Fit-for-purpose analysis of uncertainty using split-sampling evaluations

A. VAN GRIENSVEN[1], T. MEIXNER[2], R. SRINIVASAN[3] & S. GRUNWALD[4]

1 *UNESCO-IHE Water Education Institute, Department of Hydroinformatics and Knowledge Management, PO Box 3015, 2601 DA Delft, The Netherlands*
a.vangriensven@unesco-ihe.org

2 *Department of Hydrology and Water Resources, University of Arizona, Tucson, Arizona 85721, USA*

3 *Spatial Sciences Laboratory, Texas A&M University, 1500 Research Parkway, Suite B223, College Station, Texas 77845, USA*

4 *Soil and Water Science Department, University of Florida, 2169 McCarty Hall, Gainesville, Florida 32611, USA*

Abstract An uncertainty assessment method for evaluating models, the Sources of UNcertainty GLobal Assessment using Split SamplES (SUNGLASSES), is presented, which assesses predictive uncertainty that is not captured by parameter or other input uncertainties. The method uses the split sample approach to generate a quantitative estimate of the fit-for-purpose of the model, thus focusing on the purpose for which the model is used. It operates by comparing the output to be used for decision making to its observed counterpart and the associated uncertainty. The described method is applied on a Soil Water Assessment Tool (SWAT) model of Honey Creek, a tributary of the Sandusky catchment in Ohio, USA. Water flow and sediment loads are analysed. In this case study the uncertainty estimated by the proposed method is much larger than the typically estimated parameter uncertainty.

Key words uncertainty; modelling; fit-for-purpose; catchment

Analyse d'incertitude liée à la satisfaction des objectifs à l'aide d'évaluations par subdivision d'échantillon

Résumé Une méthode d'évaluation de l'incertitude, appelée "Sources of UNcertainty GLobal Assessment using Split SamplES" (SUNGLASSES), dédiée à l'évaluation de modèles, est présentée. Elle évalue l'incertitude prédictive qui n'est pas capturée par les incertitudes des paramètres et des autres informations d'entrée. La méthode utilise l'approche de la subdivision de l'échantillon afin de générer une estimation quantitative de la satisfaction des objectifs du modèle, insistant ainsi sur la vocation du modèle. Elle procède par comparaison entre la sortie destinée à l'aide à la décision et l'observation et l'incertitude correspondantes. La méthode décrite est appliquée à une modélisation avec le modèle Soil Water Assessment Tool (SWAT) du ruisseau Honey Creek, affluent du bassin de Sandusky dans l'Ohio, Etats-Unis. L'écoulement et la production sédimentaire sont analysés. Dans cette étude de cas, l'incertitude estimée par la méthode proposée est bien supérieure à l'incertitude typiquement estimée pour un paramètre.

Mots clefs incertitude; modélisation; satisfaction d'objectif; bassin versant

1 INTRODUCTION

Model uncertainty analysis aims at a quantitative assessment of the reliability of model outputs. Many water quality modelling applications used to support policy and land management decisions lack this information and thereby lose credibility (Beck, 1987). Several sources of modelling unknowns and uncertainties result in the fact that model predictions are not a certain value, but should be represented with a confidence range of values (Kuczera, 1983a,b; Beven, 1993; Gupta *et al.*, 1998; Vrugt *et al.* 2003). These sources of uncertainty are often categorized as input uncertainties (such as errors in rainfall or pollutant sources inputs), model structure/model hypothesis uncertainties (uncertainties caused by inappropriateness of the model to reflect reality or the inability to identify the model parameters) and uncertainties in the observations used to calibrate/validate the model outputs.

Over the last decade, model uncertainty analysis has been investigated by several research groups from a variety of perspectives. Most of these methods have typically focused on model parametric uncertainty and do not address overall model predictive uncertainty, which encompasses uncertainty introduced by data errors (in input and output observations), model structural

Copyright © 2008 IAHS Press

errors and uncertainties introduced by the likelihood measure or objective function used to develop a model and its particular application to a single location (Kuczera & Mroczkowski, 1998; Thiemann *et al.*, 2001; Gupta *et al.*, 2003). It is important to note that proper assessment of model prediction uncertainty is somewhat of an unattainable goal and that questions about the informativeness of data and model structural error are typically best assessed in a comparison mode such as one model structure is superior in a specific situation as opposed to a wholesale accounting of the size of model structural error (e.g. Gupta *et al.*, 1998). This problem of not being able to quantitatively account for model structural error and errors introduced during the model calibration process has been a continuing source of problems and has generally prohibited the use of robust statistical methods for assessing uncertainty since these methods typically assume that the structural form of the model is correct, and that only model parameters need to be adjusted to properly match a computational model to the observations (Beven & Young, 2003; Gupta *et al.*, 2003). It is well known that hydrological models, particularly those of the rainfall–runoff process and even more so for models of water quality, are not perfect models and thus the assumption that the model is correct does not hold for the application of hydrological models (e.g. see Beven, 1993; Mroczkowski *et al.*, 1997; Boyle *et al.*, 2001; Meixner *et al.*, 2002).

A fundamental necessity noted by many is that the model must be evaluated using data not used for model calibration (Klemeš, 1986), also called the split sample methodology. Typically this split sample approach is conducted using one half of a data set to calibrate the model, and the second half of the time series to evaluate the calibration data set. The split sample methodology is not without flaws. It is well-known that a model typically performs worse during an evaluation time period than during the calibration period, and if a model performs almost as well during the evaluation period, it is generally accepted that the model is at least an acceptable representation of the natural system it represents (e.g. Meixner *et al.*, 2000).

Singh (1988) discusses the problem of model calibration at length and particularly notes that the model calibration problem has several fundamental attributes. First, model calibration starts with the problem that the data with which the model is being calibrated have some errors associated with them. Next, Singh (1988) notes that model calibration typically over-compensates for the data error and that the standard error (i.e. the difference between the simulations and observations) of the estimate ends up being smaller than it should be. When the calibrated model is then taken to another time period for evaluation, the standard error of prediction is generally larger than the original standard error of the data since the model was overly tuned to the observations for the calibration period. Singh notes that, while the standard error of the data and of the estimate can be quantified using standard methods, there is no formalized methodology for estimation of the standard error of the prediction, which we are most interested in. This problem remains to this day.

The framework established by Singh (1988) proves useful as we think about the problem of estimating model predictive uncertainty. Since most methods estimate the standard error of the estimate, they misleadingly represent only the reduced uncertainty level indicated by Singh (1988). Given the fundamental interest in knowing the uncertainty of model predictions, as opposed to estimates during the calibration period, it should prove useful to investigate methods that can assess the uncertainty of predictions. The discussion above would indicate that using the split sample approach and an assessment of model performance during the evaluation period would be useful for estimating the overall model predictive uncertainty.

Many researchers have noted the problem that parameter uncertainty was much smaller than expected for the level of trust we should have in model predictions (Beven & Freer, 2001; Thiemann *et al.*, 2001; Freer *et al.*, 2003). Here a methodology—Sources of UNcertainty GLobal Assessment using Split SamplES (SUNGLASSES)—is presented that uses a split sample approach to estimate overall model predictive uncertainty, and these results are compared to those garnered using a previously developed parametric uncertainty method based on statistical approaches, ParaSol (Parameter Solutions) (van Griensven & Meixner, 2007). The SUNGLASSES and ParaSol approaches are then compared using the commonly used river basin water quality model, the Soil Water Assessment Tool (SWAT).

2 METHODS

The ParaSol method is an optimization and statistical uncertainty method that assesses model parameter uncertainty. On top of ParaSol, SUNGLASSES uses all parameter sets and simulations performed by ParaSol to re-estimate uncertainty using a split-sampling procedure. Additional sources of uncertainty are detected by means of an evaluation period in addition to the calibration period.

2.1 Description of ParaSol

The method "ParaSol" (parameter solutions) (van Griensven & Meixner, 2007) is developed to perform the optimization and a model parameter uncertainty analysis for complex models. Distributed (water quality) models, typically have a high number of parameters, high parameter correlations, several output variables and a complex structure leading to multiple minima in the objective function response surface. The ParaSol method calculates the objective function (OF) based on model outputs and observation time series for a selected variable and aggregates several fitting criteria to a global optimization criterion (GOC). ParaSol minimizes the OF or a GOC using the SCE-UA algorithm and performs uncertainty analysis with a choice between two statistical concepts: χ^2 statistics that are discussed below and Bayesian statistics (van Griensven & Meixner, 2007).

2.1.1 The shuffled complex evolution (SCE-UA) algorithm The SCE-UA algorithm is a global search algorithm for the minimization of a single function (Duan *et al.*, 1992). It combines the direct search method of the simplex procedure with the concept of a controlled random search of Nelder & Mead (1965), a systematic evolution of points in the direction of global improvement, competitive evolution (Holland, 1975) and the concept of complex shuffling. In a first step (zero-loop), the algorithm selects an initial "population" by random sampling throughout the feasible parameters space for *P* parameters to be optimized (delineated by given parameter ranges).

The SCE-UA has been widely used in watershed model calibration and other areas of hydrology such as soil erosion, subsurface hydrology, remote sensing and land surface modelling (Duan, 2003). It was generally found to be robust, effective and efficient (Duan, 2003). The SCE-UA has also been applied with success on SWAT for the calibration of the hydrological parameters (Eckardt & Arnold, 2001) and of the hydrological and water quality parameters (van Griensven & Bauwens, 2003).

2.1.2 Objective functions Within an optimization algorithm it is necessary to minimize or optimize a function that replaces the expert perception of curve-fitting during manual calibration. There are a wide array of possible error functions to choose from and many reasons to pick one *versus* another (for some discussions on this topic see Gupta *et al.*, 1998; Legates & McCabe, 1999). This study used a sum of the squares of the errors (SSE) that is similar to the mean square error method (MSE), and aims to match a simulated series to a measured time series:

$$\mathrm{SSE} = \sum_{n=1}^{N} \left[y_{n,\mathrm{sim}} - y_{n,\mathrm{obs}} \right]^2 \tag{1}$$

where N is the number of pairs consisting of the simulation $y_{n,\mathrm{sim}}$ and the corresponding observation $y_{n,\mathrm{obs}}$.

2.1.3 Global optimization criterion Since the SCE-UA minimizes a single function, it cannot be applied directly for multi-objective optimization. There are several methods available in the literature to aggregate objective functions to a global optimization criterion (Madsen, 2003; van Griensven & Bauwens, 2003) for multi-objective calibration, but they do not provide uncertainty analysis.

Based on Bayesian theory, a GOC is defined by the following equation (van Griensven & Meixner, 2007):

Copyright © 2008 IAHS Press

$$\text{GOC} = \sum_{m=1}^{M} \frac{\text{SSE}_m N_m}{\text{SSE}_{m,\min}} \tag{2}$$

where SSE$_m$ is the sum of the squared errors for variable m; N_m is the number of observations of variable m; and SSE$_{m,\text{ min}}$ is the minimum value that was found for the objective function SSE$_{m.}$ for all performed simulations (see below). For details of ParaSol and how equation (2) is derived, the reader is referred to van Griensven & Meixner (2007).

The probability $p()$ that the parameter vector $\boldsymbol{\theta}$ is the true one—or the likelihood of the parameter vector $\boldsymbol{\theta}$—consisting of the P parameters (θ_1, θ_2, …, θ_P) when conditioned by the observation $y_{n,\text{obs}}$ can be related to the GOC according to:

$$p(\theta \mid Y_{\text{obs}}) \propto \exp[-\text{GOC}] \tag{3}$$

Thus the sum of the squares of the residuals for a given variable gets a weight that is equal to the number of observations divided by the minimum of SSE for that variable. The minima of the individual objective functions are not initially known. To solve this problem, an update is performed for the minima of the objective functions after each loop in the SCE-UA optimization using the newly gathered information within the loop. The main advantage of using equation (2) to calculate the GOC is that it allows for a global uncertainty analysis considering all components of the objective function as described below.

Note that, in all cases, the probability for the entire parameter space is to be equal to one. Therefore, a rescaling or normalization is performed in order to assign absolute probabilities through a weighting factor that is equal to the integration of equation (3) over the entire parameter space (Box & Tiao, 1973).

2.1.4 Uncertainty analysis method The uncertainty analysis divides the simulations that have been performed by the SCE-UA optimization into "good" simulations and "not good" simulations, similarly to the GLUE methodology (Beven & Binley, 1992). The simulations gathered by SCE-UA are very valuable as the algorithm samples over the entire parameter space with a focus of solutions near the optimum/optima.

The ParaSol algorithm uses a threshold value for the objective function (GOC) to select the "good" simulations by considering all the simulations that give an objective function below this threshold. The threshold value can be defined by χ^2 statistics where the selected simulations correspond to the confidence region (CR). For a single objective calibration, the SCE-UA will find a parameter set $\boldsymbol{\theta}$* consisting of the P free parameters (θ^*_1, θ^*_2, …, θ^*_P), that corresponds to the minimum of the OF($\boldsymbol{\theta}$*). A criterion will be used for selecting the "good" parameter sets that have an objective function higher than the minimum but below the criterion, using the equation:

$$c_{\text{ParaSol}} = \text{OF}(\boldsymbol{\theta}^*)\left(1 + \frac{\chi^2_{P,0.975}}{N-P}\right) \tag{4}$$

whereby the $\chi^2_{P,0.975}$ gets a higher value for more free parameters P.

For multi-objective calibration, the selections are made using the GOC of equation (2) that normalizes the sum of the squares for the total of observations N_T, equal to the sum of N_1, …, N_m, …, N_M observations. A threshold for the GOC is calculated by:

$$c_{\text{ParaSol}} = \text{GOC}(\boldsymbol{\theta}^*)\left(1 + \frac{\chi^2_{P,0.975}}{\sum_{m=1}^{m} N_m - P}\right) \tag{5}$$

thus all simulations with GOC(θ) < c_{ParaSol} are deemed acceptable. The uncertainty bounds for the model outputs are computed by propagating all the parameter sets, that were simulated during the

Copyright © 2008 IAHS Press

ParaSol optimisation, that pass the threshold (GOC < c_{ParaSol}) by taking the minimum and maximum output values of all these simulations.

2.2 Description of SUNGLASSES

In order to get a stronger evaluation of the prediction power of a model, the SUNGLASSES method is designed to assess predictive uncertainty that is not captured by parameter uncertainty. The goal of the method is to have a good fit to observations (expressed by the SSEs) and a proper assessment of systematic errors that may lead to over- or underestimations of the outputs to be used for decision making. The latter is evaluated by assessing the increases in model prediction errors when simulations are done outside the calibration period by using a split sample strategy, whereby the evaluation period is used to re-evaluate the assessed uncertainties on the model outputs. The assessment of the prediction power is hereby based on a fit-to-purpose criterion that is related to the sort of decision the model is being used for (for instance the total exported pollution loads, or the percentage of time oxygen is below a threshold value). In case a systematic under- or overestimation is found for this evaluation criteria for all the ParaSol results, an update (increase) of the threshold *c* is done whereby more simulations will be selected according to the inequality: GOC(θ) < $c_{\text{SUNGLASSES,}}$ whereby $c_{\text{ParaSol}} < c_{\text{SUNGLASSES}}$. The parameter $c_{\text{SUNGLASSES}}$ is not defined by an equation, but by the minimum threshold value that gives the smallest uncertainty ranges on the output values possible and includes both under- and overestimations of the decision variables for both the calibration and validation periods. These final uncertainty ranges depend on: the GOC composed of objective functions representing a fit to observations, on the one hand, and, on the other, a fit-to-purpose evaluation criterion (to be related to decision making) for defining a new value *c* that is used to estimate uncertainty bounds of the model outputs. The GOC is used to assess the degree of error on the process dynamics, while the evaluation criterion defines a threshold on the GOC in such a way that no systematic over- or underestimation of the output values is allowed.

In brief, SUNGLASSES is applied by the following procedure:

1. Split the data sets into two parts—calibration and validation data set—typically, two periods in time with equal length.
2. Define a calibration objective (e.g. minimisation of the mean-squared-error or global optimisation criterion) and perform a calibration followed by parameter uncertainty method. Statistical methods can be used to define a threshold considering parameter uncertainty. In this paper, ParaSol was used to define such a threshold.
3. Define what outputs are of interest for model-based decision making (in this case mass flux model bias since the total erosion volume is of important for erosion control).
4. Compute the uncertainty bounds for the model outputs for both calibration and validation period by propagating all the acceptable parameter sets that pass the threshold for parameter uncertainty defined in Step 2.
5. Verify whether there is a bias in the calibration/validation period with regard to the model output to be used for decision making, as defined in Step 3.
6. If there is no bias, no action is needed and the uncertainty bounds for SUNGLASSES are equal to the uncertainty bounds of ParaSol.
7. If there is a bias, a threshold $c_{\text{SUNGLASSES}} > c_{\text{ParaSol}}$ needs to be defined. The SUNGLASSES method operates by ranking the GOCs (Fig. 1) and by gradually increasing the threshold until the corresponding uncertainty bounds on the model outputs, computed by propagating all the simulations with GOC lower than the threshold, contain simulations with both under- and overestimations of the model bias. Thus, the simulation do not systematically under- or overestimate the observed value.

This methodology is flexible in the sense that different combinations of objective functions can be used within the GOC. Also, alternatives for the bias as the criterion for the model

Copyright © 2008 IAHS Press

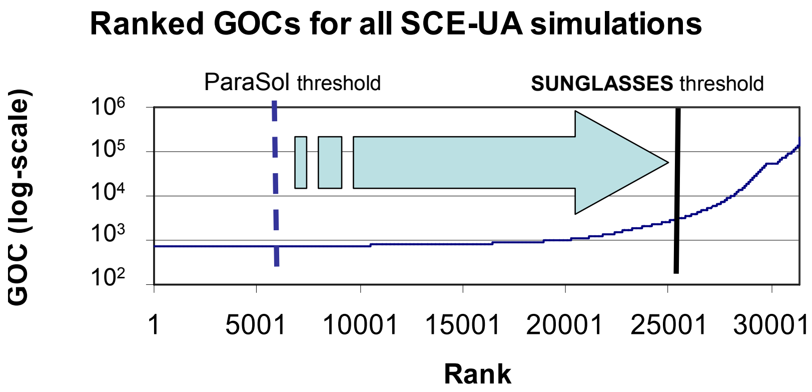

Fig. 1 Selection of good parameter sets using a threshold imposed by ParaSol or by SUNGLASSES.

evaluation period are possible depending on the model outputs to be used for decision making. Examples of alternative criteria are the percentage of time a certain output variable is higher or lower than a certain threshold (being common for water quality policy) or the maximum value or the value of a certain model prediction percentile (often important for flood control).

3 APPLICATION ON A SWAT MODEL

The SUNGLASSES method was programmed within SWAT (to be part of SWAT2005) and was applied on a small and simple catchment model for evaluation purposes of the methodology and for comparison purposes to the ParaSol method.

3.1 SWAT

The Soil and Water Assessment Tool (SWAT) (Arnold *et al.*, 1998) is a semi-distributed and semi-conceptual program that calculates water, nutrient and pesticide transport at the catchment scale on a daily time step. It represents hydrology by interception, evapotranspiration, surface runoff (SCS curve number method, USDA Soil Conservation Service, 1972), soil percolation, lateral flow and groundwater flow and river routing (variable storage coefficient method, Williams, 1969) processes. The nutrient, erosion, crop and pesticide processes are based on the GLEAMS (Leonard *et al.*, 1987), CREAMS (Knisel, 1980) and EPIC (Williams *et al.*, 1984) modelling tools. The catchment is divided into sub-basins, river reaches and hydrological response units (HRUs). While the sub-basins can be delineated and located spatially, the further subdivision into HRUs is performed in a stochastic way by considering a certain percentage of sub-basin area for each combination of soil and land-use classes, without any specified location in the sub-basin.

3.2 Parameter change options for SWAT

In the ParaSol algorithm, as implemented in SWAT2005, parameters affecting hydrology or pollution can be changed either in a lumped way (over the entire catchment), or in a distributed way (for selected sub-basins or HRUs). They can be modified by replacement, by addition (for an absolute change) or by a multiplication (for a relative change). A relative change means that the parameters, or several distributed parameters simultaneously, are changed by a certain percentage. However, a parameter is never allowed to go beyond the predefined parameter range. For instance, all soil conductivities for all HRUs can be changed simultaneously over a range of –50 to +50% of their initial values that are different for the HRUs according to their soil type. This mechanism allows for a lumped calibration of distributed parameters while keeping their relative physical meaning (soil conductivity of sand will be higher than soil conductivity of clay).

Copyright © 2008 IAHS Press

3.3 Honey Creek model description

Honey Creek is a sub-basin within the Sandusky River watershed (Ohio, USA) within the Erie Watershed and Great Lakes basin. The SWAT model for Honey Creek was abstracted from the SWAT model of the Sandusky that was provided by the University of Florida to the research group at the University of California, Riverside (van Griensven *et al.*, 2006). It covers an area of 338 km^2. To facilitate large numbers of simulations, a minimalist model structure was chosen consisting of one sub-basin, represented by five HRUs, a river reach and a point source. Daily observations during the years 1998–1999 were used to calibrate the model. These consisted of 661 flow observations and 518 sediment concentration observations.

The model for this basin is used to develop sediment management plans. Therefore, calibrations and uncertainty analysis are applied on daily series of flows and sediment loads. A sensitivity analysis was to select the 10 most important parameters for flow and sediments (van Griensven *et al.*, 2006) (Table 1). The distributed parameters are changed in a lumped way by considering a single relative change that is applied on all. A more detailed description of the SWAT model is provided in van Griensven *et al.* (2006).

Table 1 Parameters used in calibration, with sensitivity rank according to SSE for the daily observed flows (*Q*) and the sediment concentrations (SS).

Parameter	Description		*Q*	SS
SMFMX	Maximum melt rate for snow during (mm °C^{-1} d^{-1})	Lumped	2	17
ALPHA_BF	Baseflow alpha factor (d).	Lumped	8	1
ch_k2	Channel conductivity (mm/h)	Distributed	5	14
USLE-P	USLE equation support practice (*P*) factor.	Distributed	No effect	4
CN2	SCS runoff curve number for moisture condition II.	Distributed	3	2
sol_awc	Available water capacity of the soil layer (mm/mm soil).	Distributed	10	3
surlag	Surface runoff lag coefficient	Lumped	1	7
SFTMP	Snowfall temperature (°C)	Lumped	15	6
SMTMP	Snowmelt base temperature (°C)	Lumped	7	5
Sol_z	Soil depth	Lumped	9	10

3.4 Objective functions

The purpose for which SWAT was applied to the Honey Creek catchment was to estimate sediment export from the catchment. Therefore, the joint calibration included the SSE for the streamflow and the SSE for sediment concentrations with a Box-Cox transformation to reduce the heteroscedastic nature of the residuals (Box & Cox, 1982). The GOC thus represents errors associated with both flow and water quality variables.

3.5 Evaluation criterion

Based on the assumption that the model purpose was to assess global fluxes of sediment load at the outlet of the creek, the evaluation criterion was described by the model bias on the mass flux that was calculated as:

$$\mathrm{BIAS} = 100\,\frac{\sum_{n=1}^{N}\mathrm{SIM}_n - \sum_{n=1}^{N}\mathrm{OBS}_n}{\sum_{n=1}^{N}\mathrm{OBS}_n} \tag{6}$$

for N the number of pairs (simulation, observation), SIM_n the simulation at day n and OBS_n the observation of day n. The bias was calculated for the water flow and the sediment loads in the calibration and validation period.

Copyright © 2008 IAHS Press

3.6 Results

3.6.1 Entire parameter space Before evaluating model bias, the bias results were assessed for the case when all parameters were allowed to vary freely between the *a priori* parameter ranges. This situation represents the absolute maximum degree of uncertainty that the final results could have. The characteristics for the entire parameter space are calculated based on all simulations that were performed during the ParaSol optimization. It was shown that the model bias on the water volumes can vary from an underestimation of almost 100% (i.e. nearly no water leaves the system) to an overestimation of as big as three times of the total volume (Table 2). For the sediment loads, the overestimation can be as much as 32 times the total loads. These results clearly show the risk of using model results that have not been conditioned with observations.

Table 2 Minimum and maximum error on global balance (in percentage).

		FLOW	Sediment loads
1998–1999	Min	–98.60	–99.81
	Max	263.81	1501.90
2000–2001	Min	–97.37	–99.81
	Max	302.49	3171.40

3.6.2 ParaSol and SUNGLASSES confidence space A total of 34 669 simulations were performed to minimize the GOC. According to ParaSol, all the simulations with GOC smaller then c_{ParaSol} were accepted and defined an uncertainty range for the outputs as in Fig. 2. The ParaSol results, using χ^2-statistics for 97.5% confidence probability, show clear bias for the sediment load predictions during the calibration period: between –23% and –27% and an opposite bias for the validation period (34–43%) (Fig. 2). Here we see an underestimation of the sediment loads in the calibration period and an overestimation in the validation period. This means that the bias depends on the period of observations. It is also clear that the uncertainty method within ParaSol does not foresee this strong bias and that the zero-bias is not captured. This result is probably due to the compromises that have to be made between the different objective functions. The uncertainty threshold should hence be updated according to the SUNGLASSES methodology. This leads to higher uncertainty bounds that include the zero bias of sediment transport for both calibration and validation period (Fig. 2).

An application of SUNGLASSES shows that the sediment load calculations can have an overestimation of up to 167%. This result means that the model, when calibrated on a period of two years, is not performing well and is thus highly uncertain in assessing total mass fluxes. The confidence ranges for the time series give much wider bounds for SUNGLASSES that capture more of the observations as well (Fig. 4). For instance, the missed observation at the beginning of

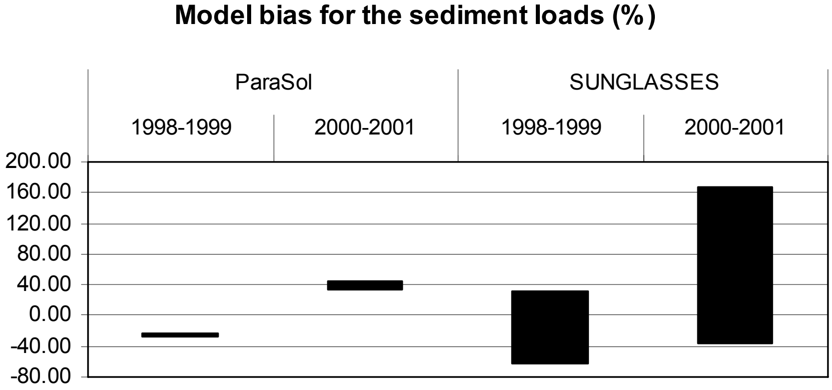

Fig. 2 Uncertainty intervals for the sediment load calculations according to ParaSol and SUNGLASSES.

Copyright © 2008 IAHS Press

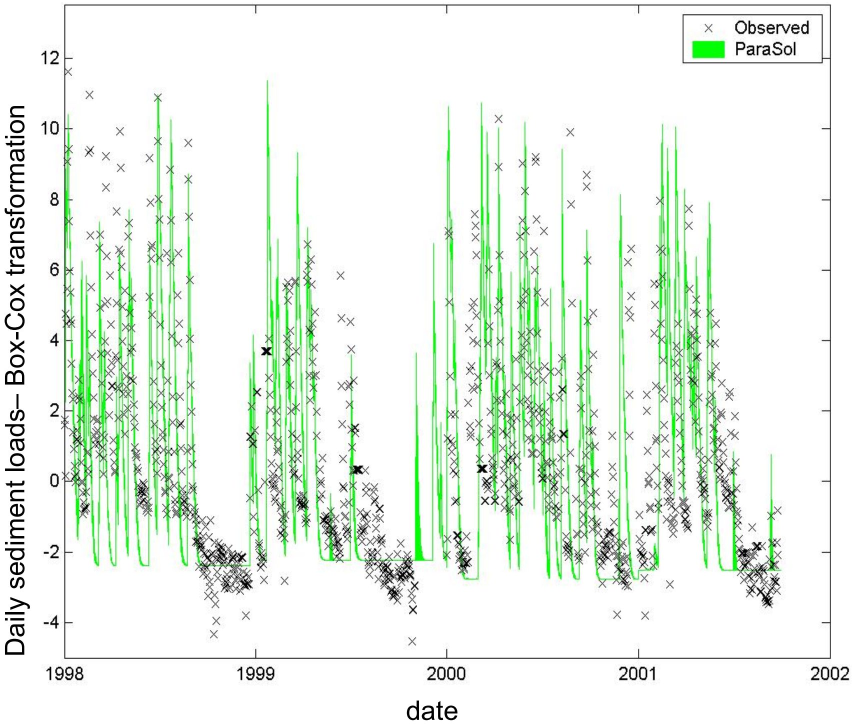

Fig. 3 Uncertainty intervals for the time series of the daily sediment loads according to ParaSol.

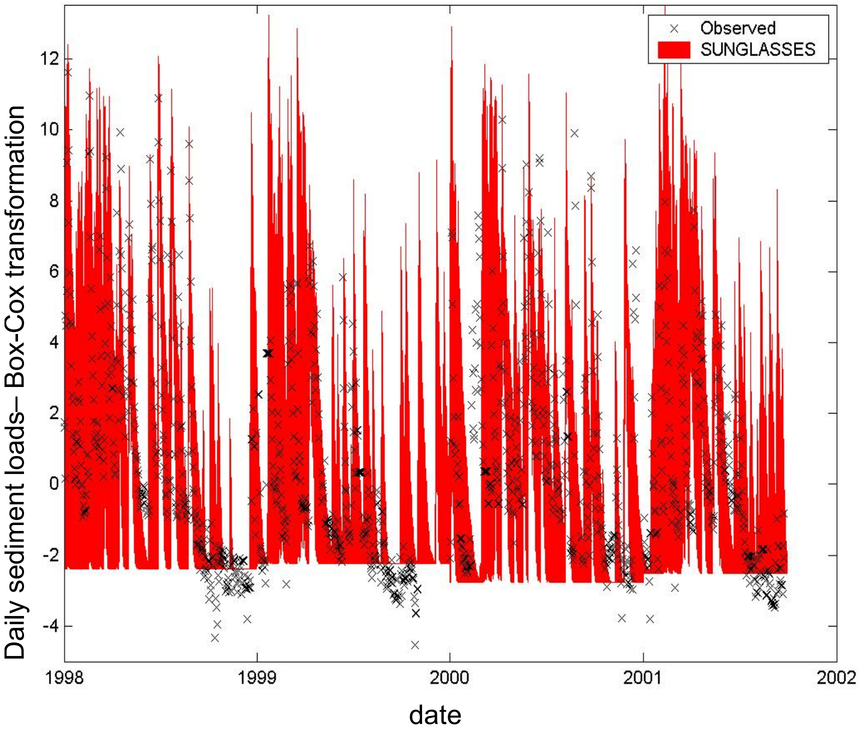

Fig. 4 Uncertainty intervals for the time series of the daily sediment loads according to SUNGLASSES.

1999 was not captured in ParaSol (Fig. 3), while it was captured with SUNGLASSES (Fig. 4). We therefore conclude that SUNGLASSES gives an overall more liberal estimation of the confidence regions (Fig. 5).

Copyright © 2008 IAHS Press

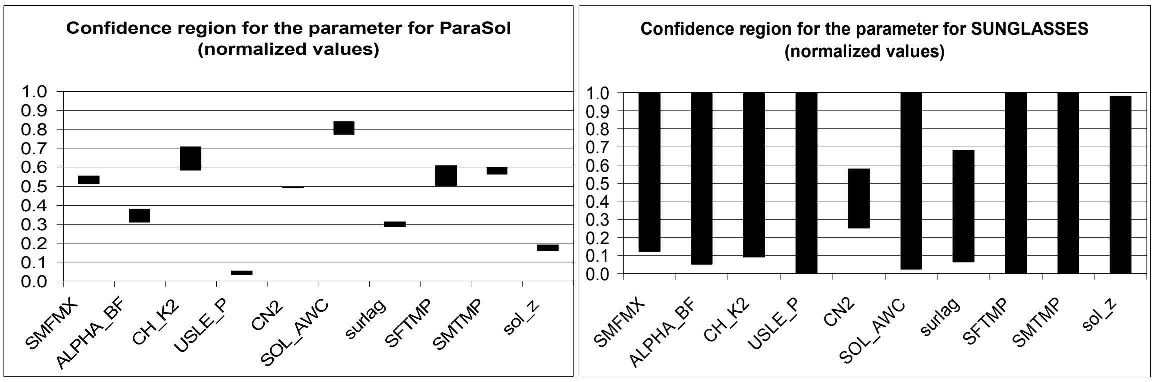

Fig. 5 Uncertainty intervals for the parameters according to ParaSol and SUNGLASSES.

4 DISCUSSION

Both ParaSol and SUNGLASSES give uncertainty assessments, but they use different approaches and represent different sources of uncertainties. ParaSol is a typical statistical method to assess parameter uncertainties such as other methods like SCEM-UA (Vrugt *et al.*, 2003), and BaRE (Thiemann *et al.*, 2001). SUNGLASSES does not have these underlying statistics; it assesses overall uncertainty but does not explain where the uncertainty comes from. It is thus a diagnostic method indicating that something is wrong rather than quantifying the specific source of the uncertainty. To identify the specific uncertainty source the model user would need to employ additional analysis or methods.

The results show an important drawback in traditional statistical uncertainty methods: these do account for the number of observations, but do not consider whether these data also cover the variability of the system necessary to identify the overall process parameters. This problem leads to a bad assessment of indicators (such as global mass fluxes) that might be used in decision making. SUNGLASSES reveals such problems by evaluating predictions outside the calibration period. SUNGLASSES thus gives a joint evaluation of the model and the parameter identification procedure. When either of these fails the test, more selections of parameter combinations and much wider uncertainty ranges will be obtained.

Our results thus indicate that the uncertainty assessment with ParaSol on the SWAT model fails or is at least incomplete. The big question is, why are the ParaSol results so biased? This problem occurs with the ParaSol method in particular, and with most parameter uncertainty methods in general because these methods are built on assumptions that are often not fulfilled, for example:

(a) **The data represent seasonal and long time variability.** In case of failure, the χ^2 statistics do not account for long-term climatic patterns and their affect on the system: such variability may cause a model bias for the validation period, even though the model was not biased for the calibration period. Since catchment hydrology is subject to strong variability in climate, calibration and validation periods will typically have different characteristics and distributions for the forcing weather inputs, and so will the flow observations. Longer time series could be used so that more system variability could be accounted for. Previously, some have found that approximately 10 years of data is needed to properly calibrate a hydrological model (Yapo *et al.*, 1996). These results are likely site and climate specific. It is unknown what period of water quality data would be sufficient.

(b) **The dynamic inputs (such as weather) are correct or represent the spatial and temporal variability of the system.** The spatial distribution of precipitation gauges is typically coarse compared to the spatial variability of the rainfall events (Willems & Berlamont, 2002). This difference may cause errors in model outputs that do not follow a normal distribution. Strange patterns of errors may occur, especially when rainfall data outside the modelling area is used:

Copyright © 2008 IAHS Press

rainfall storms may be recorded that did not happen in the area of modelling, leading to a runoff peak that was not observed. While most of the model applications do not consider these input uncertainties, there are few examples (e.g. Krzysztofowicz, 2002) that incorporate the input uncertainties in modelling.

(c) **The model structure and hypothesis (the process equations) are correct and represent the real world.** Model errors are considered by using multiple models in a Bayesian framework (Hoeting *et al.*, 1999; Montanari & Brath, 2004), by regionalized sensitivity analysis (Osidele & Beck, 2001), the GLUE methodology (Beven & Binley, 1992) or with a Pareto analysis (Meixner *et al.*, 2002; Gupta *et al.*, 2003). Such methods allow for a relative evaluation of model structures and/or a discrimination of models while a quantification of the errors associated with model uncertainty is not straightforward.

(d) **The model is unbiased or the model errors are randomly distributed.** The unbiased nature of the model is a generally accepted assumption in statistically based uncertainty analysis. This assumption is because parameter calibration is generally able to remove the model bias. However, for a variety of reasons, the complete removal of bias is not necessarily true. First, the objective functions typically used, such as sum of the squares, are not always unbiased error estimators if the errors are not normally distributed. Second, split sample evaluations often reveal large biases during validation periods (as shown in Figs 2 and 5). Third, even when a longer period of data is used for calibration, this will still lead to biased results for sub-periods due to parameters that have different optimal values for different time periods. Several studies have revealed periodic, often seasonal, preference of calibrated parameters for certain values (Freer *et al.*, 2003) and a necessity to compromise as a consequence. These studies suggest that at least some parameters cannot be generalized for any data set, no matter how long. In order to fulfill the assumptions of unbiased and randomly distributed residuals, it is thus necessary to improve the modelling tools to prevent having a preference of the parameters based on a specific period of data.

The trade-off between different fitting criteria may lead to a model bias in the calibration period, which happened in the case presented here. Such trade-offs are caused by model structural errors (Meixner *et al.*, 2002; Gupta *et al.*, 2003), and have to be solved by a better mathematical representation of the system.

(e) **Observations are typically considered as being correct.** Most observation databases only provide the observations without giving additional information on the uncertainty of measurements. Meta-databases providing background information on these measurements are thus needed. Also, some observations, such as flow data, typically have a heteroscedastic distribution of errors as opposed to the generally assumed normal distribution (Sorooshian & Dracup, 1980). These differences can be corrected for.

(f) **The residuals are independent identically distributed with normal distribution.** All previous sources of uncertainty may lead to residuals that show periodical biases or do not follow a random pattern. It is often observed that the distribution of model errors is not normal, but having a heteroscedastic pattern (e.g. Sorooshian, 1980; Sorooshian & Dracup, 1980) or auto-correlated pattern (e.g. Sorooshian & Dracup, 1980). Other distributions or transformation functions have been used to account for heteroscedastic error situations in hydrology (Sorooshian, 1980; Thiemann *et al.*, 2001) or auto-correlations (Sorooshian & Dracup, 1980), but are less popular as such distributions incorporate case dependent (and time period dependent!) constants that have to be defined. However, other sources of uncertainty (e.g. model structure and input uncertainty) will prevent errors from being consistent with certain known and well-understood distributions.

In brief, hydrological models typically do not fulfil these assumptions. Apparently, these problems cannot be neglected as it leads to a wide increase of the uncertainty bounds under SUNGLASSES.

Copyright © 2008 IAHS Press

While several of these problems can be considered in the uncertainty analysis by paying more attention to the assumptions that underlay the statistics and methods of comparison or quantification of these uncertainties are available, some sources of uncertainty remain at the present state-of-the-art inaccessible and thus hard to consider. A global evaluation of the overall procedure, such as performed by SUNGLASSES, is thus needed to improve the reliability of the model results so that decision makers can have a firm understanding of the degree to which they can trust model results.

5 CONCLUSIONS

The results show an important drawback in traditional statistical uncertainty methods: these do account for the number of observations, but do not consider whether these data also cover the variability of the system necessary to identify the overall process parameters, or for model errors (model structural errors or errors in process descriptions). This problem leads to wrong assessments of indicators (like global mass balances) that might be used in decision making. SUNGLASSES reveals these problems by evaluating predictions of importance. The SUNGLASSES results show more selections of parameter combinations and much wider uncertainty ranges.

Our purpose here is to assess predictive uncertainty and develop methodologies that will help decision makers understand how uncertain their models are so that they can put the proper level of trust in computational models of the environment as they move forward to make decisions. The preliminary results indicate that the main concern should be about the uncertainty associated with model structural error and data errors, and less so on model parametric uncertainty. As others have noted, the inability of a model to simulate a second period of time when it has been calibrated with an earlier period of data should be cause for rejecting a model as non-behavioural (not acceptable) noting that the model obviously has flaws and needs to be improved (Freer *et al.*, 2003). While the notion of rejecting such a model is an appropriate one in a scientific context, it is less acceptable in a policy application context. Since models certainly need improvement, there is always a point at which we must apply our models as they stand, to actually make decisions (Grayson & Blöschl, 2001). For these decisions to be good and properly informed, they must incorporate all the sources of uncertainty, including model structural and data errors. For this reason, instead of taking the approach of rejecting model simulations during a comparison period as non-behavioural, SUNGLASSES incorporates the evaluation period to help set the threshold of model acceptability in order to incorporate elements of model structural uncertainty and data uncertainty into overall model predictive uncertainty. While SUNGLASSES is by no means the only methodology that can be used to assess model predictive uncertainty, the split sample approach it takes bears further investigation as a way to incorporate data errors and model structural errors into predictive uncertainty estimation.

Acknowledgements Support for this work was provided by the National Science Foundation through a CAREER award to T. Meixner (EAR-0094312). The research on the Sandusky River catchment was supported by the Florida Agricultural Experiment Station and approved for publication.

REFERENCES

Arnold J. G., Srinivasan, R., Muttiah, R. S. & Williams, J. R. (1998) Large area hydrologic modelling and assessment part I: model development. *J. Am. Water Resour. Assoc.* **34**(1) 73-89.

Bard, Y. (1974) *Non Linear Parameter Estimation*. Academic Press, New York, USA.

Beck, M. B. (1987) Water quality modelling: a review of the analysis of uncertainty. *Water Resour. Res.* **23**(6), 1393–1441.

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

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

Copyright © 2008 IAHS Press

Beven, K. & Freer, J. (2001) Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. *J. Hydrol.* **249**, 11–29.

Beven, K. & Young, P. (2003) Comment on "Bayesian recursive parameter estimation for hydrologic models" by M. Thiemann, M. Torsset, H. Gupta & S. Sorroshian. *Water. Resour. Res.* **39**(5), COM 1-1–COM 1-4.

Box, G. E. P. & Cox, D. R. (1982) An analysis of transformations revisited, rebutted. *J. Am. Statist. Assoc.* **77**(377), 209–210.

Boyle, D. P., Gupta, H. V., Sorooshian, S., Koren, V., Zhang, Z. & Smith, M. (2001) Toward improved streamflow forecasts: value of semidistributed modelling. *Water Resour. Res.* **37**, 2749–2759.

Duan, Q., Gupta, V. K. & Sorooshian, S. (1992) Effective and efficient global optimization for conceptual rainfall–runoff models. *Water Resour. Res.* **28**, 1015–1031.

Duan, Q., Sorooshian, S., Gupta, H. V., Rousseau, A. N. & Turcotte, R. (2003) *Advances in Calibration of Watershed Models*. American Geophysical Union, Washington DC, USA.

Eckhardt, K. & Arnold. J. G. (2001) Automatic calibration of a distributed catchment model. *J. Hydrol.* **251**, 103–109.

Freer, J., Beven, K. & Peters. N. E. (2003) Multivariate seasonal period model rejection within the Generalised Likelihood Uncertainty Estimation procedure. In: *Calibration of Watershed Models* (ed. by Q. Duan, H. V. Gupta, S. Sorooshian, A. N. Rousseau & R. Turcotte), 69–88. American Geophysical Union, Washington DC, USA.

Grayson, R. B. & Blöschl, G. (2001) Spatial modelling of catchment dynamics. In: *Spatial Patterns in Catchment Hydrology: Observations and Modelling* (ed. by R. B. Grayson & G. Blöschl), 51–81. Cambridge University Press, Cambridge, UK.

Gupta, H., Thiemann, M. Trosset, M. & Sorooshian, S. (2003) Reply to comment by K. Beven & P. Young on "Bayesian recursive parameter estimation for hydrologic models". *Water Resour. Res.* **39**(5), COM 2-1–COM 2-5.

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

Hoeting, J. A., Madigan, D., Raftery, A. E. & Volinsky, C. T. (1999) Bayesian model averaging: a tutorial. *Statist. Sci.* **14**(4), 382–417.

Holland, J. H. (1995) *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, Michigan, USA.

Klemeš, V. (1986) Dilettantism in hydrology: transition or destiny? *Water Resour. Res.* **22**, 177S–188S.

Knisel, W. G. (1980) CREAMS, a field scale model for chemicals, runoff and erosion from agricultural management systems. USDA Conservation Research Rep. no. 26.

Krysztofowicz, R. (2002) Bayesian system for probabilistic river stage forecasting. *J. Hydrol.* **268**, 16–40.

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

Kuczera, G. (1983b) Improved parameter inference in catchment models: 2. Combining different kinds of hydrologic data and testing their compatibility. *Water Resour. Res.* **19**, 1163–1172.

Kuczera, G. & Mroczkowski, M. (1998) Assessment of hydrologic parameter uncertainty and the worth of multiresponse data. *Water Resour. Res.* **34**, 1481–1489.

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

Leonard, R. A. & Wauchope, R. D. (1980) The pesticide submodel. In: *CREAMS: A Field-Scale Model for Chemicals, Runoff, and Erosion from Agricultural Management Systems* (ed. by W. G. Knisel), Chapter 5, 99–112. Conservation Research Report no 26. US Department of Agriculture.

Madsen, H. (2003) Parameter estimation in distributed hydrological catchment modelling using automatic calibration with multiple objectives. *Adv. Water Res.* **26**, 205–216.

Meixner, T., Brown, A. D. & Bales, R. C. (1997) Modelling of biogeochemistry of two alpine watersheds in the Sierra Nevada, California. *EOS Trans. Am. Geophys. Union* **78**, S173–S174.

Meixner, T., Bales, R. C., Williams, M. W., Campbell, D. H. & Baron, J. S. (2000) Stream chemistry modelling of two watersheds in the Front Range, Colorado. *Water Resour. Res.* **36**, 77–87.

Meixner, T., Bastidas, L. A., Gupta, H. V. & Bales, R. C. (2002) Multi-criteria parameter estimation for models of stream chemical composition. *Water Resour. Res.* **38**(3), 9-1–9-9.

Montanari, A. & Brath, A. (2004) A stochastic approach for assessing the uncertainty of rainfall–runoff simulations. *Water Resour. Res.* **40**, W01106, doi:10.1029/2003WR002540,.

Mroczkowski, M., Raper, G. P. & Kuczera, G. (1997) The quest for more powerful validation of conceptual catchment models. *Water Resour. Res.* **33**, 2325–2335.

Nelder, J. A. & Mead, R. A. (1965) Simplex method for function minimization. *Comput. J.* **7**, 308–313.

Osidele, O. O. & Beck, M. B. (2001) Identification of model structure for aquatic ecosystems using regionalized sensitivity analysis. *Water Sci. & Technol.* **43**(7), 271–278.

Singh, V. P. (1988) *Hydrologic Systems—Rainfall–Runoff Modelling*. Prentice Hall, Englewood Cliffs, New Jersey, USA.

Sorooshian, S. (1980) Parameter estimation of rainfall–runoff models with heteroscedastic streamflow errors: the non-informative data case. *J. Hydrol.* **52**, 127–138.

Sorooshian, S. & Dracup, J. A. (1980) Stochastic parameter estimation procedures for hydrologic rainfall–runoff models: Correlated and heteroscedastic error cases. *Water Resour. Res.* **16**, 430–442.

Sorooshian, S. & Gupta, V. K. (1983) Evaluation of maximum likelihood parameter estimation techniques for conceptual rainfall–runoff models: influence of calibration data variability and length on model credibility. *Water Resour. Res.* **19**(1), 251–259.

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

USDA Soil Conservation Service (1983) *National Engineering Handbook*, Section 4 *Hydrology*, Ch. 19.

van Griensven A. & Bauwens W. (2001) Integral modelling of catchments. *Water Sci. & Technol.* **43**(7), 321–328.

van Griensven A., Meixner, T., Grunwald, S., Bishop, T. & Srinivasan, R. (2006) A global sensitivity analysis method for the parameters of multi-variable watershed models. *J. Hydrol.* **324**(1–4), 10–23.

van Griensven, A. & Bauwens, W. (2003) Multi-objective auto-calibration for semi-distributed water quality models. *Water Resour. Res.* doi:10.1029/2003WR002284.

Copyright © 2008 IAHS Press

van Griensven, A, & Meixner, T. (2007) A global and efficient multi-objective auto-calibration and uncertainty estimation method for water quality catchment models. *J. Hydroinf.* **9(**4), 277–291 doi:10.2166/hydro.2007.104.

Vrugt, J. A., Gupta, H. V., Bouten, W. & Sorooshian S. (2003) A shuffled complex evolution metropolis algorithm for estimating posterior distribution of watershed model parameters. In: *Calibration of Watershed Models* (ed. by Q. Duan, S. Sorooshian, H. V. Gupta, A. N. Rousseau & R. Turcotte). American Geophysical Union, Washington DC, USA. doi:10.1029/006WS07.

Willems, P. & Berlamont, J. (2002) Accounting for the spatial rainfall variability in urban modelling applications. *Water Sci. & Technol.* **45**(2), 105–112.

Williams, J. R. (1969) Flood routing with variable travel time or variable storage coefficients. *Trans. Am. Soc. Agric. Engrs* **12**(1), 100–103.

Williams, J. R., Jones, R. W. C. A. & Dyke, P. T. (1984) A modelling approach to determining the relationship between erosion and soil productivity. *Trans. Am. Soc. Agric. Engrs* **27**(1), 129–144.

Yapo, P. O., Gupta, H. V. & Sorooshian, S. (1996) Automatic calibration of conceptual rainfall–runoff models: sensitivity to calibration data. *J. Hydrol.* **181**, 23–48.

Received 24 April 2007; accepted 19 May 2008

Copyright © 2008 IAHS Press