Uncertainty assessment of integrated distributed hydrological models using GLUE with Markov chain Monte Carlo sampling

Blasone, Roberta-Serena; Madsen, Henrik; Rosbjerg, Dan

Published in:
Journal of Hydrology

Link to article, DOI:
10.1016/j.jhydrol.2007.12.026

Publication date:
2008

Document Version
Early version, also known as pre-print

Uncertainty assessment of integrated distributed hydrological models using GLUE with Markov chain Monte Carlo sampling

Roberta-Serena Blasone[1], Henrik Madsen[2] and Dan Rosbjerg[1]

(1) *Institute of Environment & Resources, Technical University of Denmark, Bygningstorvet 115, DK-2800 Kgs. Lyngby, Denmark*
(2) *DHI Water · Environment · Health, Agern Allé 5, DK-2970 Hørsholm, Denmark*

Submitted

ABSTRACT

In recent years, there has been an increase in the application of distributed physically-based integrated hydrological models. Despite this, many questions regarding how to properly calibrate and validate distributed models and assess the uncertainty of the estimated parameters and the spatially-distributed responses are still quite unexplored.
Especially for complex models, rigorous parameterization, reduction of the parameter space and use of efficient and effective algorithms are essential to facilitate the calibration process and make it more robust. Moreover, for these models multi-site validation must complement the usual time validation. In this study we illustrate, through an application, a comprehensive framework for multi-criteria calibration and uncertainty assessment of distributed physically-based, integrated hydrological models. A revised version of the generalized likelihood uncertainty estimation (GLUE) procedure based on Markov chain Monte Carlo sampling is applied in order to improve the performance of the methodology in estimating parameters and posterior output distributions. The description of the spatial variations of the hydrological processes is accounted for by defining a measure of model performance that includes multiple criteria and spatially distributed information. An initial sensitivity analysis is conducted on the model to reduce the problem of overparameterization and to increase the robustness of the approach. It is demonstrated that the employed methodology increases the identifiability of the parameters of complex hydrological models and results in satisfactory multi-variable simulations and uncertainty estimates. However, the parameter uncertainty alone cannot explain the total uncertainty at all the sites, due to the additional uncertainty that is added when distributed data are not properly included in the model calibration. This study also indicates that properly distributed information of discharge is crucial in model calibration and validation.

Keywords: uncertainty assessment; integrated distributed hydrological model; multi-objective calibration; generalized likelihood uncertainty estimation; MIKE-SHE; Markov chain Monte Carlo.

1 INTRODUCTION

In recent years, there has been an increase in the application of physically-based, distributed, integrated hydrological models, such as MIKE-SHE (Graham and Butts, 2006), SWAT (Arnold et al., 1998; Neitsch et al., 2002) and TOPMODEL (Beven, 1995). One of the main reasons for this trend is the availability of more powerful computer resources, which allow using these models also in applications that were considered prohibitive few years ago. However, distributed models still lack the extensive investigations that have been conducted for lumped, conceptual rainfall-runoff (RR) models. Therefore, many questions regarding how to properly calibrate and validate distributed models and assess the uncertainty of the estimated parameters and the spatially-distributed responses are still quite unexplored. Refsgaard (1997) addressed some of the issues related to the increased difficulties in parameterisation, calibration and validation of distributed, integrated models compared to lumped RR models.

The higher complexity of physically-based, distributed models, compared to lumped RR models, arises because these models try to describe, with different degree of details, a spatial representation of the different physical phenomena occurring within the catchment. As a consequence, proper set up, calibration and validation of distributed, integrated models require a huge amount of geological, topographic, meteorological and hydrological data representing the various variables of interest.

The higher complexity of distributed models is also reflected in the larger number of parameters they include in comparison to RR models, which must be estimated in order to get satisfactory simulations of the system behaviour. Therefore, especially for complex models, rigorous parameterization and reduction of the parameter space are essential to facilitate the calibration process and make it more robust. Overparameterisation must be avoided to ensure a higher degree of credibility to the subsequent model prediction (Andersen et al., 2001). In this respect, Refsgaard (1997) suggested to assess the parameter values from field data as much as possible and to fix spatial patterns of parameters to simplify the calibration process. The dimensionality of the parameter space can also be reduced by means of sensitivity analysis (SA) on the model response. Through SA the parameters that are non-essential in influencing the model response (and can be fixed to their prior values) can be distinguished from those that have a strong impact on the model outputs (and should be included in the calibration and subsequent assessment of uncertainty). This is a common procedure in calibration of hydrological models that has previously been employed for

distributed models by Muleta and Nicklow (2005), Christiaens and Feyen (2001; 2002) and Mertens et al. (2004). An alternative way to decrease the number of calibration parameters is to fix the values of some of them to those estimated by a simpler model, as done, for example, by Sonnenborg et al. (2003), who used the calibration results found under steady-state conditions to constrain the parameter space of a transient model.

Refsgaard (1997) also pointed out the additional difficulties of validation of distributed models than of RR models. Not only time validation should be performed, but also validation of internal variables, i.e. a multi-site validation. Moreover, if the model is used with different discretization scales, the dependency of the processes on the modelling scale should also be tested (Vazquez et al., 2002; Vazquez and Feyen, 2007). Data requirements can be an obstacle for a proper validation, since, even if distributed measurements are usually available to physically characterize the catchment, spatially-distributed time series of observations of all the variables of interest are rare. This is one of the main reasons why, also in the cases where space-validation was conducted, distributed models have usually been calibrated and validated against discharge data only (Andersen et al., 2001; McMichael et al., 2006; Engeland et al., 2006). This method is inconsistent with spatially distributed modelling and it constitutes a limit to the performance of such models (Rosso, 1994). According to the knowledge of the authors, only few studies used multi-variable and multi-site data to calibrate and validate an integrated, distributed model (Refsgaard, 1997; Vazquez et al., 2002; Madsen, 2003).

Madsen (2003) proposed a framework for use of multiple criteria to measure the model performance, which is crucial in calibration and validation of distributed models. In particular, the proposed method allows to include the different variables of interest (multi-variable criteria), their spatial variability (multi-site criteria), and the different error functions applied for evaluating the performance of the simulated variables (multi-response criteria).

Due to their increased complexity, physically-based, distributed models require numerical codes, which are more complex and more computational demanding than those of lumped, RR models. Moreover, the larger dimensionality of the parameter space increases the number of model runs needed to calibrate and assess the uncertainty of distributed models. Applying more efficient calibration and uncertainty assessment procedures can limit the computational burdens. Global calibration techniques, such as the Shuffled Complex Evolution (SCE) algorithm by Duan et al. (1992), have been successfully applied in some calibration studies of distributed models (Madsen, 2003; Mertens et al., 2004; Blasone et al., 2007a). To reach convergence, global methodologies require a larger number of model runs than local gradient-based techniques, which have also been employed in conjunction with this type of models (Sonnenborg et al., 2003; Blasone et al., 2007a). On the other hand, it has been

demonstrated that local procedures have a high probability of converging to suboptimal solutions when they are applied to integrated, distributed models (Blasone et al., 2007a).

The Generalized Likelihood Uncertainty Estimation (GLUE) technique by Beven and Binley (1992) is an alternative procedure, which has been extensively used for simultaneous calibration and uncertainty assessment of different distributed models (Lamb et al., 1998; Freer et al., 2004; Mertens et al., 2004; Muleta and Nicklow 2005; Cho and Beven, 2006; McMichael et al., 2006). Compared to other methods, GLUE has the advantages of being easy to implement and, at the same time, of allowing the simultaneous assessment of the total uncertainty present in all the components of the modelling process. Moreover, GLUE allows a flexible definition of a function of model performance (likelihood function), which is capable of including several variables in model calibration and uncertainty assessment. This feature is particularly valuable for integrated, distributed models, for which the uncertainty of multi-variable, multi-site and multi-response criteria can be assessed.

The main drawbacks of the GLUE technique are the subjectivity involved in definition of the uncertainty (likelihood function, threshold on defining the behavioural solutions) and the huge number of model simulations required by the initial sampling of the solutions. The latter feature is only of concern when running computationally expensive models, such as integrated, distributed hydrological models, since it limits the maximum number of simulations to be run (McMichael et al., 2006). The stochastic nature of the sampling scheme employed in GLUE (usually the Latin Hypercube Sampling, LHS, approach) can deplete considerably the goodness of the statistics inferred from the retained solutions, if the initial sampling of the parameter space is not dense enough to include many good solutions. Blasone et al. (2007b) have recently demonstrated that using a Markov chain Monte Carlo (MCMC) sampling scheme in combination with GLUE improves the efficiency and effectiveness of the methodology. The properties of the revised GLUE procedure, which have been tested only for RR models, may prove to be particularly favourable in applications to complex, computationally expensive models.

In this study we illustrate a comprehensive framework for multi-criteria calibration and uncertainty assessment (UA) of physically-based, distributed, integrated, hydrological models. The UA method employed is the revised version of the GLUE procedure introduced by Blasone et al. (2007b), in which a MCMC method, the Shuffled Complex Evolution Metropolis (SCEM-UA) algorithm by Vrugt et al. (2003) is used as sampler of the prior parameter distributions. The proposed framework for UA is particularly designed for complex models that include a huge number of parameters and for which different types of observations are available for calibration. Despite the choices made for implementing the procedure have been tailored on a specific case study, the methodology employed can be generally used in calibration and uncertainty

assessment of distributed and complex models. The approach can be summarized in the following steps:

1) model setup, parameterization and choice of calibration parameters;
2) definition of performance criteria, aggregation and choice of likelihood function;
3) implementation of GLUE using a MCMC sampler;
4) model validation in time and space.

The case study considered is the Danish catchment of Karup River, for which an integrated, distributed, hydrological model is set up using the MIKE SHE modelling system (Graham and Butts, 2006). An initial SA is conducted on the model to reduce the problem dimensionality. The uncertainty of spatially distributed multi-variable model responses, river discharge and groundwater levels, is simultaneously assessed. A flexible objective (likelihood) function, capable of including and equally balancing different calibration criteria (i.e. multi-variable and multi-site data) is used to include multiple information in the uncertainty assessment. The availability of a large database for the Karup catchment allows conducting time and space validation of the results of the different model outputs.

This paper is structured as follows. Chapter 2 briefly introduces the techniques used for uncertainty assessment, GLUE and SCEM-UA, and the revised GLUE procedure. Chapter 3 describes the specific case study employed: the Karup catchment, the MIKE-SHE model setup and parameterization, and the SA employed to reduce the dimensionality of the parameter space. In Chapter 4 the methodology used to aggregate multiple objectives in the calibration and uncertainty assessment is presented. The results are shown and discussed in Chapter 5, while Chapter 6 summarizes the main conclusions drawn by the study.

2 GLUE WITH SCEM-UA SAMPLING

The Generalized Likelihood Uncertainty Estimation (GLUE) method (Beven and Binley, 1992) is a procedure for calibration and uncertainty assessment based on Monte Carlo (MC) simulations. Since its first introduction, it has been extensively applied in hydrology and environmental modelling to estimate the uncertainty associated with model outputs and parameter estimates (among others Beven and Binley, 1992; Freer et al., 1996; Lamb et al., 1998; Montanari, 2005). The reasons for the success of GLUE, compared to other methods, mainly reside in the fact that this technique allows assessing the global uncertainty present in the various modelling elements (i.e. input data, model structure and parameter error) in a way that is conceptually simple (it does not require any prior assumption on the error structure) and easy to implement.

The method is based on running a large number of model simulations with different parameter sets and inferring statistics on the outputs and parameter distributions based on the set of simulations showing the closest fit to the observations. The parameter sets used for the MC simulations are randomly sampled from the prior distribution of the parameters, chosen by the modeller based on his/her knowledge of the system. A function evaluating the model performance, the likelihood function, is defined and used to distinguish between behavioural (i.e. acceptable) and non-behavioural (i.e. non-acceptable) solutions. The behavioural solutions are selected by choosing a threshold on the likelihood function or as a percentage of the accepted solutions from the sample. The likelihood functions of the accepted solutions are then rescaled similarly to a probability measure, so that their cumulative sum equals 1. The distribution function of each parameter is obtained based on the likelihood associated with the various parameter values. The same is done with the model output at every time step of the simulation. From the output distribution, the output estimate is normally computed as the median and its uncertainty bounds are estimated as percentiles of the distribution, usually chosen as 5th and 95th in the majority of GLUE applications. It must be underlined that these are the percentiles of the behavioural solutions used to infer the parameter posteriors and the uncertainty bounds. The latter are not expected to include the same percentage of observations, unless by chance. More details on GLUE can be found in Beven and Binley, (1992), Freer et al. (1996) and Montanari (2005), among others.

The GLUE methodology allows including multiple sources of information in the likelihood function, thus being very flexible in this respect. This property is particularly useful for integrated, distributed models for which observations of multiple variables at different sites are available and the errors are computed using different likelihood functions. Multiple criteria can be accounted for in model calibration and UA in different ways. The most common aggregation method used in GLUE applications (Freer et al., 1996; Lamb et al., 1998) is that of performing Bayesian updating, i.e. by further conditioning the likelihood function, L, when data of different types are available:

$$L(\theta_i \mid Y_{1,2}) = L(\theta_i \mid Y_2) \cdot L(\theta_i \mid Y_1) / C \qquad (1)$$

where $L(\theta_i \mid Y_{1,2})$ is the posterior likelihood function of the parameter set θ_i obtained after conditioning on the observed variables Y_1 and Y_2, $L(\theta_i \mid Y_1)$ is the prior likelihood of the parameter set θ_i, calculated using the observation set Y_1, and $L(\theta_i \mid Y_2)$ is the likelihood measure calculated with the observations Y_2. C is a scaling constant, which guarantees that the cumulative sum of $L(\theta_i \mid Y_{1,2})$ over all the behavioural parameter sets θ_i equals unity. This composition rule can also be applied to update likelihood functions of different time intervals when new observations become available.

Another possible way of aggregating different information into one likelihood function is as weighted sum of several criteria. This is one of the methodologies most often used in multi-criteria calibration (Madsen, 2003; Sonnenborg et al., 2003; Mertens et al., 2004, van Griensven and Meixner, 2006) and also applied in this study. The different criteria can also be evaluated separately, without being aggregated in one likelihood function. In that case, the model performance can be assessed either by considering thresholds on the single criteria (Freer et al., 2004; Sahoo et al., 2005; Muleta and Nicklow, 2005) or by ranking the parameter sets according to the Pareto criterion (Madsen, 2003; Engeland et al., 2006). According to the knowledge of the authors, the latter criterion has never been used in conjunction with GLUE to discriminate behavioural solutions; however, this might be a future area of research, as also anticipated by Engeland et al. (2006).

It has been demonstrated by Blasone et al. (2007b) that the computational efficiency and the generation of statistically representative results by GLUE can be improved if the initial random sampling scheme of the prior parameter distributions is substituted by a MCMC sampler. This is because MCMC methods are particularly designed for converging to the region of highest posterior probability. Therefore, more statistically valid estimates of the parameters and the simulation uncertainty can be inferred, as more behavioural solutions are retained to estimate their posterior distributions. Moreover, the posterior parameter distributions can be adequately approximated with fewer simulations than with a LHS or MC sampling scheme.

The MCMC method used by Blasone et al. (2007b) is the Shuffled Complex Evolution Metropolis (SCEM-UA) algorithm (Vrugt et al., 2003), which is also used in the GLUE approach implemented in this study. SCEM-UA is a variation of the Shuffled Complex Evolution (SCE) algorithm (Duan et al., 1992), in which the downhill Simplex method to search for better solutions is substituted by the Metropolis algorithm (Metropolis et al., 1953). Other features that regulates the evolution of the solutions of the SCEM-UA procedure are the controlled random search (Price, 1987), competitive evolution (Holland, 1975) and complex shuffling (Duan et al., 1992). A complete description of the SCEM-UA algorithm can be found in Vrugt et al. (2003).

3 THE CATHCMENT AND THE MODEL

3.1 The catchment

The data used in this study are from the catchment of Karup River, which is located in the western part of Denmark. The Karup River and about 20 tributaries drain an area of around 440 km^2 (see Figure 1). The geology of the catchment is quite homogeneous and characterized by sandy soils with high permeability. The hydrological processes in the Karup catchment are mainly groundwater-driven and the river regime is dominated by baseflow. The aquifer is unconfined and it has a thickness ranging between 90 m at the

upstream part of the catchment and 10 m in the western and central areas. The unsaturated zone has a depth between 25 m at the catchment boundaries and 1 m along the river.

A long record of hydro-meteorological measurements is available for the Karup catchment, which was used in previous studies by Refsgaard (1997) and Madsen (2003), among others. This hydrological database includes rainfall measurements from nine stations (daily values), runoff at the river outlet and at three internal subcatchments (daily values), groundwater elevation data from 35 wells (recorded every 15 days) and temperature (daily values). To conduct a space validation of the results, the discharge and groundwater elevation data have been divided into a calibration dataset (discharge at river outlet plus 17 wells) and a validation dataset (discharge at the three subcatchment stations plus 18 wells). The calibration and validation sites are displayed in the map of the Karup catchment in Figure 1. The data used in this study covers the period from 1 June 1971 to 1 January 1978. The first period of 3.5 years is used to calibrate the model, while the following 3-year period is employed for validation. An initial warm-up period of 2.5 years is used to reduce the impact on model performance due to non-optimal initial conditions.

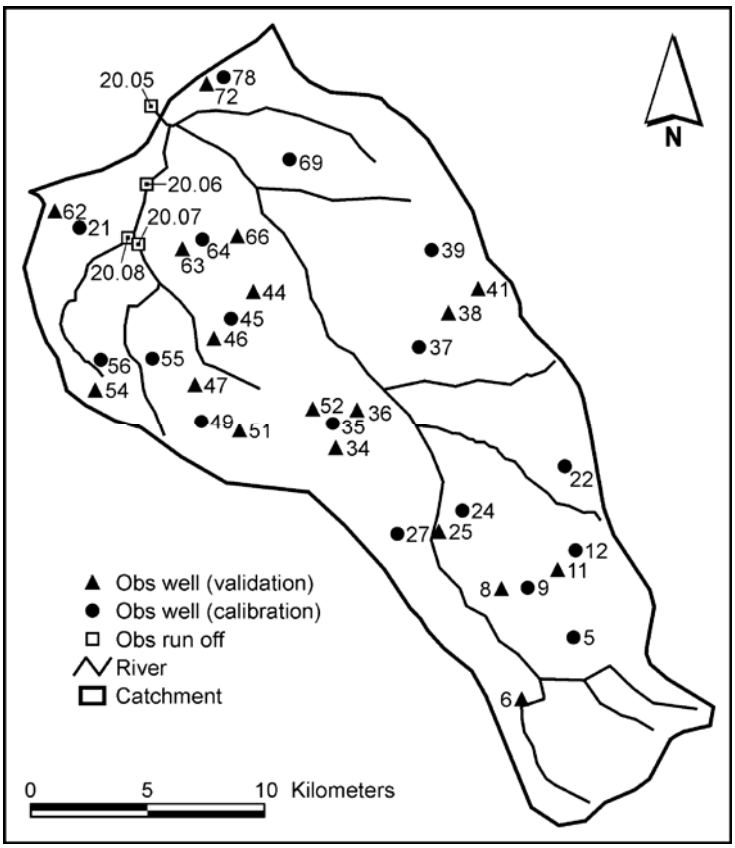

Figure 1. Karup catchment. Location of discharge gauging stations and wells.

3.2 Model setup, parameterization and sensitivity analysis

3.2.1 Model setup

The MIKE SHE modelling system (Graham and Butts, 2006) is used in this work to set up an integrated, spatially distributed, physically-based model of the Karup catchment, defined with a horizontal computational grid of 1 km x 1 km. The major catchment processes, which occur at the surface (overland flow and river runoff), in the unsaturated zone (evapotranspiration and infiltration to the aquifer) and in the saturated zone (groundwater flow and recharge) are described as well as their interactions. The components included in the model and their key parameters are shown in Table 1. Due to the high degree of complexity of this model, the number of parameters to assess is potentially huge. Since model oveparameterization create severe methodological problems for the calibration, validation and uncertainty assessment, the number of free parameters must be reduced. A way of doing so is by assessing parameter values from field data, where available, or by fixing spatial patterns of parameters, as done in the model set up of the Karup catchment.

Table 1. Model components: process description, parameterisation and parameters found to be sensitive.

Component	*Process description*	*Parameterisation*	*Sensitive*
River system	Muskingum routing scheme; river-aquifer interaction regulated by a leakage coefficient	Uniform for all branches: bed resistance and *LeakCoef*	*LeakCoef*
Overland flow	2D Saint-Venant equations	Uniform for all cells: Manning no.	
Drainage system	Linear reservoir description for each cell	Uniform for all cells: *DrnCoef*, *DrnLev*	*DrnCoef, DrnLev*
Evapotranspiration	Kristensen and Jensen (1975) model	Distributed using four vegetation classes: LAI, root depth and evapotranspiration parameters	
Unsaturated zone	Two-layers model: Yan and Smith (1994) description of interception, ponding, infiltration, evapotranspiration and groundwater recharge	Uniform in each layer: infiltration capacity, *ETdepth*, and soil moisture contents *SWCwp*, *SWCfc* and *SWCsat*	*SWCwp, SWCfc, SWCsat of upper layer, ETdept*
Saturated zone	Groundwater flow is described by the 2D Boussinesq equation	Distributed using 5 soil types: *Kh*, *Kv*, horizontal and vertical soil conductivities of the soil type	*Kh1, Kv1, Kh3, Kv3*
Snow melt accumulation	Degree-day approach	Uniform threshold temperature and degree-day coefficient	

Overland flow in the catchment is generated according to evaporation and infiltration processes along the flow path and has an additional contribution, if the first upper centimetres of the soil are saturated. The diffusive wave approximation of the Saint-Venant equations is used to describe the routing of surface runoff down-gradient towards the river system. The drainage system of the Karup catchment includes both natural and artificial drainage. The drainage flow is modelled using a linear reservoir description, which for each cell requires a drainage level, *DrnLev*, and a time constant (drainage coefficient), *DrnCoef*, that regulate how much and how fast water is drained. Both these parameters are assumed uniformly distributed in the catchment. The river system collects both the overland and the saturated zone flow. The Muskingum routing scheme is used for river flow routing. The water exchange between river and saturated zone is accounted for by a leakage coefficient, *LeakCoef*, which is assumed to be uniform for all river branches. The model also includes a description of the snowmelt process by using a simple degree-day approach.

The unsaturated flow in MIKE SHE is modelled as one-dimensional, vertical flow. Due to the high computational time required by implementing the subsurface processes using Richards' equation, the unsaturated zone is represented by a simple two-layer water balance model. The two-layer water balance method is based on a formulation presented by Yan and Smith (1994), which is suitable when the water table is shallow and the groundwater recharge is primarily influenced by evapotranspiration in the root zone. If sufficient water is available in the root zone, the water will be available for evapotranspiration. The module includes the processes of interception, ponding, infiltration, evapotranspiration and groundwater recharge. The model outputs comprise estimates of actual evapotranspiration and groundwater recharge. The input for the model includes the characterisation of the vegetation cover and the soil physical properties. The vegetation is described in terms of leaf area index (LAI) and root depth. The types of vegetation distributed in the catchment are classified according to the following land use and vegetation typologies: agriculture (57%), forest (18%), heath (7%) and wetland (18%). The soil properties include infiltration capacity and soil moisture contents at the wilting point (*SWCwp*), field capacity (*SWCfc*) and saturation (*SWCsat*). Capillarity rise occurs if the groundwater table rises above a defined thickness of the upper soil layer (*ETdepth*). All the unsaturated soil parameters are assumed uniformly distributed in the catchment.

The geological description of the saturated zone is defined with a vertical scale of 10 m. Five main soil types have been identified in the area by the Danish National Water Resources model (Henriksen et al., 2003) and are shown in Table 2. One specific soil type is assigned to each grid element, thus defining the geology of the area. The different soil types are characterized by specific hydrogeological parameters. The horizontal, *Kh*, and the vertical, *Kv*, hydraulic conductivities are assumed to be linked by a constant anisotropy factor. Thus, only *Kh* is varied, while *Kv* is set equal to one

tenth of the respective horizontal conductivity. The saturated zone is represented by a two-dimensional model defined as one computational layer. The hydraulic properties of each model grid cell are computed by considering the characteristics of all the different types of soil present in the earth column represented by the particular cell.

Table 2. Soil types used in the parameterization of the saturated zone.

Soil code	*Soil name*	*Description*
1	Melt water sand	Quarternary and Post-Glacial sand and gravel
2	Clay	Glacial, Inter-Glacial and Post-Glacial clay and silt
3	Quartz sand	Miocene, medium to coarse grained sand and gravel
4	Mica sand	Miocene, fine to medium grained sand
5	Mica clay/silt	Pre-Quarternary clay and silt

3.2.2 Sensitivity Analysis

Not all the parameters have a large impact on the model output. Thus, the problem dimensionality can be further reduced, if these parameters, also defined as non-sensitive, are fixed to their prior values, while only those significantly affecting the model response, the sensitive parameters, are assessed. To discriminate between sensitive and non-sensitive parameters, a sensitivity analysis (SA) is performed using several methods, as suggested by Christiaens and Feyen, (2002). These tests are applied to a sample of 400 parameter sets, which is obtained using the LHS technique. The sensitivity of two model responses is analysed separately; these are: the root mean square error of the discharge at the catchment outlet, $RMSE^q$, and the sum of the root mean square errors of the groundwater elevations of the 17 wells used for model calibration, $RMSE^h$, which are defined in the following in Eqs. (2) and (3). The Simlab software (Joint Research Centre of the European Commission, 2004) is used to analyse the data. The global SA tests applied are based on: visual analysis of scatterplots, Kolmogorov-Smirnov test, computation of Pearson product moment correlation coefficient (PEAR), Spearman coefficient (SPEA), Standardised Regression Coefficients (SRC) and Partial Correlation Coefficients (PCC). The last two measures are calculated also on the ranked parameters and output variables. The results of these different SA tests are quite homogeneous for the two criteria considered. Therefore, the sensitive parameters can be unambiguously determined, while the others are fixed to their previously manually-calibrated values.

For the saturated zone, the sensitive parameters are the hydraulic conductivities of two soil types, melt water sand, *Kh1*, and quartz sand, *Kh3*. In the unsaturated zone, the upper layer, which describes the evapotranspiration processes, is found to significantly affect the model response through two parameters defining the water contents *SWCfc* and *SWCsat* of the soil profile and the thickness of the layer where capillary rise can occur, *ETdepth*. For the surface water processes, the drainage level,

DrnLev, the drainage coefficient, *DrnCoef*, and the *LeakCoef*, are found to be sensitive. All these parameters and their prior ranges are shown in Table 3, together with the tied parameters.

Table 3. Parameters subject to calibration and tied parameters. * $Kv = 0.1 \cdot Kh$

Model component	*Parameter*	*Range*	*Units*	
River system	leakage coefficient, *LeakCoef*	$1 \cdot 10^{-8}$ - $1 \cdot 10^{-6}$	[s^{-1}]	calibrated
Drainage system	drainage level, *DrnLev*	–1.3 - –8	[m]	calibrated
	drainage constant, *DrnCoef*	$1 \cdot 10^{-8}$ - $1 \cdot 10^{-6}$	[s^{-1}]	calibrated
Unsaturated zone	soil water content at saturation, *SWCsat*	0.35 - 0.45	[0-1]	calibrated
	soil water content at field capacity, *SWCfc*	0.1 - 0.35	[0-1]	calibrated
	ET surface depth, *ETdepth*	0 - 3	[m]	calibrated
Saturated zone (hydraulic conductivity)	horizontal, soil 1, *Kh1*	$5 \cdot 10^{-5}$ - $5 \cdot 10^{-3}$	[m/s]	calibrated
	vertical, soil 1, *Kv1*	$5 \cdot 10^{-6}$ - $5 \cdot 10^{-4}$	[m/s]	tied*
	horizontal, soil 3, *Kh3*	$1 \cdot 10^{-4}$ - $1 \cdot 10^{-2}$	[m/s]	calibrated
	vertical soil 3, *Kv3*	$1 \cdot 10^{-5}$ - $1 \cdot 10^{-3}$	[m/s]	tied*

It might be argued that we claim to use a complex, distributed integrated model but end up calibrating only eight parameters, a number comparable to lumped conceptual models. It is important, in this respect, to distinguish between model complexity (given by the physical description of the processes, as well as by their spatial representation) and the number of parameters to calibrate and assess (i.e. the degree of freedom of the system). Even if the problem dimensionality is reduced, the higher complexity of our model as compared to lumped conceptual models is still preserved. In fact, apart from the simplified conceptual representation of the unsaturated zone, all the modelled hydrological processes are physically-based and spatially distributed. The spatial variability of the geology is explicitly accounted for. Moreover, the simplification of the problem is conducted following a rigorous methodology (Refsgaard, 1997). Where possible, the problem is simplified based on the knowledge of the catchment characteristics and field data (as, for example, by assuming uniformly distributed parameters, or by using soils maps to define zones with similar hydrological properties). Analysing the model sensitivity to the parameters (i.e. by the SA) is the second step in order to obtain a rigorous parameterization procedure.

4 IMPLEMENTATION OF GLUE: PERFORMANCE CRITERIA, AGGREGATION, LIKELIHOOD FUNCTION

4.1 Calibration criteria

Since the model couples groundwater with surface water processes, which both are of interest in the study, river discharge as well as groundwater table elevations are used for calibration and uncertainty assessment. The function used to represent the fit of the model simulations to the observations is the root mean squared error, RMSE. In this

way, two criteria are defined, the RMSE of the discharge, $RMSE^q$, and the aggregated RMSE of the wells, $RMSE^h$, which is the sum of the RMSE of the groundwater levels at the 17 wells used for calibration:

$$RMSE^q = \sqrt{\frac{\sum_t \left(q_t^{sim} - q_t^{obs}\right)^2}{N_{tq}}} \tag{2}$$

$$RMSE^h = \sum_j \left[\sqrt{\frac{\sum_t \left(h_t^{sim} - h_t^{obs}\right)^2}{N_{th}}} \right]_j \tag{3}$$

where q_t is the discharge (m^3/s) and h_t the groundwater elevation (m) at the observation time step t, the superscripts *sim* and *obs* indicate the simulated and observed data respectively, j is the index of the calibration well and N_{tq} and N_{th} are the total number of observations of discharge and groundwater elevation, respectively.

4.2 Objective function

In this study none of the two criteria considered is preferred to the other. Therefore, the objective function used by SCEM-UA to sample the parameter space should be a balanced aggregate of the runoff and groundwater elevation error functions. It must be noticed that the aggregated RMSE of discharge and groundwater wells have different orders of magnitude and different units of measure. Previous work conducted on the same catchment (Madsen, 2003) has demonstrated that a sharp trade-off exists between these two criteria. In this application, we aim at obtaining good model simulations of both discharge and groundwater levels. To select the most suitable aggregation method for this study, several MCMC experiments have been run using different objective functions. The results have been inspected in the Pareto space defined by the two calibration criteria. The chosen aggregate function is the one showing convergence to the region containing the more balanced solutions with respect to both calibration criteria. The aggregated objective function, F_i, associated with the i-th parameter set, θ_i, is given by:

$$F_i = \frac{RMSE^q(\theta_i)}{\min\{RMSE^q\}} + \frac{RMSE^h(\theta_i)}{\min\{RMSE^h\}} \tag{4}$$

where $\min\{RMSE^q\}$ and $\min\{RMSE^h\}$ are the minima of the $RMSE^q$ and $RMSE^h$ criteria, respectively, which are found so far in the solutions contained in the Markov chain.

The defined function has the property of automatically assigning equal weight to the criteria included in the calibration and uncertainty assessment processes, given the knowledge found so far in the calibration process on the minima of the two criteria, which are used as scaling factors for $RMSE^q$ and $RMSE^h$. In fact, $\min\{RMSE^q\}$ and $\min\{RMSE^h\}$ are varying, while the generation of solutions by SCEM-UA continues, as both criteria are progressively reduced during the sampling of the solutions. This requires recalculating F_i at the end of the SCEM-UA procedure using the obtained $\min\{RMSE^q\}$ and $\min\{RMSE^h\}$ to be able to compare the different solutions. However, within each population generation, the scaling values are the same, allowing for a homogeneous criterion for ranking of the solutions. The equal weighting property is preserved, also if more criteria are included in the definition of F_i, as long as each of them is rescaled by their respective minimum. If a different weight should be assigned to the criteria, this can easily be done by including weights in the expression of F_i. It must also be noted that $RMSE^q$ and $RMSE^h$ can assume their respective minima for different parameter sets (due to the trade-off), so that the minimum of F_i can be higher than the sum of the single minimum contributions.

4.3 Acceptance of behavioural solutions

Defining the acceptance criteria of the behavioural solutions is a critical point in the GLUE procedure, since it deeply affects the parameter distributions and the estimation of uncertainty bounds of model outputs. The general procedure in GLUE is to retain as behavioural solutions a given percentage of the sampled parameter sets (Lamb et al., 1998) or the solutions with a likelihood above a certain threshold (Freer et al, 1996; Mertens et al., 2004; McMichael et al., 2006). The latter method does not always provide uncertainty bounds capturing a satisfactory percentage of the observations. Increasing the number of retained parameter sets has the effect of including more uncertainty into the assessment, thus obtaining wider uncertainty bounds and capturing more observations (Montanari, 2005). A trade-off between the estimation accuracy of GLUE and the generation of uncertainty bounds including a high percentage of the observation was observed by Blasone et al. (2007b): larger uncertainty bounds are obtained, if more solutions are included in the behavioural set, at the cost of a decreased performance of the median GLUE output prediction. Thus, the selection of the behavioural parameter sets should be a compromise between these two tendencies.

In this study a total of 15000 parameter sets is generated by SCEM-UA. The selection of the behavioural solutions is based on both convergence of the MCMC method and on ensuring the generation of uncertainty bounds wide enough to include a large amount of the observations. The solutions visited by the MCMC sampler after convergence generate simulations very close to the observations and thus provide narrow uncertainty bounds, including relatively few observations. Therefore, the uncertainty intervals are enlarged by including more parameter sets, even if, by doing

so, the parameter variation described by the posterior distributions also accounts for other sources of uncertainty. A total of 250 solutions is in this case accepted as behavioural.

4.4 Transformation of the objective function into a likelihood function

The posterior distributions of the parameters are evaluated based on a chosen likelihood function, as it is done in the GLUE approach. The objective function described by Eq. (4) has not the properties of a likelihood function, as it can assume values outside the range [0, 1] and the sum of the retained likelihoods is not equal to 1. Therefore, a conversion is necessary before using the function F_i to infer posterior distributions of parameters and output. Similarly to what was done by Mertens et al. (2004), the likelihood function is calculated as the reciprocal of the objective function used in calibration and then rescaled by a constant factor:

$$L(\theta_i \mid Y) = \frac{1}{F_i} \cdot \frac{1}{C} \tag{5}$$

where Y indicates the observations and C is the normalizing factor introduced to ensure that the sum of the likelihood functions of the behavioural solutions equals 1:

$$C = \sum_i \frac{1}{F_i} \tag{6}$$

5 RESULTS

5.1 Posterior parameter distributions

The posterior distributions of the parameters are defined using the likelihood values associated with the behavioural solutions. It is expected that GLUE in conjunction with the SCEM-UA algorithm provides more realistic posterior parameter distributions than the classical GLUE procedure (Blasone et al., 2007b). Moreover, using the particular approach implemented in this study, it is expected to obtain unimodal distributions that uniquely define the parameters values.

The posterior distributions of the eight parameters considered are shown in Figure 2. All the parameters possess very well-defined and unimodal posterior distributions. From such distributions the parameter estimates can be unambiguously inferred as modal values, while the shape of the distributions indicate the degree of uncertainty of the estimates. Sharp and peaked distributions are associated with well identifiable parameters, while flat and/or spread distributions indicate more uncertain parameters values.

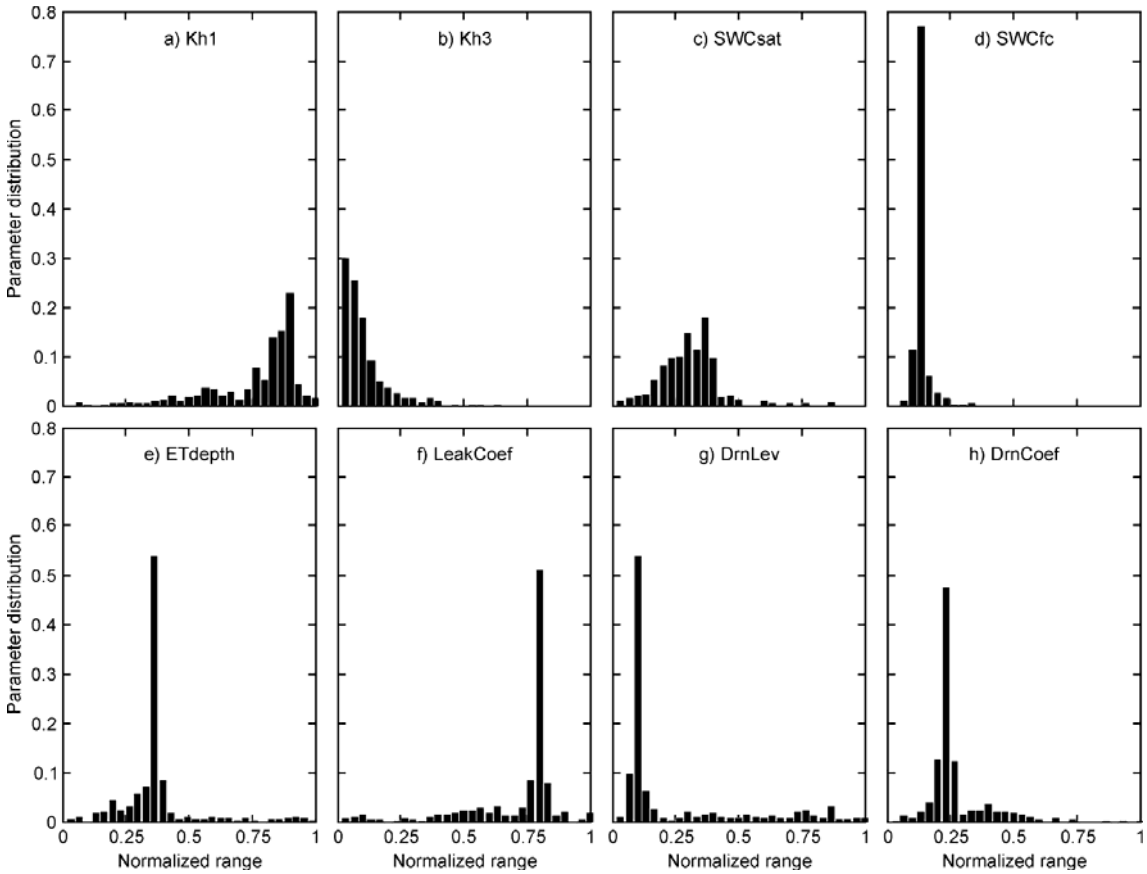

Figure 2. Parameter posterior distributions plotted in normalized range of the parameters with respect to the parameter range given in Table 2.

The posterior distributions are quite narrow for most of the parameters, indicating low uncertainty. Only three parameters *Kh1*, *Kh3* and *SWCsat*, have wider distributions. For the hydraulic conductivities this may be associated with some uncertainty in the parameterization of the saturated zone. In fact, the maps defining areas with the same hydraulic conductivity values have been determined on the basis of available knowledge of the soil types of the catchment. Therefore, the lower parameter identifiability can, in this case, be associated also with the uncertainty related to the model parameterization. For the upper layer of the unsaturated zone, the parameters are averaged values for the entire catchment, thus more uncertainty is expected in determining their values. Despite this, only the water content at saturation, *SWCsat*, shows a wide distribution, while the parameters *SWCfc* and *ETdepth* are more well-defined.

The followed approach is efficient in finding well-defined posterior parameter distributions. There are two main factors influencing this result. The first reason is the initial SA conducted on a larger parameter space, which allows excluding from the further analysis those parameters that have a relatively small impact on the model response (and are therefore less identifiable). The second reason is the use of a MCMC sampling method to search the parameter space, which results in less uncertain

parameter estimates than a random sampling approach (Blasone et al., 2007b). We do not know which of these two factors has the largest impact in this case, but we can acknowledge that their synergy has a positive effect in enhancing the parameter identifiability. This result is also confirmed by the low correlations found among the behavioral parameter sets. In fact, the maximum absolute value of the correlation coefficients is 0.57, found for the parameters *Kh1* and *DrnLev*.

5.2 GLUE estimates and uncertainty bounds

Using the behavioural parameter sets, discharge and groundwater elevation are simulated at both calibration and validation sites for the entire simulation period. The likelihood value of the parameter set generating a particular simulation is assigned to the respective model predictions. In this way the posterior distributions of river discharge and groundwater elevation at the different sites can be calculated. At each time step t of the simulation, the output estimate is obtained as the median of the distribution, as it is common practice in GLUE applications, and the uncertainty bounds are here defined as the 2.5 and 97.5 quantiles of the distribution. If these bounds are large enough to include most of the observations, it means that parameter variability alone can compensate for other sources of error, such as measurement and model structure errors, and, thus, it can account for the total output uncertainty.

This set of results allows conducting time and space validation of the GLUE posterior distributions of model outputs. First, the GLUE model predictions of discharge and groundwater levels and their associated uncertainty found for the calibration sites and the calibration period are analyzed and discussed. In the following part, the results obtained for the space and time validation are presented and compared to those obtained for the calibration data set.

5.2.1 Calibration results

5.2.1.1 Groundwater elevation

The GLUE estimates of groundwater elevation provide a quite good description of the dynamics at the majority of the locations considered. Based on a visual qualitative judgment of the results, 10 out of 17 wells show a good agreement between simulated and observed groundwater elevations in terms of dynamic description (Figure 3.a and 3.b). A satisfactory performance is obtained at 3 sites, while only at 4 sites the model prediction is poor (Figure 3.c). At the majority of the sites the observed groundwater table is either in the upper part or even above the GLUE uncertainty bounds. Therefore, the posterior probability of the model output is skewed to the bottom, as it is evident in the plots of Figure 3. On average, the GLUE median underestimates the groundwater levels of about 0.61 m.

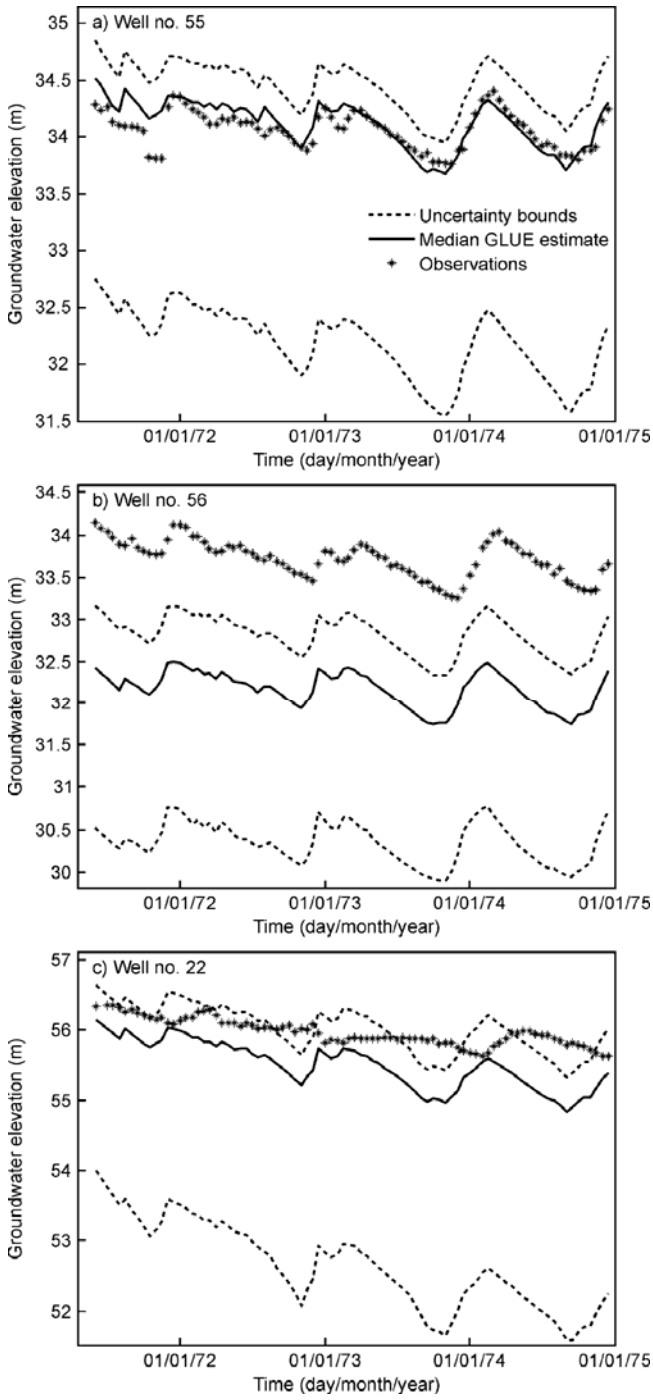

Figure 3. Calibration results. Groundwater elevation at well locations a): 55, b): 56 and c): 22. Observed data, median GLUE estimates and uncertainty bounds (2.5 and 97.5 quantiles of output distribution).

The amount of observations included in the uncertainty intervals is one of the main issues in evaluating GLUE results, since it is important that the GLUE bounds are able to account for all or most of the output variability. The percentage of measurements included inside the GLUE bounds at each of the 17 calibration locations are illustrated in Figure 4.a. At 6 out of 17 sites the uncertainty bounds do not include any of the

observations, as in the case illustrated in Figure 3.b for well no. 56, due to a large bias between observations and model simulations. At the remaining wells the GLUE bounds contain all or a part of the observations, which normally fall very close to the upper uncertainty bound (Figures 3.a and 3.c). Like the median estimates, also the GLUE uncertainty bounds follow quite well the dynamics. The average width of the bounds is 2.6 m, varying between a minimum of 1.4 m to a maximum of 3.9 m.

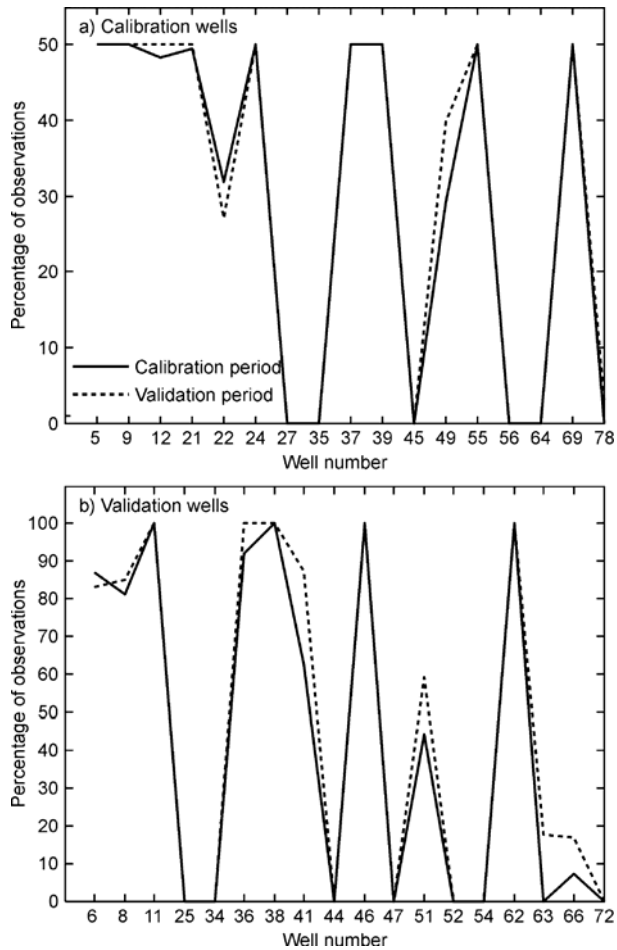

Figure 4. Percentage of observations included in the uncertainty bounds for the calibration and validation period. a): calibration wells and b): validation wells.

5.2.1.2 Discharge

The hydrograph at the catchment outlet, station 20.05, is simulated quite well by the GLUE estimate, as shown in Figure 5 for a selected period. Overestimation or underestimation trends are present, but they occur only during short time intervals. For the entire calibration period, the uncertainty bounds include a large percentage of the observations, about 82% (see Figure 6), thus describing most of the runoff variation.

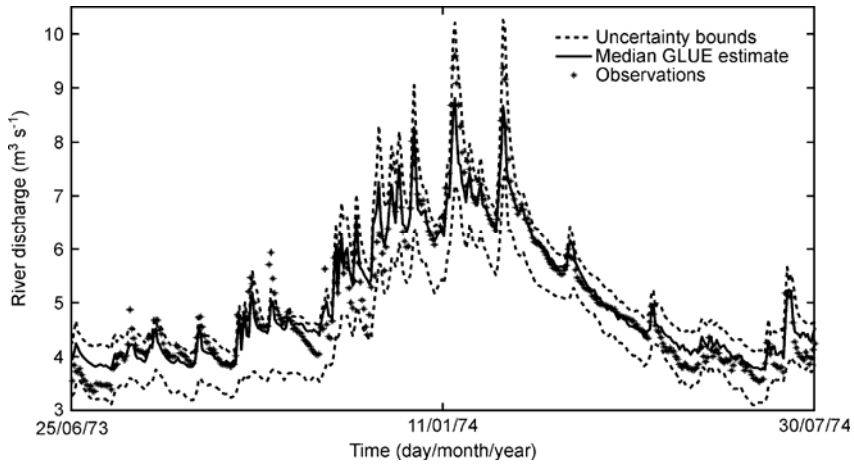

Figure 5. Calibration results for hydrological year June 1973-July 1974. Discharge at river outlet, gauging station no. 20.05. Observed data, median GLUE estimates and uncertainty bounds (2.5 and 97.5 quantiles of output distribution).

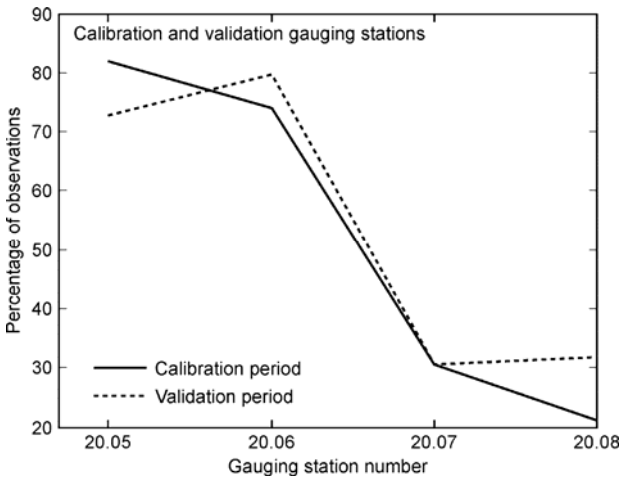

Figure 6. Percentage of observations included in the uncertainty bounds for the calibration and validation discharge gauging stations.

5.2.1.3 Comments

The calibration results of GLUE show that the MIKE SHE model is able to simulate quite well the dynamics of both groundwater table and runoff for the Karup catchment. The fact that the GLUE estimates reproduce correctly the dynamics of both variables of interest can be attributed also to the particular likelihood function employed. The method used to aggregate multiple criteria proves to be successful in generating balanced solutions, which simultaneously account equally for different variables, and it also allows including spatially distributed information.

On the other side, when looking at the GLUE estimates of the groundwater table, it is evident that the parameter uncertainty alone cannot explain the total uncertainty in simulating spatially-distributed observations. In fact, at the sites where the uncertainty bounds underestimate the observations, the results seem to be deeply

affected by the presence of other error sources, such as data and/or model errors. Significant biases in reproducing accurately the groundwater levels were found also by Madsen (2003) and Refsgaard (1997), who modelled the same catchment. This phenomenon, observed at few locations, can be the effect of the different scales of model simulations and observed data. In fact, while the measurements are point values, collected at groundwater wells, the model simulations are representative of average groundwater elevations within 1 km^2 areas. Moreover, in the Karup catchment, the groundwater table is characterized by a very high spatial gradient. For the observed groundwater levels spatial gradients up to a maximum of 3.5 m per km are seen, which is in the order of magnitude of the bias of the GLUE estimates. Thus, the bias in the model results can be attributed to the difference between model and observation scales. Moreover, it can be noticed that the scale of the uncertainty defined by the average width of the GLUE uncertainty bounds is about the same as the variation induced by the gradient of the groundwater table on a 1 km distance. This might indicate agreement between the uncertainty detected by GLUE and the scaling uncertainty present in the model.

Since the observed elevations can fluctuate a lot within 1 km^2, using a finer model grid would probably result in more precise groundwater level simulations. Another way of dealing with the bias of groundwater levels could be simply accepting the fact that it cannot be totally removed from the simulations. The calibration process could focus instead on optimizing the dynamics of the modelled responses by using an objective function that measures this, such as the variance of residuals.

5.2.2 Space validation of the results

5.2.2.1 Groundwater elevation

During the calibration period, the GLUE estimates and uncertainty bounds of the groundwater table at the validation wells show similar results as those obtained for the calibration sites. In particular, the dynamics is well reproduced at 11 out of 18 wells (see, for example, the results in Figure 7.a and 7.b), while only at 3 sites poor simulations are obtained (such as that shown in Figure 7.c for well no. 11). At the remaining 4 sites, the groundwater table dynamics is not perfectly reproduced, but it can still be considered satisfactory. Overestimation of the groundwater elevation occurs only at well no. 25, while, the underestimation of the groundwater levels previously illustrated for the calibration wells is present also at some validation sites. As a consequence, the median of the GLUE posterior distributions is shifted towards the upper interval of the uncertainty bounds also for the validation wells (Figure 7). The performance of the uncertainty bounds in including the observations is worse for the validation sites than for the calibration wells, as it can be noticed by comparing Figure 4.a to Figure 4.b. The measurements are almost completely outside the bounds at 9 out of 18 wells (as shown in Figure 7.b for well no. 54) and at only 7 locations more than

80% of the observations are contained into the uncertainty bounds (as in the cases shown in Figures 7.a and 7.c). The average width of the uncertainty intervals of the validation wells is 2.7 m, a similar value to that found for the calibration sites, but it has a larger variability, ranging between 0.9 to 4.6 m.

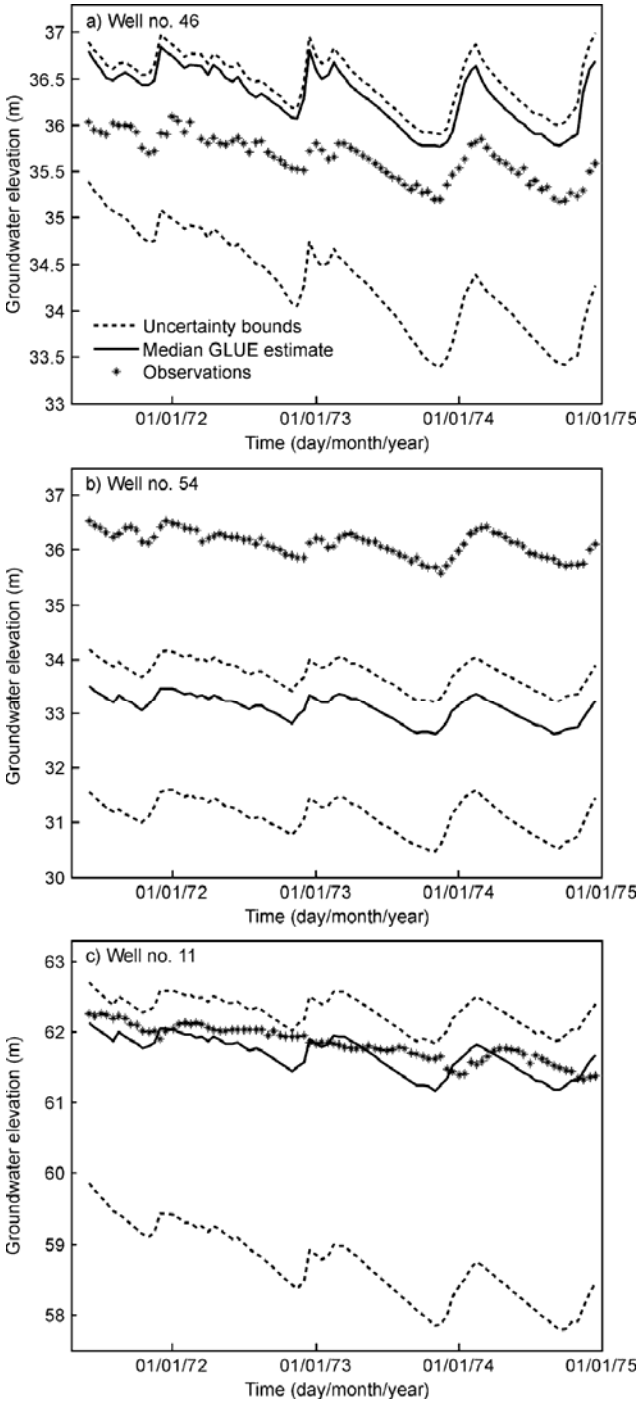

Figure 7. Space validation results. Groundwater elevation at well locations a): 46, b): 54 and c): 11. Observed data, median GLUE estimates and uncertainty bounds (2.5 and 97.5 quantiles of output distribution).

5.2.2.2 Discharge

During the calibration period the validation results of the discharge at internal river sections are not as satisfactory as those obtained for the groundwater wells, since the hydrograph shape is not always well reproduced and overestimation or underestimation of the hydrograph occurs.

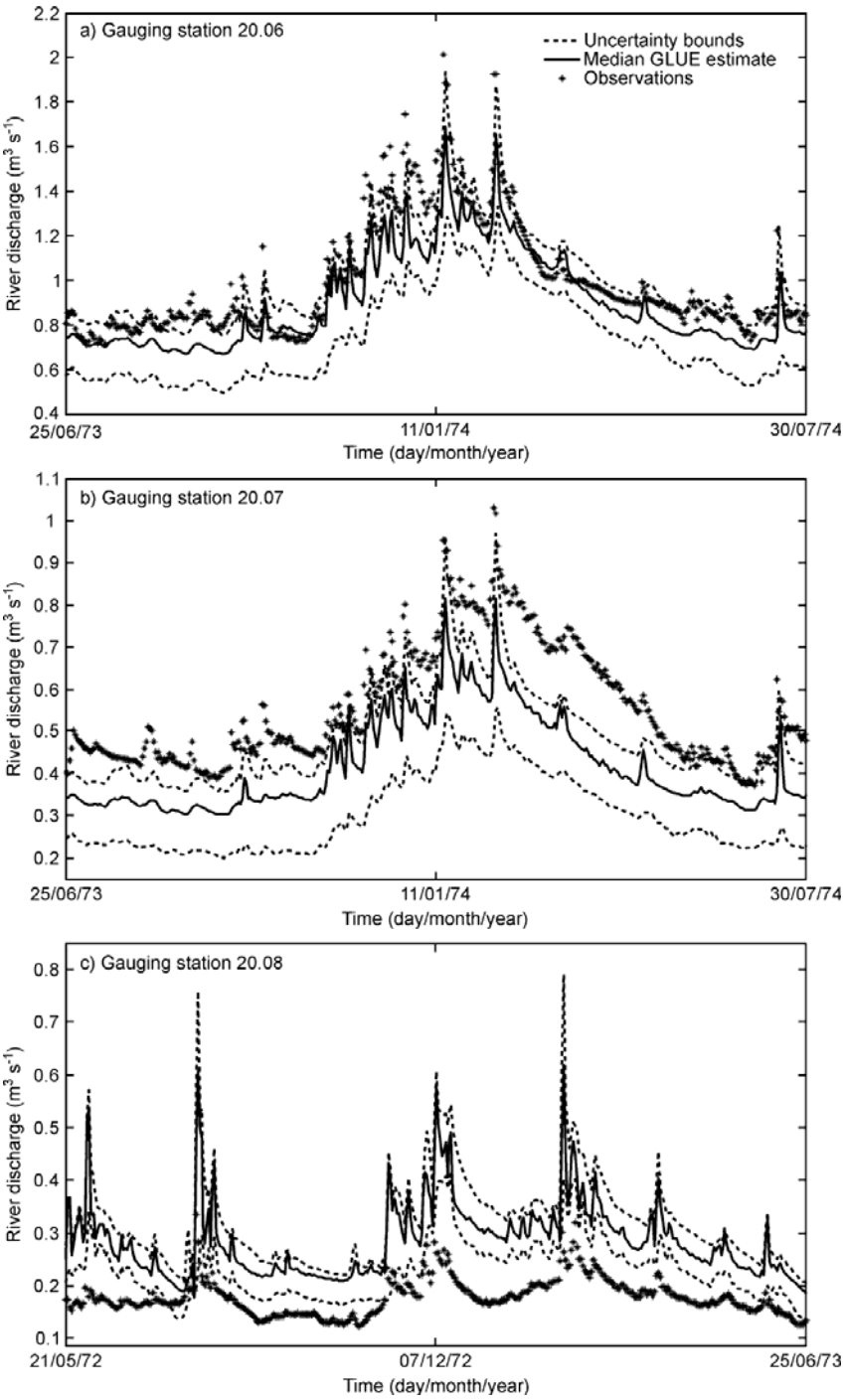

Figure 8. Space validation results for hydrological year June 1973-July 1974. Discharge at gauging stations no. a): 20.06, b): 20.07 and c): 20.08. Observed data, median GLUE estimates and uncertainty bounds (2.5 and 97.5 quantiles of output distribution).

At the gauging station closest to the river outlet, 20.06, the GLUE estimate is quite close to the observations during the first two years of the simulation, while it gets slightly worse afterwards (Figure 8.a). However, most of the measured data (around 74%) are contained within the uncertainty bounds, as shown in Figure 6, and the performance of the GLUE estimates in describing the runoff and its variation is similar to that of the station used for calibration. At the two stations at the smaller subcatchments, stations 20.07 and 20.08, the performance of the GLUE uncertainty bounds is much worse, as it can be noticed in Figure 8.b-c. and Figure 6. In particular, at station 20.08, despite the good agreement between the shape of the simulated and observed hydrographs, a severe overestimation of the runoff occurs during the first part of the calibration period and most of the measured data are below the uncertainty bounds (initial period shown in Figure 8.c). The uncertainty interval of station 20.07 includes more observations than those of station 20.08, but the GLUE estimate of the discharge follows less satisfactorily the measured one and the runoff is underestimated during the period shown in Figure 8.b.

5.2.2.3 Comments

A distributed hydrological model can be considered validated not only if it is able to produce good simulations for future conditions, but also if it is able to perform reliable predictions at internal/multi-site locations (Refsgaard, 1997). According to this definition, we can consider our model validated for the groundwater response. In fact, despite the presence of a bias in the simulations, as previously discussed, a similar satisfactory performance is achieved by the GLUE estimates and uncertainty bounds at both calibration and validation wells. The model is able to provide satisfactory spatially distributed predictions of the dynamics of groundwater levels with a performance similar to that achieved at the calibration wells. On the other hand, the ability of the uncertainty intervals of including a high percentage of observations is worse for the validation than for the calibration wells, since the bias of the simulations has a larger impact on the validation sites.

Despite the space validation of river discharge can be considered satisfactory, it is not as good as that found for the groundwater levels. The GLUE bounds, in fact, clearly overestimate or underestimate the hydrograph at the smaller subcatchments. Refsgaard (1997) observed the same phenomenon when using the discharge at the river outlet for calibration and the subcatchment stations for validation of the Karup model. Andersen et al. (2001) also found difficulties to validate discharge simulations, when using a downstream station for calibration and upstream stations for validation. This is a general problem that arises because of the different catchment scales of the sites used for calibration and validation. In fact, the downstream runoff is the sum of the contributions from the single subcatchments plus the additional contribution from the area drained by the downstream river branch. In this way, errors and biases with

opposite sign, which are present in the subcatchments simulations, may get cancelled when the total runoff is estimated. This makes it easier to obtain, in general, a good model fit of runoff at a larger basin scale than at the subcatchment scale. Using runoff gauging stations at different catchment scales to calibrate the model, the discharge simulations can be improved at both the upstream tributaries as well as at the downstream stations, as demonstrated by Andersen et al. (2001). Thus, our results could probably have been better if the error function of at least one of the smaller subcatchment gauging stations had been included into the likelihood function.

An additional consideration has to be made regarding the spatial variation of the results of the groundwater simulations. Considering both calibration and validation wells, a spatial pattern emerges for the dynamic description of the groundwater table by the GLUE procedure and the same occurs for the underestimation. The worse dynamic description is found at wells located in the south and eastern part of the catchment, while the underestimation of the groundwater table only occurs at the sites at the western side of the river. No correlation between underestimation of groundwater heads and poor dynamic description is found, as it is also confirmed by the plots in Figures 3 and 7.

5.2.3 Time validation of the results

When using an independent time series, the GLUE estimates of the uncertainty bounds and the variables of interest show similar performance to those obtained during the calibration period. Figures 4 and 6 illustrate that, at each site, the uncertainty intervals normally include the same percentage of observations during the calibration and validation periods. This result is found for both the groundwater table and the river discharge and for both calibration and validation sites. In particular, during the validation period a slight improvement in the dynamic description of the groundwater levels is observed at some calibration sites, i.e. wells 12, 22 and 24. This may be the effect of the long time influence of the initial conditions on the groundwater table simulations, which affects more the initial calibration period than the following validation. As for the time validation of the discharge at the river outlet, the GLUE estimate follows quite satisfactorily the hydrograph shape, but a slight overestimation of the runoff occurs. This also causes more observations to fall below the uncertainty bounds. The performance of the GLUE estimates for the internal gauging stations during the validation period is very similar to what is obtained for the calibration period. Actually, for station 20.08, the overestimation of the GLUE uncertainty interval is less severe and more observations are contained into the uncertainty bounds.

6 DISCUSSION AND CONCLUSIONS

This study presents an application of multi-criteria calibration and uncertainty assessment of a physically-based, distributed, integrated, hydrological model. A revised version of the GLUE procedure is applied in order to improve the performance of the methodology in estimating parameters and posterior output distributions. By using an efficient sampling scheme based on the MCMC method, more realistic parameter estimates can be obtained and the computational burden of the GLUE procedure can be reduced. The description of the spatial variations of the hydrological processes is accounted for by defining a measure of model performance that includes multiple criteria and spatially distributed information. To reduce the problem of overparameterization and to increase the robustness of the approach, a sensitivity analysis is initially conducted on the model using multiple techniques.

The employed methodology proves to be a valid tool to obtain well-defined posterior parameter distributions, from which unambiguous estimates can be inferred. Both the initial SA of the model and the MCMC sampling method contribute to reduce the uncertainty on the parameter posteriors. In particular, the SA helps in simplifying the problem, by excluding from the further analysis the parameters that do not have a large impact on the model response. The MCMC sampling method reduces the uncertainty associated with the parameter estimates, by locating solutions in the region of highest posterior probability of the parameters.

Several issues arise from this research on proper calibration and validation strategies in distributed modelling. The first one regards the importance of the aggregation function used to include multi-variable (in this case river discharge and groundwater table) and multi-site data in the model calibration and validation. A balanced aggregation of multiple criteria results in model outputs that simultaneously describe equally well the dynamics and the associated uncertainty of different variables, as demonstrated by the good performance of the discharge and groundwater level simulations obtained for the calibration sites. The spatial variability of the results can also be satisfactorily accounted for by properly aggregating multi-site information in the likelihood function, as shown by the results found for the groundwater table variable.

The second issue is related to the poorer space validation results obtained in this study for the discharge, which indicates that properly distributed information of discharge is crucial in model calibration and validation of distributed models. In particular, the locations used for calibration and validation must have similar characteristics and reproduce phenomena occurring at similar scales. While this can easily be done for variables, such as the groundwater levels, it is more problematic for the river discharge, since the simulation and observation errors of this variable are affected by the size of the associated subcatchment. For this reason, when modelling

discharge, we recommend to include spatial information from the subcatchments in the calibration process.

The results of this work also show that the biases that may occur in groundwater modelling can deteriorate the calibration results, as well as the ability of the uncertainty bounds of including the observations. This problem is likely to arise when point measurements are compared to area-averaged-simulations, as demonstrated by the simulations of groundwater elevation. This phenomenon can be particularly evident when the spatial gradient of the groundwater table is large. Thus, it is very important to properly define the size of the computational grid of the model in order to obtain more precise estimates. On the other hand, it must be remembered that a model is not a perfect representation of the physical reality, and discrepancies between observations and representation of the processes will always be present. In this respect, using distributed data to test the model results is more informative, as it allows making hypothesis on the nature of the possible calibration problems, i.e. whether they might arise because of data, model or parameter errors. If the model scale cannot be reduced, due for example to computational time or model stability constraints, the biases on the model outputs should be accepted and the calibration should focus, instead, on reproducing well the response dynamics.

As illustrated in this study, the parameter uncertainty alone cannot explain at all the sites the total uncertainty in simulating spatially-distributed variables. This is particularly true when biases are present in the model estimates, such as those arising from systematic measurement errors, as well as from the model scaling, as it is the case here for the groundwater elevations. However, the modeller should be aware that the GLUE procedure implemented in this study is just a tool to assess the errors intrinsic to the modelling process through parameter variation. The proper choices on model parameterization, parameter space sampling scheme and likelihood function aggregation method can only reduce, but cannot remove, the effect of other errors.

ACKNOWLEDGEMENTS

The authors thank Torben Dolin for formatting the figures included in this paper.

REFERENCES

Andersen, J., Refsgaard, J.C., Jensen, K.H., 2001. Distributed hydrological modelling of the Senegal River Basin – model construction and validation. Journal of Hydrology, 247, 200-214.

Arnold, J.G., Srinivasan, R., Muttiah, R.S. and 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.

Beven, K.J., Binley, A.M., 1992. The future of distributed models: Model calibration and uncertainty prediction. Hydrological Processes, 6, 279-298.

Beven, K.J., Lamb, R., Quinn, P.F., Romanowicz, R., Freer, J., 1995. Chapter 18: TOPMODEL. In: Singh, V.P. (Ed.), Computer Models of Watershed Hydrology. Water Resources Publications, Colorado, pp. 627-668.

Blasone, R.S., Madsen, H., Rosbjerg, D., 2007a. Parameter estimation in distributed hydrological modelling: comparison of global and local optimization techniques. Submitted. Under review.

Blasone, R.S., Vrugt, J.A., Madsen, H., Rosbjerg, D., Robinson, B.A., Zyvoloski, G.A., 2007b. Generalized likelihood uncertainty estimation (GLUE) using adaptive Markov chain Monte Carlo sampling. Submitted. Under review.

Christiaens, K., Feyen, J., 2002. Use of sensitivity and uncertainty measures in distributed hydrological modeling with an application to the MIKE SHE model. Water Resources Research, 38 (9), 1169. doi: 10.1029/2001WR000478.

Christiaens, K., Feyen, J., 2001. Analysis of uncertainties associated with different methods to determine soil hydraulic properties and their propagation in the distributed hydrological MIKE SHE model. Journal of Hydrology, 246 (1-4), 63-81.

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

Engeland, K., Braud, I., Gottschalk, L., Leblois, E., 2006. Multi-objective regional modelling. Journal of Hydrology, 327 (3-4), 339-351.

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, 2161-2173.

Freer, J.E., McMillan, H., McDonnell, J.J., Beven, K.J., 2004. Constraining dynamic TOPMODEL responses for imprecise water table information using fuzzy rule based performance measures. Journal of Hydrology, 291 (3-4), 254-277.

Graham, D.N., Butts, M.B., 2006. Flexible, integrated watershed modelling with MIKE SHE. In: Singh, V.P., Frevert, D.K. (Eds.), Watershed Models, pp. 245-272.

Holland, J., 1975. Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor, Michigan.

Kristensen, K.J., Jensen, S.E., 1975. A model for estimating actual evapotranspiration from potential transpiration. Nordic Hydrology. 6, 70-88.

Lamb, R., Beven, K., Myrabø, S., 1998. Use of spatially distributed water table observations to constrain uncertainty in a rainfall-runoff model. Advances in Water Resources, 22 (4), 305-317.

Madsen, H., 2003. Parameter estimation in distributed hydrological catchment modelling using automatic calibration with multiple objectives. Advances in Water Resources, 26, 205-216.

McKay, M.D., Conover, W.J., Beckman, R.J., 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics, 21, 239-245.

McMichael, C.E., Hope A.S., Loaiciga, H.A., 2006. Distributed hydrological modeling in California semi-arid shrublands: MIKE SHE model calibration and uncertainty estimation. Journal of Hydrology, 317, 307-324.

Mertens, J., Madsen, H., Feyen, L., Jacques, D., Feyen, D., 2004. Including prior information in the estimation of effective soil parameters in unsaturated zone modelling, Journal of Hydrology, 295, 251-269.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E., 1953. Equations of state calculations by fast computing machines. Journal of Chemical Physics, 21, 1087-1091.

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. doi:10.1029/2004WR003826.

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.

Neitsch, S.L., Arnold, J.G., Kiniry, J.R., Williams, J.R., 2001. Soil and Water Assessment tool (SWAT) user's manual version 2000. Grassland Soil and Water Research Laboratory, Temple, TX: ARS.

Price, W.L., 1987. Global optimization algorithms for a CAD workstation. Journal of Optimization Theory and Applications, 55 (1), 133-146.

Refsgaard, J.C., 1997. Parameterisation, calibration and validation of distributed hydrological models. Journal of Hydrology, 198, 69-97.

Rosso, R., 1994. An introduction to spatially distributed modelling of basin response. In: Rosso, R., Peano, A., Becchi, I., Bemporad, G.A. (Eds.), Advances in Distributed Hydrology. Water Resources Publications, pp. 3-30.

Sahoo, G.B., Ray, C., De Carlo, E.H., 2005. Calibration and validation of a Physically distributed hydrological model, MIKE SHE, to predict streamflow at high frequency in a flashy mountainous Hawaii stream. Journal of Hydrology, 327, 94-109.

Simlab Version 2.2, 2004. Simulation Environment for Uncertainty and Sensitivity Analysis. Joint Research Centre of the European Commission, http://simlab.jrc.it.

Sonnenborg, T.O., Christensen, B.S.B., Nyegaard, P., Henriksen, H.J., Refsgaard, J.C., 2003. Transient modeling of regional groundwater flow using parameter estimates from steady-state automatic calibration. Journal of Hydrology, 273 (1), 188-204.

van Griensven, A., Meixner, T., 2006. Methods to quantify and identify the sources of uncertainty for river basin water quality models. Water Science and Technology, 53 (1), 51-59.

Vazquez, R.F., Feyen, J., 2007. Assessment of the effects of DEM gridding on the predictions of basin runoff using MIKE SHE and a modelling resolution of 600m. Journal of Hydrology. 334 (1-2), 73-87.

Vazquez, R.F., Feyen, L., Fejen, J., Refsgaard, J.C., 2002. Effect of grid size on effective parameters and model performance of the MIKE-SHE code applied to a medium sized catchment. Hydrological Processes, 16 (2), 355-372.

Vrugt, J.A., Gupta, H.V., Bouten, W., Sorooshian, S., 2003. 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.

Yan, J., Smith, K.R., 1994. Simulation of integrated surface water and ground water systems - model formulation. Water Resources Bulletin, 30 (5), 879-890.