Appendix B

Methods for uncertainty analysis

B.1 UNCERTAINTY FRAMEWORKS

B.1.1 Frequentist

In the frequentist framework, it is assumed that the true value of a quantity (e.g., model parameter) is fixed (Johnson & Wichern, 2007). The problem is then to find a good approximation of that value as well as a region of confidence where this true value lies. This confidence interval is a measure of the uncertainty about the true value of the quantity. The distribution describes the probability for the obtained value of a variable or parameter (e.g., of variable measurements, model parameter estimates).

In general, the frequentist framework advocates the repetition of experiments in order to obtain samples of measurements or parameter estimates. Based on this sampling, one then estimates the distribution of the obtained values for the variable or parameter. For example, one can calculate the mean and variance which offer a complete characterisation of the normal distribution, assuming that the choice of a normal distribution is correct.

In the application of frequentist theory, it is assumed that by increasing the size of the sample, the estimated distribution will converge to the true distribution. For this to be true, two important conditions need to be met. First, the sampling procedure must lead to independently sampled values for the quantified variables. This is not always the case, especially when dealing with dynamic processes. Often, one has no access to repeated measurements or parameter estimates which are independent from each other. Second, the true distribution function should be able to be described by means of the fitted distribution function. This is often violated as well. Indeed, it is typical to assume the normal distribution for parameters in models that are non-linear in the parameters while the true distribution is not normal.

B.1.2 Bayesian

In the Bayesian framework, it is assumed that the quantity one seeks to identify is uncertain. Therefore, the quantified variable can take several plausible values. Each of these values will appear with different

Editor(s): Evangelia Belia, Lorenzo Benedetti, Bruce Johnson, Sudhir Murthy, Marc Neumann, Peter Vanrolleghem and Stefan Weijers
doi: 10.2166/9781780401034_0145

probability. This probability can again be described by a distribution; however, this distribution describes the probability for *the true value of a variable or parameter* (e.g., of the *true* variable value, *true* model parameters).

In the Bayesian context, one assumes a process which generates data. This process is given as a model which includes parameters and possibly input values. With such a process model, one generally describes the likelihood of measurements, y, conditional to the model parameters, p, denoted $L(y|p)$. Note that the likelihood is proportional to probability. It is not the same as probability however, as probability should sum to one while the likelihood does not in general sum to one.

The objective is to determine the likelihood (or probability) of the model parameters given some observations, written as $L(p|y)$. This describes the distribution of the parameters. To obtain this likelihood, one uses Bayes' rule (hence Bayesian statistics):

$$L(p|y) = L(y|p) \cdot L(p)/L(y) \quad \text{(Bayes' Rule)}$$

with

$$L(y) = \int L(y|p) \cdot L(p) \quad \text{(Sum rule: sum over all plausible values for } p\text{)}$$

In this formula, $L(p)$ represents the so-called prior likelihood for the parameters, in short *prior*.

By means of this prior, one includes prior information, knowledge or beliefs about the parameter into the calculations. For example, if a certain parameter cannot be negative then one constructs a prior which is zero for negative values of the parameter: $L(p)_{p<0} = 0$. $L(y)$ represents the total likelihood of the data. This is the overall likelihood for the data to have been observed for all considered values for the parameters p. In Bayes' rule, $L(y)$ is a scaling factor which makes sure that $L(p|y)$ integrates to one and thus represents a probability. If one does not scale with $L(y)$ then one can still find the parameters which maximise $L(p|y)$ since $L(y)$ is a constant. Such parameters are called the maximum likelihood parameter estimates. However, to obtain confidence limits for the parameters, one relies on the probability and should scale properly. Hence, $L(y)$ is needed for the quantification of uncertainty. In general, the calculus of $L(y)$ is difficult because there is no closed form or analytic solution for this sum/integral equation. As a result, several methods have been developed to approximate this integral.

B.2 MONTE CARLO SIMULATION

In Monte Carlo methods, the uncertainties in the model inputs and parameters are expressed as probability distributions. Multivariate samples are then obtained using a statistical sampling method, propagated through the model using simulations, and the results are analysed to develop probability distributions for the model output variables.

With sufficient sampling from an unknown distribution, the true distribution can eventually be approximated numerically. This paradigm can be put to use in a classic frequency-based statistical framework, where the true parameters are considered fixed and distributions of parameter estimates are characterised, or in a Bayesian context, where the model parameters, are considered to be uncertain and where the distributions of the parameters, not their estimates, are characterised. In the latter, a prior distribution is set up for the parameters, which reflects the expected distribution for the parameters in the absence of experimental data and/or before experimentation (hence prior).

One of the earliest documented applications of the basic Monte Carlo method was in the determination of the value of π (Hall, 1873). The term 'Monte Carlo' was coined in the 1940s by researchers working on nuclear fusion at the Los Alamos National Laboratory (Metropolis & Ulam, 1949).

Monte Carlo methods generate the solution of the integral of the product of two variables. Many problems can be formulated in this context such as finding the mean of a stochastic variable which is defined as the integral of the variable multiplied by its probability density function. Monte Carlo methods are often used to evaluate difficult multidimensional integrals with complicated boundary conditions. The problem of estimating uncertainties in simulation results can be formulated as an integration problem. For example, the mean of the model outputs is the integral of the product of the model outputs and the joint probability density function.

The basic Monte Carlo method can require a large number of samples in order to converge. The uncertainty in Monte Carlo simulations is proportional to $1/\sqrt{n}$ (Eckhardt, 1987), where n is the number of samples. This means that every decimal point of extra accuracy requires 100 times the number of samples. As a result, Monte Carlo simulations could require hundreds or thousands of simulations to converge depending on the required accuracy.

In order to reduce the number of simulations that must be run, methods have been developed to generate more efficiently the sets of random numbers required as model inputs. These include Markov chain Monte Carlo (Metropolis *et al.*, 1953), stratified sampling methods such as Latin hypercube sampling (LHS) and quasi-Monte Carlo (see Torvi & Hertzberg, 1998). Quasi-Monte Carlo methods construct deterministic sequences such as the Halton, Sobol or Hammersley sequences that share properties of random or pseudo-random sequences. These methods are found to have less error than random Monte Carlo methods and require fewer samples to converge but the advantage may be slight in large problems (Morokoff & Caflisch, 1995).

The probability density functions used for the model input variables and parameters depend on the available data. In cases where data are available, the distribution of the data can be determined using statistical techniques. For variables for which little information is known except for expected minimum and maximum values, a uniform distribution is often used. A triangular distribution is used if a most likely value and minimum and maximum values are known.

B.2.1 Random sampling and LHS

In the random sampling (RS) procedure, at each Monte Carlo run, a vector of model parameters is randomly sampled from the joint distribution of parameters. The sampling of parameters at each Monte Carlo run is independent from the previous ones. Therefore, in this sampling approach, the coverage of the entire support of distributions (used for the characterisation of model parameter uncertainty) might not be guaranteed, unless a large-enough number of Monte Carlo simulations is performed.

An alternative sampling to the RS method for exploring the support of different parameter distributions is the LHS method. In the LHS method, the range of the input variable distribution is divided into N sub-intervals (e.g., $N = 4$ in Figure B.1) with equal probabilities. One value is selected from each sub-interval and this process is repeated for all the input variables. The generated input variable values are then paired randomly to generate a sequence of input samples for use in the Monte Carlo simulations. Compared to the RS method in which different samples are generated by RS directly from the entire range of distributions, in LHS, RS is performed in each sub-interval and all sub-intervals are sampled.

Figure B.1 illustrates the result of generating four vectors of parameters in a two-dimensional parameter space generated using RS and LHS methods. As indicated in (a), in this particular realisation of four samples, generated according to the RS method, no value is sampled from sub-interval (1) of parameter θ_1 and sub-interval (2) of parameter θ_2. However, the sampling result based on the LHS method indicates that the generated values include representatives from all sub-intervals.

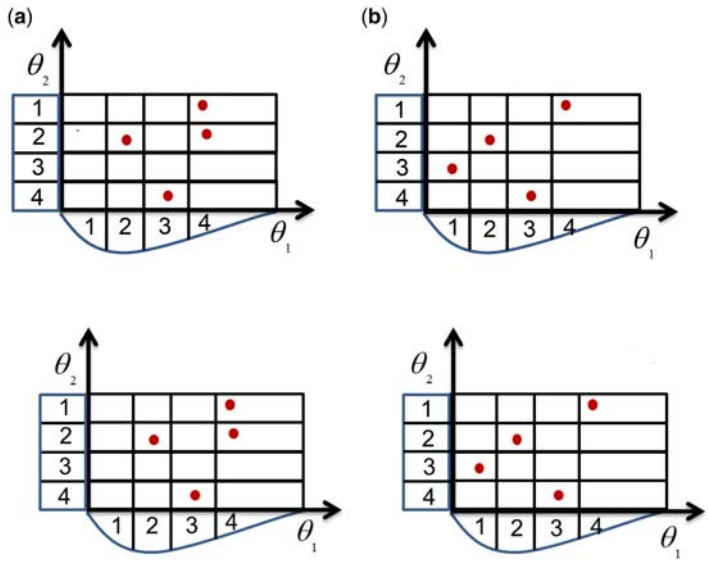

Figure B.1 (a) Schematic of a RS and (b) LHS procedures.

In general, the application of LHS could reduce the number of sampled values required to reach convergence of the output distributions (Tung & Yen, 2005). However, there could be some cases where LHS sampling would not lead to a more rapid convergence of output distribution compared to the RS sampling as the convergence also depends on the complexity of the model and its parameters.

In addition, in the RS method, the sampling of parameters at each Monte Carlo run is independent from the next one and in each run the convergence of the output distributions can be checked to determine whether more simulations are required or not. In contrast, in the LHS method, the number of Monte Carlo runs should be determined and samples generated before running any simulations. Therefore, if the selected number of Monte Carlo simulations turns out to be insufficient, the users cannot simply add more samples (like in the RS method) unless the consistency of the LHS procedure is insured.

A possible solution to increase the size of samples in the LHS method is the replicated LHS method (McKay *et al.*, 1979) in which instead of generating N number of samples using the LHS, k number of LHS designs with n number of samples each, is generated ($N = k \times n$). After the termination of each Monte Carlo simulation with n samples, the convergence of the output distributions is checked, and if more simulations are required, other n samples are generated using the LHS and Monte Carlo simulation continues using the newly generated samples. The efficiency of the repeated LHS depends on the appropriate choice of n as selecting it too large may not result in significant reductions of model runs and a value that is too small could result in inadequate coverage of the entire parameter space (Benedetti *et al.*, 2011).

B.2.2 Introducing correlations between parameters

One of the important factors in Monte Carlo simulations that could affect some of the statistical properties of the simulated output distributions is proper incorporation of possible correlation structures in the sampling of uncertain parameters. Different methods presented in the literature can be used to introduce a desired correlation structure among the sampled values (Iman & Conover, 1982; Tung & Yen, 2005). However,

some of the methods suffer from certain shortcomings and their application depends on the validity of a set of assumptions regarding the marginal distribution of the parameters (Tung & Yen, 2005).

One of the commonly used methods for introducing correlation among uncertain parameters is the method of Iman−Conover (Iman & Conover, 1982). Being independent from the type of marginal distributions, applicability to any sampling scheme (e.g., RS or LHS), and relatively simple implementation have been mentioned as the main advantages of this method (Iman & Davenport, 1982).

REFERENCES

Benedetti L., Claeys F., Nopens I. and Vanrolleghem P. A. (2011). Assessing the convergence of LHS Monte Carlo simulations of wastewater treatment models. *Water Science Technology*, **63**(10), 2219–2224. http://hdl.handle.net/1854/LU-1922233 (accessed 2021).

Eckhardt R. (1987). Stan Ulam, John Von Neumann, and the Monte Carlo Method. *Los Alamos Science*, Special Issue.

Hall A. (1873). On an experimental determination of PI. *Messenger of Mathematics*, **2**, 113–114.

Iman R. and Conover W. J. (1982). A distribution-free approach to inducing rank correlation among input variables. *Communication in Statistics – Simulation and Computation*, **11**(3), 311–334.

Iman R. L. and Davenport J. M. (1982). An Iterative Algorithm to Produce a Positive Definite Correlation Matrix From an 'Approximate Correlation Matrix'. Sandia National Laboratories, Albuquerque, NM (USA), SAND81-1376. https://doi.org/10.2172/5152227 (accessed 2021).

Johnson R. A. and Wichern D. W. (2007). Applied Multivariate Statistical Analysis, 6th edn, Pearson Prentice Hall, Upper Saddle River, NJ.

McKay M. D., Beckman R. J. and Conover W. J. (1979). Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, **21**(2), 239–245, doi: 10.1080/00401706.1979.10489755

Metropolis N. and Ulam S. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, **44**, 335–341.

Metropolis N., Rosenbluth A. W., Rosenbluth M. N. and Teller A. H. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087. https://doi.org/10.1063/1.1699114 (accessed 2021).

Morokoff W. J. and Caflisch R. E. (1995). Quasi-Monte Carlo integration. *Journal of Computational Physics*, **122**(2), 218–230.

Torvi H. and Hertzberg T. (1998). Methods of evaluating uncertainties in dynamic simulation – a comparison of performance. *Computers & Chemical Engineering*, **22**(Suppl), S985–S988.

Tung Y. K. and Yen B. C. (2005). Hydrosystem Engineering Uncertainty Analysis. McGraw-Hill Book Company, New York.