Uncertainty modelling and analysis of environmental systems: a river sediment yield example

Karel J. Keesman[a], **Jarkko J. Koskela**[b], **Joseph H. Guillaume**[c], **John P. Norton**[c], **Barry Croke**[c], **Anthony J. Jakeman**[c]

[a] *Systems and Control Group, Wageningen University, Bornse Weilanden 9, 6708 WG Wageningen, The Netherlands*
Email: karel.keesman@wur.nl
[b]*Aalto University School of Engineering, Department of Civil and Environmental Engineering, Espoo, Finland*
[c] *Fenner School of Environment and Society and National Centre for Groundwater Research and Training, The Australian National University, Canberra ACT 0200*

Abstract: Throughout the last decades uncertainty analysis has become an essential part of environmental model building (e.g. Beck 1987; Refsgaard et al., 2007). The objective of the paper is to introduce stochastic and set-membership uncertainty modelling concepts, which basically differ in the assumptions that are made with respect to the uncertainty characterization. Stochastic uncertainty modelling is most frequently applied and is characterized by probability density functions (pdf's) or simply by means and (co)variances. Typical approaches are the Bayesian and the Monte Carlo Markov Chain methods. Alternatively, a set-membership or bounded-error characterization, as opposed to a stochastic characterization, is favoured when assumptions about distribution or estimates of mean and covariance cannot be satisfactorily tested, as with small data sets or heavily structured (modelling) errors. The bounded-error characterization is in essence deterministic. Both approaches, using tools as DREAM, GLUE, exact and approximate bounding, MCSM and a pavement-based technique, were tested on a real-world example. The example, based on Wasson's (1994) sediment yield – area data and after a log-log transformation of the data, is a linear static problem with two parameters.

There is a continuing debate on the use of formal and informal uncertainty methods in hydrology (see e.g. Beven et al. 2008). GLUE has been criticized especially for the use of an informal likelihood function and a subjective choice of threshold to separate behavioural (feasible) and rejected (infeasible) parameter vectors. It has also been pointed out (Qian et al., 2003) that the inefficient sampling inherent in naive Bayesian estimation techniques of this type is very likely to yield uninformative results for the posterior distribution of the parameters. On the other hand, it has been known for a long time that assumptions about the residual errors in formal approaches are often violated, making inference unreliable (Kuczera, 1983). Studies have compared GLUE using an informal likelihood measure with a method using a formally correct likelihood function (see e.g. Freni et al., 2009). Vrugt et al. (2009a) conclude that although GLUE and a statistically formal (DREAM) Bayesian approach can give very similar results, the main advantage of a formal approach is that it allows to disentangle different contributions to total uncertainty. The set-membership approach to uncertainty modelling and analysis, as an alternative to the stochastic uncertainty modelling approaches, has a firm theoretical basis, see Walter (1990), Norton (1994, 1995), Milanese et al (1996) and Keesman (2011). In real applications, the critical point, however, is the choice of the bounds on the error vector.

The stochastic and set-membership methods were evaluated with respect to (i) Accuracy or precision of methods, (ii) Time and computational requirements, (iii) Skill requirements and (iv) Range of applicability. From an application point of view, we conclude that uncertainty analysis is of paramount importance in scenario studies of (complex) environmental systems if one wants to implement robust measures. It is also worthy to be eclectic in methods for uncertainty analysis, because any method makes assumptions and it is therefore valuable to consider uncertainty quantification under different conditions.

Keywords*:* *Uncertainty analysis, stochastic approach, set-membership approach, river sediment yield*

1. INTRODUCTION

Nowadays, uncertainty analysis is considered as an essential part of environmental model building. The objective of this paper is to present and evaluate stochastic and set-membership uncertainty modelling concepts, described below, on a real-world example of river sediment yield modelling.

Stochastic uncertainty modelling is most frequently applied and is characterized by probability density functions or simply by means and (co)variances. Given these concepts, we can compute the error propagation either from output error to uncertainty in the parameter estimates or from uncertainty in the parameter estimates to prediction uncertainty, possibly extended by structural modelling uncertainty.

With set-membership identification (see overviews by Walter, 1990; Norton, 1994, 1995; Milanese et al, 1996), the aim is not to find a single vector of optimal parameter estimates, but a set of feasible parameter-vector values that are consistent with a given model structure and data with uncertainty with specified bounds. This approach is favoured when assumptions about distribution or estimates of mean and covariance cannot be satisfactorily tested, as with small data sets or heavily structured (modelling) errors. In general, a non-connected vector set with non-convex subsets of feasible parameter vectors may result.

Our uncertainty modelling starts with the specification of error terms in the mathematical model. In the state-space modelling framework, it is usual to distinguish between process noise, that affects the dynamics of the system, and measurement noise (e.g. Keesman, 2011). However, in this paper the focus is on modelling the output error and its effect on the uncertainty in the parameter estimates and the corresponding model predictions.

The real-world example to be examined below focuses the interpolation and prediction of sediment yield as a function of river catchment area. It uses Wasson's (1994) data, collected over many catchments in Australia, and a power-law model of the relationship between sediment yield and area. After log-transformation of both variables the model becomes a linear regression with two unknown parameters.

2. DEFINITIONS AND NOTATION

Our starting point is the following general non-linear model

$$\mathbf{y} = \mathbf{F}(\vartheta) + \mathbf{e} \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^N$ contains the observed output data, $\mathbf{F}(\vartheta)$ is a non-linear vector function mapping the unknown parameter vector $\vartheta \in \mathbb{R}^m$ into a noise-free model output $\hat{\mathbf{y}}$ and $\mathbf{e}$ is the error vector. Naturally, we may interpret $\mathbf{F}(\vartheta)$ as a vector containing one or more output trajectories from a, possible temporal-spatial, simulation model.

2.1. Stochastic uncertainty modelling

Assuming that components of vector **e** are independent and identically distributed (i.i.d.) observations of the output error, with probability density function (pdf) $p(y_i|\vartheta)$, the objective function (likelihood function) is

$$L(\vartheta\,|\,\mathbf{y}) = \prod_{i=1}^{N} p(y_i|\vartheta) \tag{2}$$

Normally, the maximisation problem, within maximum likelihood estimation (MLE) procedure, is solved with a numerical optimisation algorithm. By using some Monte Carlo Markov Chain (MCMC) based sampling in the optimisation step, it is also possible to estimate the uncertainties of the parameter estimates in addition to a single parameter vector giving the "best" fit. In the stochastic uncertainty framework, it is also possible to utilise prior information about the parameter estimates, leading to a Bayesian approach.

$$L(\vartheta|\mathbf{y}) = c \cdot p(\vartheta) \prod_{i=1}^{N} p(y_i|\vartheta) \tag{3}$$

with $p(\vartheta)$ the prior pdf and c a normalising constant,

2.2. Set-membership uncertainty modelling concept

Within the set-membership framework, the information uncertainty vector **e** is assumed to be bounded in a given norm. In what follows, it is assumed that it is piece-wise bounded, so that

$$\|\mathbf{e}\|_\infty \leq \varepsilon \tag{4}$$

where ε is a fixed positive number. The related measurement uncertainty set (MUS), containing all possible model-output vectors consistent with the observed output data and uncertainty characterization (4), is defined as

$$\Omega_y := \{\, \tilde{\mathbf{y}} \in \mathbb{R}^N : \|\, \mathbf{y} - \tilde{\mathbf{y}} \,\|_\infty \le \varepsilon \} \tag{5}$$

The set

$$\Omega_\vartheta := \{\vartheta \in \mathbb{R}^m : \|\, \mathbf{y} - \mathbf{F}(\vartheta) \|_\infty \le \varepsilon \} \tag{6}$$

then defines the feasible parameter set (FPS). The set-membership estimation problem is to characterize this feasible parameter set, which is consistent with the model (1), the data (**y**) and uncertainty characterization (4). The image set related to the FPS, which is an unfalsified model output set and also called the feasible model output set (FMOS), is then defined as:

$$\Omega_{\hat{y}} := \{\, \hat{\mathbf{y}} \in \mathbb{R}^N : \hat{\mathbf{y}} = \mathrm{F}(\vartheta);\ \vartheta \in \Omega_\vartheta \} = \Omega_{\hat{y}} \cap \Omega_y \tag{7}$$

Hence, instead of trying to find an optimal value of the parameter vector as in ordinary least squares or in a statistical estimation procedure, the goal is to find the set of feasible parameter-vector values consistent with the model (1) and the data, subject to specified bounds on the error between model output and output observations (4). This approach avoids any assumptions beyond the model-output error bounds.

3. A REAL-WORLD EXAMPLE

In the real-world example we examine a data set collated by Wasson (1994) of sediment yields versus catchment area, see Figure 1. The Wasson (1994) data, was collated from numerous studies of long-term sediment yields in south east Australia. The data shows a high level of scatter – a function of (i) high inherent spatial and temporal variability of sediment yield across south east Australia; and (ii) the use of a range of different underlying methods to estimate sediment yields. Our particular interest in the analysis of this data is to identify likely upper and lower bounds of plausibility for sediment yield estimates.

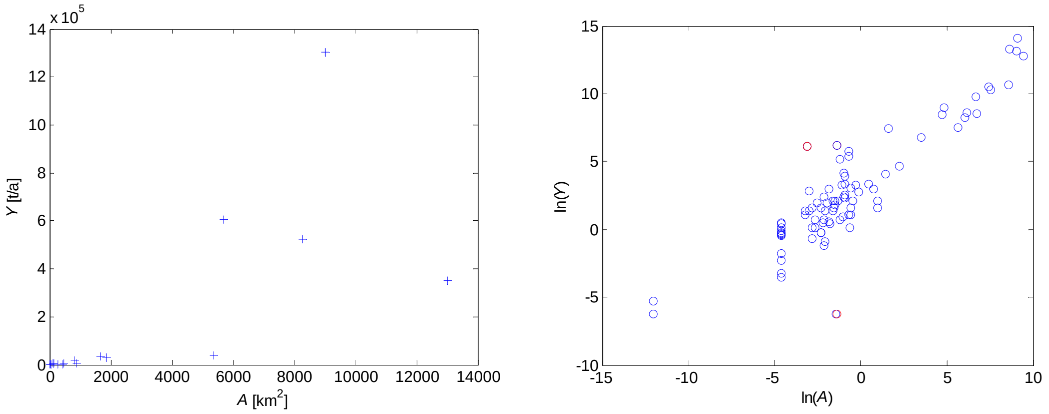

Figure 1. Data set (from Wasson, 1994), left panel: original space, and right panel: log-log space

Presume that the catchment area-sediment yield data can be described by the power law: $Y = aA^k$, where *a* and *k* are unknown parameters. After applying a natural logarithmic transformation of the power law we obtain,

$$\ln Y = \ln a + k \ln A \tag{8}$$

from which we define $\alpha_1 := \ln a$ and $\alpha_2 := k$. Hence, in terms of the general non-linear regression model for each area with index *i*: $F_i(\vartheta) = F_i\vartheta = \begin{bmatrix} 1 & \ln A_i \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}$.

3.1. Stochastic uncertainty characterization of sediment yields

Least-squares estimation: It has been shown in Keesman et al. (2010) that, using ordinary least-squares estimation and using more or less standard assumptions about the errors, that is the errors in **e** are i.i.d and

Gaussian, or at least have a symmetrical distribution, and A_i error-free, the parameters in (8) and the corresponding covariance matrix is given by

$$\hat{\alpha}_1 = 3.3125, \qquad \hat{\alpha}_2 = 0.9177; \qquad \Sigma_{\hat{\vartheta}} = \hat{\sigma}_e^2 \left(\mathbf{F}^T\mathbf{F}\right)^{-1} = \begin{pmatrix} 0.0353 & 0.0012 \\ 0.0012 & 0.0020 \end{pmatrix} \tag{9}$$

Consequently, the predicted sediment yield and the prediction error variance for a specific area A_i can be found from

$$\ln \hat{Y}_i = F_i \vartheta = [1 \quad \ln A_i] \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{bmatrix}, \quad \sigma^2_{\ln \hat{Y}i} = F_i \Sigma_{\hat{\vartheta}} F_i^T = [1 \quad \ln A_i] \Sigma_{\hat{\vartheta}} \begin{bmatrix} 1 \\ \ln A_i \end{bmatrix} \tag{10}$$

with $\hat{\alpha}_1$, $\hat{\alpha}_2$ and $\Sigma_{\hat{\vartheta}}$ given in (9).

For instance, using (10), for an area of exp(−8) = 3.35 10^{-4} km^2 we find a predicted log-transformed sediment yield of −4.03 ± 0.38, that is 0.02 ± 1.46 t/a. The calculation of the covariance matrix in (9) is mainly determined by large values in **F** and thus related to very small and very large catchment areas. This will ultimately lead to underestimated errors in the interpolated sediment yields. Hence, more realistic uncertainty modelling is needed.

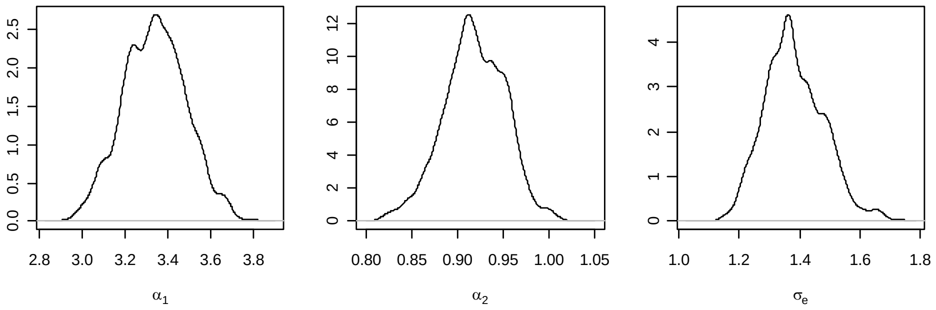

Figure 2. Marginal posterior distributions of the parameters using DREAM.

Bayesian inference: In what follows, it is assumed that residual errors of the model are independent, identically and Gaussian distributed. In Eq. 3 vague uniform priors are used for model parameters α_1 [0, 100] and α_2 [0.5, 1.5] as well as for standard deviation of the residual error σ_e [0, 10] that was inferred from the data at the same time with model residuals. Two observations were removed from the data sets, as they seemed to be outliers, see red circles in Figure 1b. Using the Differential Evolution Adaptive Metropolis algorithm (DREAM) (Vrugt et al., 2009b), Figure 2 shows the marginal posterior distributions of the variables and Figure 3 shows the resulting uncertainty bounds. It is essential to check whether the initial assumptions about the residuals are valid. In our example, the residuals pass the normality test and there is no clear heteroscedasticity in the residuals.

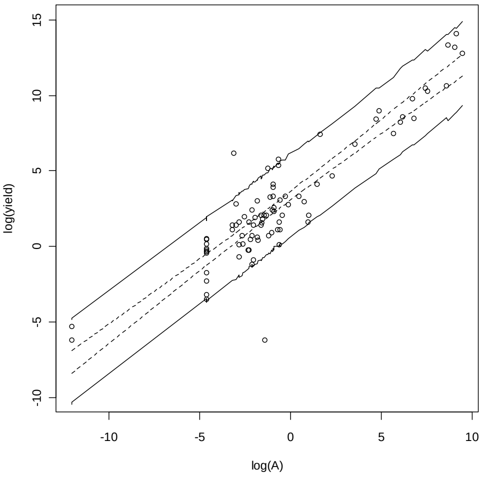

Figure 3. Uncertainty (95 %) bounds using DREAM. Solid lines present lower and upper total uncertainty bounds and dashed lines parameter uncertainty, respectively.

3.2. GLUE

In the Generalized Likelihood Uncertainty Estimation (GLUE) analysis of Beven and Binley (1992) R^2 was used as a fit criterion. 100000 parameters sets were evaluated and the threshold was selected such that best 1% (1000) of the sets was considered behavioural. Figure 4 shows the behavioural parameters values and Figure 5 shows the resulting 95 % uncertainty bounds.

3.3 Set-membership uncertainty characterization of sediment yields

In what follows, the results are shown for an error bound of 5, thus ε = 5. Within the set-membership approach, and given Eq. 4 and 8, we derive the following inequalities

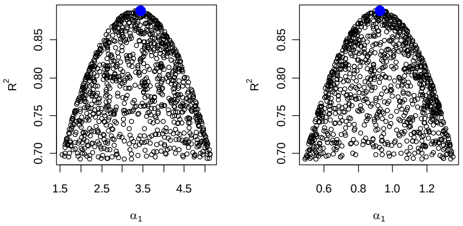

Figure 4. Behavioural parameter sets using GLUE. Blue dots show the “best” parameter values.

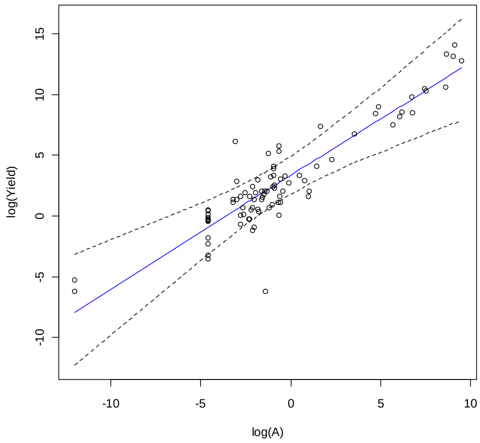

Figure 5. Uncertainty (95 %) bounds using GLUE. Dashed lines present lower and upper total uncertainty bounds and the solid blue line the median of the interpolation.

$$\ln Y_i - \varepsilon \le \alpha_1 + \alpha_2 \ln A_i \le \ln Y_i + \varepsilon, \quad \text{for} \quad i = 1,\ldots,N \tag{11}$$

Graphical solution: Notice from (11) that for each measurement pair (Y_i , A_i) two linear inequalities appear. Hence, the sharp constraints are given by: $\alpha_1 + \alpha_2 \ln A_i = \ln Y_i + \varepsilon$ and $\alpha_1 + \alpha_2 \ln A_i = \ln Y_i - \varepsilon$, for i =1, ..., N, which in this case define two lines in the two-dimensional parameter space. In Figure 7 the red lines are related to the upper bounds, while the green are related to the lower bounds.

Orthotopic outer-bounding: The hypercube aligned with each of the axis in the parameter space that tightly outer-bounds the FPS is found by solving 2m LP problems of the form

$$\begin{aligned}
&\min/\max\ f^T \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} \\
&\text{s.t. } \alpha_1 + \alpha_2 \ln A_i \le \ln Y_i + \varepsilon \\
&\quad -(\alpha_1 + \alpha_2 \ln A_i) \le -(\ln Y_i - \varepsilon) \quad \text{for} \quad i = 1,\ldots,N
\end{aligned} \tag{12}$$

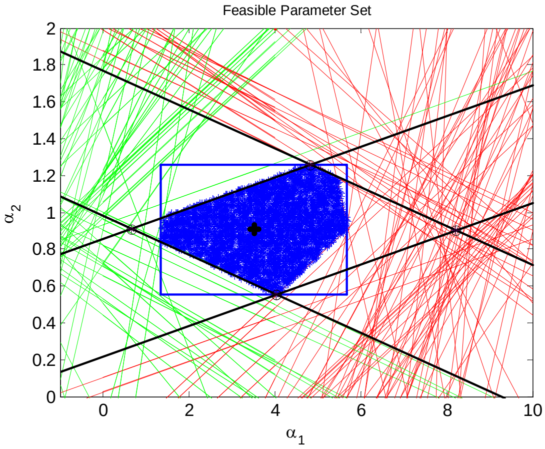

Figure 6. MCSM approximation of the feasible parameter set (blue dots), orthotopic (blue box) and parallelotopic (black parallelogram) outer-bounding set related to an error bound of 5

with f = [1 0]^T or [0 1]^T, and thus we have to solve two minimization and two maximization problems. Each of these solutions defines a vertical or horizontal line in the two-dimensional parameter space. This result is graphically presented by the blue box outer-bounding the FPS in Figure 6. Notice that this box can be rather conservative, especially when there is a strong correlation between the parameter estimates.

Parallelotopic outer-bounding: In addition to the observation that each measurement provides two parallel lines in the parameter space, it can be verified from Figure 6 that each couple of two measurements gives a parallogram in the two-dimensional parameter space. Hence, processing all $\binom{N}{2}$ combinations of measurements, which was done sequentially, and saving the one with the smallest volume led to the parallelogram define by the black lines in Figure 6.

Monte Carlo Set-Membership: Sampling the parameter space from the box $\mathcal{B}$ = [0, 10] × [0, 2] (5000 times) and evaluating the corresponding response with respect to the constraints given in (11) led to a finite set of feasible parameter vectors. In Figure 6 each feasible parameter vector from this procedure is presented by a blue dot, which all together give an inner approximation of the FPS.

As expected, the FPS indicated by blue dots and constrained by lower (green) and upper (red) bounds contains the min-max estimate (α_1 = 3.5237 and α_2 = 0.9072) and it contains the FPSs related to error bounds smaller than five. Increasing the error bound will thus lead to a larger FPS. Hence, the FPS directly reflects the larger uncertainty considered in the data.

The set-membership model output results related to an error bound of 5 are presented in Figure 7. Notice that the FMOS does contain almost all of the measurements and thus we may consider these results as appropriate for further evaluation. Notice from Figure 8 that the uncertainties increase for small and for large catchment areas (*A*). For instance, the (interpolated) bounds could be used in the identification of an erosion model, using bounded information, i.e. basically taking into account constraints instead of point measurements.

Pavement: For the pavement-based approximation we refer to Moore (1992) and Jaulin and Walter (1993). The algorithm subdivides an initial parameter hypercube (or in our case a box) into two along its longest side. It then calculates the bounds of an objective function for that box of parameters. Boxes with bounds that are completely within the required bounds are kept, and those completely outside the bounds are discarded. Boxes with bounds that cross the required bounds are added to a queue of boxes to further refine. The process keeps going until a tolerated maximum box width is reached (effectively a precision), or a number of iterations are completed. The objective function and bounds are chosen to match the other set-membership methods, that is the residuals with maximum magnitude must lie within [−5, 5], see for results Figure 8.

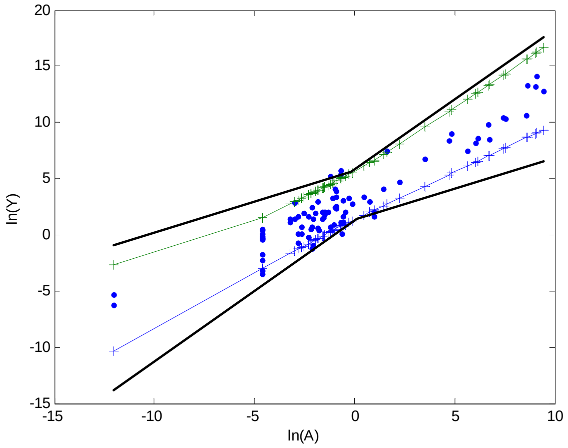

Figure 7. Feasible model output bounds from MCSM (upper bound: green +; lower bound: blue +) and from an outer-bounding box approximation related to an error bound of 5.

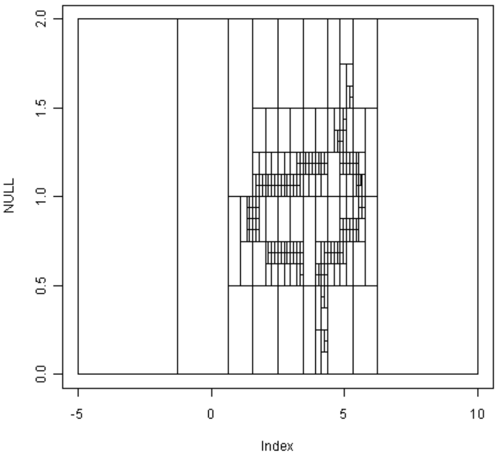

Figure 8. Collection of infeasible-indeterminate-feasible boxes as a result of a pavement algorithm.

4. DISCUSSION AND CONCLUSIONS

Both the Bayesian and set-membership estimation approaches have a solid basis, as all the assumptions are transparent and they can (and should) be checked afterwards. In real applications, the critical point, however, is the choice of the pdf's and the choice of bounds on the error vector **e**.

Accuracy: The accuracy of *DREAM/Bayes* depends on how well assumptions for example with respect to homoscedasticity of the residuals are fulfilled. The use of priors is both the strength and the weakness of the Bayesian inference method as narrow priors will dominate the results if there is not enough information in the data itself to identify the model parameters. The *GLUE* results depend on the user-defined threshold and objective function. In GLUE output uncertainty is a weighted average of results of all behavioural parameter vectors. Weights are based on likelihood measure of each behavioural set, while in MCSM the output uncertainty is directly evaluated from the feasible parameter vectors, thus without resampling.

For estimation problems linear in the parameters, thus with $\mathbf{F}(\vartheta) = \mathbf{F}\vartheta$, exact solutions in terms of vertices and edges exist for the characterization of the FPS. Paving algorithms are theoretically capable of arbitrary precision.

Computational requirements: *DREAM* is an efficient MCMC sampler, but for large problems (many parameters) may still require a long time to converge. *GLUE* is easy to use but as it uses naive MCMC it may be inefficient. Although theoretically the exact solution of linear set-membership estimation problems can be found, in practice it may be a hard problem due to the large complexity of the FPS. Hence, in practical applications approximate solutions, with limited computational requirements, may be more appropriate. A pavement algorithm requires the specification of initial parameter bounds, a way to run the model with interval parameters and obtain bounds on the objective function.

Skill requirements: *DREAM* is available in R and Matlab. However, the correct selection of the likelihood function requires good knowledge of the statistical properties of residuals and is thus case dependent. In *GLUE* there are subjective decisions that have to be made and these decisions will have an effect on the final results. For *exact/approximate bounding* tools are available in Matlab. These tools are easy to use. The key problem is the choice of appropriate error bounds. As MCSM is essentially based on simulations and a binary selection criterion, it is easy to understand. Paving algorithms are easy to understand, but a dedicated tool is required to implement the algorithm.

Range of applicability: *DREAM* and *GLUE* are both well suitable for all kind of calibration problems but time usage may be a problem for highly distributed models that need lots of time for a single function evaluation. *MCSM* is widely applicable, even to spatially distributed models with thresholds and other strong non-linearities. The basic limitation is the computational effort needed to provide a reasonable approximation of the FPS and FMOS. *Paving algorithms* are theoretically usable for any model and objective function. However, bounds on the objective function must be specified beforehand, and an appropriate method of calculating those bounds is required.

In conclusion, we can say that the uncertainty bounds given by different methods vary and thus there is a need for careful selection of the method used. Preferably, the method should have a firm theoretical basis. The user should understand this theoretical basis, so that results are not misinterpreted and abused. In the example, linear case with two parameters, there were no computational difficulties in using different methods and thus we recommend to use either Bayesian inference or an exact bounding algorithm. On the other hand, GLUE is worthy of consideration when it is difficult to fulfil residual assumptions. Indeed it could be used as a screening method to gain some initial understanding of uncertainties.

REFERENCES

Beck M. B., 1987. Water Quality Modeling: A Review of the analysis of Uncertainty. Wat Resour Res 23, 1393-1442.

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

Beven K. J., P. Smith and Freer J., 2008. So just why would a modeller choose to be incoherent? J Hydrol 354, 15-32.

Freni G., G. Mannina, G. Viviani, 2009.Urban runoff modelling uncertainty: Comparison among Bayesian and pseudo-Bayesian methods. Environmental Modelling & Software 24 1100–1111.

Jaulin, L. and E. Walter, 1993. Set inversion via interval analysis for nonlinear bounded-error estimation. Automatica, 29(4), 1053-1064.

Keesman, K.J., 2011. System Identification: an Introduction. Springer Verlag, London.

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

Milanese, M., J.P. Norton, H. Piet-Lahanier and E. Walter (Eds.), 1996. Bounding Approaches to System Identification, Plenum Press, NY.

Moore R.E., 1992. Parameter sets for bounded-error data. Math. and Computers in Simulation, 34, 113-119.

Norton, J.P., 1994. Bounded-error estimation: issue 1. Int. J. Adapt. Contr. and Sign. Process., 8(1).

Norton, J.P., 1995. Bounded-error estimation: issue 2. Int. J. Adapt. Contr. and Sign. Process., 9(1).

Qian, S. S., C. A. Stow, and M. E. Borsuk. 2003. On Monte Carlo methods for Bayesian inference. Ecol Model 159, 269-277.

Refsgaard J. C., J. P. van der Sluijs, A. L. Højberg, P. A. Vanrolleghem, 2007. Uncertainty in the environmental modelling process - A framework and guidance. Environmental Modelling & Software 22, 1543-1556.

Vrugt J. A., C. J. F. terBraak, H. V. Gupta and B. A. Robinson, 2009a. Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling? Stoch Environ Res Risk Assess 23:1011–1026 DOI 10.1007/s00477-008-0274-y.

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

Walter, E. (Ed.), 1990. Parameter identifications with error bounds. Special Issue Math. Comp. Simul., 32(5&6).

Wasson, R. J. (1994) Annual and Decadal Variation of Sediment Yield in Australia, and Some Global Comparisons, Variability in Stream Erosion and Sediment Transport (Proceedings of the Canberra Symposium, December 1994) IAHS Publication no. 224, pp. 269-279.