A New Bayesian Approach for Statistical Calibration of Computer Model

Frédérick Delay*[1], Thierry Mara[2], and Anis Younes[1,3]

[1]Laboratoire Hydrologie et Geochimie de Strasbourg, University of Strasbourg/EOST, CNRS, 1 rue Blessig 67084 Strasbourg, France
[2]PIMENT, Université de La Réunion, 15 Avenue René Cassin, BP 7151, 97715 Moufia, La Réunion, France
[3]IRD UMR LISAH, F-92761 Montpellier, France

Abstract

The validation of computer models is an essential task to increase their credibility. One of the most important exercises in the validation framework is to check whether the computer model adequately represents reality [1]. This is achieved by comparing model predictions to observation data. This exercise generally leads to model calibration because the model parameters are usually poorly known a priori (i.e. before collecting data). Good practice in calibration of computer models consists of searching for all parameter values that satisfactorily fit the data, thus determining their plausible range of uncertainty. This can be achieved in a Bayesian framework in which the prior knowledge about the model and the observed data are merged to define the joint posterior probability distribution function (pdf) of the parameters. The issue is then to assess the joint posterior pdf.

In probabilistic inverse modeling, the parameter set $\mathbf{x} = (x_1, \ldots, x_d)$ of a computer model is inferred from a set of observation data $\mathbf{y}$ using the Bayesian inference, which defines the conditional joint posterior pdf as follows:

$$p(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{x})p(\mathbf{x}), \tag{1}$$

where $p(\mathbf{x})$ is the prior density that characterizes the investigator's beliefs about the parameters before collecting the new observations, and $p(\mathbf{y}|\mathbf{x})$ is the likelihood function, which measures how well the model fits the data. The parameter set that maximizes Eq. (1), namely:

$$\mathbf{x}^{\text{MAP}} = \underset{\mathbf{x}}{\text{argmax}}\, p(\mathbf{x}|\mathbf{y}), \tag{2}$$

is called the Maximum A Posteriori (MAP) estimate of the parameters. It is the most probable parameter set given the data and can be inferred via an optimization technique. The marginal posterior pdf that characterizes the uncertainty of a single parameter is defined by the following integral:

$$p(x_i|\mathbf{y}) = \int p(\mathbf{x}|\mathbf{y})d\mathbf{x}_{-i}, \qquad \forall i = 1, \ldots, d \tag{3}$$

where $\mathbf{x}_{-i}$ represents all the parameters except x_i. Usually, the integral in Eq. (3) is evaluated by a multidimensional quadrature method or by direct summations in a large sample of $p(x_i|\mathbf{y})$ obtained, for instance, via an MCMC technique.

*Corresponding author: fdelay@unistra.fr

The inference of model parameter posterior pdf by means of Markov chain Monte Carlo (MCMC) sampling techniques [2, 3] has received much attention in the last two decades. MCMC explores the region of plausible values in the parameter space and provides successive parameter draws directly sampled from the target joint pdf. Some selection criteria are used to ensure that the successive draws in the chain improve. This means that, throughout the sampling process, probable draws with respect to the target distribution are more likely drawn. Many developments and improvements have been proposed to accelerate MCMC convergence (see [4–6]).

Recently, [7] proposed a new probabilistic approach to the inverse problem whose main idea is to maximize the joint posterior pdf of a parameter set with one selected parameter sampling successive prescribed values. This provides the so-called Maximal Conditional Posterior Distribution (MCPD) of the selected parameter. The maximal conditional posterior distribution of x_i is defined as follows:

$$\mathcal{P}(x_i) = \max_{\mathbf{x}_{-i}} \left(p(\mathbf{x}_{-i}|\mathbf{y}, x_i)\right) \times p(x_i|\mathbf{y}). \tag{4}$$

An informal definition can be given by stating that a point estimate of the MCPD is the maximal value reached by the joint pdf (Eq. (1)) for a given (prescribed) value of one parameter i.e. x_i). This maximal value, in the context of model inversion, assumes that the set $\mathbf{x}_{-i}$ maximizes (Eq. (1)), knowing that x_i is prescribed. By applying the axiom of conditional probabilities to (Eq. (4)), it can be stated that $\max\left\{p(\mathbf{x}_{-i}|\mathbf{y}, x_i)\right\} \times p(x_i|\mathbf{y}) = \max_{\mathbf{x}_{-i}}\left\{p(\mathbf{x}_{-i}, x_i|\mathbf{y})\right\}$. Therefore, the MAP estimate (when it exists) belongs to the MCPD of all parameters.

The main advantage of the recent MCPD technique is that parameter distributions can be inferred independently. Therefore, the MCPDs can be simultaneously evaluated on multicore computers (or on multiple computers). This drastically reduces the computational effort in terms of computational time units (CTU). Our presentation will develop the MCPD approach and exemplify its efficiency in solving inverse problems.

References

[1] M. .J. Bayarri, J. O. Berger, R. Paulo, J. Sacks, J. A. Cafeo, J. Cavendish, C. H. Lin, and J. Tu. A framework for validation of computer models. *Technometrics*, 49:138–154, 2007.

[2] N.-A. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1091, 1953.

[3] H. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[4] H. Haario, M. Laine, A. Mira, and E. Saksman. DRAM: Efficient adaptive MCMC. *Statistics and Computing*, 16(339–354), 2006.

[5] C. J. F. ter Braak and J. Vrugt. Differential evolution markov chain with snooker updater and fewer chains. *Statistics and Computing*, 18(4):435–446, 2008.

[6] E. Laloy and J. A. Vrugt. High-dimensional posterior exploration of hydrologic models using multiple-try DREAM$_{(\mathrm{zs})}$ and high-performance computing. *Water Resources Research*, 48:W01526, 2012.

[7] T. A. Mara, N. Fajraoui, A. Younes, and F. Delay. Inversion and uncertainty of highly parameterized models in a bayesian framework by sampling the maximal conditional posterior distribution of parameters. *Advances in Water Resources*, 76:1–10, 2015.