.*, 2005; Kuczera, 1983; Kuczera and Mroczkowski, 1998; Mroczkowski *et al.*, 1997; Schaefli *et al.*, 2007). The research

presented here uses a previously published multi-criteria formulation (Balin-Talamba *et al.*, 2010; Schaefli *et al.*, 2007) accounting for Auto-Regressive parameters; the formulation is described in Appendix A2 of this thesis. Note that the AR scheme is applied separately to the low- and high-flow regimes and this resulted in the addition of two AR parameters (ρ_L for low-flows and ρ_H for high-flows) to the set of calibrated parameters. The DREAM MCMC sampler is used for formal Bayesian inference (Vrugt *et al.*, 2009b), which is previously described in Section 2.3.1.

3.2.2 Sampling-Based Uncertainty Estimation Using Non-Converged MCMC

Even though applications of MCMC sampling with pseudo-likelihood functions have been previously reported in the literature (Blasone *et al.*, 2008b; Vrugt *et al.*, 2003a), there has been no report on evaluation of the results from non-converged MCMC samplers with formal likelihood functions. In this study, non-converged DREAM results are used to approximate the converged MCMC sampling strategy. The number of solutions taken from a given DREAM chain is defined to be consistent with the informal methods considered in this study (explained below). For example, if the informal methods use a budget of 10000 simulations, then only 10000 solutions from the initial part of the long DREAM chain are used. Afterwards, the last 1000 solutions of this set are treated as posterior solutions to derive prediction intervals. Clearly, such an approach is neither formal (as convergence has not occurred) nor informal (as it uses formal likelihood function). That is the reason why this approach is separated from formal Bayesian and informal GLUE approaches.

3.2.3 Sampling-Based Uncertainty Estimation using GLUE

The GLUE technique (Beven and Binley, 1992) is the most commonly applied method in the family of informal sampling-based methods. Sections 2.4.2.2 and 2.4.2.4 describe GLUE in details and elaborates on the strategies of implementing GLUE in the multi-criteria context. Note that this study aggregates the criteria, and upon sorting the solutions based on the aggregated creation, considers the top 10% strategy to define behavioural samples.

3.2.4 Comparison Measures

The main goal of calibration and uncertainty estimation is to assess a model predictive capability. Therefore, in order to evaluate uncertainty-based calibration techniques, it seems necessary that the focus is more on the validation time period rather than the calibration period. Nonetheless, a portion of the analyses examine differences between calibration and validation results. The comparative measures are calculated based on the results obtained using the posterior parameter sets. It should be noted that the parameter uncertainty is derived based on the envelope of model outputs using the posterior parameter

sets. Moreover, in order to derive the predictive uncertainty, the entire set of posterior parameters is first used in simulation model to derive the parameter uncertainty. Afterwards, error parameters are sampled to generate a correlated residual time series which is then added to model outputs.

To evaluate the quality of resulting model outcomes, efficiency measures such as Nash-Sutcliffe Efficiency (NSE) is used to assess model performance; NSE is calculated as Equation (9) in Chapter 2. In the multi-criteria context of this study, the scatterplot of posterior parameter sets is illustrated in bi-criteria space (*i.e.,* NSE for high- and low-flows).

In addition, the generated model outcomes using the posterior solutions derived from different techniques are used to derive the predictive uncertainty which can be assessed using a variety of measures. Among the simplest measures for comparing alternative realizations of predictive uncertainty are the reliability and sharpness measures (Yadav *et al.*, 2007). For a given prediction interval, the reliability measure is the percentage of discharge observations that are captured by the prediction interval. Reliability values are calculated by counting the number of times the observed streamflow falls within the prediction band, divided by the length of the time series. Sharpness is a measure of the prediction intervals' width relative to the hydrograph prediction bounds obtained from sampling prior feasible parameter ranges. If the posterior prediction bounds for the hydrograph form a single line, sharpness would be 100%. Whereas when the posterior prediction bounds are the same as those obtained using priori feasible parameter ranges, sharpness would be 0% (clearly undesirable). Ideally, and for a given prediction interval, the reliability should be equal to the desired interval percentage (*e.g.,* 90% of observations should be captured by a 90% prediction interval) and larger values of the corresponding sharpness measure are better than smaller values.

The Bayesian posterior predictive p-value is another measure of the predictive capacity of uncertainty-based calibration techniques (Gelman *et al.*, 2004, pp. 162–163). The Bayesian p-value is the probability that the model prediction at a particular time step could be more extreme than the observed data at that same time step. Such values may be estimated by the proportion of simulations for which the simulated value equals or exceeds the observed value. Probability distributions of p-values can be constructed from the complete series of p-value calculations. If the model output and measured data are consistent, the corresponding p-value distribution should be uniformly distributed over the interval [0,1]. This can be checked graphically using QQ-plots (Laio and Tamea, 2007; Thyer *et al.*, 2009) and deviations from the bisector (the 1:1 line) denote interpretable deficiencies (see Figure 1).

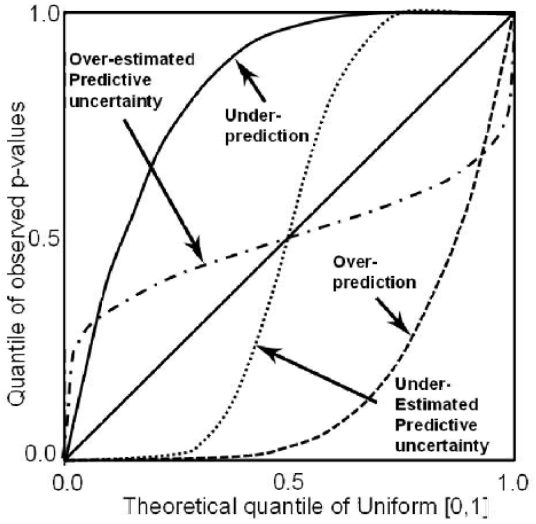

Figure 1. Schematic of the predictive QQ plot based on Thyer *et al.* (2009)

The approach of this study to compute comparative performance metrics with GLUE such as reliability, sharpness and Bayesian p-values is consistent with studies computing one or more of these metrics for GLUE results based on a pseudo-likelihood function such as Vrugt *et al.* (2008a), Yang *et al.* (2008) and Jin *et al.* (2010)

3.2.5 Case Studies

Bayesian inference is expected to result in robust expression of predictive uncertainty, as long as all assumptions are satisfied and the posterior PDFs are taken from a converged MCMC sampler. Two case-studies involving real data from two catchments are used in this study, for which the DREAM sampler is run to convergence to extract formal posterior distributions. The non-converged MCMC sampling and GLUE methods are also applied to the same problems.

One case-study applies the HYMOD hydrologic model to the Leaf River catchment. HYMOD is applied to the 1994 km^2 Leaf River watershed located north of Collins, Mississippi. This catchment has been studied intensively in the past (*e.g.,* Boyle, 2000; Sorooshian *et al.*, 1993; Thiemann *et al.*, 2001; Vrugt *et al.*, 2003b; Vrugt *et al.*, 2008a) and may be considered a standard benchmark for parameter estimation of hydrological models. Three years (1953-1955) of hydrologic data (mean areal precipitation [mm/d], potential evapotranspiration [mm/d], and streamflow [m^3/s]) are used, the first two years for model calibration and the third year as validation period. HYMOD is a relatively simple rainfall excess model (Moore, 1985) connected with a series of linear reservoirs. HYMOD requires estimation of five parameters (CMAX, B_{EXP}, ALPHA, R_Q, and R_S). Details about these parameters as well as their prior

range are provided in literature (*e.g.,* Boyle, 2000; Vrugt *et al.*, 2003b), and in Appendix D of this thesis, which provides more details on HYMOD and its application to the Leaf watershed.

The second case applied WetSpa rainfall-runoff model to the 1,131 km^2 Hornad River catchment located in Slovakia. The observations for this catchment are collected from 1991 to 2000. The first five years (1991 to 1995) are used for model calibration and the remaining data (1996 to 2000) for validation. WetSpa is a grid-based hydrologic model that simulates water and energy transfer between soil, plants and the atmosphere. WetSpa can be configured to run in semi-distributed or fully-distributed mode of which the former is chosen for this study. According to previous applications of WetSpa model to Hornad catchment (Bahremand *et al.*, 2007; Liu *et al.*, 2003; Shafii and Smedt, 2009), 11 WetSpa parameters are adjusted in the calibration process (Shafii and Smedt, 2009). More details about the structure of the WetSpa model, its parameters, and its application to the Hornad catchment is provided in Appendix D.

The multi-criteria formulation used in this study is created by splitting a single time series of responses (*e.g.,* discharges) into high- and low-flows. Following Schaefli *et al.* (2007), high-flows correspond to time steps in which the hydrograph is rising, and low-flows are defined based on the recession part of hydrograph. Separate Nash-Sutcliffe values (or formal likelihood values, in the case of MCMC sampling) are then calculated for each flow regime, yielding a bi-criteria calibration problem. Note that the computational overhead required for GLUE and DREAM are both dominated by the simulation model run time and as such, for the same number of model simulations completed, GLUE and DREAM require approximately the same computation time. The simulation model run time for HYMOD and WetSpa are 0.65 and 2.25 seconds, respectively, on a PC with 3-GHz Intel processor.

3.3 Results

For each of the case studies, the DREAM sampler is first applied to establish a converged chain of samples, and the non-converged DREAM and GLUE are then applied. Note that, as mentioned earlier, an AR-based Bayesian formulation is used in this study. Transformation and/or scaling of parameters is an important factor that can affect the difficulty of parameter estimation (Bates and Watts, 1981; Johnston and Pilgrim, 1976; Kuczera, 1983) and the convergence behaviour of MCMC samplers (Hills and Smith, 1992). For the HYMOD Leaf River and WetSpa Hornad River case studies, a series of preliminary numerical experiments are performed to explore alternative parameter transformations within the DREAM sampler. These experiments indicate that the most suitable transformations are to logarithmically transform HYMOD and WetSpa model parameters and use un-transformed auto-regressive parameters. It should also be noted that, in the formal Bayesian approach, discharges are also transformed logarithmically to stabilize the error variance.

3.3.1 HYMOD

When applied to the HYMOD Leaf River case study, the DREAM sampler converged after approximately 143000 simulations. The convergence of MCMC sampler is checked using the Gelman-Rubin convergence metric, which is also cross-checked to verify residuals normality (via inspection of a QQ-plot) and non-correlation (via inspection of the auto-correlation function). Appendix C shows the post-diagnostic checks evaluated to make sure that MCMC has converged in the HYMOD cases study. Furthermore, 1000 out of the last 10000 post-convergence samples are taken from the DREAM chain and used to derive baseline posterior parameter distributions. For the non-converged DREAM approach, a new trial of DREAM is considered up to 10000 simulations of which the last 1000 samples are used to derive corresponding posterior distributions. The GLUE method is applied using the generalized likelihood function (Equations (15-16) in Chapter 2) considering two scenarios, (i) a budget of 10000 simulations called 'GLUE Low-budget', and (ii) identical computation budget to DREAM (*i.e.,* 143000 simulations in HYMOD case study) and is called 'GLUE Full-budget'.

Figure 2 illustrates the posterior parameter information derived by the various calibration methods when applied to the HYMOD Leaf River case study. As observed in Figure 2 the posterior parameter ranges vary across methods, especially with respect to parameters R_s and R_q. Most of the ranges given by non-converged DREAM are wider than those given by converged DREAM. The difference between the location of posterior solutions derived from Bayesian inference and GLUE is not surprising, and can be explained by the fact that different likelihood functions have been used in these methods. However, comparison between these posterior ranges indicates that incorporating two additional error parameters (*i.e.,* higher complexity in comparison to informal formulation) results in a higher level of identifiability, especially for parameters R_s and R_q.

Figure 3 illustrates NSE values of the HYMOD Leaf River case study for calibration (upper panel) and validation (lower panel) period, demonstrating the results of DREAM (light points) versus non-converged DREAM and GLUE (dark points) along low and full computational budget. Conversion of DREAM likelihood values into equivalent NSE values is non-trivial because the fitted error series should also be accounted for. Proper conversion into equivalent NSE values must consider additional elements of the revised Bayesian formulation, namely, the two extra auto-regressive parameters (ρ_L and ρ_H) and the AR-based residuals term (δ_t). Thus, for a given parameter vector $\boldsymbol{\varphi}_i$ containing a model parameter set $\boldsymbol{\theta}_i$ and corresponding $\rho_{L,i}$ and $\rho_{H,i}$ auto-regressive parameters, the corresponding error variances are sampled to generate 100 different time series of error realizations. These errors are then combined with

simulated discharges and auto-regressive terms to yield 100 different NSE values for parameter vector $\boldsymbol{\varphi}_i$. The average of these NSE values is then used as the equivalent NSE value converted from the original DREAM likelihood value. Repeating this process for all parameter vectors contained in the DREAM posterior samples yields the equivalent NSE values plotted in Figure 3 for calibration and validation period.

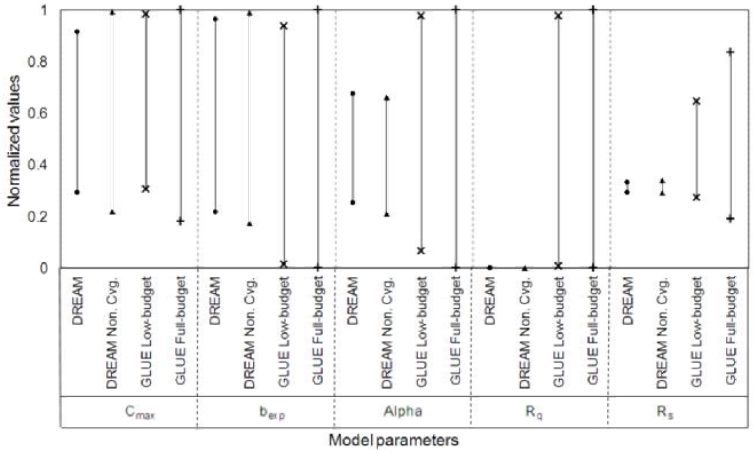

Figure 2. Posterior ranges of HYMOD parameters for the Leaf River case study; The parameter ranges correspond to 95% posterior intervals for different uncertainty-based calibration methods.

As shown in the calibration part in Figure 3 (upper panel), the results obtained from DREAM are superior (based on NSE values) to those given by other methods, and there is some overlap between the posterior sets of solutions given by converged and non-converged DREAM sampler. Note that these sets of solutions are sometimes called 'posterior clouds', as they look like a cloud in NSE space. In the validation part of Figure 3 (lower panel), the non-converged DREAM posterior cloud very closely resembles the DREAM posterior cloud. This is a good indication that much of the high-density areas of the parameter space are explored prior to the DREAM sampler satisfying the Gelman-Rubin convergence criteria. The results of GLUE in Figure 3 also indicate that regardless of the computation budget considered, the samples are located in fairly identical space in NSE space (but with different densities) both in calibration and validation period. However, GLUE with full computational budget performs slightly better considering extreme NSE values of GLUE in Figure 3. The GLUE results in calibration period showed that 8% of behavioural samples result in negative NSE values for low-flows, but since their NSE values for high-flows are high, they could rank in the top 10% of all GLUE samples. It should be pointed out that, similar to previous studies (Balin-Talamba, 2004; Lamb *et al.*, 1998), the threshold for classifying solutions as behavioural utilized the formulations in Equations (15-16) and did not take into consideration the condition of positive NSE values. This explains why there are some solutions with negative low-flows NSE values among posterior samples.

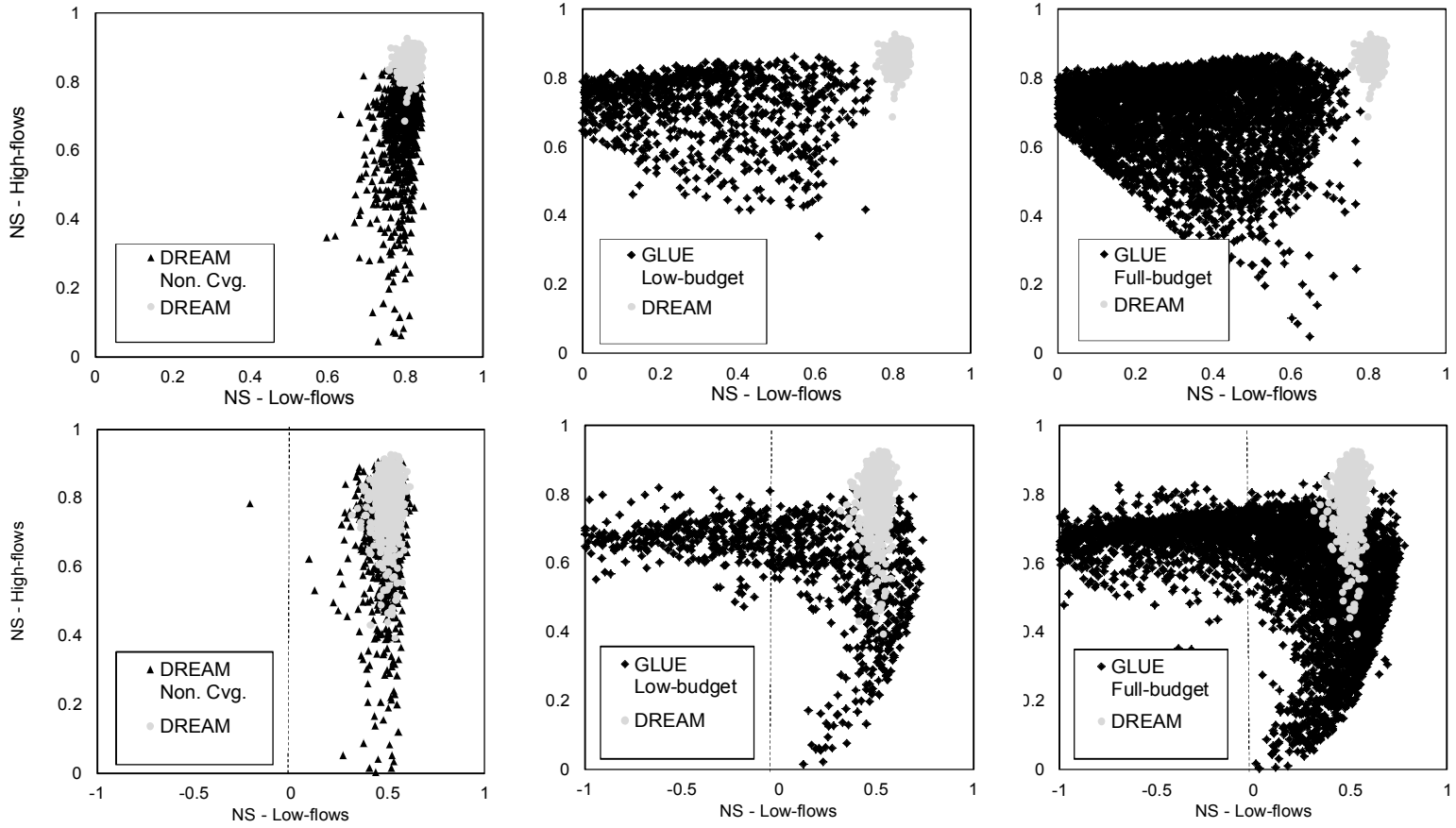

Figure 3. NSE values of low-flows (horizontal axis) and high-flows (vertical axis) in calibration (upper panels) and validation (lower panels) period for HYMOD case study, derived from DREAM (light points) versus non-converged DREAM and GLUE methods (dark points).

Figure 3 also shows that GLUE yields good performance in terms of matching the simulations with observation in validation low-flows, but not as good in high-flows compared to DREAM sampler. In contrast, the posterior cloud generated by DREAM in validation period (Figure 3 lower panel) emphasizes matching high-flows (points clustered in the 0.8 to 1.0 range for NS_{high}) at the expense of matching low-flows (points clustered around $NS_{low} = 0.5$).

Ideally, all posterior samples would generate positive NSE values in validation period for low- and high-flows. The vertical dashed lines in Figure 3 (lower panel) separates the region with positive NSE values for low-flows, and thus, the ideal region would be the right half of the scatter plots. It is observed that all posterior samples from DREAM and all but one of the non-converged DREAM posterior samples are located in this ideal region. However, almost 40% of posterior GLUE (full-budget) samples generate negative validation period NSE values for 'low-flows'. It should be pointed out that almost 92% of these samples had resulted in positive NSE values both for low- and high-flows in calibration period.

Figure 4 (left panels) illustrates the tradeoff between reliability and sharpness measures for the HYMOD Leaf River case study (only in validation period) for the various methods that are considered (*i.e.,* DREAM, non-converged DREAM, and GLUE with low and full computational budget). The reliability and sharpness values are calculated based on 95% prediction intervals on the corresponding posterior PDFs of simulated discharges. The reliability is calculated based the percentage of coverage of observations by prediction bounds, whereas sharpness is based on the amount of reduction in discharge ranges through comparison with the range of model simulations using prior parameter ranges. In order to define such prior intervals, 100000 Latin hypercube samples are taken from prior parameter ranges which are used in HYMOD to generate 100000 discharge hydrographs. The minimum and maximum of discharges at each time steps are then identified to serve as prior discharge ranges.

The HYMOD results in Figure 4 (left panel) show that the converged DREAM sampler and 'GLUE Full-budget' cannot dominate each other with respect to both reliability and sharpness. Compared to 'GLUE Full-budget', the converged DREAM results in improved sharpness both for low- and high-flows. In terms of reliability, as the goal is to generate 95% prediction intervals, both methods come fairly close to this goal given that reliabilities in validation period range from 93% to 97%. Comparison between non-converged DREAM and 'GLUE Low-budget' shows that neither of these two methods is superior to the other one with respect to both reliability and sharpness. The reliabilities of these two methods are close to 95%. The sharpness of non-converged DREAM is larger than 'GLUE Low-budget' in low-flows, and approximately the same in high-flows.

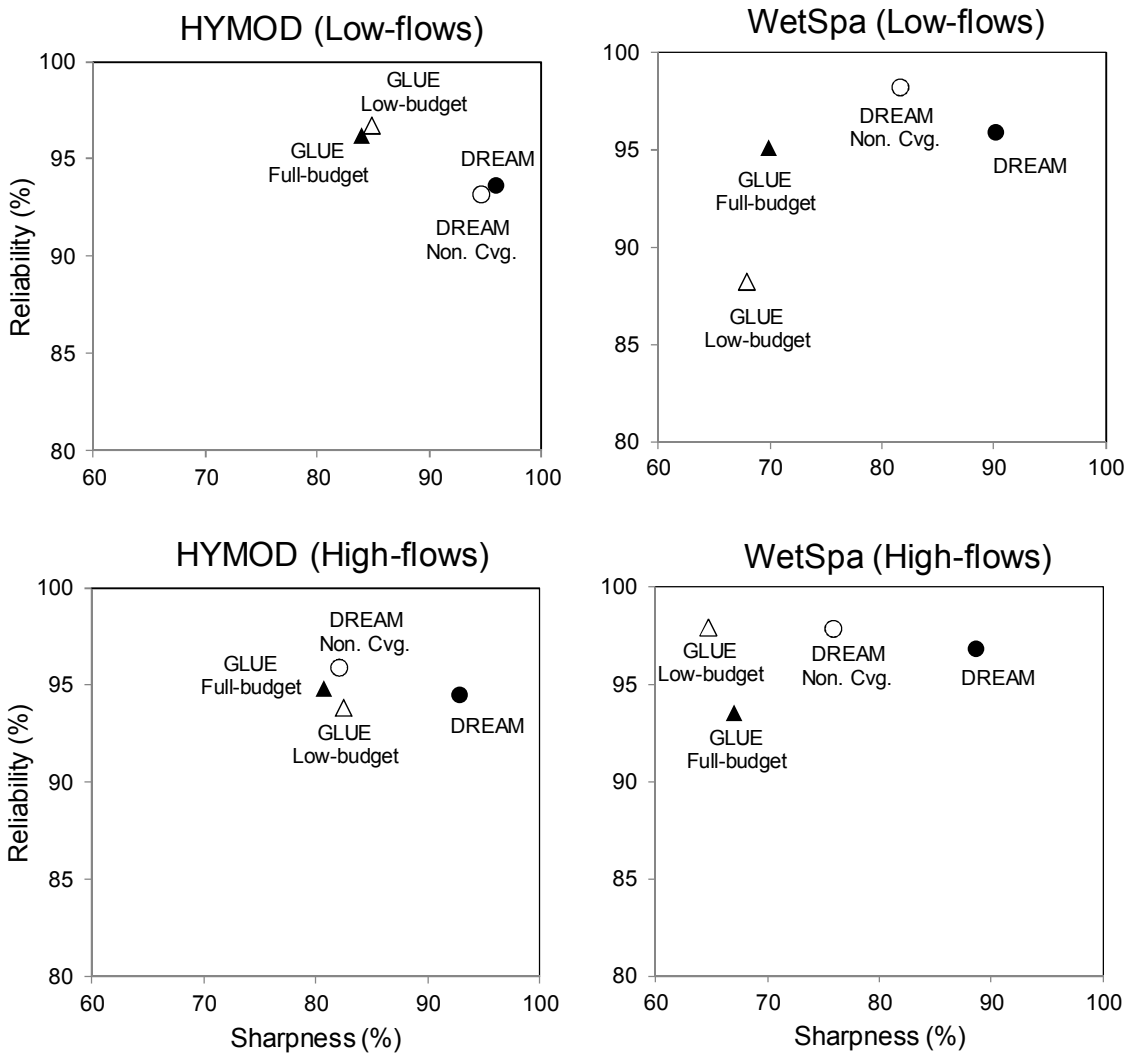

Figure 4. Validation period reliability and sharpness for low-flows (upper panels) and high-flows (lower panels) in application of different techniques (shown in different shapes) to the HYMOD and WetSpa simulation models.

Figure 5 contains Bayesian p-values for both the calibration and validation periods of the HYMOD Leaf River case study for non-converged and converged DREAM approaches. Note that the p-values are derived using the entire set of posterior solutions. Figure 5 shows that even though the p-value results for the converged and non-converged DREAM sampler are different during the calibration period, the results in validation period, however, are fairly similar. Also, both methods yield underestimation of predictive uncertainty with respect to low-flows in validation period. This might be due to the fact that standard Bayesian formulation is used without disaggregation of different sources of uncertainty, which will be discussed later in the discussion section.

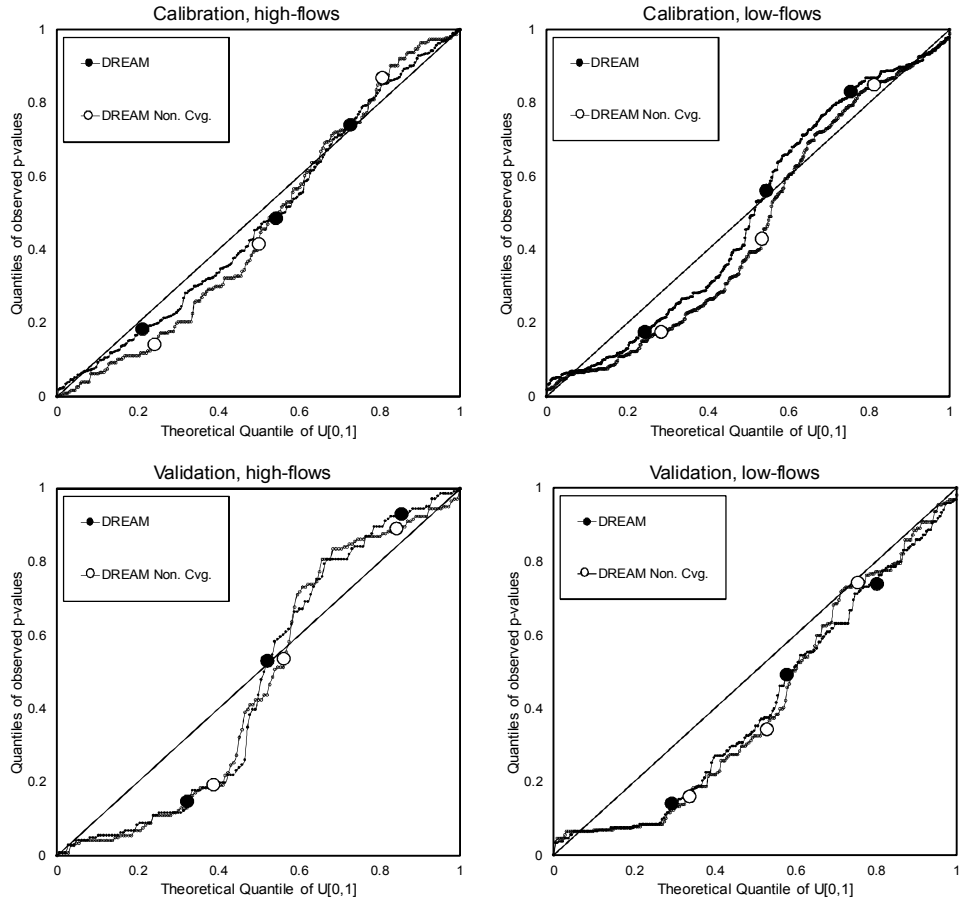

Figure 5. QQ plot of Bayesian p-values for high- and low-flows derived from converged and non-converged DREAM, for calibration (upper) and validation (bottom) periods of the HYMOD Leaf River case study.

Figure 6 illustrates the prediction bounds given by the posterior simulations of the considered calibration techniques for the validation period in HYMOD case study. The bounds shown in Figure 6 are derived in a manner similar to those given for posterior parameters of Figure 2 and are assumed to represent 95% prediction intervals. As shown in Figure 6, the converged DREAM sampler reliably covers the validation dataset. Prediction bounds of the non-converged DREAM sampler resemble those generated from the converged DREAM sampler but at the cost of larger width and larger peak flow values. Figure 6 also shows that the prediction bounds associated with 'GLUE Full-budget' are larger than those derived with 'GLUE Low-budget', but covered the observations better.

Across the various comparative measures that are evaluated in the context of the HYMOD Leaf River case study, it is observed that the formal Bayesian method (both converged and non-converged MCMC sampling) turn out to be more appropriate than informal GLUE strategy in calibration period. Once the validation period is used to evaluate the methods, the formal Bayesian inference (given the formulation of this study) results in a level of underestimation of predictive uncertainty, which would be probably solved

through more complex HBS systems, as elaborated in the discussions section. On the other hand, the GLUE methodology is only successful in partially meeting the predictive criteria in validation period. The WetSpa Hornad River real case study (next section) investigates whether these findings would hold for a more complex hydrological model (involving more uncertain parameters) applied to a different catchment.

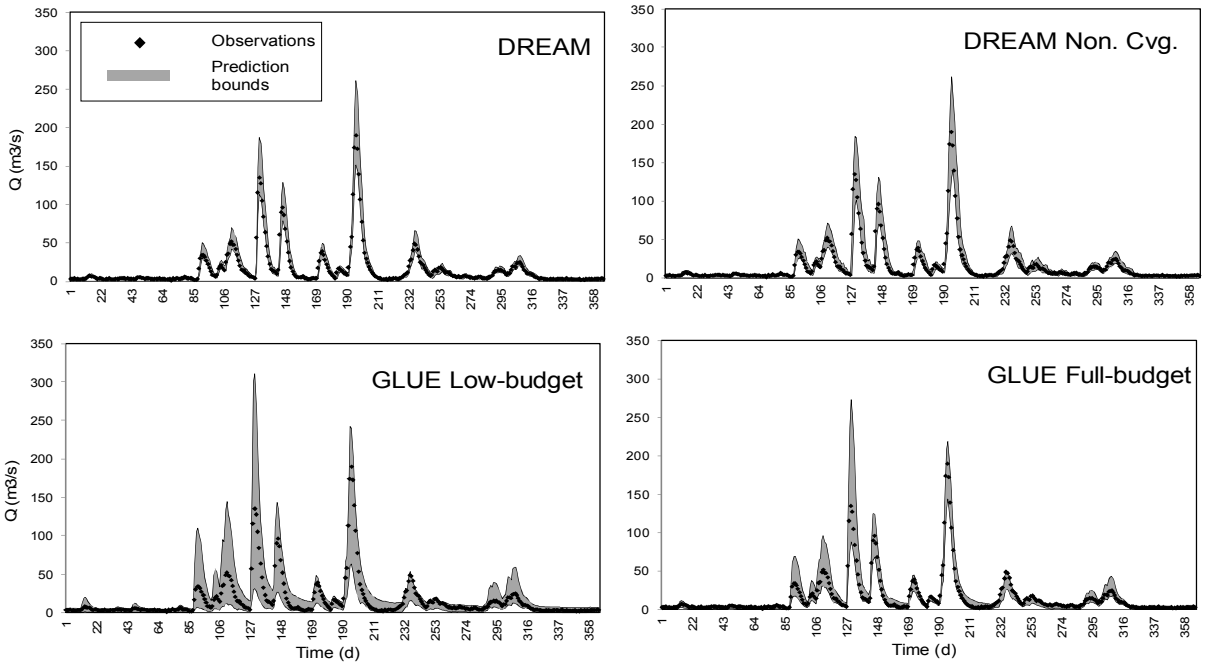

Figure 6. 95% uncertainty bounds and observations for validation period in the HYMOD case study.

3.3.2 WetSpa

For the WetSpa case study (application to Hornad River catchment), the DREAM sampler is again configured to use a formal auto-regressive Bayesian inference formulation and the method converged (based on the Gelman-Rubin statistic) after 470,000 simulations. Appendix C shows the post-diagnostic checks evaluated to make sure that MCMC has converged in the WetSpa cases study. These checks include test for normality, constant variance, and auto-correlation among residuals. As with the HYMOD studies, 10000 post-convergence DREAM samples are taken to construct the Bayesian posterior distributions. Similar to the previous case, the results of non-converged DREAM are derived based on running DREAM only up to 10000 simulations (independent trial than converged DREAM). GLUE is also applied to the WetSpa case study using low and full computational budget as described in HYMOD Leaf River case study.

Figure 7 contains normalized posterior ranges of the WetSpa model parameters generated by the various calibration methods. The first result noted in Figure 7 is that some parameters are deemed non-

identifiable (K_s, K_{GI}, and K_{RD}) by the converged DREAM sampler, as indicated by 95% posterior intervals covering almost the entire prior range. When informal likelihood functions are used (*i.e.,* GLUE), most of parameters appeared to be poorly-identifiable. However, it should be noted that the difference between the location of posterior parameter ranges and identifiability levels obtained by formal and informal methods would be explained by the difference in the likelihood functions used in these methods. It is also observed in Figure 7 that the posterior parameter ranges derived from non-converged DREAM covered those obtained from converged DREAM, and this shows how the sampler located a smaller posterior region after it converged.

Figure 8 illustrates the NSE values for calibration (upper panel) and validation (lower panel) period of the WetSpa Hornad River case study as evaluated by non-converged DREAM and GLUE (dark points), in comparison to those calculated based on the posterior solutions of the converged DREAM sampler (light points). Note that two cases are reported for GLUE, one with low and one with full computational budget. Also note that the axes in lower panel of Figure 8 are centred between ±1, the dashed lines showing the origin where both NSE values are zero. A number of GLUE solutions are not within this range and are not depicted in Figure 8. The ideal region for a given calibration method to sample from would be the upper right quadrant of validation panel where both low- and high-flow NSE values are positive. It is observed in the calibration panel that DREAM yields the best NSE values both for low- and high-flows. Given that non-converged DREAM and DREAM achieve these high NSE values, it seems the inclusion of an error term is important to achieve such high performance. The posterior cloud from non-converged DREAM overlaps substantially the converged DREAM posterior cloud, which indicates that the posterior distribution has likely been sampled from well before the Gelman-Rubin statistic indicated convergence. The results of GLUE (low and full computational budgets) also indicate that increasing the number of simulations in GLUE does not result in comparable model performance as DREAM (see distance between the location of posterior clouds). It is also observed in GLUE results (both low and full budgets) that there are a considerable number of points not located in the ideal region, that is, positive NSE values for low and high-flows or the upper right quadrant identified by dashed lines, even though they were all behavioural in the calibration period.

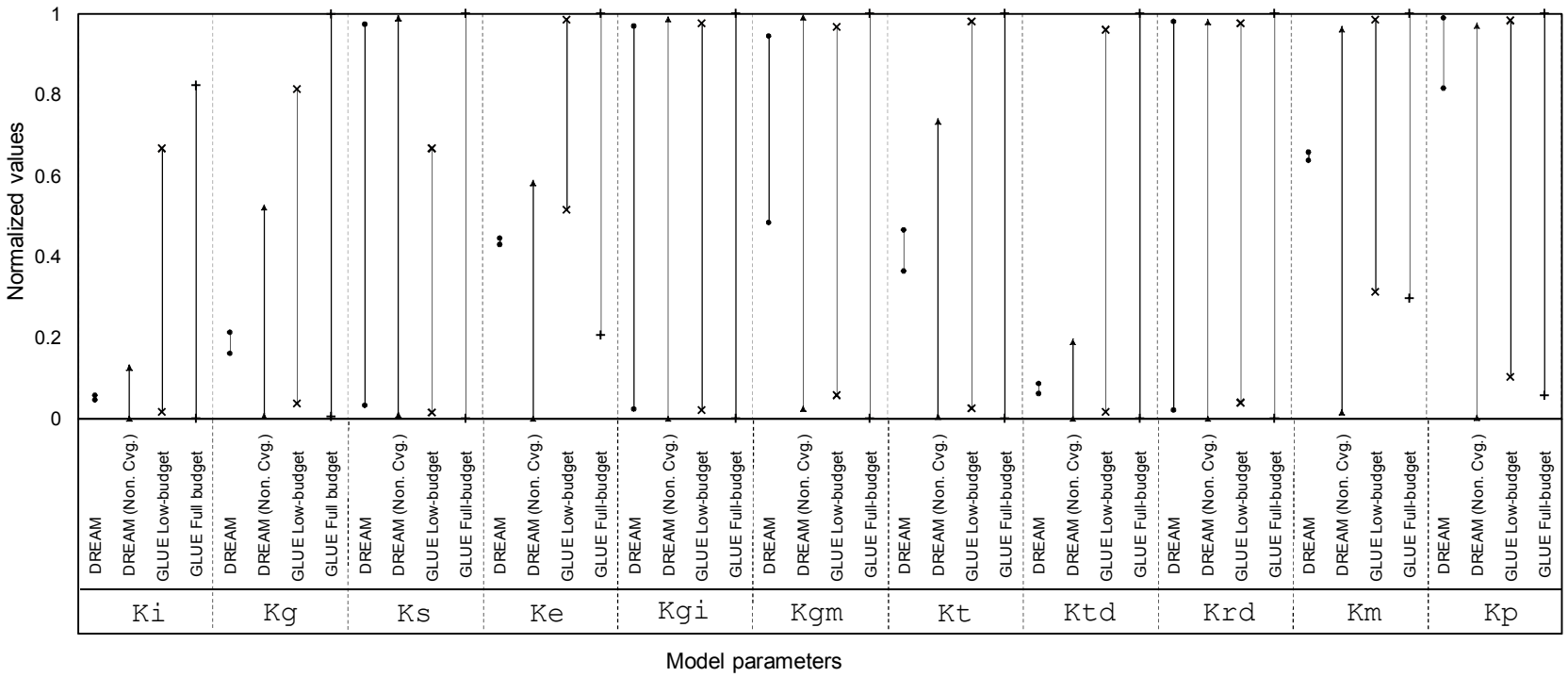

Figure 7. Posterior ranges of WetSpa parameters derived by different uncertainty-based calibration techniques.

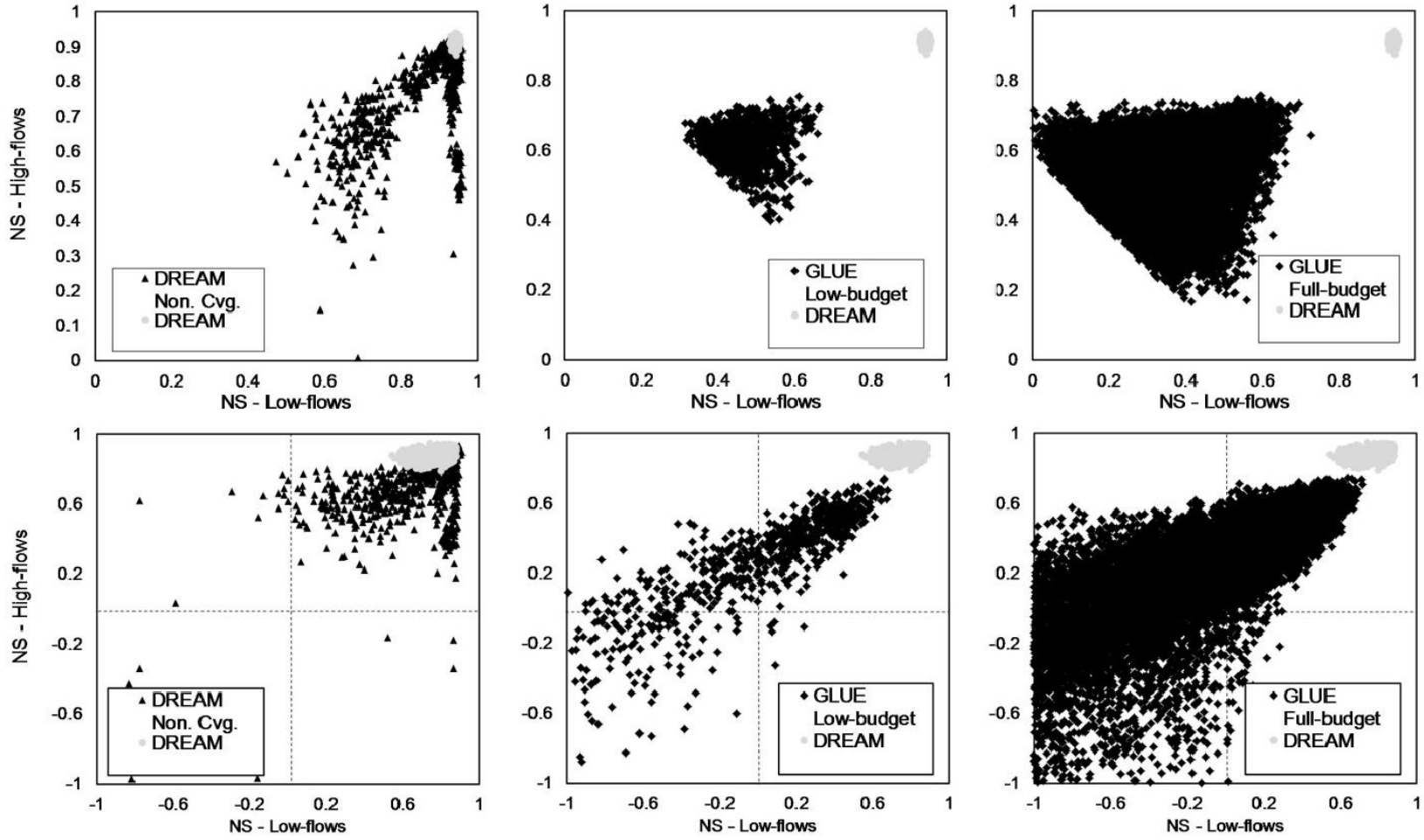

Figure 8. NSE values of low-flows (horizontal axis) and high-flows (vertical axis) in calibration (upper panel) and validation (lower panel) period for WetSpa case study, derived from DREAM (light points) versus non-converged DREAM and GLUE methods (dark points)

The sharpness and reliability measures for the validation period of the WetSpa Hornad River case study are given in Figure 4 (right panel). These measures are computed in the same manner as those for the HYMOD Leaf River case study. In terms of reliability, as the goal is to generate 95% prediction intervals, all methods come fairly close to this goal for high flows given that reliabilities in validation period ranged from 94% to 98%. The same is true for validation period low flows except that 'GLUE Low-budget' results have a slightly lower reliability of 88%. Comparing converged DREAM with 'GLUE Full-budget', it is observed that DREAM results dominate GLUE in both low-flows and high flows (*i.e.,* larger reliability and larger sharpness). In other words, DREAM generates tighter 95% prediction intervals and simultaneously improves reliability. Similarly, non-converged DREAM dominates 'GLUE Low-budget' results in high flows and practically dominates 'GLUE Low-budget' results in low flows (very similar reliabilities but significantly improved sharpness for DREAM).

Figure 9 compares the Bayesian p-value QQ plots for non-converged and converged DREAM sampling for the calibration (upper panel) and validation periods (lower panel) of the WetSpa Hornad River case study. As implied by the sigmoid shapes of their respective p-value curves, both DREAM samplers (converged and non-converged) exhibit systematic under-estimation of uncertainty for low-flows in validation period, even though the results of converged DREAM in calibration period are promising both for low-flows and high-flows. This finding is similar to results in Thyer *et al.* (2009) and the previous HYMOD case study. The under-estimation of only low-flow uncertainty by the converged DREAM procedure can be considered as indication of model structural error. This suggests that improving the low-flow modules in WetSpa may be a worthwhile enterprise. Such insight highlights the usefulness of multi-criteria Bayesian p-value separation as a post-diagnostic measure for detecting model structural deficiencies. However, it is also possible that the above-mentioned issue may be due to mis-specification of likelihood function.

Figure 10 illustrates the prediction bounds given by the posterior simulations of the considered calibration techniques for one year (1999) of the 5-year validation period (whereas Figure 4 reliability and sharpness values summarize prediction bounds over the entire 5-year period). The bounds shown in Figure 10 are derived in a manner similar to those given for posterior parameters of Figure 7 and are assumed to represent 95% prediction intervals. As shown in Figure 10, the converged DREAM sampler reliably covers the validation dataset even though the Bayesian p-value analysis indicates that the results are not perfect with respect to low-flows. Prediction bounds of the non-converged DREAM sampler resemble those generated from the converged DREAM sampler but at the cost of larger width and larger peak flow values. The prediction bounds associated with 'GLUE Full-budget' are larger than those derived with 'GLUE Low-budget', but covered the observations better.

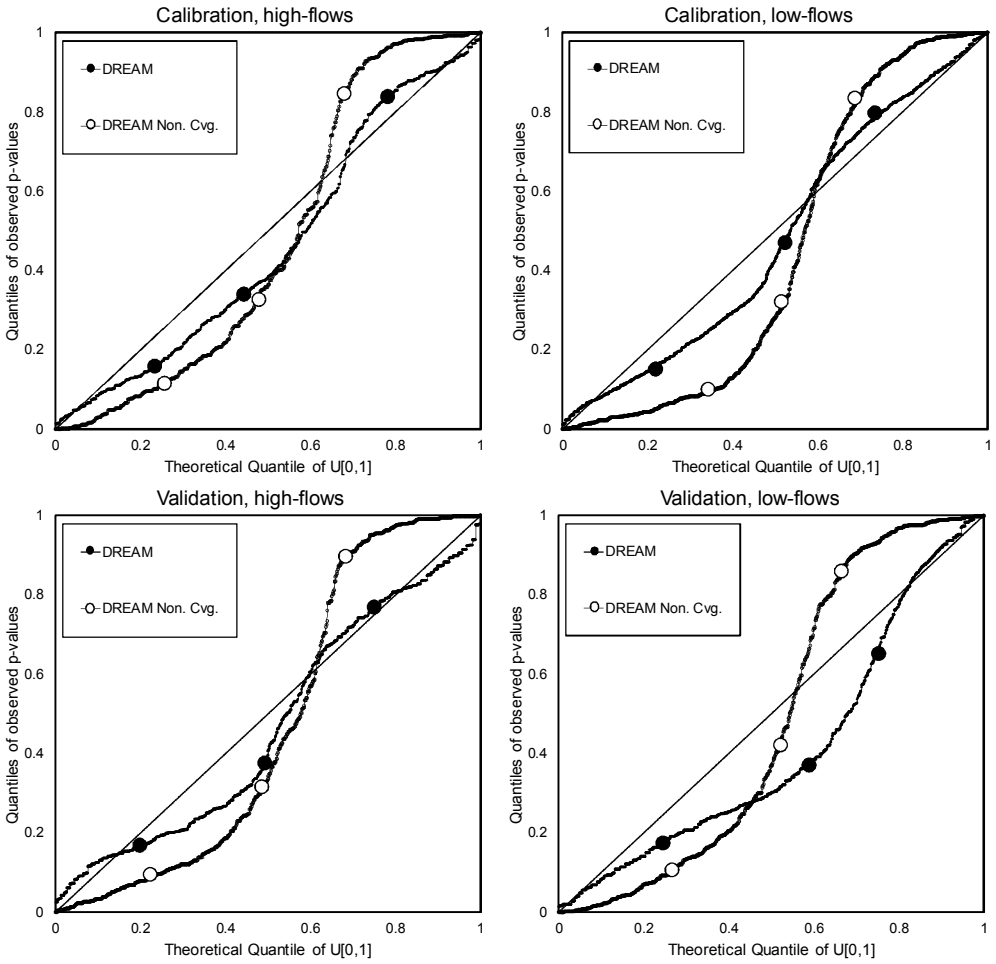

Figure 9. QQ plot of Bayesian p-values for high- and low-flows derived from converged and non-converged DREAM, for the calibration (upper) and validation (bottom) periods of the WetSpa Hornad River case study.

Across all comparative measures, the results of the WetSpa case study suggest the following conclusions: (1) the formal Bayesian inference through the standard formulation of this paper using converged DREAM yields good results with respect to almost all predictive measures, except for p-values of low-flows in validation period; (2) the non-converged DREAM sampler yields results that are nearly universally consistent with the converged DREAM sampler while requiring a fraction (2%) of the computational budget; and (3) considering the predictive measures addressed in this study, GLUE does not meet all measures as satisfactorily as formal DREAM methodology, even when the full computational budget is considered.

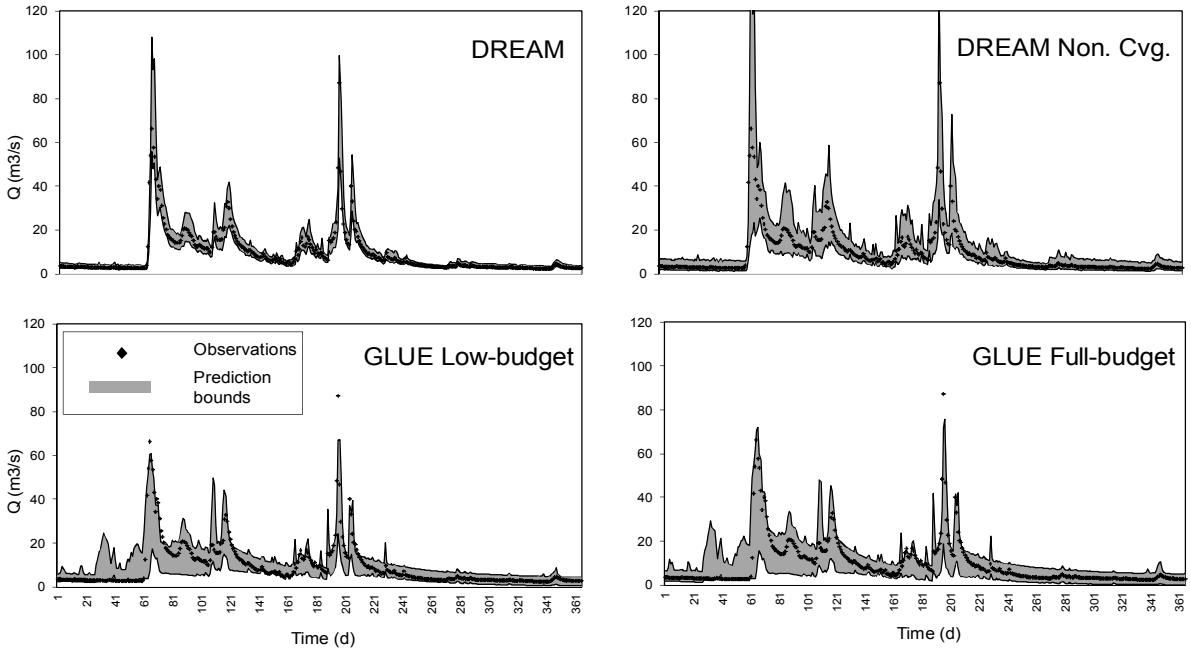

Figure 10. Prediction bounds and observations for the year 1999 of validation period for the WetSpa Hornad River case study.

3.4 Discussion

The DREAM results suggest that the Gelman-Rubin convergence criterion is too stringent since non-converged DREAM results closely approximates converged DREAM results and yet requires a fraction of the computational budget. It may also be possible to further improve the results of the non-converged DREAM sampler (*i.e.,* make it more closely approximate the converged DREAM results) by filtering out obviously low quality solutions for the calibration period (those with NSE values smaller than 0.5 in upper left panels of Figures 3 and 8). Moreover, one might think of applying alternative convergence measures. A potential hydrology-based convergence metric can be the reproduction of hydrological signatures that represent the overall hydrologic behaviour of the catchment (Gupta *et al.*, 2008; Yilmaz *et al.*, 2008). Future research should explore these and other alternative convergence measures in a multi-criteria context.

Comparison between formal and informal methods could also be viewed from the standpoint of aleatory and epistemic uncertainties recently discussed among researchers (Beven *et al.*, 2012b; Beven *et al.*, 2011; Clark *et al.*, 2012a; Montanari, 2011). Uncertainties are categorized as aleatory (also called natural uncertainty) if they are presumed to be the intrinsic randomness of a stochastic process which can be represented in terms of the probabilities of different outcomes. On the other hand, many of the errors that enter into the modelling process stem from a lack of knowledge about processes and boundary

conditions. These errors are called epistemic or limited-knowledge uncertainty. In statistical models (including Bayesian inference structures), uncertainties are accounted for by providing a representation of all of the important sources of uncertainty as aleatory (Beven *et al.*, 2011), even though in reality uncertainties could be a mixture of both aleatory and epistemic. The results reveal that validation period performance measures are generally poorer compared to calibration period which is expected to be caused by epistemic errors (Beven *et al.*, 2011). Thus, in the presence of epistemic errors, neither the standard Bayesian formulation nor the informal methods (such as GLUE) would be perfectly reliable in prediction mode. There are improved informal and formal approaches for case studies where epistemic errors are thought to be significant, *e.g.,* the use of hierarchical Bayesian structures (*e.g.,* Huard, 2008; Kuczera *et al.*, 2006; Moradkhani *et al.*, 2005; Renard *et al.*, 2010; Wei *et al.*, 2010), or the concept of 'limits of acceptability' used for identifying behavioural models in GLUE (Blazkova and Beven, 2009; Liu *et al.*, 2009). Comparison between these two more advanced formal and informal uncertainty-based calibration methods is an interesting future research avenue.

3.5 Concluding Remarks

This research evaluates the applicability of formal (Bayesian inference) and informal (GLUE) multi-criteria methods to uncertainty-based calibration in hydrological modelling. Bayesian inference is implemented through DREAM sampling based on a multi-criteria formulation. The results of non-converged DREAM are also evaluated. The results are compared with those obtained from two scenarios for GLUE, using a restricted computational budget and the full computational budget equivalent to the budget required for DREAM sampler to converge. The various methods are applied to two cases involving the 5-parameter HYMOD model and the 11-parameter WetSpa model. Results demonstrate that there can be considerable differences in prediction intervals generated by formal and informal strategies for uncertainty-based calibration. Future uncertainty quantification studies for simulation models with a large number of parameters should be aware of the potential considerable difference between the results of formal and informal strategies.

Results also demonstrate that it is advisable to consider multiple comparative measures, including traditional metrics like the Nash-Sutcliffe efficiency, when comparing alternative calibration strategies. Furthermore, it is observed that the choice of using the validation period or the calibration period for selected comparative measures would influence the analysis and as such it is recommended that future uncertainty-based calibration method comparison studies should include and largely focus on comparative performance assessment for the validation period.

In general, the Bayesian inference methodology performs well (in comparison with other methods) along all comparative measures except for low-flows in validation period considering the same computational budget, *i.e.,* DREAM validation period prediction intervals are simultaneously tighter and more reliable than corresponding GLUE intervals. In case of limited computational budget (only 10000 simulations in this study), non-converged MCMC sampling using DREAM proves to be fairly consistent with formal Bayesian inference. This indicates the potential value of utilizing formal MCMC sampling results before convergence as a promising alternative to informal methods such as GLUE.

The results obtained through application of Bayesian inference to the two cases of this paper indicated under-estimation of predictive uncertainty for low-flows in the validation period. A standard Bayesian formulation is applied, which lumps all uncertainties into a single additive error term. More recently, Renard *et al.* (2010; 2011) have shown that consideration of rainfall and model structural uncertainties outside of the error term used in Bayesian formulation yields more reliable estimation of the predictive uncertainty for all runoff ranges, as opposed to the typical Bayesian formulation in our paper. Application of hierarchical Bayesian structures to the case studies of this paper is currently being investigated.

There are many ways to formulate and conduct GLUE analyses, and to some extent DREAM calibration experiments. Our experiments require a number of subjective decisions and as such our results are conditional on these decisions. However, it is believed that the subjective decisions made here are consistent with the decisions others have made in the literature. For example, although it is possible to apply GLUE using a formal likelihood function (Romanowicz *et al.*, 1994), the literature suggests that it is relatively uncommon, and thus, that approach is not examined here. GLUE is used with an informal generalized likelihood function in this study because the objective of the study was to assess its performance as an informal method. It may be possible that applying informal methods such as GLUE using formal likelihood functions would improve their performance, but this is not the focus of this research. Future comparative studies systematically varying such subjective decisions would be valuable.

Chapter 4
Addressing Subjective Decision-Making Inherent in GLUE-based Multi-Criteria Rainfall-Runoff Model Calibration

Summary

Informal uncertainty estimation methodologies are among the simple and practical approaches to address uncertainties in hydrological modelling. Chapter 2 describes a number of informal methods commonly used in the hydrological literature. In Chapter 3, GLUE as a candidate informal method is compared with formal Bayesian inference in the multi-criteria context to explore the applicability of this informal approach from the uncertainty quantification perspective. This chapter evaluates the impact of subjective decisions involved in GLUE on the uncertainty bounds of the model output. Different behavioural solutions identification strategies are assessed, and moreover, two new objective strategies of identifying behavioural solutions are developed. GLUE is applied to multi-criteria calibration and uncertainty estimation of three hydrologic models with different number of parameters. To demonstrate the impact of GLUE's subjectivity in behavioural solution identification on the model output uncertainty, GLUE is also compared with two other informal multi-criteria calibration and uncertainty estimation methods (Pareto optimization using the AMALGAM algorithm and DDS Approximation of Uncertainty or DDS-AU) in multiple numerical experiments. Results indicate that, for a robust implementation of GLUE, the modeller needs to consider multiple behavioural solution identification strategies, compare them, and select the most appropriate strategy. Moreover, the proposed objective identification approaches prove to be among the selected approaches in most of the case studies. Section 4.1 provides a brief introduction, Section 4.2 elaborates on the material and methods, results are illustrated in Section 4.3, and finally, Section 4.4 highlights the concluding remarks drawn in this study. This chapter is a mirror of the following submitted article (first round of review comments was received and modest changes were applied) with minor changes to increase its consistency with the body of the thesis and to avoid redundant material. Changes are only made in the Summary (abstract), Section 4.1, and Section 4.2. Moreover, references are unified at the end of the thesis.

Shafii, M., Tolson, B., Matott, L.S., (2014), Addressing Subjective Decision-Making Inherent in GLUE-based Multi-Criteria Rainfall-Runoff Model Calibration, *Journal of Hydrology*, In Review.

4.1 Introduction

The GLUE uncertainty estimation is an extension to Monte Carlo Simulation (MCS), and belongs to the family of importance sampling algorithms. GLUE is based on the concept of 'equifinality' and classifies Monte Carlo samples as having produced model output that is either 'behavioural' (*i.e.,* plausible, given the data and one's knowledge of the system) or 'non-behavioural'. As such, GLUE's implementation comes at the cost of some subjective decisions such as the use of informal likelihood measures and the threshold for filtering behavioural solutions. Previous GLUE studies have reported on the impact of varying the likelihood function definition on the model output uncertainty (Beven and Freer, 2001; Freer *et al.*, 1996) and on the impact of the behavioural threshold on prediction intervals (Montanari, 2005).

The present study is designed to evaluate the impact of varying the behavioural solution identification strategy (with a constant likelihood function definition) on the uncertainty bounds of the model outcome and two new identification strategies are proposed. To better demonstrate the impact of behavioural solution identification strategies in GLUE on the model output uncertainty, this study also compares GLUE with two other informal uncertainty estimation methodologies. The main focus is on multi-criteria calibration cases, but the concepts addressed in this study can be also applied to single-criterion calibration cases.

4.2 Material and Methods

The problem addressed in this study is multi-criteria calibration and uncertainty estimation in rainfall-runoff modelling using the GLUE methodology. A typical multi-criteria model calibration process can involve multiple likelihood functions used for different sets of measurements, *e.g.,* discharge, sediment, snow water equivalent, etc. However, even in the case of a model with only one output flux to be simulated, the model evaluation may still be considered to be inherently multi-criteria (Gupta *et al.*, 1998). The multi-criteria numerical experiments in this study consider both cases (explained in the case studies section).

4.2.1 Classical Strategies of Finding Behavioural Solutions

Section 2.4.2.4 describes different possible strategies of finding behavioural solutions in multi-criteria applications of GLUE. The first strategy considered in this study is to consider the top 10% of the population as behavioural solutions after aggregating the criteria. The second strategy is based on NSE values; the first approach in this strategy classifies those parameter sets with positive NSE for both responses as behavioural, and in the second approach in this strategy, those parameter sets with NSE

greater than 0.5 for both responses are considered behavioural. Note that, to rescale the weight of each sample in the second strategy, sum of likelihood values is applied (more information in Section 2.4.2.4).

4.2.2 Objective Behavioural Solution Identification Strategies

To alleviate the subjectivity in GLUE, the literature has provided some solutions. Blasone *et al.* (2008b) suggest that, instead of using a pre-defined threshold to identify the behavioural set, the uncertainty bounds should be tuned so that they exhibit appropriate coverage of observations. However, the width of prediction intervals has not been considered in the study by Blasone *et al.* (2008b). More recently, the concept of 'limits of acceptability' has been suggested in Beven (2006) as an improvement to the standard GLUE approach, which has shown to be a promising way to objectively identify behavioural models (Blazkova and Beven, 2009; Liu *et al.*, 2009). This concept functions to use observation error to define limits of acceptability, and reject those models of which the predictions fall outside of such limits.

The behavioural solutions identification strategies in this study is based on simultaneous consideration of two metrics, reliability and sharpness defined in Yadav *et al.* (2007). The reliability measure is the percentage of discharge observations that are captured by the prediction interval throughout the calibration period. Reliability values are calculated by counting the number of times the observed streamflow falls within the prediction band, divided by the length of the time series. Sharpness is a measure of the prediction intervals' width relative to the hydrograph prediction bounds obtained from uniform sampling of prior feasible parameter ranges, *i.e.,* before calibration. If the posterior prediction bounds for the hydrograph form a single line, sharpness would be 100%. When the posterior prediction bounds are the same as those obtained using priori feasible parameter ranges, sharpness would be 0% (clearly undesirable). Thus, larger values of the corresponding sharpness measure are better than smaller values such that both reliability and sharpness are to be maximized.

The idea behind the proposed strategy (*i.e.,* simultaneous consideration of reliability and sharpness for identification of behavioural solutions) is the fact that the model output uncertainty spread should be ideally as small as possible and consistent with observations so that the PDF of model output is as sharp as possible (Gneiting *et al.*, 2007). Thus, reliability and sharpness need to be considered at the same time, which makes the process of identifying behavioural solutions a multi-criteria decision-making problem with reliability and sharpness being the two criteria. Note that the proposed approach of this study is a generalization of the coverage-based approach in Blasone *et al.* (2008b) to provide a more formal expression of their coverage concept. Furthermore, Blasone *et al.* (2008b) only consider reliability when selecting behavioural solutions, and not the width of the model output bounds. Such an approach may result in bounds that are too wide in cases where the simulation model is not conceptually consistent with

the real behaviour of the catchment, *i.e.,* difficult to cover low- and high-flows simultaneously. Furthermore, the multi-criteria calibration problem in Blasone *et al.* (2008b) is based on criteria aggregation. However, our study considers both Pareto domination-based (Section 4.2.2.1) and aggregation-based (Section 4.2.2.2) multi-criteria calibration.

4.2.2.1 Pareto Rank-Based Behavioural Solution Identification

The first behavioural solution identification strategy proposed in this study is to consider multi-criteria sorting of the parameter sets based on the concept of Pareto dominance, and select the behavioural solutions based on a Pareto rank-based threshold. This approach does not aggregate the likelihood measures. Instead, the concept of Pareto ranking (Goldberg, 1989) is used to assign a 'front rank' to each parameter set, and identify the optimal number of behavioural solutions based on such a ranking scheme. The proposed rank-based strategy is inspired by multi-criteria sensitivity analysis of Rosalem *et al.* (2012) who propose a rank-based fully multiple-criteria implementation of the Sobol′ variance-based sensitivity analysis approach (Sobol, 2001). To comprehensively explain the approach of this study, a hypothetical example of only two likelihood functions ($L1$ and $L2$) is explained for a GLUE experiment based on N parameter sets sampled from prior distributions. This approach can be certainly expanded to more than two likelihood functions. The proposed strategy consists of the following four steps:

Step 1: Assign Front Rank to Each Parameter Set

The N parameter sets are all sorted based on the concept of non-domination and assign a Pareto rank r to each parameter set (see Gupta *et al.*, 1998) on the basis of simultaneous maximization of two likelihood functions $L1$ and $L2$. Figure 11 illustrates a hypothetical case with a number of parameter sets which are sorted and assigned a rank. Note that the Pareto front solutions in Figure 11 would get rank 1 (also shown in Figure 11), and other parameter sets are assigned a rank between 2 and $r_{\max}$ based on non-domination sorting. Solutions having the same rank are shown linked by a dotted line in Figure 11, and in this figure, $r_{\max} = 5$. Note that, in the numerical experiments of this study, likelihood 1 and 2 are two NSE measures calculated either for low- and high-flows in some case studies or for discharge and sediment transport in other cases.

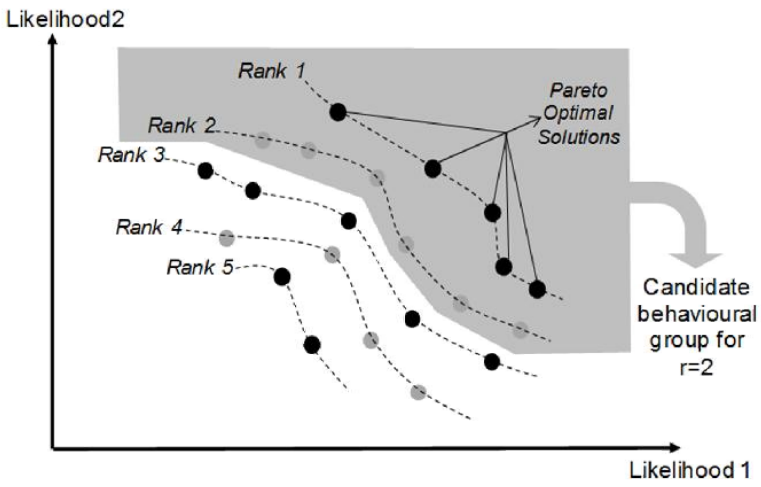

Figure 11. Solutions ranking and Pareto front solutions for a hypothetical bi-criteria example.

Step 2: Create Rank-to-r Candidate Behavioural Groups

The rank-to-r candidate behavioural group contains all parameter sets of which the assigned front rank is less than or equal to r (*i.e.,* front ranks 1, 2, …, r). Thus, when r increases, the number of parameter sets in the candidate behavioural group r increases, as well. In Figure 11, the parameter sets in the candidate behavioural group with $r = 2$ are shown in the shaded area.

Step 3: Create Distance-to-Ideal Values versus Front Rank Number

Once different candidate behavioural groups are identified, the sharpness (S_i in %) and reliability (R_i in %) of each group are calculated for each response i. Then, the distance between (S_i, R_i) and the ideal point $(100,100)$ or D_i is calculated for each likelihood function i. Afterwards, the Distance-To-Ideal (DTI) metric is calculated for each rank number (or candidate behavioural group) by Equation (18):

$$DTI_j = \sqrt{\sum_{i=1}^{M} D_i^2} \qquad j = 1,2,...,r_{\max} \qquad (18)$$

where M is the number of likelihood functions (2 in our case studies) and $r_{\max}$ is the maximum rank number identified after sorting N solutions based on non-domination concept. The outcome of this step is a vector of DTI values where the size of the vector being equal to $r_{\max}$.

Step 4: Select a Front Rank Threshold to Identify Behavioural Solutions

Once all of the parameter sets are analysed and DTI at each rank is calculated for candidate behavioural groups, the minimum value for DTI metric is identified and the corresponding front rank is selected (called r_{best}). Finally, the parameter sets contained in the candidate behavioural group corresponding to r_{best} are identified as behavioural. Obviously, the number of behavioural solutions which are objectively determined this way will be different for different case studies. Figure 12 schematically shows the structure of the Pareto rank-based approach described above.

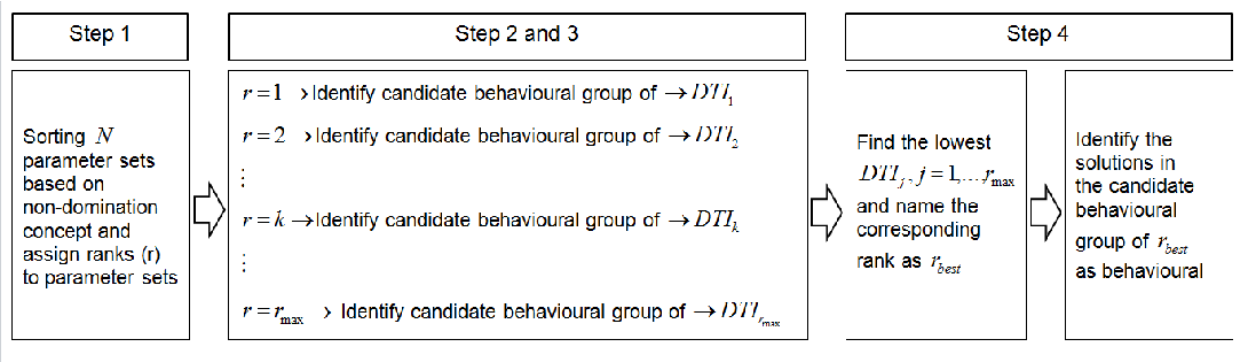

Figure 12. Schematic diagram of Pareto rank-based behavioural solution identification strategy.

In step 3 of GLUE (see Section 2.3.2.1) when each behavioural group is identified, a single weight should be assigned to each parameter set for the rescaling step of GLUE and extracting the uncertainty intervals of the model outcome (and the corresponding reliability and sharpness values). The weighting scheme should account for all informal likelihood functions involved, and at the same time, it needs to assign more weight to parameter sets with higher overall likelihood values. Note that all the solutions on a particular front rank will have equal importance from the multi-criteria sorting perspective. One weighting approach can be based on sum of performance criteria calculated for all responses. The sum of NSE is used as an example metric, which means, the parameter sets with higher values of aggregated NSE criterion will get a higher weight, and other parameter sets will be weighted accordingly.

4.2.2.2 Optimal criteria-aggregation-based behavioural threshold identification

As an alternative to the new approach described above, a second approach is proposed that employs the concept of minimizing DTI in order to identify behavioural solutions more objectively in the criteria aggregation-based schemes. The weighted criterion is sorted from best to worst. Starting with the best parameter set in the behavioural set, parameter sets are added one by one to behavioural set, and the resulting DTI values are calculated after adding each parameter set. After all possible behavioural sets

are analysed, the one with the minimum DTI value is selected to define the behavioural solutions. In this way, instead of a subjective threshold (for example the top 10%), it is possible to more objectively identify the optimal percentage from the possible parameter sets.

There are two main differences between the rank-based behavioural solution identification strategy and the aggregation-based approach. First, the rank-based strategy uses NSE criterion to filter behavioural solutions whereas the criteria aggregating strategy aggregates all criteria to do so. Moreover, the rank-based strategy is not conditioned on criteria aggregation for identifying behavioural solutions.

4.2.3 Comparison Framework

Each strategy for identifying behavioural solutions results in a set of solutions with particular values for reliability and sharpness. These values can be compared graphically (see Figure 13 for hypothetical example) for a comparison among the strategies. Figure 13(a) is called a Sharpness-Reliability ($SH-R$) graph and Figure 13(b) a DTI graph throughout this document. The horizontal axis in $SH-R$ graphs is sharpness (in percent) and the vertical axis is reliability (also in percent). Considering only the optimal parameter set (the parameter sets with highest performance) as behavioural, the output single hydrograph results in sharpness of 100% and reliability at or very close to 0% (point 'F' in Figure 13(a) called the sharpest point). A set of parameter sets uniformly sampled from the prior parameter ranges yields reliability very close to 100% at the cost of sharpness equal to 0% (point 'A' in Figure 13(a) called the most reliable point).

The results of each GLUE behavioural identification strategy falls somewhere on $SH-R$ graph between the sharpest and the most reliable points (see hypothetical points 'B', 'C', 'D', and 'E' in Figure 13(a)) and these points can be graphically compared. For example, point 'D' is better than (or dominates) point 'C' as it has better values for both reliability and sharpness. However, some points such as 'B', 'D' and 'E' cannot be clearly deemed better or worse than one another. One way of comparing such points (usually called non-dominated solutions in the multi-criteria context) is to evaluate the DTI metric (note that the ideal point in Figure 13(a) has coordinates of 100% and 100%). This distance metric is illustrated on DTI graph in Figure 13(b) for the hypothetical problem. The point 'D' has a lower DTI than points 'B' and 'E', so it is preferred. It is worth noting that there are other multi-criteria methods for decision making, and the DTI metric that is proposed in this study is only one of the possible approaches for making such decisions.

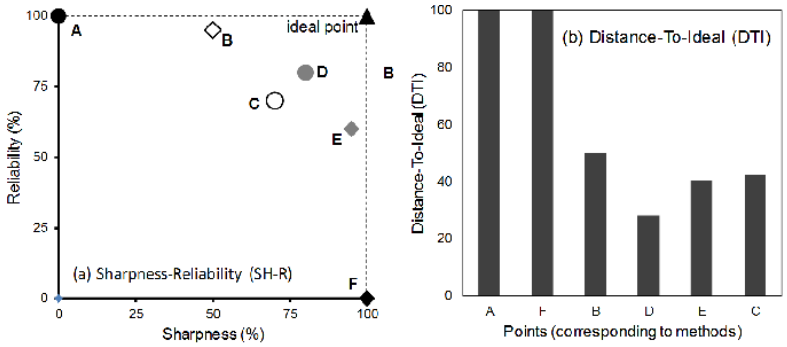

Figure 13.. SH-R (a) and DTI (b) graphs for a hypothetical calibration problem. Points A and F bound the results of informal calibration approaches where A is the most reliable point (parameter sets sampled from prior) and F is the sharpest point (only a single optimal point). Points B through E would correspond to alternative informal approaches.

In single-criterion calibration cases, the *DTI* metric can be calculated as explained above to account for two decision criteria (reliability and sharpness). In multi-criteria cases, however, the total number of decision criteria will be two times the number of responses. For instance, if two responses are involved (*e.g.,* discharge and sediment), there will be four decision criteria, namely, reliability and sharpness of both discharge and sediment. The *DTI* metric in such a case will be the Euclidean distance from the ideal point in four-dimensional space, which is Equation (18) with M=2.

Uncertainties in rainfall-runoff modelling reduce the usefulness of models, and therefore, it is desirable to develop systematic, flexible, and consistent methods to estimate the reliability of rainfall-runoff models under uncertainties (Melching *et al.*, 1990). Calibration is generally a process of progressive model identification aiming at sequentially reducing such uncertainties while constraining the model to be structurally and functionally consistent with the available information about the catchment (Gupta *et al.*, 2005a). Traditionally, once the calibration is performed, the identified model(s) are evaluated through the split-sample technique where models are tested outside of the calibration time period (often called validation period) to assure that models can make accurate predictions (Klemeš, 1986; Refsgaard and Knudsen, 1996; Tsang, 1991). This study also compares different informal calibration and uncertainty estimation techniques in the validation period.

4.2.3.1 Comparing GLUE with Other Informal Methods

In order to assess the impact of the subjectivity involved in GLUE on the uncertainty bounds of the model outcome, this study compares GLUE with two other informal approaches for multi-criteria calibration and uncertainty estimation in rainfall-runoff modelling. These methods are Pareto optimization (Gupta *et al.*, 1998) using the AMALGAM optimization algorithm (Vrugt and Robinson, 2007), and approximation of uncertainty using DDS or simply DDS-AU (Tolson and Shoemaker, 2008). Section 2.4.2.1 describes the

Pareto optimization uncertainty assessment method and the corresponding calibration formulation. This study applies AMALGAM (Vrugt and Robinson, 2007) for Pareto optimization, which is also described in Section 2.4.2.1. Interested readers may refer to Vrugt and Robinson (2007) for more details about the structure of AMALGAM and its parameters setting.

Section 2.4.2.2 elaborates on criteria-aggregation schemes for uncertainty estimation and the optimization formulation used in this method. DDS-AU implementation in this study belongs to this class of methods. DDS-AU is applied using the DDS optimization algorithm (Tolson and Shoemaker, 2007), which is a novel and simple stochastic single-solution based heuristic global search algorithm that is developed for the purpose of finding good global solutions (as opposed to globally optimal solutions) within a specified maximum function (or model) evaluation limit. The only algorithm parameter to set in the DDS algorithm is the scalar neighbourhood size perturbation parameter (r) that defines the random perturbation size standard deviation as a fraction of the decision variable range. A default value of the r parameter is recommended as 0.2 (and used in this study).

In GLUE, a traditional behavioural solution identification strategy is considered, as well as the proposed two new objective identification strategies (described in Section 4.2.2.1 and 4.2.2.2). Each method is applied to the three case studies along different computational budgets. When the envelope of model outputs associated with the behavioural solutions is derived, *i.e.,* uncertainty propagation (McIntyre *et al.*, 2002), reliability and sharpness are calculated over the validation period. (*e.g.,* Franz and Hogue, 2011; Vrugt *et al.*, 2008a; Yang *et al.*, 2008). Note that the concept of behavioural solutions holds in both Pareto optimization and DDS-AU methods. In the former approach, Pareto front parameter sets are behavioural, and in the latter, the best parameter set found in each DDS trial is behavioural provided the objective function value is equal or better than some subjective threshold.

4.2.4 Case Studies

Three case-studies are considered in this study. The first case, HYMOD rainfall-runoff model, and the second case study, WetSpa model, are explained in details in Section 3.2.5. In these models, the discharge time series is divided into low- and high-flows based on the method described in Schaefli *et al.* (2007). High-flows correspond to time steps in which the hydrograph is rising, and low-flows are defined based on the recession part of hydrograph. In the third case study, SWAT simulation model (Neitsch *et al.*, 2001) is used for predictions of discharge in the 37-km^2 Cannonsville Reservoir in upstate New York. This catchment has been studied in the past by Tolson (2005) and (Tolson and Shoemaker, 2007) using SWAT to predict flow, sediment and phosphorus. SWAT is a spatially distributed continuous simulation model for predicting flow, sediment, nutrient and other contaminant transport. Tolson and Shoemaker

(2007) solve a three-objective calibration problem for the Town Brook watershed by aggregating objective functions associated with streamflow, phosphorous and sediment prediction accuracy. In this study, the same calibration problem is solved except that SWAT is calibrated based only on two responses, discharge and sediment transport. There were 26 SWAT parameters to be adjusted via calibration. Appendix D provided more details on these parameters (*e.g.,* description and prior range), and elaborates on the SWAT implementation to this case study. The simulation period for discharge was from January 1st 1996 to September 30th 2000 with a 639-day warm-up period resulting in a 1096-day period. This period is divided in two parts, the calibration period from October 1st 1997 to the end of 1998, and the validation period from January 1st 2000 to September 30th 2000. For the sediment transport, the simulation period is October 1st 1998 to September 30th 2000 in which the sup-period October 1st 1998 to the end of 1999 was used for calibration, and the period January 1st 2000 to September 30th 2000 for validation. In the SWAT case study, two responses are considered, discharge and suspended sediment transport.

4.3 Results

4.3.1 Comparison among Multiple Behavioural Solution Identification Strategies

The strategies of identifying behavioural solutions are notated as follows: 'NSE>0' accepting all parameter sets with NSE greater than zero for both responses; 'NSE>0.5' accepting all parameter sets with NSE greater than 0.5 for both responses; 'top 10%' sorting the parameter sets based on the aggregated likelihood and use the top 10% of the sample set as behavioural; 'Rank-based' accepting behavioural solutions according to the rank-based strategy explained in Section 4.2.2.1; and 'Opt. Perc.' sorting the parameter sets based on weighted criterion and identify the optimal percentage from the top solutions based on DTI metric. Figure 14 illustrates the reliability and sharpness values in calibration period associated to different strategies for the three case studies. The results shown in Figure 14 correspond to 100,000 simulations for each case study. Figure 14 demonstrates that different strategies for identifying behavioural solutions yield a range of points on $SH-R$ graph. The results show that the 'Rank-based' and 'Opt. Perc.' strategies compare favourably well with other strategies. These two approaches are also located very close to each other on $SH-R$ graph in almost all case studies. Two other computational budgets with lower number of simulations are also considered (10,000 and 50,000). Showing $SH-R$ graphs for all cases, especially in multi-criteria context, might not be useful. Thus, to provide better tools to do comparison among different strategies, the DTI metric is calculated for all case studies along the three computational budgets. Figure 15 depicts the obtained DTI values where the horizontal axis shows different approaches and bar colours represent different computational budgets.

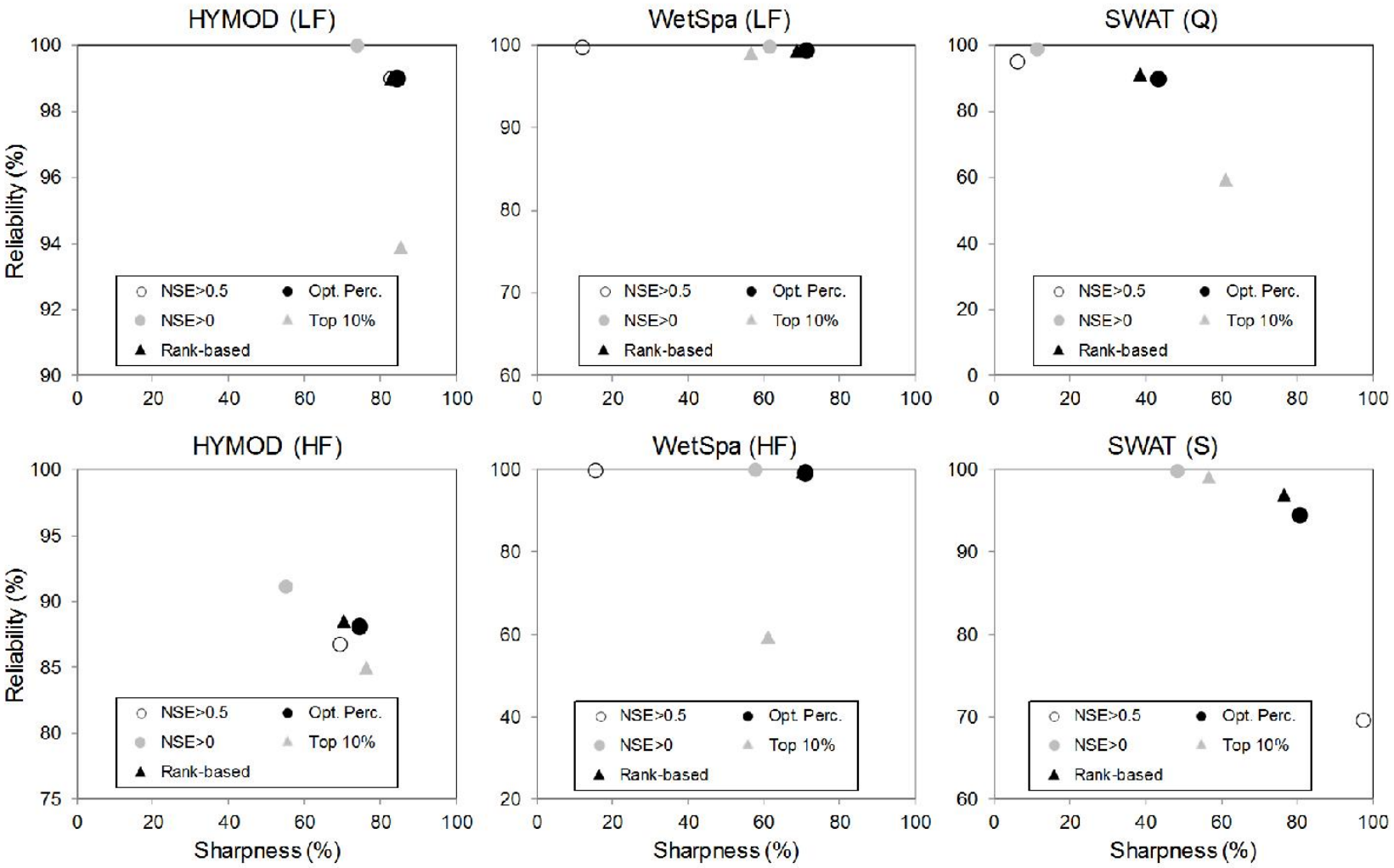

Figure 14. Reliability and sharpness associated with different strategies for identifying behavioural solutions in three case studies along 100000 simulations (calibration period).

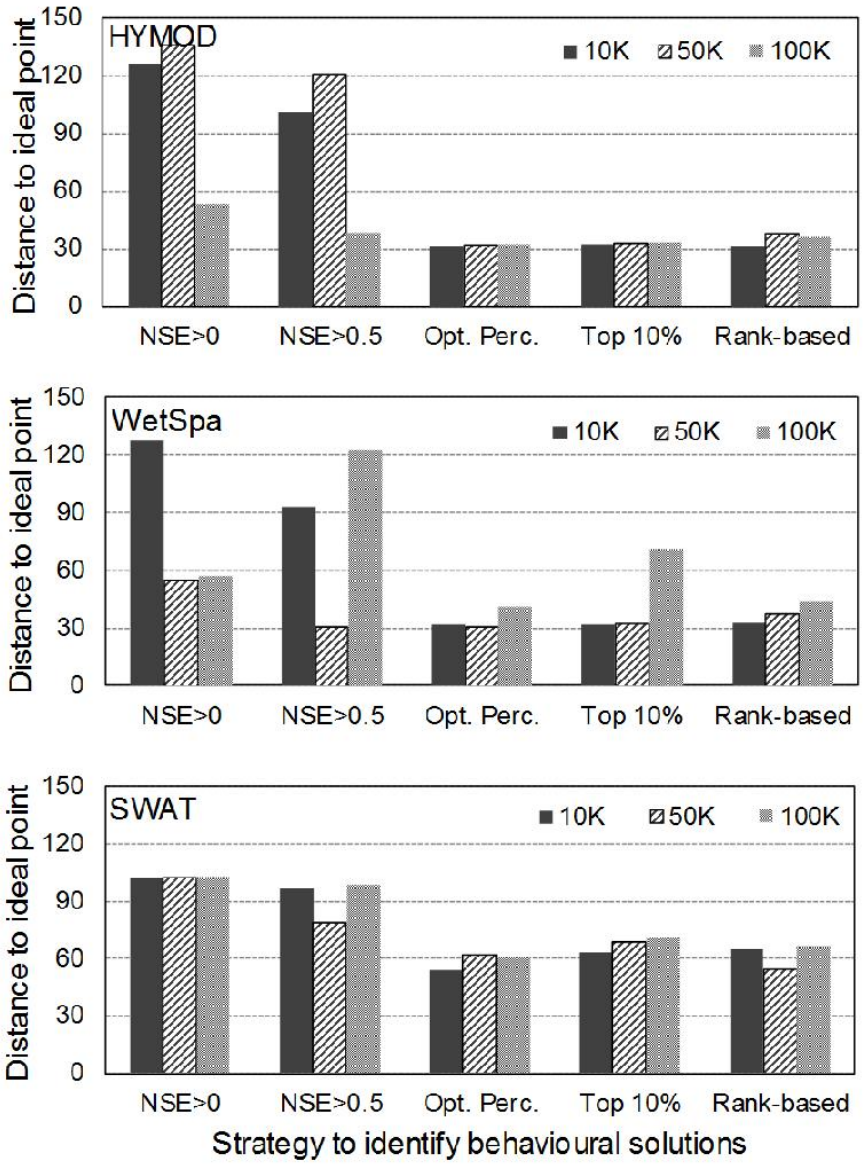

Figure 15. Distance-to-ideal (DTI) values for different strategies of identifying behavioural solutions in GLUE, obtained in different case studies along different computational budget.

For each experiment shown previously, the strategy which has the lowest *DTI* value is identified and shown in Table 1 which shows that either the 'Opt. Perc.' or 'Rank-Based' behavioural solution identification strategy turns out to be the best approach in almost all cases. More importantly, these results highlight that there is no single identification strategy which yields the minimum *DTI* for all cases. This suggests that the common practice in GLUE studies of subjectively choosing a single identification strategy (without comparing alternatives) is not robust. Instead, a more robust application of GLUE would consider multiple candidate behavioural solution identification strategies and our *DTI* metric provides

one multi-criteria decision-making approach to choose the best identification strategy. The fact that one identification strategy is selected over many alternatives functions to reduce the subjective nature of the behavioural solution identification process. It is critical to note here that the implementation of multiple identification strategies does not involve re-running the hydrologic model and it only requires post-processing the simulation model results.

Table 1. The behavioural solution identification strategy with lowest DTI value, identified for each experiment

Computational budget	Case study		
	HYMOD	WetSpa	SWAT
10K	Rank-based	Opt. Perc.	Opt. Perc.
50K	Opt. Perc.	NSE>0.5	Rank-based
100K	Opt. Perc.	Opt. Perc.	Opt. Perc.

4.3.2 Comparison of GLUE with Other Informal Methods

GLUE is compared with Pareto optimization and DDS-AU methods for the three case studies along the same computational budgets. The informal likelihood function used as the calibration objective in DDS-AU to evaluate models' performance is NSE, *i.e.,* DDS-AU is run considering sum of NSE over all responses as the objective function. The numbers of trials are 100, 200, and 300 for the 10000-, 50000-, and 100000-simulation budgets, respectively. Then, the top parameter set of each trial is selected as a behavioural solution as long as it satisfies the condition that the aggregated criterion is greater than or equal to 1.0 (note that the best possible value for aggregated objective function would be 2.0). In AMALGAM, NSE values calculated over different responses are considered as the objective functions, and the parameter settings suggested in Vrugt and Robinson (2007) are used. The population size is set to 100, 200, and 300 for the three computational budgets mentioned above. Then, the Pareto optimal solutions are classified as behavioural, and used the hydrographs associated to these solutions to derive the uncertainty bounds of the model outcome.

In GLUE, two sets of results are used for each experiment, one being the best strategy obtained in the previous part of the study (*i.e.,* Table 1), and the other one being the top 10% as a traditional strategy of identifying behavioural solutions in GLUE. Comparison between GLUE considering these two strategies and other informal methods can help us assess the impact of GLUE's subjectivity on decision making. First, the $SH-R$ graph is generated based on the behavioural parameter sets obtained from different techniques applied to different case studies (not shown here). The results showed that different methods located their results in different regions on the $SH-R$ graph, but most of them were non-dominating

with respect to each other. This fact required using the DTI metric for a better comparison. Figure 16 shows the DTI values for all case studies (HYMOD the top, WetSpa the middle, and SWAT the lower graph). Different bar colours represent different methods, and the horizontal axis shows three computational budgets considered (10000, 50000, 100000 simulations). It should be pointed out that the values shown on the vertical axes in Figure 16 represent the Euclidean distance between the behavioural sets and the ideal point in the four-dimensional decision space (considering reliability and sharpness of two responses as the four criteria). Note that the method with lowest DTI value would be preferred.

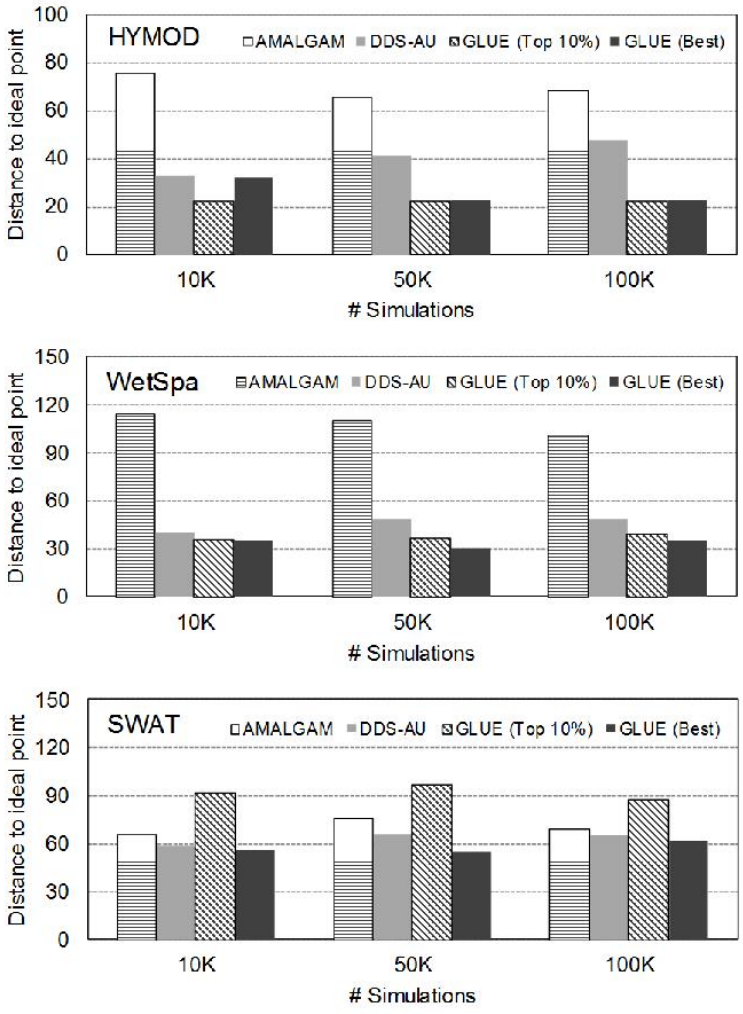

Figure 16. Distance-to-ideal values for different case studies obtained from different informal methods along different computational budgets in the validation period; this metric should be as low as possible.

It is observed in Figure 16 that for HYMOD and WetSpa as examples of low-dimensional calibration problems, both behavioural solution identification strategies in GLUE result in smaller DTI values. In

the SWAT case as a high-dimensional calibration problem, it is observed that the top 10% strategy in GLUE results in larger DTI values than DDS-AU and AMALGAM. On the other hand, when the best behavioural solution identification strategy is used in GLUE (*i.e.,* meaning that every case study may have a different strategy), the resulting DTI values are better than DDS-AU and AMALGAM. This indicates that if the subjective decisions in GLUE are properly overcome, GLUE may still remain a good candidate for low- and high-dimensional modelling cases.

Figure 17 illustrates the observed hydrograph as well as the 95% uncertainty intervals of the model outcome obtained from different informal methodologies in the SWAT case study after 100,000 simulations. As explained earlier, the 'GLUE (BEST)' approach corresponds to the best strategy of identifying behavioural solutions in GLUE, while 'GLUE Top10%' is related to the top 10% of parameters sets after sorting them based on Equation (18) (traditional approach). Figure 17 also shows the results of DDS-AU and AMALGAM. The calibration panel in Figure 17 shows that GLUE (especially the 'Best' approach) results in fairly wider prediction intervals in comparison to DDS-AU and AMALGAM. This is in fact the reason why GLUE yields more reliable predictions in the validation period, especially for peaks. Comparison between the results of 'GLUE (Best)' and 'GLUE Top10%' in Figure 17 also shows that the former managed to better cover peaks and low-flows. This highlights the importance of subjective decision involved in GLUE implementation and the impact on the resulting uncertainty bounds of the model outcome.

It is worth noting that a key reason GLUE gives a better DTI value than DDS-AU might be because only a single behavioural solution identification strategy is used in DDS-AU. The results might change if multiple DDS-AU configurations are considered, the same way as the study did for GLUE. This, however, remains an open question for future research, and this study only focuses on evaluating if the uncertainty bounds generated by GLUE might improve if novel and objective strategies are adopted for identifying behavioural solutions.

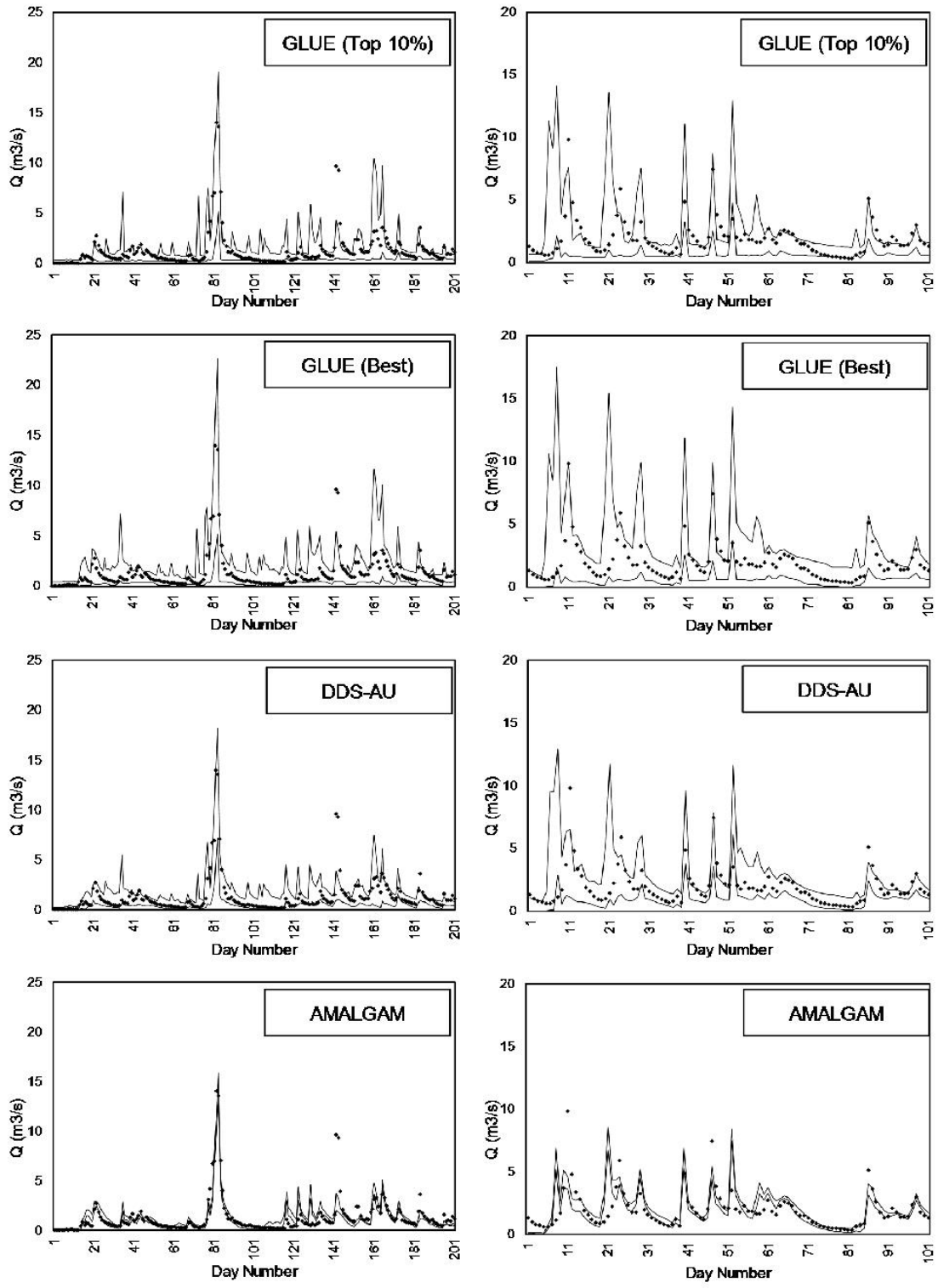

Figure 17. Hydrographs over a portion of calibration period (left panel) and validation period (right panel) in SWAT case study after 100000 simulations, obtained from different methods; Observations are shown by dots while the 95% prediction intervals are depicted by solid lines.

4.3.3 Discussion

This study shows how different strategies of identifying behavioural solutions in GLUE can be evaluated using $SH-R$ and DTI graphs. The proposed methodology for comparison among different behavioural solution classification strategies in GLUE provides the decision-maker with the opportunity to more objectively choose the most appropriate option. GLUE is considered as a candidate method, because it requires a number of subjective decisions. The proposed behavioural solution identification strategies can be, however, implemented in informal methods of uncertainty estimation where the uncertainty of the model outcome is described by a suite of behavioural solutions. As an example, the idea can be used to identify the best behavioural set from the solutions obtained from DDS-AU. Also, DTI can be implemented in ParaSOL (van Griensven and Meixner, 2007) as an additional tool for selecting behavioural solutions, and also for comparing different approaches of identifying behavioural solutions in that methodology.

The concept of DTI (calculated based on reliability and sharpness of predictions) is suggested in this study as a diagnostic to evaluate and compare informal calibration and uncertainty estimation methods in hydrological modelling. As such, the DTI metric is something that would normally be computed after a set of candidate behavioural solutions are identified. However, it is expected that it can also be offered as a convergence metric for informal sampling-based uncertainty estimation methods. If the objective of an informal calibration and uncertainty estimation is translated into an attempt to approach (as much as possible) the ideal point in $SH-R$ graph, DTI can serve as a guiding metric to assess the convergence of whatever sampling procedure is used to generate behavioural solutions. Besides GLUE, DTI can be used as a stopping criterion for sequential sampling strategies such as SUFI (Abbaspour *et al.*, 2004) or SMC (Jeremiah *et al.*, 2011). Moreover, for informal implementations of MCMC samplers in calibration of hydrologic models (Blasone *et al.*, 2008a; Vrugt *et al.*, 2003a), DTI can identify a subset of the entire chain as behavioural solutions.

The objective of this study is to demonstrate the importance of behavioural solution identification strategies in GLUE, and develop new objective strategies that alleviate the subjectivity in identifying behavioural solutions. Two other strategies are noted in Section 4.2.2 for behavioural solution identification (Blasone *et al.*, 2008b; Liu *et al.*, 2009). Although this study did not compare the proposed two new strategies against these, it is not suggested that the approaches in Blasone *et al.* (2008b) or Beven (2006), or any other more traditional strategy, be replaced by the strategies of this study. Instead, the results put emphasis on the need to consider multiple behavioural identification strategies, as opposed to using only one.

As pointed out in this study, DTI is one option (among others) to be utilized for objective identification of behavioural solutions. It might be argued that DTI-based identification strategy is still subjective, *i.e.,* subjectivity in GLUE is moved from the selection of the informal likelihood measure and threshold to the selection of the metric. However, it should be noted that the subjective decisions which are still involved in the proposed framework are more closely aligned to how predictive models should be used in the real-world (*e.g.,* easier and more transparent for practitioners to explicitly decide on tradeoffs between reliability and sharpness than it is for them to pick some arbitrary NSE threshold). Thus, the present study complements the previous research papers on other subjective decisions of GLUE towards the development of a more proper GLUE uncertainty estimation framework for the practitioners and hydrological modellers who still apply this methodology.

4.4 Concluding remarks

This study develops new strategies for objective identification of behavioural solutions in the GLUE methodology, evaluates the results obtained from these strategies against other behavioural solution identification strategies used previously in GLUE, and finally compares GLUE with two other informal multi-criteria calibration and uncertainty estimation methods. The proposed behavioural solution identification strategies are based on simultaneous optimization of reliability and sharpness of prediction intervals. The first strategy is based on Pareto ranking of parameter sets, while the second strategy considers a ranking scheme on the basis of criteria-aggregation. The results indicate the subjectivity involved in behavioural solution identification strategies would impact the uncertainty bounds of the model outcome, and more importantly, no single strategy shows to be the best in all case studies. Nevertheless, the new objective strategies proposed here are the best options in the case studies investigated here (5-parameter HYMOD to 26-parameter SWAT hydrologic model calibration problems)

The proposed strategies of identifying behavioural solutions in GLUE are not meant to replace other strategies proposed and used by other researchers. Instead, it is believed that these should be added to the set of strategies available to modellers using GLUE. It would eventually be the user who should choose the best option based on the relative importance given to reliability, sharpness, or other factors involved in the decision-making process. The decision in this study is based on simultaneous optimization of reliability and sharpness, and the $SH-R$ and DTI graphs are meant to facilitate better decision-making. Since the results show variability of prediction intervals depending on the subjective definition of behavioural solutions, in order to have a robust implementation of GLUE, it is recommended that future studies consider multiple behavioural solution identification strategies and choose the best one according to the modeller's priorities.

In the second part of the paper where GLUE is compared with optimization-based approaches (Pareto optimization and DDS-AU), the impact of behavioural solution identification strategies in GLUE on the uncertainty bounds of the model outcome is evaluated. The reliability and sharpness in the validation period are considered (the time frame of which the data are not used for model calibration) and the methods are compared visually (based on $SH-R$ graph) and quantitatively (based on DTI). The results demonstrate that GLUE is robust in comparison to the other approaches for both low- and high-dimensional calibration problems provided that in the calibration phase, different behavioural solution identification strategies are first compared and the best one is then selected.

Future research is needed to evaluate the utility of DTI in other sampling-based uncertainty estimation methods as a convergence metric to stop the sampling process. Future studies should also utilize the ideas in this study to address or reduce the subjective decisions inherent in other informal methods for describing the uncertainty of the hydrologic model outcome.

Chapter 5
Calibration and Uncertainty Estimation of Hydrologic Models Using Multi-Criteria Optimization and Hydrological Signatures

Summary

The simulated outcome of a calibrated hydrologic model should be hydrologically consistent with the measured response data. Therefore, hydrologic modellers first calibrate models to optimize statistical performance metrics, *e.g.,* the Nash-Sutcliffe Efficiency measure, and second, 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 multi-criteria optimization context. This chapter develops a multi-criteria model calibration and uncertainty estimation methodology based on the optimization of hydrological consistency that is dependent on performance metrics as well as a level of acceptability for each signature. Different calibration approaches and formulations are introduced and compared with the traditional formulations found in the literature for four hydrologic models in seven catchments. An informal uncertainty estimation approach is also developed in this chapter where the samples obtained in the many-objective signature-based calibration process are employed to derive the model output uncertainty bounds. The results reveal that a Pareto dominance-based formulation yields the highest level of consistency among all formulations, and the findings prove insensitive to the type of optimization algorithm. Section 5.1 provides the introduction, Section 5.2 elaborates on the material and methods, results are illustrated in Section 4.3, and finally, Section 5.4 highlights the concluding remarks drawn in this study.

5.1 Introduction

Consider a time series of N streamflow observations, $Y_1, Y_2, ..., Y_N$ (or $\mathbf{Y}$ in vector notation) that is to be used to calibrate a deterministic hydrologic model represented generically as $\hat{\mathbf{Y}} = h(\mathbf{X}, \boldsymbol{\theta})$ where $h(\,)$ indicates evaluation of the hydrological model to simulate the various output of interest given input data $\mathbf{X}$ and a corresponding parameter vector ($\boldsymbol{\theta}$). The residual vector defines the difference between the observed and simulated system's behaviour $E_t(\theta) = \hat{Y}_t - Y_t,\ t = 1,..,N$. The calibration problem is to find a parameter set whose model output best resembles the observed system's behaviour. In the context of calibration and uncertainty estimation of hydrologic models, the classical statistical approach (Bates and Watts, 1988; Seber and Wild, 1989) minimizes the sum of squared residuals to find the most appropriate parameter set. Statistical methods also confront the comparison of simulations and observations based on a Bayesian viewpoint (Box and Tiao, 1973) and derive 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 is the implementation of informal measures to evaluate the consistency of model behaviour in comparison to the observations where these informal measures replace formal likelihood functions such as sum of squared errors. Example strategies include generalized likelihood approaches (Beven, 2006), the application of fuzzy set theory (Seibert and McDonnell, 2002), multi-objective calibration (*e.g.,* Boyle *et al.*, 2000; Gupta *et al.*, 1998; Shafii and Smedt, 2009; Vrugt *et al.*, 2003a), etc.

The average or aggregate measures of model data similarity such as sum of squared errors or Nash-Sutcliffe measure may not provide the modeller with enough power to make a meaningful comparative evaluation of the consistency in model form, the equations, and function, representations such as storage or discharge (Gupta *et al.*, 2005a; Gupta *et al.*, 1998; Wagener and Gupta, 2005). 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 modelling components that, when assumed to be functioning properly, explain the discrepancy between the simulated and observed behaviour (Gupta *et al.*, 2008). Gupta *et al.* (2008) indicate that such diagnostics can be either correlative, established by observing a correlative relationship, or causal, using the underlying theory to investigate possible causes of an observed system's response. As an example of the latter, if there are any problems in simulation of overland flow processes by models, the infiltration or saturation excess components of the model might be at fault.

To develop rigorous diagnostic-based model evaluation methods, numerous studies have employed hydrological signatures that reflect the functional behaviour of the catchment 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 precipitation-runoff data to identify a number of indices or signatures to quantify relevant aspects of the system behaviour 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. Literature shows a variety of signatures proven to be useful for model evaluation, the 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 baseflow index estimated from soil types (Bulygina and Gupta, 2009), the peak distribution (Sawicz *et al.*, 2011), etc.

Hydrological signature applications can be categorized as either signature-based model selection or signature-based model calibration. In signature-based model selection (elaborated on in Section 5.1.1), signatures are employed to make a selection among candidate model structures for a given catchment. These model structures are calibrated independent from 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 5.1.2) 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.

5.1.1 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 certain signatures, inter-annual variability of runoff, mean monthly variation of runoff, and the flow duration curve (FDC) in a downward approach for identification of the model components required at different modelling 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 a proper structure for hydrological modelling at a given catchment. Note that the diagnostic tests developed in McMillan (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 multiple-criteria model performance evaluation strategy applied to a set of 764 catchments across the United States, aiming at searching for 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.

5.1.2 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. This finding makes hydrological signatures suitable for calibrating an existing model structure from the hydrological perspective. For instance, Yadav *et al.* (2007) demonstrate that runoff ratio (*i.e.,* average

annual runoff divided by average annual precipitation) 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.

Because an acceptable level of a hydrological signature may be achieved at the expense of other signatures, a robust approach for model calibration requires multiple signatures to fully characterize the system's behaviour. Simultaneous consideration of many hydrological signatures may, however, be challenging due to computational limitations in multi-criteria optimization associated with many objective functions (Pokhrel *et al.*, 2012). As a result, rather than a full multi-criteria analysis considering a large number of signatures, hydrologists either use a limited set of signature-based measures (van Werkhoven *et al.*, 2009) or conduct traditional calibration (optimization-based approaches) first, and then, use consistency tests to filter consistent 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, a number of hydrological signatures are employed to identify the most appropriate subset of these parameter sets.

Alternatively, researchers use Monte Carlo Simulations (MCS) to identify consistent parameter sets (resulting in signature values contained in a predefined range) 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 modelling (Beven, 2006; Liu *et al.*, 2009). More recently, Vrugt *et al.* (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) define hydrological consistency as the reproduction of multiple signatures given a particular model structure and a single parameter set. An ideal signature-based model calibration would use numerous hydrological signatures to guide the parameter search towards regions of hydrological consistency in the search space. The literature shows that a few studies 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 translated into exact numbers). Yilmaz *et al.* (2008) utilize signature information to progressively constrain the ranges of parameters that are found to be sensitive to the signature indices, reducing the prior ranges to the ranges associated to the best signature values. Vrugt and Sadegh (2013) consider four signatures and build informal likelihood measures based

on these signatures. They then use an iterative Monte Carlo sampling scheme to sample parameter sets that show hydrologic consistency according to these signature-based informal likelihood measures. A promising and seemingly unexplored approach to signature-based model calibration would be to apply multi-objective optimization concepts with the explicit objective of optimizing the hydrological consistency.

5.1.3 Research Outline

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

Multiple sampling formulations are introduced and these are also compared with the two most common signature-based formulations in the literature on the basis of hydrological consistency. Our formulations are shown to handle as many as 15 signatures simultaneously and can produce more 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. Furthermore, the present research explores the potential of uncertainty estimation using the parameter sets generated by the sampling algorithms through an uncertainty estimation scheme that is fairly similar to signature-based implementations of GLUE (Beven, 2006; Liu *et al.*, 2009; Winsemius *et al.*, 2009). The uncertainty estimation part of the proposed methodology is detailed in Section 5.2.3.

5.2 Methodology

In signature-based model calibration, the modeller 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. Each sampled parameter set is a potential solution to the calibration problem and satisfies a certain number of hydrological signatures. All solutions evaluated in the calibration process are retained in an archive. Figure 18 illustrates the results for a hypothetical calibration problem using three different sampling-based calibration approaches; the horizontal and vertical axes represent the proportion of satisfied

signatures and sampled solutions, respectively. Proportions on the horizontal and vertical axes are computed based on the following equations, respectively:

$$P_{signatures} = k/K \quad (19)$$

$$P_{solutions} = m/M \quad (20)$$

where P is the proportion, k is the number of satisfied signatures, K is total number of signatures considered, m is the number of solutions satisfying a particular number of signatures, and M is total number of solutions in the archive. Given K signatures and M evaluated parameter sets, the modeller requires a number of the M solutions to satisfy all K signatures, or in the completely ideal case, all M solutions satisfy all K signatures (approach A in Figure 18). Satisfying all signatures might not be achievable due to imperfections in the sampling algorithm or model structure, approaches B and C are more realistic, B being more consistent than C. Modellers are obviously interested in approaches towards the most consistent case and in particular, approaches that maximize the number of signatures satisfied.

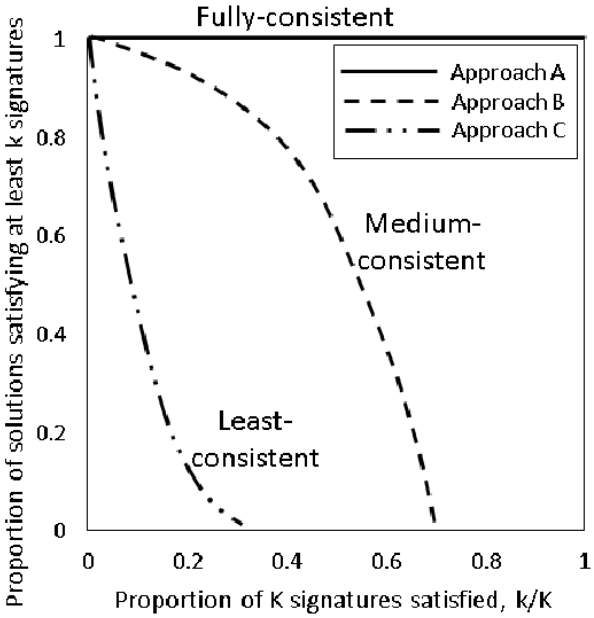

Figure 18. Conceptual graph showing fully consistent to the least consistent sets of results obtained in a hypothetical calibration problem by three alternative calibration approaches.

The numerical experiments in this study consider 13 signatures that are described in Appendix B, as well as two additional criteria based on standard goodness of fitness measures (described below). Each signature is translated into a measure that quantifies the difference between simulated and observed signature values. This measure is subsequently coupled with a given acceptability threshold to define a binary calibration criterion, 1 means that the signature-based criterion is satisfied and 0 means it is not

satisfied. In this thesis, satisfaction of a signature-based measure and satisfaction of a criterion are used interchangeably. Note that plausible values for all signature-based measures fall in the range $[-100, 100]$. The proposed methodology (described in the following sections) 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 employs multi-criteria optimization concepts, and thus, it can also implement standard performance metrics such as NSE formulated as Equation (9) in this thesis.

All signature-based calibration approaches are evaluated with respect to their ability to optimize the hydrological consistency defined in Section 5.2.1. The two types of approaches to calibration, and multiple formulations of each, are described in Section 5.2.2. The strategy of uncertainty estimation is elaborated on in Section 5.2.3, followed by Section 5.2.4 that describes the seven rainfall-runoff modelling case studies.

5.2.1 Hydrological Consistency Measure

The consistency measure is based on the number of satisfied signature-based measures as well as the value of these measures, and is designed to be maximized. Given a parameter set θ, assume that M_i is the signature-based measure for signature i (M_i is the difference between simulated and observed value of the signature), and Th_i is the acceptability threshold for signature-based measure i. Then, a metric (called acceptability metric or ACC) is calculated for each signature as follows:

$$ACC_i = \begin{cases} 1, & M_i \le Th_i \\ \dfrac{M_{i,\max} - M_i}{M_{i,\max} - Th_i}, & \text{Otherwise} \end{cases} \qquad (21)$$

where $M_{i,\max}$ is the maximum plausible value for M_i. If the difference between simulated and observed signature values is within the desired threshold ($M_i \le Th_i$), the acceptability metric is assigned the value of 1. Otherwise, a number in the range $(0-1)$ is assigned to the acceptability metric where $M_{i,\max}$ and Th_i function to normalize M_i. For non-satisfied signature-based measures, the closer *ACC* is to 1, the closer the measure is to the satisfaction level.

When the acceptability metric (*ACC*) is calculated for all *i* signatures, the resulting values are sorted in a descending order where ACC_1 indicates the highest *ACC* value and ACC_i indicates the lowest *ACC*

value. Next, as shown in Equation (22), the overall consistency measure ($C_{overall}$) is calculated as the sum of the highest $n^* + 1$ ACC values, where n^* is the number of satisfied signature-based measures, ACC=1. The resulting sum is considered as the consistency value of parameter set θ. This consistency measure can be maximized to obtain the most hydrologically consistent parameter set. Note that this approach quantifies the degree of inconsistency based on the proximity of the closest non-satisfied signature to its corresponding threshold. This is a natural way to scale between integer counts of the number of satisfied signatures.

$$C_{overall} = \sum_{j=1}^{n^*+1} ACC_j \qquad (22)$$

Note that the above strategy of calculating the consistency measure is independent from the type of signatures, and the modeller only needs to identify the acceptability threshold and a maximum plausible value for each signature. Because the modeller may have multiple (often numerous) signatures of interest, the consistency metric can be implemented in the calibration process using different approaches, which are described in the following section.

5.2.2 Signature-Based Calibration Approaches

Consistent 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 statistical 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 5.2.2.1) and 2) utilize a parameter sampling scheme that does adapt to the degree of hydrologic consistency of previously sampled parameter sets (Approach 2 described in Section 5.2.2.2).

5.2.2.1 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) MCS using the prior parameters range, called A1-MC, and (ii) sampling using a multi-criteria optimization algorithm, called A1-Opt. The optimization formulation in the latter approach is as follows:

$$\max_{\theta \in \Theta} F(\boldsymbol{\theta}) = \begin{bmatrix} NSE_{LF} \\ NSE_{HF} \end{bmatrix} \quad (23)$$

where $F(\boldsymbol{\theta})$ is the vector of objective functions, NSE_{LF} and NSE_{HF} are the NSE of low- and high-flows, respectively, $\boldsymbol{\theta}$ is an arbitrary parameter set, and Θ is the feasible parameters space. Once the sample set is obtained, all solutions are then post-processed to evaluate the simulated and observed signature values, and calculate the consistency metric accordingly. Note that even though we do not implement the exact same set of signatures or calibration objectives as reported in any one of the aforementioned studies, our approach to the calibration problem is the same.

5.2.2.2 Approach 2: Parameter Search Guided by Hydrologic Consistency

The second approach does not consider post-processing and instead employs hydrological signatures throughout the calibration process, specifically during parameter optimization. Multiple formulations are considered in this approach based on either criteria aggregation or Pareto-based multi-criteria optimization. Note that the main goal of the proposed methodology is to implement many signature-based criteria in the calibration process.

Single Objective Formulation: A2-SO

The first formulation, called A2-SO, aggregates statistical and hydrological measures into the consistency metric defined by Equation (22), and applies a single-criterion optimization algorithm for its optimization. The optimization formulation in this case is as follows:

$$\max_{\theta \in \Theta} F(\boldsymbol{\theta}) = C_{overall} \quad (24)$$

where $F(\boldsymbol{\theta})$ is the objective function equal to the overall consistency, $\boldsymbol{\theta}$ is an arbitrary parameter set, and Θ is the feasible parameters space. To implement statistical measures such as NSE in this formulation, a suitable acceptability threshold must be defined. NSE is used in our numerical experiments, and different acceptability thresholds are considered (0.5 and 0.7). Considering two NSEs for low- and high-flows plus 13 hydrological signatures, the calibration problem involves 15 binary criteria aggregated into the single consistency measure. To sample parameter sets from the search space in A1-SO, the DDS optimization algorithm (Tolson and Shoemaker, 2007) is applied.

Multi-Objective Formulation 1: A2-MO1

Another formulation is considered in the second approach, called A2-MO1, where two criteria are defined. The first criterion is the average of continuous statistical measures, which is the NSE of low- and

high-flows in this study. The second criterion is the consistency metric defined by Equation (22) that aggregates only the signature-based measures (excludes NSEs). The optimization formulation in this case is as follows:

$$\max_{\theta \in \Theta} F(\boldsymbol{\theta}) = \begin{bmatrix} C_{overall} \\ \\ \overline{NSE} \end{bmatrix} \qquad (25)$$

where $F(\boldsymbol{\theta})$ is the vector of objective functions, $C_{overall}$ is the overall consistency, $\overline{NSE}$ is the average NSE of low- and high-flows, $\boldsymbol{\theta}$ is an arbitrary parameter set, and Θ is the feasible parameters space. This formulation yields a standard bi-criterion optimization problem which is solved using a multi-criteria optimization algorithm; in this study, AMALGAM (Vrugt and Robinson, 2007) is used.

Multi-Objective Formulation 2: A2-MO2

Another formulation is developed in the second approach, called A2-MO2. This formulation conducts calibration through Pareto-based multi-criteria optimization where a separate objective is defined for each signature and each statistical measure. This formulation considers the same 15 criteria used in formulation A2-SO, however, the continuous statistical measures such as NSE are used directly as objectives (to be maximized).

The multi-criteria 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 non-dominated (or Pareto solutions) tends to become large, and this issue is termed dominance resistance which is known to slow the search (Farina and Amato, 2004; Fleming *et al.*, 2005; Purshouse and Fleming, 2007). Moreover, Teytaud (2007) shows that for large numbers of objectives (≥10), the rule for selecting candidate solutions during optimization may not be more effective than random search.

With more than 10 signatures (15 in our case), we assume that a multi-criteria calibration formulated as a multi-objective optimization problem with 10 or more continuous value objectives (to minimize M_i for all i=1, …, 13 signature-based measures and to maximize two NSEs; 15 criteria in total) will cause the above-mentioned problems for many candidate multi-objective algorithms. Indeed, preliminary tests using this many continuous objectives demonstrated very poor algorithm performance. Hadka and Reed (2011) suggest this finding is likely limited to algorithms that do not implement epsilon-dominance (such as AMALGAM). Nonetheless, solving such a many objective problem in this way can require an incredibly large computational budget. Thus, we do not consider the optimization of many continuous criteria in this study, and instead, approach the many-objective calibration problem by essentially

converting the continuous objectives into binary objectives utilizing the level of acceptability for each signature-based measure (satisfied signature according to Equation (21) is given a value of 1 whereas non-satisfied signatures are assigned a value of 0 instead of a fractional value). A binary discrepancy measure (Hornberger and Spear, 1981) defines a range of criterion values considered as successful achievement of the criterion. Using the aforementioned conversion, the optimization formulation in this case will be as follows:

$$\max_{\theta \in \Theta} F(\boldsymbol{\theta}) = \left[B_1, B_2, \cdots, B_{13}, NSE_{LF}, NSE_{HF} \right]^T \qquad (26)$$

where $F(\boldsymbol{\theta})$ is the vector of objective functions, B_i is the binary criteria associated with signature i (*i.e.,* $B_i = 1$ when $ACC_i = 1$ and $B_i = 0$ when $ACC_i \neq 1$; ACC_i is defined in Equation (21)), NSE_{LF} and NSE_{HF} are the NSEs of low- and high-flows, respectively, $\boldsymbol{\theta}$ is an arbitrary parameter set, and Θ is the feasible parameters space. This formulation is then optimized by a multi-criteria optimization algorithm, AMALGAM in this research.

Note that the binary nature of the signature-based criteria is hypothesized to make the high-dimensional calibration problem of A2-MO2 more tractable for algorithms like AMALGAM, because the optimization algorithm optimizes the criteria up to the acceptability level, and not unconditionally. This hypothesis was supported by a preliminary experiment, which showed that when 15 continuous maximization criteria were considered and 10000 realizations of these criteria were generated. Upon sorting these realizations using non-dominated sorting, almost 70% of the entire set of realizations showed to be a Pareto front solution. However, converting 15 criteria to 15 binary criteria reduced the number of Pareto front solutions to less than 1%. As a result, the issues in many-objective optimization (Farina and Amato, 2004; Fleming *et al.*, 2005) are at least partially mitigated by converting to binary criteria.

The main goal in A2-MO2 is to apply Pareto sorting to direct the search towards the most consistent regions in the parameters space (obtaining near-optimal NSE values as well as satisfying as many signature-based measures as possible). Note that the proposed formulation is inspired from Komuro *et al.* (2006) whereby an ecological process model is developed using multi-criteria assessment. To gain proper insight into multi-criteria optimization, interested readers are referred to the state-of-the-art papers in the hydrological literature (some recent review papers are Efstratiadis and Koutsoyiannis, 2010; Reed *et al.*, 2013).

Figure 19 illustrates a hypothetical case with one continuous maximization criterion in the range (0,1), as well as four binary criteria where black and white squares represent 1 and 0, respectively. In Figure 19,

Solution S1 is superior to solution S3 because it has a larger value of the continuous criterion, satisfies the measures involved in criteria B and D (these two are also satisfied by S3), and satisfies an extra criterion, E. Similarly, S1 is superior to S4. In multi-criteria optimization terminology, solution S1 dominates solutions S3 and S4 and eliminates these two solutions from consideration as potential solutions. Figure 19 also indicates that solution S1 and S2 do not dominate one another because there is at least one criterion that is satisfied by one of them, but not the other one. Solutions S1 and S2 are called non-dominated solutions and are added to the so-called Pareto set. Traditional multi-criteria calibration uses the non-domination concept to sort the entire population. Upon termination of the optimization algorithm, the solutions in the Pareto set are considered as the calibration results and are used for further analyses.

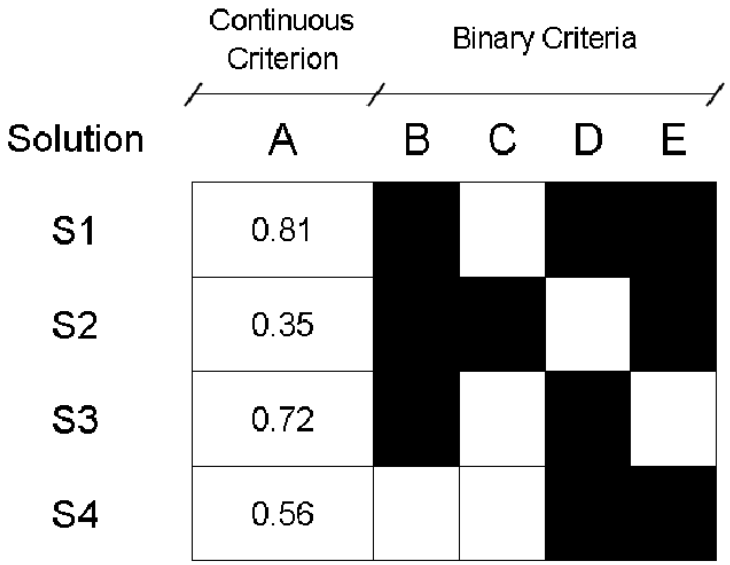

Figure 19. Illustration of a set of solutions (S1-S4) with one continuous criterion (A) and four signature-based criteria (B-E); black and white squares represent 1 and 0 in binary criteria.

Section 2.4.2.1 describes how multi-criteria optimization algorithms (including AMALGAM which is used in this chapter) utilize the selection process to generate and evaluate candidate solutions. When the binary criteria was in the optimization process using AMALGAM, it was observed that the standard selection metrics performed identically as random selection, and did not push the search towards regions of high consistency effectively. This failure was predictable as these metrics have not been designed for many-objective optimization with binary criteria. To overcome this issue the consistency metric described in Equation (22) was defined as the selection metric to direct the search. Thus, when AMALGAM compares two solutions for selection and generation of new solutions, the one with larger consistency is retained and the other one is discarded. This change in the algorithm selection procedure is considered a key difference between the proposed methodology and the standard multi-criteria optimization.

In summary, A2-MO2 applies AMALGAM to a multi-criteria calibration problem with 15 criteria in total, 2 NSEs (continuous) and 13 binary signature-based criteria. The proposed formulation does not aggregate the criteria, and instead, employs the Pareto dominance concept considering 15 criteria. Furthermore, AMALGAM uses the consistency measure as the selection metric to guide the search space towards high consistency regions in the search space.

5.2.3 Uncertainty of Model Outcome

For predictive hydrologic models to be useful and defensible, proper consideration of all uncertainty sources, parameter values, meteorological inputs, and other non-calibrated forcing functions, is essential (Wagener and Gupta, 2005). As a result, research in the past decades has emphasized comprehensive statistical model assessment procedures either using frequentist approaches (*e.g.,* Bates and Watts, 1988; Refsgaard *et al.*, 2005; Reichert, 1997; Seber and Wild, 1989), or Bayesian inference (*e.g.,* Bates and Campbell, 2001; Kavetski *et al.*, 2002; Kuczera, 1983; Thiemann *et al.*, 2001; Vrugt *et al.*, 2003b) to elucidate appropriate probability density functions (PDF) for parameters and model outcomes. Alternatively, a variety of non-statistical, or informal, uncertainty estimation methods have also been developed in the hydrological literature, which provide the model output uncertainty as an approximation to the formal predictive uncertainty generated by statistical methods. Many researchers and practitioners have applied these techniques for model calibration and uncertainty estimation (Abebe *et al.*, 2010; Beven and Binley, 1992; Breinholt *et al.*, 2013; Liu *et al.*, 2012; Mannina, 2011; Setegn *et al.*, 2010; Tolson and Shoemaker, 2008; van Griensven and Meixner, 2007; Zhang *et al.*, 2012).

More recently, the notion of "limits of acceptability" has been suggested to identify the behavioural solutions in GLUE considering certain limits calculated based on the observation errors (Beven, 2006; Liu *et al.*, 2009). In the signature-based GLUE implementations (Winsemius *et al.*, 2009), the limits of acceptability are defined based on the satisfaction of signature-based measures, the parameter sets that can satisfy all signature-based measures are identified as behavioural. Subsequently, the band of the model output encompassed by all behavioral parameter sets is derived, and is referred to as "plausibility intervals" (Winsemius *et al.*, 2009).

The present study generates plausibility intervals using parameter sets sampled by one trial of an optimization algorithm to quantify the model output uncertainty. Given N parameter sets sampled during the calibration, a number of these solutions, say K, might satisfy all signature-based measures, and can be used to derive the plausibility intervals. However, when the computational budget is limited (in most of case studies in this paper, only 5000 samples are considered), it may not be possible to find a satisfactory number of solutions satisfying all signature-based measures. Thus, we suggest sorting the sampled

solutions with respect to hydrological consistency, and identify the behavioural solutions using a proper identification strategy. Chapter 4 (Shafii *et al.*, 2014a) concludes that GLUE applications should consider multiple behavioural solution identification strategies, and select the best one with respect to both reliability (coverage of observations) and sharpness (width of the uncertainty bound). This recommendation is assumed to apply to other informal approaches to uncertainty estimation, such as utilizing the multiple solutions identified in any of the non-GLUE based approaches. As in Chapter 4, the best identification strategy is taken to be the one associated to the minimum distance to ideal (DTI) metric in the reliability-sharpness space, formulated as:

$$DTI = \sqrt{(100 - R)^2 + (100 - S)^2} \qquad (27)$$

where R is the reliability and S is the sharpness value of a potential set of behavioural solutions for the model calibration period. In each case study, multiple identification strategies (described in Table 2) are applied, and the best one (with lowest DTI) is selected. Note that some of the strategies in Table 2 (based on NSE and DTI metrics) are consistent with previous GLUE applications in the literature (*e.g.,* Arabi *et al.*, 2007; Blazkova *et al.*, 2002; Dean *et al.*, 2009; Freer *et al.*, 1996; Gallart *et al.*, 2007; Lamb *et al.*, 1998; Mitchell *et al.*, 2009; Muleta and Nicklow, 2005; Shafii *et al.*, 2014a), and the consistency-based strategies are developed in this study.

When all identification strategies are applied to each case study and the best one is selected, the corresponding behavioural solutions are used to generate the envelope of model simulations in the validation, rather than calibration, period to characterize the model output uncertainty.

Table 2. Multiple behavioural solution identification strategies considered in the uncertainty estimation part of the proposed signature-based methodology

Strategy	Behavioural Solution Identification Process
NSE-Based	- Select the solutions with NSE greater than Th_{NSE} *
NSE-Based-Top10%	- Sort all solutions based on average NSE of low- and high-flows - Select the top 10% (500) solutions
CNS-Based-Top10%	- Sort all solutions based on the consistency measure - Select the top 10% (500) solutions set
CNS-NSE-Based-Top10%	- Identify the solutions with NSE greater than Th_{NSE} * - Sort the identified solutions based on the consistency measure - Select the top 500 solutions
DTI-Based **	- Sort all solutions based on the consistency measure - Round down the consistency values (resulting in the number of signatures satisfied by each solution, called N_S) - Considering the largest value, $N_{S,\max}$, identify the solutions that can satisfy at least $N_{S,\max}$ signatures; add these solutions to a set $S_{N_{S,\max}}$ - Calculate reliability, sharpness, and DTI of $S_{N_{S,\max}}$ - Repeat the previous two steps for $N_{S,\max}-1$, $N_{S,\max}-2$, …, 0 - Find the lowest DTI value - Select the set associated to the lowest DTI as the set of behavioural solutions

* Different thresholds are used in the case studies of this paper; see the results section.

** More details about the DTI-based strategy is provided in Chapter 4 (Shafii *et al.* (2014a)).

5.2.4 Case Studies

Calibration of four rainfall-runoff models is considered in this study, ranging from lumped to distributed with low to high number of parameters. The first model is HYMOD, a 5-parameter rainfall excess lumped model (Moore, 1985) 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.*, 2003b). 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 validation period in both watersheds. Appendix D provides more information about HYMOD, its parameters, and the catchments studied in this chapter.

The second model is HBV, originally developed by Swedish Meteorological and Hydrological Institute (SMHI) in the 1970s (Bergström, 1976). Recently, Aghakouchak and Habib (2010) have 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 scale 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, two years of data are used as the calibration period, and one year as the validation period. The HBV model used in this chapter, its parameters, and the catchments are detailed in Appendix D.

The third rainfall-runoff model used in this study is a semi-distributed 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 five years of data for calibration and one year for validation, and the Baron River watershed in the United States using four years of data for calibration and one year for validation. Application of WetSpa to these catchments is previously studied in the literature (Bahremand *et al.*, 2007; Liu *et al.*, 2003; Safari *et al.*, 2009; Shafii and Smedt, 2009). The first and second WetSpa case studies require 11 and 6 parameters, respectively, to be estimated in the calibration stage. More information on the WetSpa model and the catchments are provided in Appendix D.

The forth model is SWAT2000 simulation model (Neitsch *et al.*, 2001), called 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). We use SWAT 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 to be estimated for this catchment (Tables 2 and 3 in Tolson and Shoemaker (2007)). The simulation period consists of 92 days as the warm-up period, two years as the calibration period, followed by 274 more days as the validation period. Appendix D details the SWAT model, its parameters, and the catchment in the study.

5.3 Results

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, 1000 simulations (Yilmaz *et al.*, 2008) to three 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 the calibrated models can satisfy at least a set of signatures. Note that identical computational budget was considered for all formulations. In Monte Carlo experiments, a larger computational budget (*i.e.,* increased by one order of magnitude or 50000 simulations) is also considered in each trial.

Three different scenarios are considered in defining the acceptability levels. Given that the plausible values for all signature-based measures fall in the range $[-100\,\%, 100\,\%]$, the first scenario uses the interval $[-20\,\%, 20\,\%]$ as the acceptability interval for all signature-based measures. In this scenario, the acceptability threshold for signature-based measures is shown by ±20%. Moreover, in terms of the threshold for NSE (in cases where NSE is translated into a binary criterion), 0.5 is used in the case of ±20%. Similarly, the second and third scenarios consider threshold combinations ±10% and 0.7, and ±5% and 0.7, respectively. The last scenario applies the strictest thresholds (the most difficult problem) in the consistency-based calibration problem. When the sampling process terminates, the entire set of solutions evaluated in each trial is post-processed to evaluate criteria satisfaction rates in that trial (how many solutions satisfy 15 criteria, how many solutions satisfy 14 criteria, etc.).

Figure 20 illustrates the average results of such an analysis 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 18. The vertical axis in Figure 20 shows the average proportion of samples in five trials, and the horizontal axis demonstrates the proportion of satisfied criteria. For example, in the top left panel of Figure 20 (HBV (IL) 20%), A1-Opt formulation could only satisfy 87% (or 13) of the 15 criteria (13 signature-based measures and 2 NSEs) at best and about 0.7% (or 35) of the solutions evaluated during the search satisfied 13 signatures. Note that in these 35 solutions, not all solutions would satisfy the same 13 signatures. The vertical axis is logarithmic to show the difference among small proportions more properly. The left, middle, and right columns of graphs show the results for the first (±20%), second (±10%), and third (±5%) set of acceptability thresholds.

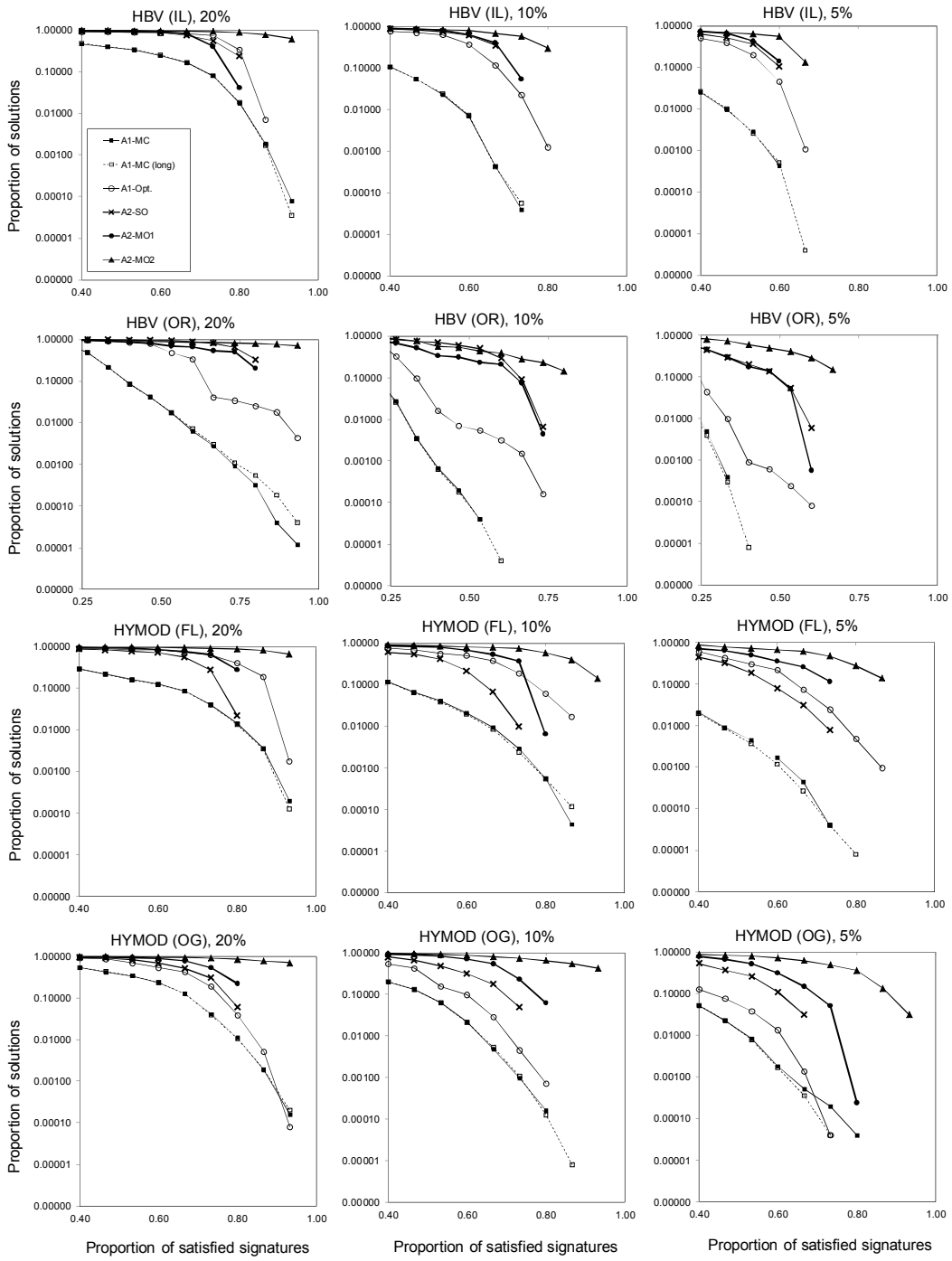

Figure 20. Proportion of solutions versus proportion of satisfied measures obtained in all case studies in different formulations. Case studies include HBV applied to two catchments named OR and IL, HYMOD in two catchments named FL and OG, WetSpa in two catchments named HO and BA, and SWAT in a catchment named TO. Formulations include: A1-MC and A1-MC (long): Monte Carlo simulations and signature-based post-process along 5000 and 50000 simulations, respectively; A1-Opt.: calibration using AMALGAM with two criteria, NSEs of low-and high-flows, followed by signature-based post-processing; A2-SO: criteria-aggregated single-objective calibration using DDS; A2-MO1: bi-criterion calibration using AMALGAM with two criteria, average NSE and the consistency measure; A2-MO2: Pareto-based multi-criteria calibration using AMALGAM with 15 criteria.

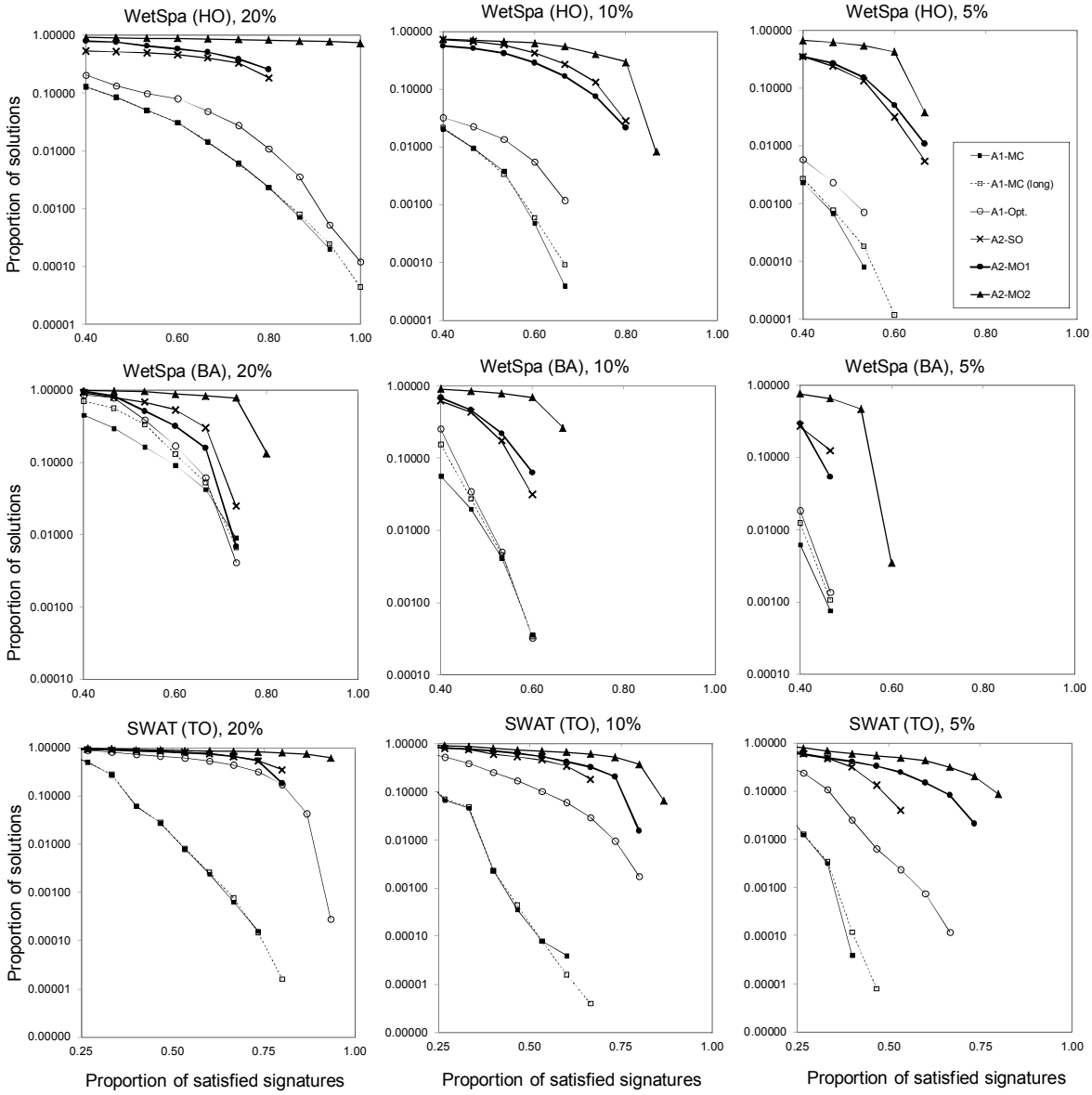

Figure 20. Continued.

Figure 20 shows that A2-MO2, considering continuous NSE and binary signature-based criteria, yields the largest proportion of measures satisfied by a higher number of solutions 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 proposed formulation. Moreover, it is observed in Figure 20 that both A2-SO and A2-MO1 perform more promising than the approach A1 in all cases. Overall, this finding indicates that, in comparison to the traditional approach of signature-based

calibration, 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 21 shows the maximum number of satisfied signature-based measures 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 criteria in total, 13 signature-based measures and 2 NSEs). Each graph in Figure 21 is associated to one case study. Figure 21 shows that A2-MO2 yields the highest consistency among all formulations. This finding indicate that when satisfaction of signature-based measures is considered explicitly in the search for highly consistent parameter sets, the high consistency regions in the search space can be explored more effectively. Figure 21 also demonstrates that the single-objective approach shows substantially higher variability of results in some cases (HYMOD-FL and WetSpa-HO), which might be due to aggregation of criteria in this approach. So while the single objective approach may do relatively well compared to the multi-objective approaches in one trial, the approach may be quite poor in another trial.

Figure 22 provides the results of evaluating the impact of tying NSE and signature-based measures on parameters consistency for two example case studies (SWAT-TO 5%, and HYMOD-FL 5%). In the results shown in Figure 22, the signature-based measures are separated from NSE metrics. The horizontal axis in Figure 22 shows different number of hydrological signatures from 1 to 13. The vertical axis shows the proportion of solutions that satisfy the NSE metric; how these proportions are calculated is explained as follows. For each number of hydrological signatures, say *K*, results are determined by post-processing the evaluated parameter sets (among 5000) that can satisfy at least K hydrological signatures. The proportion of these solutions satisfying K signatures that also satisfy the NSE metrics (here this is taken to be an NSE greater than or equal 0.7 in case of 5% acceptability threshold), is then determined and used as the Y coordinate of each point. For example, in the left panel and looking at the results of A1-Opt., the first point on the line indicates that 68% of the solutions that satisfy at least 1 signature-based measure will satisfy the NSE metric as well.

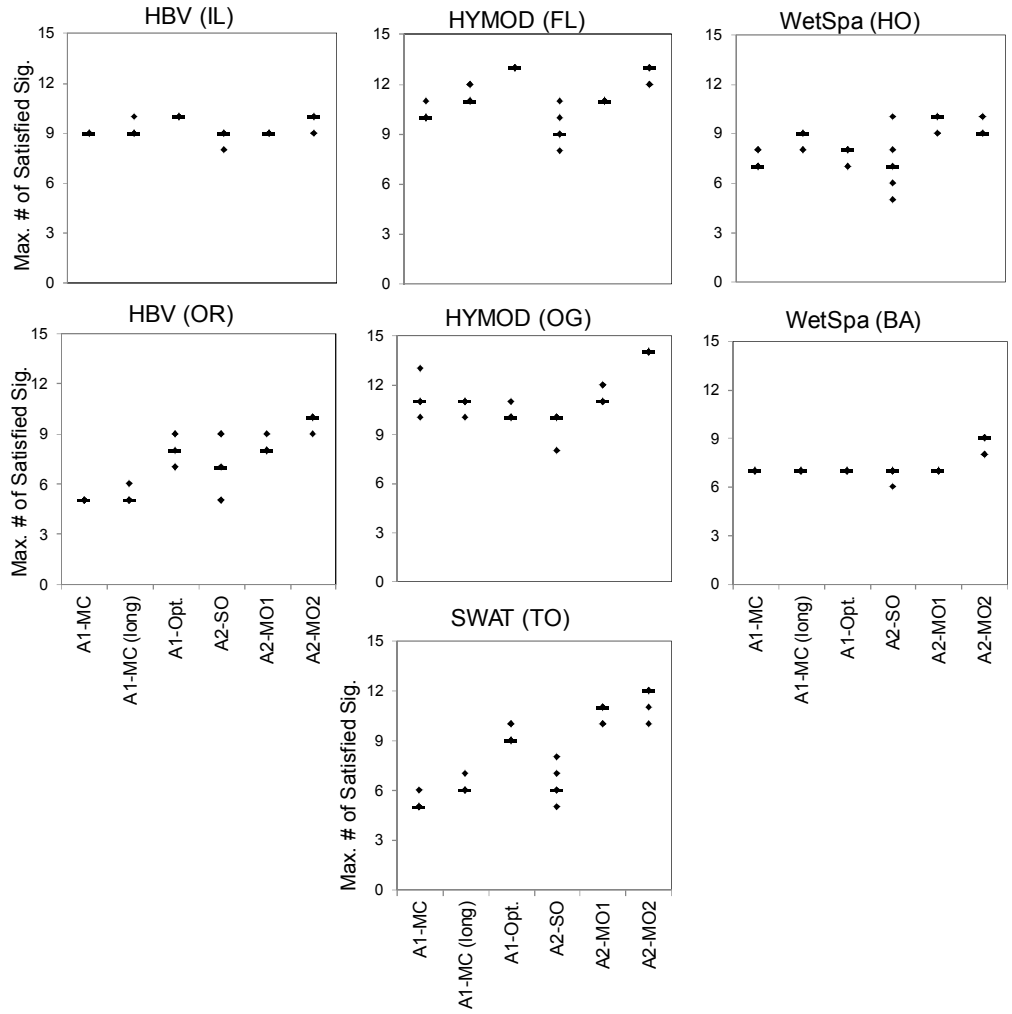

Figure 21. Maximum number of satisfied criteria obtained in five independent trials (solid points) in all case studies and the median of these results (horizontal bar), given the acceptability threshold of 5%. Case studies and formulations are described in the caption of Figure 20.

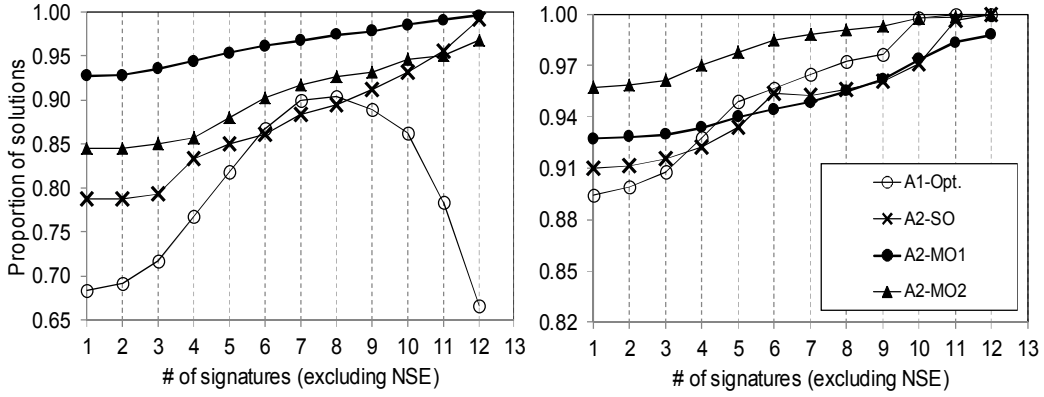

Figure 22. Proportion of solutions satisfying NSE-based measures versus number of satisfied signature-based measures from 1 to 13 indicating increase of hydrological consistency. Formulations are described in the caption of Figure 20. Left panel is SWAT-TO 5% case study and right panel is HYMOD-FL 5% case study.

The left panel in Figure 22 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. As shown in the right panel in Figure 22, approach A1 can yield increasing proportions as well. Overall, all the results show that the approach A1 results in variations similar to the left panel in eight case examples out of 21, seven case studies and three acceptability thresholds in each case. In contrast, the formulations A2-SO, A2-MO1, and A2-MO2 demonstrate near-monotonically increasing proportions in all case studies. The A2 formulations thus effectively tie NSE and hydrological signature-based measures during calibration so that both improve together. These overall results indicate that a multi-objective approach to maximizing NSE followed by post-processing can result in a proportion of parameter sets that may appear consistent based on hydrologic signatures but show relatively poor statistical performance.

Figure 23 illustrates the consistency measure on the vertical axis (Equation (22)) 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 panels correspond to ±20% (top), ±10% (middle), and ±5% (bottom) acceptability thresholds. The two metrics shown in Figure 23 are explicitly used in A2-MO1 as the two calibration criteria. For other formulations, first these two metrics are calculated for all 5000 samples, and then, the solutions are sorted based on the non-domination sorting scheme with respect to these two criteria. To make the figure visually presentable, Figure 23 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-MO2 considering a large computational budget, Figure 23 also demonstrates the Pareto front solutions obtained after 50000 simulations.

Figure 23 demonstrates that NSE values obtained in formulation A1-Opt (which is solely focused on optimizing the NSE values) are just slightly improved over the maximum NSE values of other formulations. In contrast, A1-Opt cannot find consistency measures that are comparable to that of A2-MO2. Figure 23 also shows that increasing the computational budget in A2-MO2 makes it possible to get more consistent results with relatively higher NSE values.

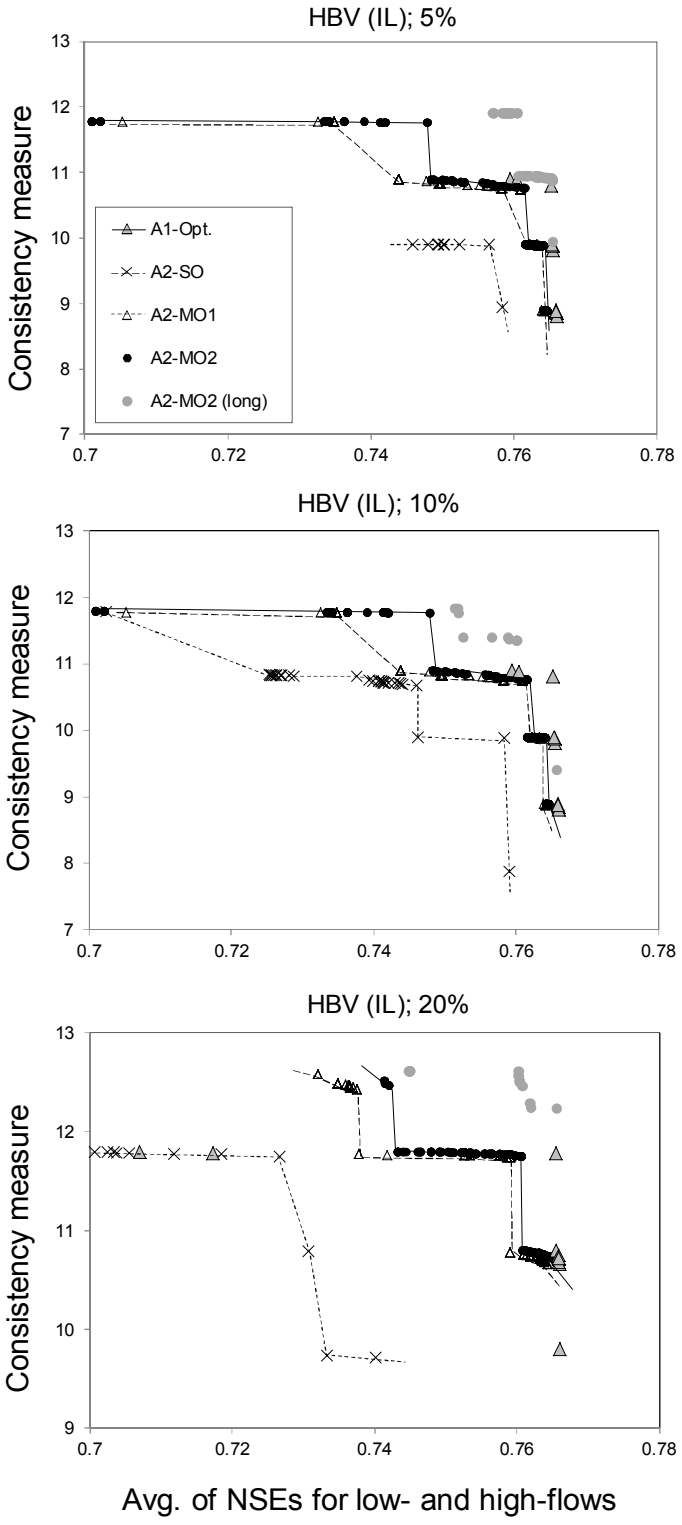

Figure 23. Consistency measure versus average NSE for low- and high-flows obtained from different formulations in the HBV case study applied to the catchment in Illinois (IL). Three acceptability levels are ±20% (top panel), ±10% (middle panel), and ±5% (bottom panel). Formulations are described in the caption of Figure 20. A2-MO2 (long) is the same as A2-MO2, but with the computational budget of 50000 simulations.

5.3.1 Sensitivity to Optimization Algorithms

The results illustrated previously are based on application of one set of optimization algorithms in different formulations. To investigate the sensitivity of the findings to the algorithms performance, 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) within the approach A1, and PADDS (Asadzadeh and Tolson, 2013) to calibration of HBV in the Illinois catchment (IL) within A2-MO1 and A2-MO2. Figure 24 illustrates the proportion of satisfied criteria versus the average proportion of samples in five trials. The upper panel in Figure 24 shows the results of the WetSpa case study in A2-SO. The lower panel in Figure 24 demonstrates the results of the HBV case study in A2-MO1 and A2-MO2.

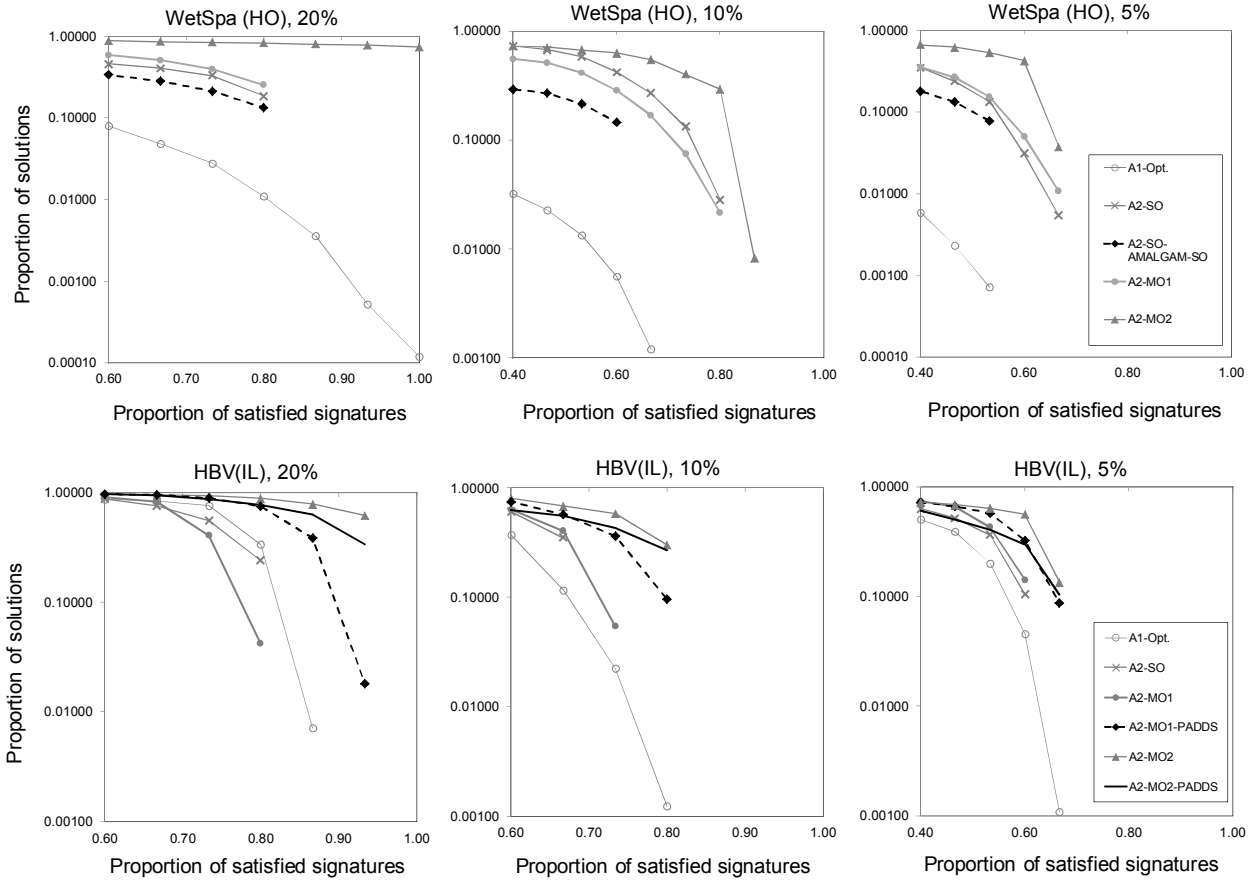

Figure 24. Proportion of solutions versus proportion of satisfied measures obtained in the HBV (upper panel) and WetSpa (lower panel) case studies. Formulations are described in the caption of Figure 20. New optimization algorithms applied in the experiments include AMALGAM-SO in A2-SO, PADDS in A2-MO1 and A2-MO2.

Figure 24 indicates that, in the WetSpa case study, the application of AMALGAM-SO does not improve over DDS results (results are somewhat degraded). Moreover, in the HBV case study, PADDS and AMALGAM perform fairly similarly in A2-MO2, whereas PADDS performs more satisfactorily than AMALGAM in A2-MO1. Nevertheless, in both cases, A2-MO2 still shows to be the most promising approach among the formulations considered in this study. Hence, the overall conclusions do not change when one alternative algorithm is considered, and the full multi-criteria formulation (A2-MO2) remains as the most promising formulation with respect to hydrological consistency.

5.3.2 Uncertainty of Model Outcome

This study does not aim to develop a new uncertainty estimation methodology, and instead, the main objective is to show that the proposed sampling-based calibration method has the ability of accommodating uncertainty estimation. Following the uncertainty estimation approach described in Section 5.2.3, this section provides the uncertainty bounds associated to the results of A2-MO2 based on the signature threshold of ±5%. Multiple behavioural solution identification strategies (Table 2) are applied and the one with the lowest DTI is selected in each case study. Note that, in the 'NSE-Based' strategy, we consider the following NSE thresholds for both flows and log-flows: 0.5 in HBV and HYMOD, 0.25 in SWAT, and 0 in WetSpa case studies. The selected identification strategies (based on DTI metric) are as follows: 'NSE-Based' in HBV-IL, HYMOD (both catchments), and WetSp (both catchments); 'NSE-Based-Top10%' in HBV-OR; and 'CNS-NSE-Based-Top10%' in the SWAT case study.

Figure 25 illustrate the model output uncertainty in validation period for all seven case studies. Table 3 provides the sharpness and reliability values associated to the uncertainty bounds generated by the proposed methodology, and the bounds associated to the Pareto front solutions obtained after sorting the entire solutions set based on non-domination sorting scheme with respect to NSE of flows and log-flows. In multi-criteria calibration, the Pareto front solutions are used to reflect uncertainty propagation through the model (Gupta *et al.*, 1998). Pareto uncertainty defines the minimum uncertainty in the parameters that can be achieved without stating a subjective relative preference for minimizing one specific objective function at the expense of another (Vrugt *et al.*, 2003a).

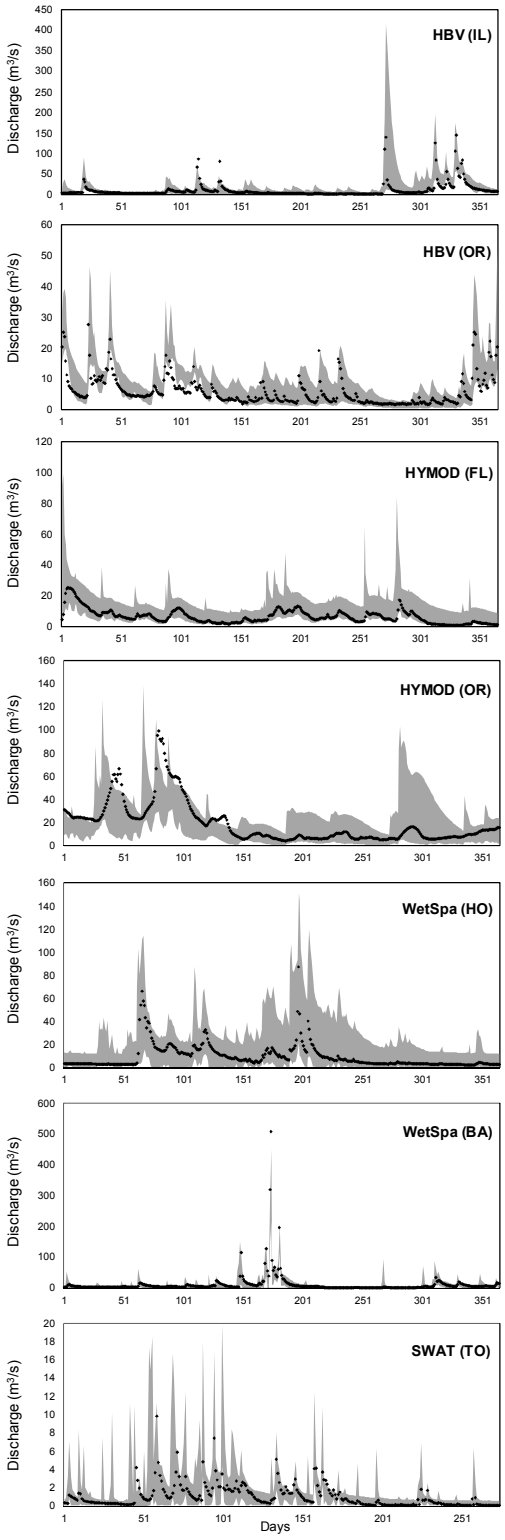

Figure 25. Model output uncertainty bounds in validation period associated to the behavioural solutions identified in the proposed methodology in all seven case studies.

Table 3. Sharpness(SHP) and reliability (REL) associated to the uncertainty bounds generated by the proposed methodology, and the Pareto optimal solutions (with respect to NSE of flows and log-flows)

	Proposed		Pareto	
	SHP	REL	SHP	REL
HBV-IL	5.6	95.6	92.1	30.1
HBV-OR	16	82.5	89.3	6
HYMOD-FL	75.6	98	99.2	1.1
HYMOD-OG	79.9	83.3	96.2	23
WetSpa-HO	39.5	94.5	93	13.2
WetSpa-BA	16.4	78.4	82	31
SWAT-TO	10.2	95.6	82.6	36.9

Table 3 shows that whereas the Pareto uncertainty proves to be too narrow in the validation period (expected as the resulting bounds are related only to the optimal Pareto front), the bounds generated by the proposed method yield high coverage. Table 3 shows that for reasonable uncertainty bounds to be created, the entire search history of A2-MO2 must be used rather than just the Pareto solutions.

5.4 Concluding remarks

This study develops a multi-criteria calibration methodology where both statistical metrics such as NSE and hydrological signature-based measures are implemented as explicit objectives in the optimization formulation of the calibration problem. Different formulations are provided and multiple optimization algorithms are applied to take samples from parameter space. The formulations include single-criterion optimization whereby all criteria including NSE and signature-based measures are aggregated, Pareto-based multi-criteria optimization implementing individual criteria without any aggregation, and bi-criterion optimization where the first and second criteria aggregate statistical and hydrological measures, respectively. These formulations are compared with the traditional signature-based model evaluation in the literature that post-processes the calibration result against hydrological signatures. The comparison among calibration formulations is on the basis of hydrological consistency of the parameter sets quantified in terms of a consistency measure. The proposed calibration formulations as well as the traditional formulation are applied to calibration of four rainfall-runoff models in seven catchments using a set of optimization algorithms.

The results show that the full multi-criteria Pareto-based model calibration yields the highest level of hydrological consistency in all case studies in comparison to Monte Carlo sampling and traditional

calibration approaches that only use statistical metrics. In other words, consideration of hydrologic signatures throughout the calibration experiment enhances the chance of finding fully hydrologically consistent parameter sets. Moreover, this formulation results in a larger number of consistent parameter sets. Superiority of this formulation in comparison to traditional formulation is more significant in calibration problems with stricter levels of acceptability (*i.e.,* 5%). This finding highlights the advantage of multi-criteria optimization concepts in calibration of hydrologic models towards a higher level of hydrological consistency.

This study also evaluates the sensitivity of the obtained results to the type of optimization algorithms by using different single- and multi-criterion 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 hydrological consistent remains the same when an alternative algorithm is selected. Therefore, the findings depend more on the formulation as opposed to the optimization algorithm.

This study also describes the uncertainty of model output using the results of signature-based model calibration where the parameter sets sampled by the optimization algorithm during the calibration stage are employed. The uncertainty bounds generated by the proposed method are compared with those associated to the Pareto optimal solutions (after non-domination sorting based on NSE of low- and high-flows). While the latter show to be too narrow with low reliability, the former cover the observations in the validation period much more satisfactorily. The uncertainty estimation method proposed in this study is only to show the potential of using the samples taken in the calibration process to quantify the model output uncertainty. To better characterize the predictions uncertainty will be the focus of future research. More specifically, research should explore uncertainty assessment in the presence of both epistemic and aleatory uncertainty, recently discussed among researchers (Beven *et al.*, 2012b; Beven *et al.*, 2011; Clark *et al.*, 2012a; Montanari, 2011). Aleatory uncertainty is due to randomness of a stochastic process, whereas epistemic uncertainty stems from lack of knowledge about processes and boundary conditions.

Recently, and as mentioned in the introduction part of the paper, 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 modelling. In the present paper, we did not implement the ABC approach, because the ABC concept is very recent to hydrology and is only demonstrated on a few lumped modelling (models with very small computational burden) case studies. The traditional methods we implemented are more established as they have been applied in multiple 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.

Moreover, because the signature-based model calibration constrains the model output from the hydrological perspective, it provides the modeller with an insight into the performance of different modules in the simulations 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 signature-based criteria are considered (similar to this study), the modeller 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 deficiencies.

Chapter 6
Summary, Conclusions, and Recommendation for Future Work

This thesis proposes improvements to informal uncertainty estimation methods for multi-criteria hydrologic model calibration. The improvements are applied to the GLUE method, however, they can be applied to other informal methods that quantify the model output uncertainty utilizing a set of behavioural solutions. Conclusions from each chapter below are based on experiments using two to seven hydrologic model case studies applying two to four different hydrologic models (including lumped and spatially distributed models). Therefore, the findings should extend to other case studies and other models. The following sections outline the conclusions as well as recommendations for future work. Note that chapters 1 and 2 provide an overview of the literature as well as the review of uncertainty estimation techniques.

6.1 Summary and Concluding Remarks

Chapter 3 compares formal (Bayesian inference) and informal (GLUE) multi-criteria calibration and uncertainty estimation methods in hydrological modelling with respect to computational considerations as well as the uncertainty quantification point of view. GLUE is implemented considering two scenarios, using a restricted computational budget and the full computational budget. Furthermore, MCMC sampling with a restricted computational budget (*i.e.,* non-converged sampling) is also considered as an alternate informal uncertainty estimation method. The following conclusions are drawn in Chapter 3:

- The results demonstrate that it is advisable to consider multiple comparative metrics (measuring the predictive capacity) when comparing alternative calibration strategies. Furthermore, the choice of using the validation or calibration period for selected comparative measures would influence the analysis, and thus, it is recommended that future comparative studies should include and largely focus on comparative performance assessment in the validation period.
- The results also reveal that, in the multi-criteria context, there can be considerable differences in prediction intervals generated by Bayesian inference and GLUE applied to uncertainty estimation of rainfall-runoff models. Future uncertainty quantification studies for simulation models with a large number of parameters and considering multiple responses (or criteria) should be aware of the potential differences between the results of formal and informal methods.
- In general, considering the same computational budget, the Bayesian inference methodology performs well in comparison with GLUE (*e.g.,* prediction intervals are simultaneously tighter and more reliable than corresponding GLUE intervals) with respect to all comparative measures except for low-flows in the validation period. The failure of Bayesian methods in the satisfaction of assumptions associated with low-flows suggests that using more complex hierarchical

Bayesian structures may be warranted, which is outside the scope of this thesis. In case of a limited computational budget, non-converged MCMC sampling proves to be fairly consistent with formal Bayesian inference. This indicates the potential value of utilizing formal MCMC sampling results before convergence as a promising alternative to informal methods such as GLUE.

Chapter 4 is the first study to evaluate the impact of behavioural solution identification strategies on the model output uncertainty estimation in GLUE. New strategies for the identification of behavioural solutions are proposed (*i.e.,* based on simultaneous optimization of reliability and sharpness of the prediction intervals). GLUE is also compared with two other informal multi-criteria calibration and uncertainty estimation methods (*i.e.,* Pareto optimization and DDS-AU) with respect to the reliability and width of the model output uncertainty bounds. Chapter 4 yields the following conclusions:

- The results indicate that the subjectivity involved in behavioural solution identification strategies impacts the uncertainty of model outcome, and more importantly, no single strategy is shown to be the best in all case studies.
- Thus, to have a robust implementation of GLUE, it is recommended that future studies consider multiple behavioural solution identification strategies and choose the best one based on calibration period performance assessed according to the subjective modeller's priorities.
- Results also reveal that the new objective strategies proposed in the research are among the best options in all case studies investigated in this thesis. Thus, it is recommended that the two new strategies for identification of behavioural solutions strategies of this research be considered among the set of behavioural solution identification strategies in future GLUE applications.

Chapter 5 develops new many-objective calibration formulations and sampling strategies where both standard performance metrics, *e.g.,* Nash-Sutcliffe Efficiency (NSE), and hydrological signature-based model evaluation measures are explicitly considered at all phases of the calibration process to enhance the chance of obtaining more hydrologically consistent parameter sets. The proposed approaches are compared with the traditional signature-based model calibration approaches in the literature. The comparison is on the basis of hydrological consistency of the parameter sets, which is quantified in terms of a consistency measure. Chapter 5 also demonstrates an approach for the estimation of model output uncertainty adapted for the consistency-based calibration formulation proposed in the study. The behavioural parameter sets sampled by the optimization algorithm during the calibration stage are employed to generate model output uncertainty bounds. To identify the behavioural solutions, multiple

strategies are considered (as suggested in the previous chapter), and the most appropriate strategy is used to quantify model output uncertainty in validation period. Specific conclusions are as follows:

- The proposed calibration approach utilizes the optimization concepts to provide the modeller with the opportunity of parameter adjustment considering many criteria (15 in this study). Such a many objective model calibration is novel in the hydrological modelling literature.
- The results show that the full multi-criteria Pareto-based model calibration (*i.e.,* closely reproducing each measured hydrologic signature as a separate objective function) yields the highest level of hydrological consistency in all case studies, especially in calibration problems where the modeller has strict levels of acceptability for the calibration criteria. The results also demonstrate that the findings are consistent for two alternative optimization algorithms. Thus, it is suggested that future signature-based model calibration studies utilize hydrological signatures directly during parameter adjustment process, and furthermore, avoid criteria aggregation, and instead, implement Pareto-based multi-criteria optimization to calibrate the simulation model.
- Comparison between the proposed and traditional signature-based model calibration (via optimization of statistical measures) reveal that, unlike the former approach, the latter approach (*i.e.,* when statistical and hydrological measures are not tied together) can result in a proportion of parameter sets that may appear consistent based on hydrologic signatures but show relatively poor statistical performance.
- Comparison between the uncertainty bounds generated by the proposed uncertainty estimation approach and those obtained based on the parameter sets that belong to the Pareto front (*i.e.,* upon non-domination sorting against NSE of flows and log-flows) reveals that the latter are too narrow with very low reliability or coverage of observations in the validation period. In contrast, the bounds of the proposed method are wider and properly cover the observations in the validation period.

To demonstrate different aspects of the methods explored in this research, comparisons were conducted with a minimum of 10,000 model simulations and thus relatively fast hydrologic simulation models were utilized. Since this many simulations may be implausible for some computationally intensive simulation models, such as fully distributed hydrologic models, more testing of these methods at lower computational budgets and for fully distributed model calibration problems is recommended.

Recommendations for Future Research

In the comparison between converged and non-converged MCMC sampling (Chapter 3), the results suggest that the Gelman-Rubin convergence criterion is too stringent since non-converged MCMC sampling results closely approximate converged sampling results and yet require a fraction of the computational budget. Therefore, alternative convergence measures need to be considered in parallel with the Gelman-Rubin metric. A potential hydrology-based convergence metric can be the reproduction of hydrological signatures, as explored in Chapter 5. Because these signatures represent the link between the model and overall hydrologic behaviour of the catchment, sampling can be terminated provided that certain hydrological indices are reproduced properly. Future research should explore such signature-based measures and other alternative convergence metrics as the stopping criteria for MCMC sampling used for uncertainty estimation of hydrologic models.

The results in Chapter 3 indicate a performance difference between the calibration and validation period. The literature shows that this difference might be due to epistemic uncertainties involved in hydrological modelling (Beven *et al.*, 2012b; Beven *et al.*, 2011; Clark *et al.*, 2012a; Montanari, 2011), which makes both standard Bayesian formulation and the informal GLUE methodology not perfectly reliable in prediction mode in the presence of epistemic uncertainty. Therefore, future research should explore the application of more proper formal uncertainty estimation approaches, *e.g.,* the use of hierarchical Bayesian structures (*e.g.,* Huard, 2008; Kuczera *et al.*, 2006; Moradkhani *et al.*, 2005; Renard *et al.*, 2010; Wei *et al.*, 2010), or more appropriate identification of behavioural models in GLUE, *e.g.,* the concept of 'limits of acceptability' (Blazkova and Beven, 2009; Liu *et al.*, 2009) in cases with epistemic uncertainties. Moreover, the comparison between such approaches should be the topic of future studies.

Chapter 4 uses the concept of distance to ideal point (DTI) (calculated based on reliability and sharpness of predictions) as a diagnostic to evaluate and compare informal calibration and uncertainty estimation methods. It is also expected that DTI can serve as a convergence metric for informal sampling-based uncertainty estimation methods. If the objective of an informal uncertainty estimation methodology is translated into an attempt to approach (as much as possible) the ideal point in sharpness-reliability space, future studies should investigate using DTI as a guiding metric to assess the sampling convergence. Example methods that can benefit from this idea include sequential sampling strategies such as SUFI (Abbaspour *et al.*, 2004) or SMC (Jeremiah *et al.*, 2011), and furthermore, informal implementations of MCMC samplers (Blasone *et al.*, 2008a; Vrugt *et al.*, 2003a).

Chapter 5 develops a calibration and uncertainty estimation methodology where many signature-based criteria can be implemented. Because the proposed method uses the signature-based measures during the

calibration phase, it provides the modeller with the opportunity of quickly evaluating the model structure, *i.e.,* through monitoring which signatures are not satisfied along simulations. Therefore, the proposed method can be expanded in future studies to an interactive calibration framework where the modeller can modify certain modules in the simulation model based on observations of signature-based measures as the calibration algorithm is running.

Appendix

A1 – Review of AR-Based Bayesian Inference Procedure

To avoid biasing parameter estimates, auto-correlation of the model errors must be explicitly accounted for during parameter estimation. One approach to remove such autocorrelation is to use a first-order Auto-Regressive (AR) scheme for the error series:

$$\varepsilon_i = \rho\varepsilon_{i-1} + \delta_i \quad i = 1,...,N \qquad (A1)$$

where $\varepsilon_0 = 0$, ρ is the first-order correlation coefficient, and δ_i is the remaining error, which is prescribed to have a zero mean and constant variance σ_δ^2. Accordingly, the AR-based regression formulation is as follows:

$$Y_i = (Y_i^{sim} | \boldsymbol{\theta}, \mathbf{X}) + \rho\varepsilon_{i-1} + \delta_i \qquad (A2)$$

where Y_t and Y_i^{sim} are the observed and simulated values for the model response at time step i, $\boldsymbol{\theta}$ is the model parameters vector, $\mathbf{X}$ is the model inputs vector, ρ is the lag-one AR parameter, $\varepsilon_i = \left(Y_i - Y_i^{sim}(\boldsymbol{\theta}, \mathbf{X})\right)$ is the residual between observation and model prediction at time step i (and $\varepsilon_0 = 0$), and δ_i is random error term:

$$\delta_i(\boldsymbol{\theta}, \rho \,|\, \mathbf{Y}) = \varepsilon_i(\boldsymbol{\theta} \,|\, \mathbf{Y}) - \rho\varepsilon_{i-1}(\boldsymbol{\theta} \,|\, \mathbf{Y}) \quad i = 1,...,N \qquad (A3)$$

Sorooshian and Dracup (1980) incorporated this AR scheme into the formulation of a revised log-likelihood expression:

$$lp(\boldsymbol{\theta}, \rho, \sigma_\delta^2 | \mathbf{Y}) \propto \ln(p(\boldsymbol{\theta}, \rho)) - \frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln\left(\frac{{\sigma_\delta}^{2N}}{1-\rho^2}\right) - \frac{1}{2}(1-\rho)^2 {\sigma_\delta}^{-2}\varepsilon_1(\boldsymbol{\theta} \,|\, \mathbf{Y})^2 - \frac{1}{2}{\sigma_\delta}^{-2}\sum_{i=2}^{N}\delta_i(\boldsymbol{\theta}, \rho \,|\, \mathbf{Y})^2 \qquad (A4)$$

The AR-based error variance, σ_δ^2, may also be estimated using a chi-squared distribution, but with the scale calculated as:

$$s = \frac{1}{N}\left(\varepsilon_1^2(1-\rho^2) + \sum_{i=2}^{N}{\delta_i}^2\right) \qquad (A5)$$

Implementation of the AR scheme within an MCMC sampler is straightforward - a Jeffrey's prior is assumed for σ_δ^2 and this parameter is integrated out, moreover, a uniform prior is assumed for ρ and this hyper-parameters is appended to the parameter vector utilized within the sampler. Once σ_δ^2 is integrated out, the resulting joint posterior distribution of $\boldsymbol{\theta}$ and ρ would be

$$p(\boldsymbol{\theta},\rho|\mathbf{Y}) \propto \left(\varepsilon_1^2(1-\rho^2)+\sum_{i=2}^{N}\delta_i{}^2\right)^{-\frac{N}{2}} \qquad (A6)$$

A2 – AR-Based Bayesian Formulation in the Multi-Criteria Context

The AR-based scheme can be implemented in the multi-criteria context, as well. This section briefly describes the formulation; details can be found in previous studies (Balin-Talamba *et al.*, 2010; Schaefli *et al.*, 2007). Considering J responses, individual error series will be as follows:

$$\delta_i \sim N(0,\sigma_j^2) \qquad (A7)$$

with σ_j^2 being the residual variance for response j, here considered unknown and should be estimated, *i.e.,* J parameters (representing error variance for J responses) need to be estimated in the Bayesian inference methodology. Under the assumption of multiple and statistically independent responses, the combined statistical likelihood function for multiple responses is simply the product of the individual likelihood functions:

$$\begin{aligned} l_{multiple} &= \prod_{j=1}^{J} l_j(\boldsymbol{\theta},\rho,\sigma_j^2,\mathbf{X}) \\ &= \prod_{j=1}^{J} \frac{1}{\left(\sqrt{2\pi}\right)^{t_j}.\sigma_j^{t_j}}.\exp\left(-\frac{\sum_{i=1}^{t_j}\delta_{j,i}{}^2}{2\sigma_j^2}\right) \end{aligned} \qquad (A8)$$

where $\delta_{j,i}=\varepsilon_{j,i}-\rho\varepsilon_{j,i-1}$ for observation set j and time step i (note that $\varepsilon_{j,0}=0$), respectively; J is the number of observation sets, and t_j is the number of time steps for each observation set j. In order to derive the posterior distribution of parameters, a bounded uniform prior distribution is considered for $\boldsymbol{\theta}$ over prior feasible range, and the prior distribution of error variance is also considered to be Jeffrey non-informative distribution as follows:

$$p(\sigma_j^2) \propto 1/\sigma_j^2 \quad \text{for} \quad 0<\sigma_j^2<\infty \qquad (A9)$$

Using such prior distributions enables us to integrate out the error variances, and the Bayesian formulation results in the joint posterior distributions from which the marginal distribution of model parameters and error variances can be estimated conditioned on the observed data **Y**.

Alternatively, MCMC sampling can be used to directly take samples from the posterior distributions, all of which are contained in the chain. In MCMC implementations, the acceptance/rejection criterion ratio (between posterior densities of the new candidate and old current samples) is used to accept/reject the candidate to be added to the chain. In the multi-criteria Bayesian formulation, let $\sigma^2_{j,current}$ and $\sigma^2_{j,candidate}$ be the error variance of the current and candidate solutions, respectively, which are estimated based on the residuals after running the simulation model. Also assume the quantity $S_j = 0.5\sum_{i=1}^{t_j} \delta_{j,i}{}^2$, such that $S_{j,current}$ and $S_{j,candidate}$ be the values for the current and the candidate solutions, respectively. The final form of the acceptance/rejection criterion can then be shown as follows:

$$\alpha = \prod_{i=1}^{J} \exp\left[\left(\frac{1}{\sigma^2_{j,current}} + \frac{1}{\sigma^2_{j,candidate}}\right).\left(S_{j,current} - S_{j,candidate}\right)\right] . \left(\frac{S_{j,candidate}}{S_{j,current}}\right)^{\frac{t_j}{2}} \quad (A10)$$

B – Signature-Based Measures Applied in Chapter 5

This Appendix presents the mathematical formulations of 13 hydrological signature measures used in this study. Note that the numerical experiments in this study also consider two more criteria that are NSE (Equation (9)) of flows and log-flows. These signature measures are calculated using time series of observed flow (QO), log-transformed observed flow (qO), simulated flow (QS), and log-transformed simulated flow (qS):

1. The water balance diagnostic signature measure represented as the percent bias in overall runoff ratio considering flow (%BiasRR) and log-transformed flow (%Biasrr) calculated as follows:

$$\%BiasRR = \frac{\sum_{t=1}^{N}(QS_t - QO_t)}{\sum_{t=1}^{N} QO_t} \times 100 \quad (B1)$$

$$\%Biasrr = \frac{\sum_{t=1}^{N}(qS_t - qO_t)}{\sum_{t=1}^{N} qO_t} \times 100 \quad (B2)$$

2. The FDC-based diagnostic signature measures including the following three metrics:

(i) the percent bias in FDC mid-segment slope (%BiasFMS):

$$\%BiasFMS = \frac{[\log(QS_{m1}) - \log(QS_{m2})] - [\log(QO_{m1}) - \log(QO_{m2})]}{[\log(QO_{m1}) - \log(QO_{m2})]} \times 100 \quad (B3)$$

where m1 and m2 are the lowest and highest flow exceedance probabilities (0.2 and 0.7 respectively) within the mid-segment of FDC.

(ii) the percent bias in FDC high-segment volume (%BiasFHV):

$$\%BiasFHV = \frac{\sum_{h=1}^{H}(QS_h - QO_h)}{\sum_{h=1}^{H} QO_h} \times 100 \quad (B4)$$

where h = 1, 2,. . .H are flow indices for flows with exceedance probabilities lower than 0.02.

(iii) the percent bias in FDC low-segment volume (%BiasFLV):

$$\%BiasFLV = -1.\frac{\sum_{l=1}^{L}[(\log(QS_l) - \log(QS_L)] - \sum_{l=1}^{L}[(\log(QO_l) - \log(QO_L)]}{\sum_{l=1}^{L}[(\log(QO_l) - \log(QO_L)]} \times 100 \quad (B5)$$

where l = 1,2,. . .L is the index of the flow value located within the low-flow segment (0.7–1.0 flow exceedance probabilities) of the flow duration curve, L being the index of the minimum flow.

3. The discharge spectral diagnostic measures including the following seven metrics:

(i) the percent bias in the mean discharge (%BiasMeanQ):

$$\%BiasMeanQ = \frac{\overline{QS} - \overline{QO}}{\overline{QO}} \times 100 \quad (B6)$$

where $\overline{QS}$ and $\overline{QO}$ are average simulated and observed flow.

(ii) the percent bias in the discharge variance (%BiasVarQ):

$$\%BiasVarQ = \frac{Var(QS) - Var(QO)}{Var(QO)} \times 100 \quad (B7)$$

where $Var(QS)$ and $Var(QO)$ are the variance of simulated and observed flow.

(iii) the percent bias in the median discharge (%BiasMedQ):

$$\%BiasMedQ = \frac{M(QS) - M(QO)}{M(QO)} \times 100 \quad (B8)$$

where $M(QS)$ and $M(QO)$ are the median of simulated and observed flow.

(iv) the percent bias in the peak discharge (%BiasPeakQ):

$$\%BiasPeakQ = \frac{P(QS) - P(QO)}{P(QO)} \times 100 \quad (B9)$$

where $P(QS)$ and $P(QO)$ are the maximum of time series of simulated and observed flow,

(v) the percent bias in the lag-1 autocorrelation function (%BiasLag1):

$$\%BiasLag\ 1 = \frac{\rho(1)_S - \rho(1)_O}{\rho(1)_O} \times 100 \quad (B10)$$

where $\rho(1)_S$ and $\rho(1)_O$ are the lag-1 autocorrelation coefficient of simulated and observed flow.

(vi) the percent bias in the mean log-transformed discharge (%BiasMeanLogQ):

$$\%BiasMeanLogQ = \frac{\overline{qS} - \overline{qO}}{\overline{qO}} \times 100 \quad (B11)$$

where $\overline{qS}$ and $\overline{qO}$ are average simulated and observed log-flow,

(vii) the percent bias in the variance of log-transformed discharge series (%BiasVarLogQ):

$$\%BiasVarLogQ = \frac{Var(qS) - Var(qO)}{Var(qO)} \times 100 \quad (B12)$$

where $Var(qS)$ and $Var(qO)$ are the variance of simulated and observed log-flow.

4. The monthly discharge diagnostic signature measures calculated as follows (%BiasMonthQ):

$$\%BiasMonthQ = Max(D_M(i)), i = 1,2,...,12 \quad (B13)$$

where $D_M(i) = \frac{M_S(i) - M_O(i)}{M_O(i)} \times 100$ with $M_S(i)$ and $M_O(i)$ being the average monthly simulated and observed flow in month i.

C – Post-diagnostics of the Bayesian results in Chapter 3

Figures C1 and C2 show the post-diagnostic checks for the results of Bayesian inference in the HYMOD and WetSpa case studies, respectively, in Chapter 3. The top, middle, and bottom rows illustrate the sample auto-correlation function, residuals, and QQ plot of the residuals, respectively. The left and right columns provide the results of low- and high-flows, respectively. These results are associated to the parameter set with the highest posterior density (*i.e.,* posterior mode), which is obtained after 143,000 simulations in the HYMOD case study and after 470,000 simulations in the WetSpa case study. The results show that the assumptions are satisfied and MCMC has successfully converged in both case studies.

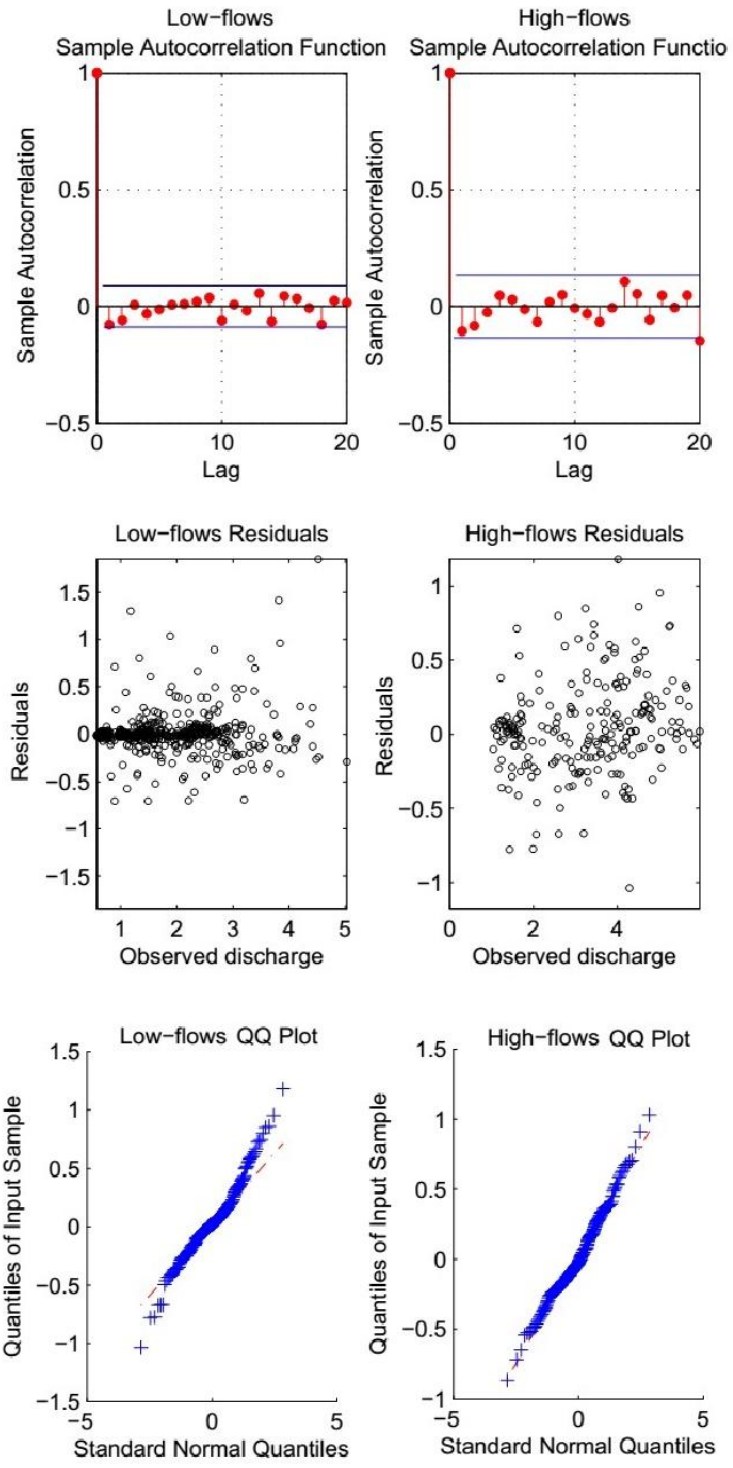

Figure C1. Post-diagnostic checks for the result of Bayesian inference of HYMOD in Chapter 3, including the sample auto-correlation function (top row), scatterplot (middle row), and QQ plot (bottom row) of the residuals. Left and right panels show the results of low- and high-flows, respectively.

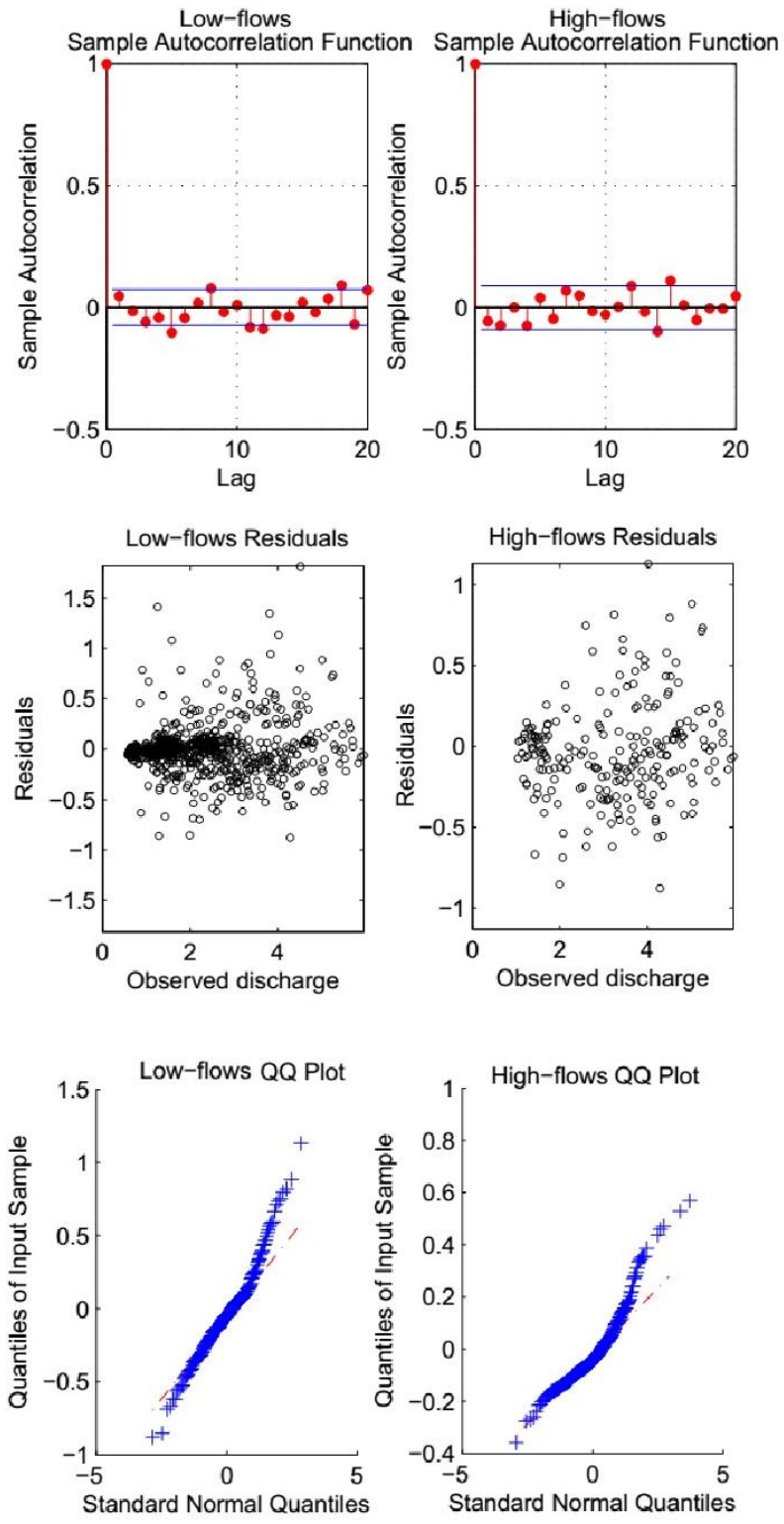

Figure C2. Post-diagnostic checks for the result of Bayesian inference of WetSpa in Chapter 3, including the sample auto-correlation function (top row), scatterplot (middle row), and QQ plot (bottom row) of the residuals. Left and right panels show the results of low- and high-flows, respectively.

D – Details on hydrologic models and the case studies used in the thesis

HYdrologic MODel (HYMOD)

HYMOD has a simple model structure consisting of a simple two-parameter rainfall excess model connected with two series of linear reservoirs (three, identical, for the quick and a single reservoir for the slow response) in parallel as a routing component (Figure D1). The model structure requires five parameters, which are described in Table D1, along with their prior ranges for the calibration purposes.

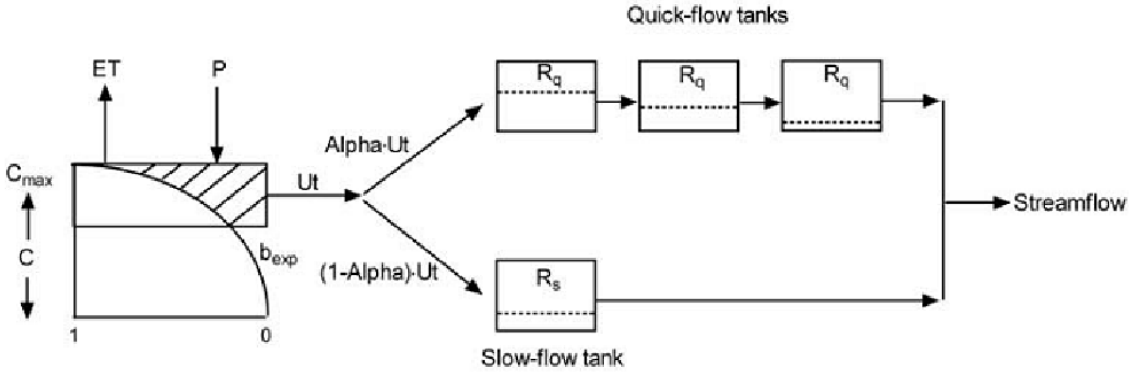

Figure D1. Conceptual structure of HYMOD; taken from Vrugt *et al.* (2003).

Table D1. HYMOD parameters and their prior range.

Parameter	Descritption	Unit	Prior Range
C_{max}	Maximum storage capacity	[mm]	1-500
b_{exp}	Degree of spatial variability of the soil moisture capacity within the catchment	[ ]	0.1-2
Alpha	Distributing factor on the flow between the two series of reservoirs	[ ]	0.1-0.99
K_q	Residence time of the quick reservoirs	[d]	0.1-0.99
K_s	Residence time of the slow reservoirs	[d]	0-0.1

The soil-moisture accounting module of HYMOD utilizes a storage capacity distribution function for the storage elements of the catchment. In this module, the storage elements of the catchment are distributed according to a probability density function defined by the maximum soil moisture storage (C_{max}), and the distribution of soil moisture stores. C_{max} represents the capacity of the largest soil moisture store, while the shape parameter (b_{exp}) describes the degree of spatial variability of the stores. Thus, the soil moisture storage capacity, c, is calculated as follows:

$$f(c) = 1 - \left(1 - \frac{c(t)}{C_{\max}}\right)^{b_{\exp}} \qquad 0 \le c(t) \le C_{\max} \qquad (D1)$$

Evaporation from the soil moisture store occurs at the rate of the potential evaporation estimates using the Hamon approach. Following evaporation, the remaining rainfall and snowmelt are used to fill the soil moisture stores. Subsequently, the rainfall excess model is sent to the routing module. The routing module divides the excess rainfall using split parameter (Alpha) and routes these through parallel conceptual linear reservoirs meant to simulate the quick and slow flow response of the system. The flow from each reservoir is controlled by the quick flow residence time (K_q) and the slow flow residence time (K_s). The simulated streamflow is therefore the addition of the outputs from each of these reservoirs.

In this thesis, HYMOD is applied to three catchments, *i.e.,* Leaf watershed in Chapter 3, and two MOPEX catchments in Florida and Georgia in Chapter 5. In all these cases studies, model calibration is conducted considering five parameters in Table D1. In all these three modeling case studies, parameters are lumped (*i.e.,* constant over the entire catchment).

The Leaf River watershed is located north of Collins, Mississippi. This catchment has been studied intensively in the past (*e.g.,* Boyle, 2000; Sorooshian *et al.*, 1993; Thiemann *et al.*, 2001; Vrugt *et al.*, 2003b; Vrugt *et al.*, 2008a) and may be considered a standard benchmark for parameter estimation of hydrological models. Three years (1953-1955) of hydrologic data (mean areal precipitation [mm/d], potential evapotranspiration [mm/d], and streamflow [m3/s]) are used, the first two years for model calibration and the third year as validation period. The MOPEX watersheds (Duan *et al.*, 2006) are located in the United States, 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 validation period in both watersheds.

Water and Energy Transfer between Soil, Plants and Atmosphere (WetSpa)

WetSpa is a grid-based distributed hydrologic model for water and energy transfer between soil, plants and atmosphere, which was originally developed in Wang *et al.* (1996) and further adapted for flood prediction on hourly time step (*e.g.* Liu *et al.*, 2003; Shafii and De Smedt, 2009). The model predicts peak discharges and hydrographs, which can be defined for any numbers and locations in the channel network, and can simulate the spatial distribution of basin hydrological variables. The hydrologic processes considered in the model are precipitation, interception, depression storage, surface runoff, snowmelt, infiltration, evapotranspiration, interflow, percolation, and groundwater drainage. More recently, Safari *et al.* (2009) have provided a complete list of equations that characterize different hydrological processes in WetSpa. Figure D2 schematically depicts different model components of WetSpa.

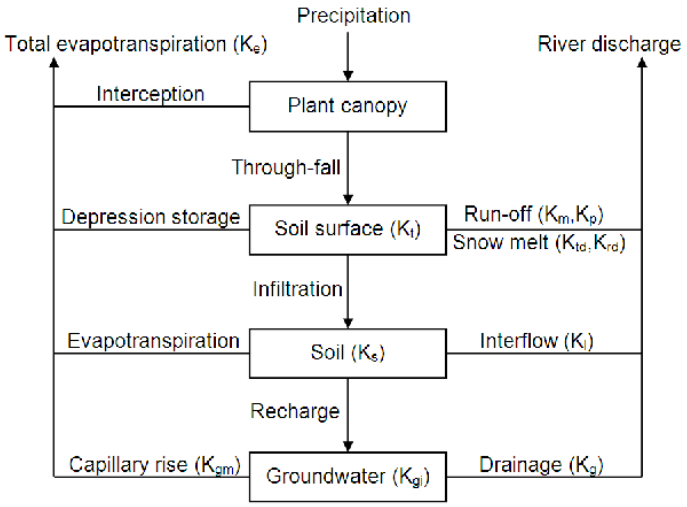

Figure D2. Components of WetSpa simulation model; taken from Shafii and De Smedt (2009).

This model potentially involves a large number of model parameters to be specified during the model setup. Most of these parameters can be assessed from field data, *e.g.* hydrometeorological observations, maps of topography, soil types, land use, etc. However, some model parameters (*i.e.* 11 parameters) are of a more conceptual nature and cannot be directly assessed in field. Hence, these parameters have to be determined through a calibration process. Table D2 provides list of these 11 parameters, along with their initial feasible range. Moreover, impact of different parameters on WetSpa hydrological processes is already shown on Figure D2.

Table D2. Calibration parameters of WetSpa model with their initial ranges.

Description	Parameter	Units	Prior range
Interflow scaling factor	K_i	-	0-10
Groundwater recession coefficient	K_g	d^{-1}	0-0.05
Initial soil moisture factor	K_s	-	0-2
Correction factor for PET	K_e	-	0-2
Initial groundwater storage	K_{gi}	mm	0-500
Groundwater storage scaling factor	K_{gm}	mm	0-2000
Base temperature for snowmelt	K_t	^oC	-1-1
Temperature degree-day coefficient	K_{td}	mm $^oC^{-1}d^{-1}$	0-10
Rainfall degree-day coefficient	K_{rd}	$^oC^{-1}$ d^{-1}	0-0.05
Surface runoff coefficient	K_m	-	0-5
Rainfall scaling factor	K_p	mm	0-500

The simulated streamflow is calculated in WetSpa as the sum of surface, and groundwater discharges. Among these 11 global parameters of WetSpa model, 3 parameters are employed in groundwater module (*i.e.* K_g, K_{gi}, K_{gm}), and the other 8 parameters are associated with surface runoff generating modules. In this thesis, WetSpa is applied to two case studies, *i.e.,* the Hornad catchment in Chapters 3, 4, and 5, and the Baron catchment in Chapter 5.

The Hornad river is located in Slovakia. The drainage area of the river up to Margecany station is 1,131 km2. Margecany is a subwatershed in the Hornad catchment. Figure D3 shows the Hornad catchment, the topography of Margecany subcatchment, and location of gauging stations. Details on the catchment's data and characteristics are provided in Bahremand *et al.* (2007). In summary, the observations for this catchment are collected from 1991 to 2000. The first five years (1991 to 1995) are used for model calibration and the remaining data (1996 to 2000) for validation. In this case study, all 11 WetSpa parameters (Table D2) should be adjusted in the calibration process.

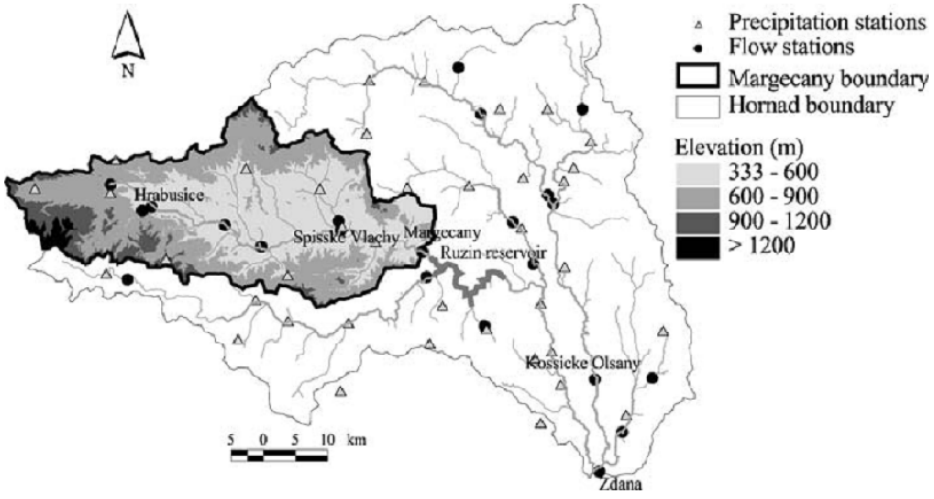

Figure D3. Hydrologic network of Hornad catchment, topography of Margecany subcatchment, and location of gauging stations

The Baron basin is one of the Distributed Model Intercomparison Project (DMIP) catchments (Smith and Gupta, 2012). Figure D4 shows the locations of the United States Geological Survey (USGS) streamflow gauging stations in the Baron basin. Table D3 also provides the information on the gauging stations. Note that the Baron basin contains two subbasins, *i.e.,* stations 2 and 3 in Figure D4. Safari *et al.* (2009) describe the data that are used in the hydrological modeling process of the Baron catchments using WetSpa model. These data are obtained from the DMIP website (http://www.nws.noaa.gov/oh/hrl/dmip/2/index.html).

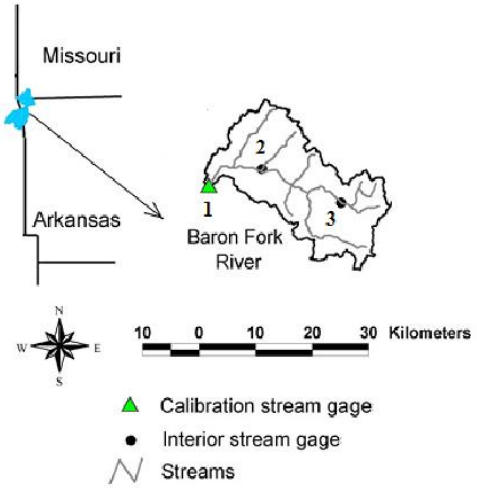

Figure D4. Location of Baron watershed and stream gages; taken from Safari *et al.* (2009). The IDs of the gage stations are given in Table D3.

Table D3. Location, ID, and basin area of the Baron watershed's and its sub-basins' stream gages as shown in Figure D4.

ID	Location	Area (km2)
1	Eldon, OK	795
2	Peacheater Creek at Christie, OK	65
3	Dutch Mills, AR	105

All spatial model parameters are automatically derived using Geographic Information Systems (GIS) tools and are not adjusted via calibration. The model implementation is based on 30m spatial resolution and 1h time–step. Among the 11 WetSpa parameters that need to be adjusted in the calibration process, there are only six parameters to be tuned in the Baron basin. These parameters are K_m, K_p, K_e, K_{gm}, K_i, and K_g. There is no snowmelt process in this basin; therefore, K_t, K_{td}, and K_{rd} are not used. Furthermore, based on preliminary manual calibration, initial soil moisture factor (K_s) and initial groundwater storage ($K_{gi)}$ are set to 1 and 100 mm, respectively.

Soil and Water Assessment Tool (SWAT)

The Soil and Water Assessment Tool version 2000 (SWAT2000) is a spatially distributed continuous simulation model for predicting flow, sediment, nutrient and other contaminant transport. SWAT2000, which is also called SWAT in this thesis, is designed to compute long-term runoff and nutrient export from rural watersheds, especially those dominated by agriculture (Arnold *et al.*, 1998). The model is maintained by the Agricultural Research Service of the U.S. Department of Agriculture (USDA) and distributed by the U.S. Environmental Protection Agency (EPA) for nonpoint source modeling. Recently,

Tolson (2005) and Tolson and Shoemaker (2007) applied SWAT to Cannonsville Reservoir in upstate New York to predict flow, sediment and phosphorus. The Cannonsville Reservoir is one of New York City's (NYC's) largest drinking water reservoirs and is located in Delaware County in the Catskill region of Upstate NY. This watershed is divided into 43 sub-basins. The locations of the measured flow and continuous water quality stations are provided in Figure D5.

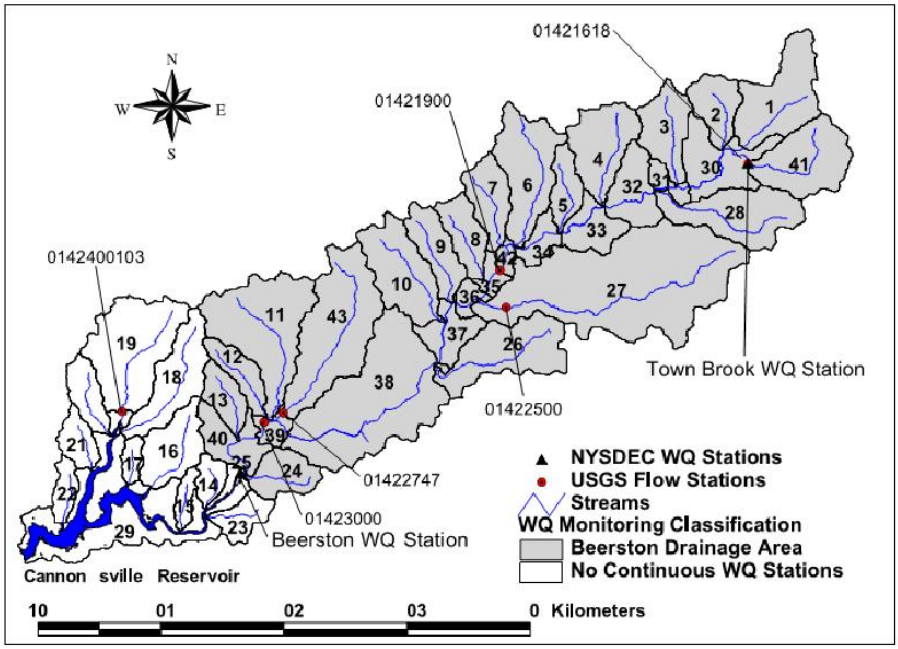

Figure D5. Sub-basins and continuous flow and water quality (WQ) monitoring stations in the Cannonsville Watershed; taken from Tolson and Shoemaker (2007)

In SWAT, the watershed is discretized into sub-basins and these sub-basins are typically further discretized into Hydrologic Response Units (HRUs). HRUs are formed as unique soil and land use combinations that are not necessarily contiguous land parcels. As Tolson and Shoemaker mention, the Cannonsville watershed is represented by 758 HRUs in total in SWAT implementation. Tolson and Shoemaker (2007) address SWAT's calibration in the Cannonsville watershed. Due to high computational time of a single SWAT run in this watershed, Tolson and Shoemaker (2007) created a single sub-basin model for the Town Brook subwatershed (37 km2 drainage area) to allow for a higher number of model evaluations in the automatic calibration trials. The Town Brook case study is also used in this thesis. Among the calibration formulations developed in Tolson and Shoemaker (2007), a 26-dimensional calibration formulation is used in this thesis. The description of the 26 SWAT parameters adjusted via calibration as well as their prior ranges are listed in Table D4.

Table D4. SWAT calibration parameters and their prior ranges

	Parameter	Description	Units	Prior range
1	SFTMP	Snow fall temperature	^{o}C	-5 – 5
2	SMTMP	Snowmelt temperature threshold	^{o}C	-5 – 5
3	SMFMX	Melt factor for snow	mm $H_2O/^{o}C$ day	1.5 – 8
4	TIMP	Snowpack temperature lag factor	-	0.01 – 1
5	SURLAG	Surface runoff lag coefficient	-	1 – 24
6	GW_DELAY	Groundwater delay time	days	0.001 – 500
7	ALPHA_BF	Base flow alpha factor	-	0.001 – 1
8	GWQMN	Threshold groundwater depth for return flow	mm	0.001 – 500
9	LAT_TIME	Lateral flow travel time	days	0.001 – 180
10	ESCO	Soil evaporation compensation factor	-	0.01 – 1
11	CN2_f	Runoff curve number multiplicative factor	-	0.75 – 1.25
12	AWC_f	Available water capacity range factor	-	0 – 1
13	Ksat_f	Saturated hydraulic conductivity range factor	-	0 – 1
14	DepthT_f	Soil profile total depth range factor	-	0 – 1
15	APM	Tributary channel peak rate adjustment sediment routing factor	-	0.5 – 1.5
16	PPERCO	Phosphorous (P) percolation coefficient	10 m^3/Mg	10 – 17.5
17	PHOSKD	P soil partitioning coefficient	m^3/Mg	100 – 200
18	CMN	Rate factor for humus mineralization of active organic P	-	0.0001 – 0.003
19	UBP	Plant P uptake distributions parameter	-	0.1 – 100
20	LAT_SED	Sediment concentration in lateral and groundwater flow	mg/L	0.1 – 22.8
21	ERGORGP	P enrichment ratio for loading with sediment	-	1 – 5
22	SLSUBBSN_F	Average slope length	m	0.5 – 1.5
23	SL_SOIL_f	Slope length for lateral subsurface flow	m	0.5 – 1.5
24	CLAY_f	Soil layer content range factor	-	0 – 1
25	ROCK_f	Soil layer rock content range factor	-	0 – 1
26	MUSLEadj	Erosion under snow cover adjustment parameter	-	0 – 1

The simulation period for discharge was from January 1st 1996 to September 30th 2000 with a 639-day warm-up period resulting in a 1096-day period. This period is divided in two parts, the calibration period from October 1st 1997 to the end of 1998, and the validation period from January 1st 2000 to September 30th 2000. For the sediment transport, the simulation period is October 1st 1998 to September 30th 2000 in which the sup-period October 1st 1998 to the end of 1999 was used for calibration, and the period January 1st 2000 to September 30th 2000 for validation.

Hydrologiska Byråns Vattenbalansavdelning (HBV)

HBV is originally developed by Swedish Meteorological and Hydrological Institute (SMHI) in the 1970s (Bergström, 1976). Recently, Aghakouchak and Habib (2010) have 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 thesis. The principal model structure and process representations are presented in this section. HBV can be used as a fully-distributed or a semi-distributed model by dividing the catchment into sub-basins; however, in the current study a simplified spatially-lumped version of the model is used for teaching purposes. In a lumped model, it is assumed that the study area (watershed) is one single unit (zone) and the parameters do not change spatially across the watershed.

The HBV model consists of four main modules: (1) Snowmelt and snow accumulation module; (2) Soil moisture and effective precipitation module; (3) Evapotranspiration module; (4) Runoff response module. The model can run at a daily or monthly time step; the required input data include time series of precipitation and temperature observations at each time step, and long-term estimates of mean monthly temperature and potential evapotranspiration rates. The model includes a module that processes the input precipitation either as rainfall or snow, based on the input temperature at each time step. The rainfall and snowmelt (if it exists) are then processed in the soil moisture module where the effective rainfall that contributes to the surface runoff is evaluated. The remaining part of the rainfall contributes to the soil moisture storage which itself can be evaporated as long as there is enough water content in the subsurface. The main output of the model is discharge runoff at the outlet of the watershed, which has three components: surface runoff, interflow (contribution from near surface flow) and baseflow (contribution from groundwater flow). Figure D6 illustrates the general processes of the simplified educational version of the HBV model. Details on the formulation of each module is provided in Aghakouchak and Habib (2010). This model has 10 parameters to be adjusted during calibration. These parameters and their prior ranges are described in Table D5. Note that this version of HBV is lumped. In this thesis, HBV is applied to two catchments in the MOPEX experiment, one in Illinois with USGS station code of 07196500 and one in Oregon with the code of 11501000. Two years of data are used as the calibration period, and one year as the validation period.

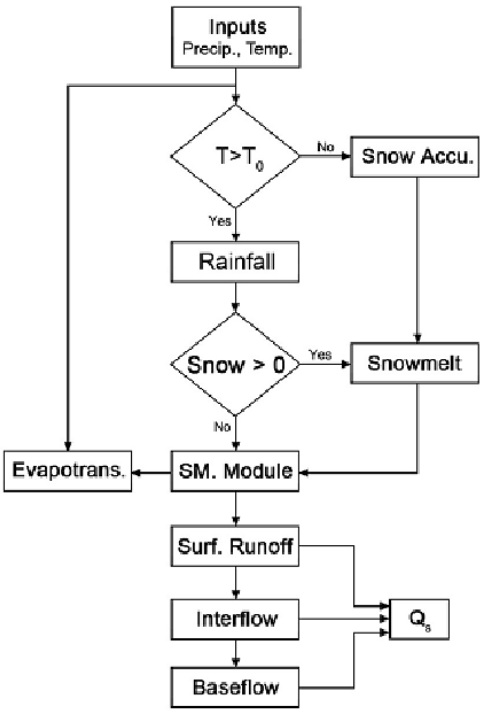

Figure D6. General processes of the HBV model used in this thesis; taken from Aghakouchak and Habib (2010).

Table D5. HBV model parameters and their prior range.

	Parameter	Description	Units	Prior range
1	FC	Field capacity	mm	100 – 200
2	PWP	Soil permanent wilting point	mm	90 – 180
3	DD	Degree-day factor	mm/^{o}C/day	3 – 7
4	K0	Near surface flow storage coefficient; upper reservoir	day^{-1}	0.05 – 0.2
5	K1	Interflow storage coefficient; upper reservoir	day^{-1}	0.01 – 0.1
6	K2	Base flow storage coefficient; lower reservoir	day^{-1}	0.01 – 0.05
7	Kp	Percolation storage coefficient	day^{-1}	0.01 – 0.05
8	L	Surface flow generation threshold	mm	2 – 5
9	β	Effective rainfall shape coefficient	-	1 – 7
10	C	Evapotranspiration parameter	^{o}C^{-1}	0.01 – 0.07

References

Abbaspour, K.C., Johnson, C.A., van Genuchten, M.T. (2004), "Estimating Uncertain Flow and Transport Parameters Using a Sequential Uncertainty Fitting Procedure", *Vadose Zone Journal*, 3(4): 1340-1352.

Abebe, N.A., Ogden, F.L., Pradhan, N.R. (2010), "Sensitivity and uncertainty analysis of the conceptual HBV rainfall–runoff model: Implications for parameter estimation", *Journal of Hydrology*, 389(3–4): 301-310.

Aghakouchak, A., Habib, E. (2010), "Application of a Conceptual Hydrologic Model in Teaching Hydrologic Processes", *International Journal of Engineering Education*, 26(4): 963-973.

Ajami, N.K., Duan, Q., Sorooshian, S. (2007), "An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction", *Water Resources Research*, 43: W01403.

Arabi, M., Govindaraju, R.S., Hantush, M.M. (2007), "A probabilistic approach for analysis of uncertainty in the evaluation of watershed management practices", *Journal of Hydrology*, 333(2-4): 459-471.

Arnold, J.G., Srinivasan, R., Muttah, R.S., Williams, J.R. (1998), "Large area hydrologic modeling and assessment part I: model development", *Journal of the American Water Resources Association*, 34 (1): 73-89.

Asadzadeh, M., Tolson, B. (2013), "Pareto archived dynamically dimensioned search with hypervolume-based selection for multi-objective optimization", *Eng Optimiz*, 45(12): 1489-1509.

Bahremand, A., De Smedt, F., Corluy, J., Liu, Y., Poorova, J., Velcicka, L., Kunikova, E. (2007), "WetSpa Model Application for Assessing Reforestation Impacts on Floods in Margecany–Hornad Watershed, Slovakia", *Water Resources Management*, 21(8): 1373-1391.

Balin-Talamba, D. (2004), "Hydrological behaviour through experimental and modeling approaches; Application to the Haute-Mentue catchment, PhD Thesis", Swiss Federal School of Technology of Lausanne.

Balin-Talamba, D., Parent, E., Musy, A. (2010), "Bayesian multiresponse calibration of TOPMODEL: Application to the Haute-Mentue catchment, Switzerland", *Water Resources Research*, 46: W08524.

Bates, B.C., Campbell, E.P. (2001), "A Markov Chain Monte Carlo Scheme for parameter estimation and inference in conceptual rainfall-runoff modeling", *Water Resources Research*, 37(4): 937-947.

Bates, D.M., Watts, D.G. (1981), "Parameter transformations for improved approximate confidence regions in nonlinear least squares", *The Annuals of Statistics*, 9(6): 1152-1167.

Bates, D.M., Watts, D.G. (1988), "*Nonlinear regression analysis and its applications*". Wiley, NY.

Bergström, S. (1976), "Development and application of a conceptual runoff model for Scandinavian catchments", Norrköping.

Beven, K. (2006), "A manifesto for the equifinality thesis", *Journal of Hydrology*, 320(1-2): 18-36.

Beven, K., Binley, A. (1992), "The future of distributed models: model calibration and uncertainty prediction", *Hydrological Processes*, 6(3): 279-298.

Beven, K., Binley, A. (2013), "GLUE: 20 years on", *hydrological Processes*: n/a-n/a.

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

Beven, K., Smith, P., Westerberg, I., Freer, J. (2012a), "Comment on "Pursuing the method of multiple working hypotheses for hydrological modeling" by P. Clark *et al.*", *Water Resources Research*, 48.

Beven, K., Smith, P.J., Westerberg, I., Freer, J. (2012b), "Comment on "Pursuing the method of multiple working hypotheses for hydrological modeling" by P. Clark *et al.*", *Water Resources Research*, 48: W11801.

Beven, K., Smith, P.J., Wood, A. (2011), "On the colour and spin of epistemic error (and what we might do about it)", *Hydrol. Earth Syst. Sci.*, 15: 3123-3133.
Beven, K.J. (2009), "Comment on "Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling; by Jasper A. Vrugt, Cajo J. F. ter Braak, Hoshin V. Gupta and Bruce A. Robinson.", *Stoch Environ Res Risk Assess*, 23: 1059-1060.
Beven, K.J., Smith, P.J., Freer, J. (2008), "So just why would a modeller choose to be incoherent? ", *J. Hydrology*, 354: 15-32.
Blasone, R.-S., Madsen, H., Rosbjerg, D. (2008a), "Uncertainty assessment of integrated distributed hydrological models using GLUE with Markov chain Monte Carlo sampling", *J. Hydrology*, 353: 18– 32.
Blasone, R.S., Vrugt, J.A., Madsen, H., Rosbjerg, D., Zyvoloski, G.A., Robinson, B.A. (2008b), "Generalized likelihood uncertainty estimation (GLUE) using adaptive Markov Chain Monte Carlo sampling", *Adv. Water Res.*, 31: 630-648.
Blazkova, S., Beven, K. (2009), "A limits of acceptability approach to model evaluation and uncertainty estimation in flood frequency estimation by continuous simulation: Skalka catchment, Czech Republic", *Water Resources Research*, 45: W00B16.
Blazkova, S., Beven, K.J., Kulasova, A. (2002), "On constraining TOPMODEL hydrograph simulations using partial saturated area information", *Hydrological Processes*, 16(2): 441-458.
Box, G.E.P., Cox, D.R. (1964), "An Analysis of Transformations", *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2): 211-252.
Box, G.E.P., Tiao, G.C. (1973), "*Bayesian Inference in Statistical Analysis*". Addison-Wesley, Boston, MA.
Boyle, D.P. (2000), "Multicriteria calibration of hydrological models", *Ph.D. Dissertation Thesis*, Univ. of Ariz.
Boyle, D.P., Gupta, H.V., Sorooshian, S. (2000), "Toward improved calibration of hydrological models: Combining the strengths of manual and automatic methods", *Water Resources Research*, 36(12): 3663– 3674.
Breinholt, A., Grum, M., Madsen, H., Örn Thordarson, F., Mikkelsen, P.S. (2013), "Informal uncertainty analysis (GLUE) of continuous flow simulation in a hybrid sewer system with infiltration inflow – consistency of containment ratios in calibration and validation?", *Hydrol. Earth Syst. Sci.*, 17(10): 4159-4176.
Bulygina, N., Gupta, H. (2009), "Estimating the uncertain mathematical structure of a water balance model via Bayesian data assimilation", *Water Resources Research*, 45.
Carrillo, G., Troch, P.A., Sivapalan, M., Wagener, T., Harman, C., Sawicz, K. (2011), "Catchment classification: hydrological analysis of catchment behavior through process-based modeling along a climate gradient", *Hydrol Earth Syst Sc*, 15(11): 3411-3430.
Choi, H.T., Beven, K. (2007), "Multi-period and multi-criteria model conditioning to reduce prediction uncertainty in an application of TOPMODEL within the GLUE framework", *J. Hydrology*, 332((3-4)): 316-336.
Christensen, S., Cooley, R.L. (1999), "Evaluation of confidence intervals for a steady-state leaky aquifer model", *Advances in Water Resources*, 22(8): 807-817.
Clark, M.P., Kavetski, D., Fenicia, F. (2011), "Pursuing the method of multiple working hypotheses for hydrological modeling", *Water Resources Research*, 47.
Clark, M.P., Kavetski, D., Fenicia, F. (2012a), "Reply to comment by K. Beven *et al.* on "Pursuing the method of multiple working hypotheses for hydrological modeling"", *Water Resources Research*, 48: W11802.
Clark, M.P., Kavetski, D., Fenicia, F. (2012b), "Reply to comment by K. Beven *et al.* on "Pursuing the method of multiple working hypotheses for hydrological modeling"", *Water Resources Research*, 48(11): W11802.

Coello Coello, C. (2005), "Recent Trends in Evolutionary Multiobjective Optimization", In: Abraham, A., Jain, L., Goldberg, R. (Eds.), *Evolutionary Multiobjective Optimization*. Advanced Information and Knowledge Processing. Springer London, pp. 7-32.

Dean, S., Freer, J., Beven, K.J., Wade, A.J., Butterfiel, D. (2009), "Uncertainty assessment of a process-based integrated catchment model of phosphorus", *Stoch Environ Res Risk Assess*, 23: 991–1010.

Deb, K., Pratap, A., Agarwal, S., Meyarivan, T. (2002), "A fast and elitist multiobjective genetic algorithm: NSGA-II", *IEEE Transactions on Evolutionary Computation*, 6(2): 182-197.

Del Moral, P., Doucet, A., Jasra, A. (2012), "An adaptive sequential Monte Carlo method for approximate Bayesian computation", *Stat Comput*, 22(5): 1009-1020.

Dilks, D.W., Canale, R.P., Meier, P.G. (1992), "Development of Bayesian Monte-Carlo techniques for water-quality model uncertainty", *Ecological Modelling*, 62: 149-162.

Dotto, C.B.S., Mannina, G., Kleidorfer, M., Vezzaro, L., Henrichs, M., McCarthy, D.T., Freni, G., Rauch, W., Deletic, A. (2012), "Comparison of different uncertainty techniques in urban stormwater quantity and quality modelling", *Water Research*, 46(8): 2545-2558.

Duan, Q., Schaake, J., Andreassian, V., Franks, S., Goteti, G., Gupta, H.V., Gusev, Y.M., Habets, F., Hall, A., Hay, L., Hogue, T., Huang, M., Leavesley, G., Liang, X., Nasonova, O.N., Noilhan, J., Oudin, L., Sorooshian, S., Wagener, T., Wood, E.F. (2006), "Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops", *Journal of Hydrology*, 320(1-2): 3-17.

Duan, Q., Sorooshian, S., Gupta, V.K. (1992), "Effective and efficient global optimization for conceptual rainfall-runoff models", *Water Resources Research*, 28(4): 1015– 1031.

Eder, G., Sivapalan, M., Nachtnebel, H.P. (2003), "Modelling water balances in an Alpine catchment through exploitation of emergent properties over changing time scales", *hydrological Processes*, 17(11): 2125-2149.

Efstratiadis, A., Koutsoyiannis, D. (2010), "One decade of multi-objective calibration approaches in hydrological modelling: a review", *Hydrological Sciences*, 55(1): 58-78.

Emmerich, M., Beume, N., Naujoks, B. (2005), "An EMO algorithm using the hypervolume measure as selection criterion", *Evolutionary Multi-Criterion Optimization*, 3410: 62-76.

Engeland, K., Xu, C.Y., Gottschalk, L. (2005), "Assessing uncertainties in a conceptual water balance model using Bayesian methodology", *Hydrological Sciences*, 50(1): 45-63.

Euser, T., Winsemius, H.C., Hrachowitz, M., Fenicia, F., Uhlenbrook, S., Savenije, H.H.G. (2013), "A framework to assess the realism of model structures using hydrological signatures", *Hydrol. Earth Syst. Sci.*, 17: 1893-1912.

Farina, M., Amato, P. (2004), "A fuzzy definition of" optimality" for many-criteria optimization problems", *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 34(3): 315-326.

Fenicia, F., Savenije, H.H.G., Matgen, P., Pfister, L. (2007), "A comparison of alternative multiobjective calibration strategies for hydrological modeling", *Water Resources Research*, 43(3).

Feyen, L., Vrugt, J., Ó Nualláin, B., van der Knijff, J., De Roo, A. (2007), "Parameter optimization and uncertainty assessment for large scale streamflow simulation with the LISFLOOD model", *J. Hydrology*, 332: 276–289.

Fleming, P.J., Purshouse, R.C., Lygoe, R.J. (2005), "Many-objective optimization: An engineering design perspective", *Evolutionary multi-criterion optimization*. Springer, pp. 14-32.

Franz, K.J., Hogue, T.S. (2011), "Evaluating uncertainty estimates in hydrologic models: borrowing measures from the forecast verification community", *Hydrol. Earth Syst. Sci.*, 15: 3367–3382.

Freer, J., Beven, K.J., Ambroise, B. (1996), "Bayesian estimation of uncertainty in runoff prediction and the value of data: an application of the GLUE approach ", *Water Resources Research*, 32(7): 2161-2173.

Freer, J., McMillan, H., McDonnell, J.J., Beven, K.J. (2004), "Constraining Dynamic TOPMODEL responses for imprecise water table information using fuzzy rule based performance measures.", *J. Hydrology*, 291: 254-277.

Freni, G., Mannina, G. (2009), "Bayesian approach for uncertainty quantification in water quality modelling:The influence of prior distribution", *J. Hydrology*, 392: 31–39.
Gallart, F., Latron, J., Llorens, P., Beven, K. (2007), "Using internal catchment information to reduce the uncertainty of discharge and baseflow prediction", *Adv. Water Res.*, 30(4): 808-823.
Garen, D.C., Burges, S.J. (1981), "Approximate error bounds for simulated hydrographs", *Journal of the Hydraulics Division (ASCE)*, 107(HY11): 1519-1534.
Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B. (2004), "*Bayesian Data Analysis*". Chapman & Hall/CRC, Boca Raton, Florida.
Geman, S., Geman, D. (1984), "Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images", *Transactions on Pattern Analysis and Machine Intelligence*, 6: 721–741.
Gneiting, T., Balabdaoui, F., Raftery, A.E. (2007), "Probabilistic forecasts, calibration and sharpness", *Journal of the Royal Statistical Society Series B*, 69(Part 2): 243–268.
Goldberg, D.E. (1989), "*Genetic Algorithms in Search, Optimization and Machine Learning*". Addison-Wesley, Reading, MA.
Gupta, H.V., Beven, K.J., Wagener, T. (2005a), "Model calibration and uncertainty estimation", In: Hoboken, N.J. (Ed.), *Encyclopedia of Hydrological Sciences*. John Wiley, pp. 1 –17.
Gupta, H.V., Beven, K.J., Wagener, T. (2005b), "Model calibration and uncertainty estimation", *Encyclopedia of Hydrologic Sciences*. John Wiley and Sons.
Gupta, H.V., Sorooshian, S., Yapo, P.O. (1998), "Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information", *Water Resources Research*, 34(4): 751-764.
Gupta, H.V., Wagener, T., Liu, Y. (2008), "Reconciling theory with observations: Elements of a diagnostic approach to model evaluation", *Hydrological Processes*.
Haario, H., Saksman, E., Tamminen, J. (2001), "An adaptive Metropolis algorithm", *Bernoulli*, 7(2): 223-242.
Hadka, D., Reed, P. (2011), "Diagnostic Assessment of Search Controls and Failure Modes in Many-Objective Evolutionary Optimization", *Evolutionary Computation*, 20(3): 423-452.
Hastings, W.K. (1970), "Monte Carlo sampling methods using Markov chains and their applications", *Biometrika*, 57(1): 97-109.
Helton, J.C., Davis, F.J. (2003), "Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems", *Reliability Engineering & System Safety*, 81(1): 23-69.
Hills, S.E., Smith, A.F.M. (1992), "Parameterization issues in Bayesian inference", In: J. M. Bernardo, J.B., A. P. Dawid and A. F. M. Smith (Ed.), *Bayesian Statistics 4*. Oxford University Press, pp. 227-46.
Hingray, B., Schaefli, B., Mezghani, A., Hamdi, Y. (2010), "Signature-based model calibration for hydrological prediction in mesoscale Alpine catchments", *Hydrolog Sci J*, 55(6): 1002-1016.
Hong, B., Strawderman, R.L., Swaney, D.P., Weinstein, D.A. (2005), "Bayesian estimation of input parameters of a nitrogen cycle model applied to a forested reference watershed, Hubbard Brook Watershed Six", *Water Resources Research*, 41.
Honti, M., Stamm, C., Reichert, P. (2013), "Integrated uncertainty assessment of discharge predictions with a statistical error model", *Water Resources Research*, 49(8): 4866-4884.
Hornberger, G.M., Spear, R.C. (1981), "Approach to the preliminary analysis of environmental systems", *Journal of Environmental Management*, 12(1): 7-18.
Huard, D., and A. Mailhot (2008), "Calibration of hydrological model GR2M using Bayesian uncertainty analysis", *Water Resources Research*, 44: W02424.
Hutton, C., Kapelan, Z., Vamvakeridou-Lyroudia, L., Savić, D. "Application of Formal and Informal Bayesian Methods for Water Distribution Hydraulic Model Calibration", *Journal of Water Resources Planning and Management*, 0(0): 04014030.
Hutton, C., Kapelan, Z., Vamvakeridou-Lyroudia, L., Savić, D. (2014), "Application of Formal and Informal Bayesian Methods for Water Distribution Hydraulic Model Calibration", *Journal of Water Resources Planning and Management*, 0(0): 04014030.

Jeremiah, E., Sisson, S., Marshall, L., Mehrotra, R., Sharma, A. (2011), "Bayesian calibration and uncertainty analysis of hydrological models: A comparison of adaptive Metropolis and sequential Monte Carlo samplers", *Water Resources Management*, 47(W07547).

Jin, X., Xu, C.-Y., Zhang, Q., Singh, V.P. (2010), "Parameter and modeling uncertainty simulated by GLUE and a formal Bayesian method for a conceptual hydrological model", *Journal of Hydrology*, 383: 147-155.

Johnston, P.R., Pilgrim, D.H. (1976), "Parameter optimization for Watershed models", *Water Resources Research*, 12(3): 477-486.

Jothityangkoon, C., Sivapalan, M., Farmer, D.L. (2001), "Process controls of water balance variability in a large semi-arid catchment: downward approach to hydrological model development", *Journal of Hydrology*, 254(1-4): 174-198.

Kavetski, D., Franks, S.W., Kuczera, G. (2002), "Confronting Input Uncertainty in Environmental Modeling", In: Duan, Q., Gupta, H.V., Sorooshian, S., Rousseau, A.N., Turcotte, R. (Eds.), *Calibration of Watershed Models*. AGU, Washington, DC, pp. 49-68.

Kavetski, D., Kuczera, G., Franks, S.W. (2006), "Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory", *Water Resources Research*, 42, W03407, doi: 10.1029/2005WR004368.

Kennedy, J., Eberhardt, R.C., Shi, Y. (2001), "*Swarm Intelligence*". Morgan Kaufmann, San Francisco(CA).

Kirchner, J.W. (2006), "Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology", *Water Resources Research*, 42(3).

Klemeš, V. (1986), "Operational testing of hydrological simulation models", *Hydrological Sciences*, 31(1): 13–24.

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

Krzysztofowicz, R. (1999), "Bayesian Forecasting via Deterministic Model", *Risk Analysis*, 19(4): 739-749.

Kuczera, G. (1983), "Improved parameter inference in catchment models: 1. Evaluating parameter uncertainty", *Water Resources Research*, 19(5): 1151-1162.

Kuczera, G. (1988), "On the validity of first-order prediction limits for conceptual hydrologic models", *J. Hydrology*, 103: 229-247.

Kuczera, G., Kavetski, D., Franks, S., Thyer, M. (2006), "Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters", *Journal of Hydrology*, 331(3-4): 161–177.

Kuczera, G., Mroczkowski, M. (1998), "Assessment of hydrologic parameter uncertainty and the worth of multiresponse data", *Water Resources Research*, 34(6): 1481–1489.

Kuczera, G., Parent, E. (1998), "Monte Carlo assessment of parameter uncertainty in conceptual catchment models: the Metropolis algorithm", *Journal of Hydrology*, 211(1-4): 69-85.

Laio, F., Tamea, S. (2007), "Verification tools for probabilistic forecasts of continuous hydrological variables", *Hydrol. Earth Syst. Sci.*, 11(4): 1267- 1277.

Laloy, E., Fasbender, D., Bielders, C.L. (2010), "Parameter optimization and uncertainty analysis for plot-scale continuous modeling of runoff using a formal Bayesian approach", *J. Hydrology*, 380: 82–93.

Lamb, R., Beven, K.J., Myrabø, S. (1998), "Use of spatially distributed water table observations to constrain uncertainty in a rainfall–runoff model", *Adv. Water Res.*, 22(4): 305–317.

Legates, D.R., McCabe, G.J. (1999), "Evaluating the use of ‘‘goodness-of-fit’’ measures in hydrologic and hydro-climatic model evaluation", *Water Resources Research*, 35: 233-241.

Li, L., Xia, J., Xu, C.-Y., Singh, V.P. (2010a), "Evaluation of the subjective factors of the GLUE method and comparison with the formal Bayesian method in uncertainty assessment of hydrological models", *Journal of hydrology*, 390: 210-221.

Li, Z., Shao, Q., Xu, Z., Cai, X. (2010b), "Analysis of parameter uncertainty in semi-distributed hydrological models using bootstrap method: A case study of SWAT model applied to Yingluoxia watershed in northwest China", *Journal of Hydrology*, 385(1–4): 76-83.

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

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

Liu, Y., Zhang, J., Wang, G., Liu, J., He, R., Wang, H., Liu, C., Jin, J. (2012), "Quantifying uncertainty in catchment-scale runoff modeling under climate change (case of the Huaihe River, China)", *Quaternary International*, 282(0): 130-136.

Liu, Y.B., Gebremeskel, S., De Smedt, F., Hoffmann, L., Pfister, L. (2003), "A diffusive transport approach for flow routing in GIS-based flood modeling", *Journal of Hydrology*, 283(1-4): 91-106.

Ljung, L. (1987), " *System Identification-Theory for the User*". Prentice Hall, Englewood Cliffs, NJ.

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

Mannina, G. (2011), "Uncertainty Assessment of a Water-Quality Model for Ephemeral Rivers Using GLUE Analysis", *Journal of Environmental Engineering*, 137(3): 177-186.

Mantovan, P., Todini, E. (2006), "Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology", *J. Hydrology*, 130((1 -2)): 368- 381.

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

Martin, C., Ayesa, E. (2010), "An Integrated Monte Carlo Methodology for the calibration of water quality models", *Ecological Modelling*, 221(22): 2656-2667.

Martinez, G.F., Gupta, H.V. (2010), "Toward improved identification of hydrological models: A diagnostic evaluation of the "abcd" monthly water balance model for the conterminous United States", *Water Resources Research*, 46.

Martinez, G.F., Gupta, H.V. (2011), "Hydrologic consistency as a basis for assessing complexity of monthly water balance models for the continental United States", *Water Resources Research*, 47.

Matott, L.S., Babendreier, J.E., Purucker, S.T. (2009), "Evaluating uncertainty in integrated environmental models: A review of concepts and tools", *Water Resources Research*, 45.

McIntyre, N., Wheater, H., Lees, M. (2002), "Estimation and propagation of parametric uncertainty in environmental models", *Journal of Hydroinformatics*, 4: 177-198.

McMillan, H., Clark, M.P. (2009), "Rainfall-runoff model calibration using informal likelihood measures within a Markov chain Monte Carlo sampling scheme", *Water Resources Research*, 45: W04418.

McMillan, H., Krueger, T., Freer, J. (2012), "Benchmarking observational uncertainties for hydrology: rainfall, river discharge and water quality", *hydrological Processes*, 26(26): 4078-4111.

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

Melching, C.S., Yen, B.C., Wenzel, H.G. (1990), "A reliability estimation in modeling watershed runoff with uncertainties", *Water Resources Research*, 26(10): 2275-2286.

Mengersen, K.L., Tweedie, R.L. (1996), "Rates of convergence of the Hastings and Metropolis algorithms", *Ann. Stat.*, 24: 101-121.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E. (1953), "Equation of State Calculations by Fast Computing Machines", *The Journal of Chemical Physics*, 21(6): 1087-1092.

Mitchell, S., Beven, K., Freer, J. (2009), "Multiple sources of predictive uncertainty in modeled estimates of net ecosystem CO2 exchange", *Ecological Modelling*, 220: 3259–3270.

Montanari, A. (2005), "Large sample behaviors of the generalized likelihood uncertainty estimation (GLUE) in assessing the uncertainty of rainfall-runoff simulations", *Water Resources Research*, 41: W08406.

Montanari, A. (2011), "Interactive comment on “On the colour and spin of epistemic error (and what we might do about it)” by K. Beven *et al.*", *Hydrol. Earth Syst. Sci. Discussion*, 8: C2885–C2891.
Montanari, A., Brath, A. (2004), "A stochastic approach for assessing the uncertainty of rainfall-runoff simulations", *Water Resources Research*, 40: W01106.
Montanari, A., Toth, E. (2007), "Calibration of hydrological models in the spectral domain: An opportunity for scarcely gauged basins?", *Water Resources Research*, 43(5).
Moore, R.J. (1985), "The probability-distributed principle and runoff production at point and basin scales", *Hydrological Sciences*, 30(2): 273–297.
Moradkhani, H., Sorooshian, S., Gupta, H.V., Houser, P.R. (2005), "Dual state-parameter estimation of hydrological models using ensemble Kalman filter", *Adv. Water Resour.*, 28: 135-147.
Morgan, M.G., Small, M. (1992), "*Uncertainty: a guide to dealing with uncertainty in quantitative risk and policy analysis*". Cambridge University Press.
Mroczkowski, M., Raper, G.P., Kuczera, G. (1997), "The quest for more powerful validation of conceptual catchment models", *Water Resources Research*, 33: 2325- 2335.
Muleta, M.K., Nicklow, J.W. (2005), "Sensitivity and uncertainty analysis coupled with automatic calibration for a distributed watershed model", *Journal of Hydrology*, 306: 127-145.
Nash, J.E., Sutcliffe, J.V. (1970), "River flow forecasting through conceptual models: Part 1 - A discussion of principles", *Journal of Hydrology*, 10(3): 282-290.
Neitsch, S.L., Arnold, J.G., Kiniry, J.R., Williams, J.R. (2001), "Soil and Water Assessment Tool user’s manual version 2000", *U.S. Dept. of Agric. Agric. Res. Serv.*, Temple, Tex.
Nott, D.J., Marshall, L., Brown, J. (2012), "Generalized likelihood uncertainty estimation (GLUE) and approximate Bayesian computation: What’s the connection?", *Water Resour. Res.*, 48(W12602).
Pfannerstill, M., Guse, B., Fohrer, N. (2014), "Smart low flow signature metrics for an improved overall performance evaluation of hydrological models", *Journal of Hydrology*, 510(0): 447-458.
Pokhrel, P., Yilmaz, K.K., Gupta, H.V. (2012), "Multiple-criteria calibration of a distributed watershed model using spatial regularization and response signatures", *Journal of Hydrology*, 418: 49-60.
Purshouse, R.C., Fleming, P.J. (2007), "On the evolutionary optimization of many conflicting objectives", *Evolutionary Computation, IEEE Transactions on*, 11(6): 770-784.
Qian, S.S., Stow, C.A., Borsuk, M.E. (2003), "On Monte Carlo methods for Bayesian inference", *Ecological Modelling*, 159: 269-277.
Razavi, S., Tolson, B., Matott, L.S., Thomson, N., MacLean, A., Seglinieks, F. (2010), "Reducing the Computational Cost of Automatic Calibration through Model Pre-Emption", *Water Resources Research*, 46(11): W11523.
Reed, P.M., Hadka, D., Herman, J.D., Kasprzyk, J.R., Kollat, J.B. (2013), "Evolutionary multiobjective optimization in water resources: The past, present, and future", *Advances in Water Resources*, 51: 438-456.
Refsgaard, J.C. (1997), "Parameterisation, calibration and validation of distributed hydrological models", *J. Hydrology*, 198: 69-97.
Refsgaard, J.C., Knudsen, J. (1996), "Operational validation and intercomparison of different types of hydrological models", *Water Resour. Res.*, 32(7): 2189- 2202.
Refsgaard, J.C., Nilsson, B., Brown, J., Klauer, B., Moore, R., Bech, T., Vurro, M., Blind, M., ,, Castilla, G., Tsanis, I., Biza, P. (2005), "Harmonised techniques and representative river basin data for assessment and use of uncertainty information in integrated water management", *Environment Science Policy*, 8(3): 267-277.
Refsgaard, J.C., van der Sluijs, J.P., Hojberg, A.L., Vanrolleghem, P.A. (2007), "Uncertainty in the environmental modelling process - A framework and guidance", *Environmental Modelling & Software*, 22(11): 1543-1556.
Reichert, P. (1997), "On the necessity of using imprecise probabilities for modelling environmental systems", *Water Science and Technology*, 36(5): 149-156.
Renard, B., Kavetski, D., Kuczera, G., Thyer, M. (2010), "Understanding predictive uncertainty in hydrologic modeling:The challenge of identifying input and structural errors", *2010*, 46: W05521.

Renard, B., Kavetski, D., Leblois, E., Thyer, M., Kuczera, G., Franks, S.W. (2011), "Toward a reliable decomposition of predictive uncertainty in hydrological modeling : Characterizing rainfall errors using conditional simulation", *Water Resour. Res.*, 47: W11516.

Roberts, G.O., Tweedie, R.L. (1996), "Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms", *Biometrika*, 83: 95-110.

Romanowicz, R.J., K.J. Beven, Tawn, J. (1994), "Evaluation of predictive uncertainty in nonlinear hydrological models using a Bayesian approach", In: Turkman, V.B.a.K.F. (Ed.), *Statistics for the Environment 2, Water Related Issues*, pp. 297-315.

Rosolem, R., Gupta, H.V., Shuttleworth, W.J., Zeng, X., de Gonçalves, L.G.G. (2012), "A fully multiple-criteria implementation of the Sobol′ method for parameter sensitivity analysis", *J. Geophys. Res.*, 117: D07103.

Safari, A., De Smedt, F., Moreda, F. (2009), "WetSpa model application in the Distributed Model Intercomparison Project (DMIP2)", *Journal of Hydrology*, In Press, Corrected Proof.

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

Schaefli, B., Talamba, D.B., Musy, A. (2007), "Quantifying hydrological modeling errors through a mixture of normal distributions", *J. Hydrology*(332): 303– 315.

Schoups, G., Vrugt, J.A. (2010), "A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors", *Water Resources Research*, 46: W10531.

Seber, G.A., Wild, C.J. (1989), "*Nonlinear Regression*". John Wiley and Sons, New York (NY).

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

Setegn, S.G., Srinivasan, R., Melesse, A.M., Dargahi, B. (2010), "SWAT model application and prediction uncertainty analysis in the Lake Tana Basin, Ethiopia", *Hydrological Processes*, 24(3): 357-367.

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

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

Shafii, M., Tolson, B., Matott, L.S. (2014a), "Addressing Subjective Decision-Making Inherent in GLUE-based Multi-Criteria Rainfall-Runoff Model Calibration", *Journal of Hydrology*, In Review.

Shafii, M., Tolson, B., Matott, L.S. (2014b), "Uncertainty-based multi-criteria calibration of rainfall-runoff models: a comparative study", *Stoch Env Res Risk A*, 28: 1493-1510.

Shamir, E., Imam, B., Gupta, H.V., Sorooshian, S. (2005), "Application of temporal streamflow descriptors in hydrologic model parameter estimation", *Water Resources Research*, 41(6).

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

Smith, M.B., Gupta, H.V. (2012), "The Distributed Model Intercomparison Project (DMIP) – Phase 2 experiments in the Oklahoma Region, USA", *Journal of Hydrology*, 418–419(0): 1-2.

Smith, T.J., Marshall, L.A. (2008), "Bayesian methods in hydrologic modeling: A study of recent advancements in Markov chain Monte Carlo techniques", *Water Resources Research*, 44: W00B05.

Sobol, I.M. (2001), "Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates", *Math. Comput. Simul.*, 55(1-3): 271-280.

Sorooshian, S., Dracup, J.A. (1980), "Stochastic parameter estimation procedures for hydrologie rainfall-runoff models: Correlated and heteroscedastic error cases", *Water Resources Research*, 16(2): 430-442.

Sorooshian, S., Duan, Q., Gupta, V.K. (1993), "Calibration of rainfall-runoff models: Application of global optimization to the Sacramento Soil Moisture accounting model", *Water Resources Research*, 29: 1185– 1194.

Spear, R.C., Hornberger, G.M. (1980), "Eutrophication in Peel Inlet, II: Identification of critical uncertainties via generalized sensitivity analysis", *Water Research*, 14(1): 43-49.

Starrfelt, J., Kaste, O. (2014), "Bayesian uncertainty assessment of a semi-distributed integrated catchment model of phosphorus transport", *Environmental Science: Processes & Impacts*.

Stedinger, J.R., Vogel, R.M., Lee, S.U., Batchelder, R. (2008), "Appraisal of the generalized likelihood uncertainty estimation (GLUE) method", *Water Resources Research*, 44: W00B06.

Storn, R., Price, K. (1997), "Differential Evolution - A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces", *Journal of Global Optimization*, 11(4): 341-359.

Tarantola, A. (2005), "*Inverse Problem Theory and Methods for Parameter estimation*", *Soc. Ind. Appl. Math.*, Philadelphia, Pa.

ter Braak, C., Vrugt, J. (2008), "Differential Evolution Markov Chain with snooker updater and fewer chains", *Stat Comput*, 18(4): 435-446.

Teytaud, O. (2007), "On the hardness of offline multi-objective optimization", *Evolutionary Computation*, 15(4): 475-491.

Thiemann, M., Trosset, M., Gupta, H.V., Sorooshian, S. (2001), "Bayesian recursive parameter estimation for hydrologic models", *Water Resources Research*, 37(10): 2521-2535.

Thyer, M., Renard, B., Kavetski, D., Kuczera, G., Franks, S.W., Srikanthan, S. (2009), "Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis", *Water Resources Research*, 45: W00B14.

Tierney, L. (1994), "Markov chains for exploring posterior distributions", *Ann. Stat.*, 22(4): 1701- 1728.

Tierney, L., Mira, A. (1999), "Some adaptive Monte Carlo methods for Bayesian inference", *Stat. Med.*, 18: 2507-2515.

Todini, E. (2008), "A model conditional processor to assess predictive uncertainty in flood forecasting", *International Journal of River Basin Management*, 6(2): 123-137.

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

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

Tolson, B.A., Shoemaker, C.A. (2008), "Efficient prediction uncertainty approximation in the calibration of environmental simulation models", *Water Resources Research*, 44: W04411.

Troutman, B.M. (1983), "Runoff prediction errors and bias in parameter estimation induced by spatial variability of precipitation", *Water Resources Research*, 19(3): 791-810.

Troutman, B.M. (1985), "Errors and Parameter Estimation in Precipitation-Runoff Modeling. 1. Theory", *Water Resources Research*, 21(8): 1195-1213.

Tsang, C.F. (1991), "The modelling process and model validation", *Groundwater*, 29(6): 825- 831.

van Griensven, A., Meixner, T. (2007), "A global and efficient multi-objective auto-calibration and uncertainty estimation method for water quality catchment models", *J. Hydroinformatics*, 9(4): 227-291.

van Straten, G., Keesman, K.J. (1991), "Uncertainty propagation and speculation in projective forecasts of environmental change: a lake-eutrophication example", *J. Forecasting*, 10: 163-190.

van Werkhoven, K., Wagener, T., Reed, P., Tang, Y. (2009), "Sensitivity-guided reduction of parametric dimensionality for multi-objective calibration of watershed models", *Advances in Water Resources*, 32(8): 1154-1169.

von Neumann, J. (1951), "Various techniques used in connection with random digits. Monte Carlo methods", *Natl. Bur. Stand.*, 12: 36-38.

Vrugt, J., Robinson, B. (2007), "Improved evolutionary optimization from genetically adaptive multimethod search", *PNAS*, 104(3): 708-711.

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

Vrugt, J.A., Gupta, H.V., Bouten, W., Sorooshian, S. (2003b), "A Shuffled Complex Evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters", *Water Resources Research*, 39(8): 1201, doi:10.1029/2002WR001642.

Vrugt, J.A., Robinson, B.A., Hyman, J.M. (2009a), "Self-Adaptive Multimethod Search for Global Optimization in Real-Parameter Spaces", *IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION*, 13(2): 243-259.

Vrugt, J.A., Sadegh, M. (2013), "Toward diagnostic model calibration and evaluation: Approximate Bayesian computation", *Water Resources Research*, 49(7): 4335-4345.

Vrugt, J.A., ter Braak, C.J.F., Diks, C.G.H., Higdon, D., Robinson, B.A., Hyman, J.A. (2009b), "Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling", *International Journal of Nonlinear Sciences and Numerical Simulation*, 10(3): 273-290.

Vrugt, J.A., ter Braak, C.J.F., Gupta, H.V., Robinson, B.A. (2008a), "Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling?", *Stoch Environ. Res. Risk Assess.*, 44: 1-16.

Vrugt, J.A., ter Braak, C. J. F.,, Clark, M.P., Hyman, J.M., Robinson, B.A. (2008b), "Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation", *Water Resour. Res.*, 44: W00B09.

Wagener, T., D. P., B., M. J., L., H. S., W., H. V., G., S., S. (2001), "A framework for development and application of hydrological models", *Hydrol. Earth Syst. Sci.*, 5(1): 13– 26.

Wagener, T., Gupta, H.V. (2005), "Model identification for hydrological forecasting under uncertainty", *Stochastic Environ. Res. Risk. Assess.*, 19.

Wagener, T., Montanari, A. (2011), "Convergence of approaches toward reducing uncertainty in predictions in ungauged basins", *Water Resources Research*, 47.

Wang, Z.-M., Batelaan, O., De Smedt, F. (1996), "A distributed model for water and energy transfer between soil, plants and atmosphere (WetSpa)", *Physics and Chemistry of The Earth*, 21(3): 189-193.

Wei, W., Clark, J.S., Vose, J.M. (2010), "Assimilating multi-source uncertainties of a parsimonious conceptual hydrological model using hierarchical Bayesian modeling", *J. Hydrology*, 394: 436–446.

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

Winsemius, H.C., Schaefli, B., Montanari, A., Savenije, H.H.G. (2009), "On the calibration of hydrological models in ungauged basins: A framework for integrating hard and soft hydrological information", *Water Resources Research*, 45.

Wood, E.F. (1976), "An analysis of the effects of parameter uncertainty in deterministic hydrologic models", *Water Resources Research*, 12(5): 925-932.

Yadav, M., Wagener, T., Gupta, H. (2007), "Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins", *Advances in Water Resources*, 30 (8): 1756–1774.

Yang, J., Reichert, P., Abbaspour, K.C., Yang, H. (2007), "Hydrological modelling of the Chaohe Basin in China: Statistical model formulation and Bayesian inference", *J. Hydrology*, 340: 167– 182.

Yang, J., Reichert, P., Abbaspour, K.C., Yang, H., Xia, J. (2008), "Comparing Uncertainty Analysis Techniques for a SWAT Application to the Chaohe Basin in China", *J. Hydrology*, 358: 1-23.

Yapo, P.O., Gupta, H.V., Sorooshian, S. (1998), "Multi-objective global optimization for hydrologic models", *Journal of Hydrology*, 204(1-4): 83-97.

Yilmaz, K.K., Gupta, H.V., Wagener, T. (2008), "A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model", *Water Resources Research*, 44: W09417.

Zhang, X., Hörmann, G., Fohrer, N., Gao, J. (2012), "Parameter calibration and uncertainty estimation of a simple rainfall-runoff model in two case studies", *Journal of Hydroinformatics*, 14(4): 1061-1075.