Random sampling from joint probability distributions defined in a bayesian framework

Thierry A. Mara, Marwan Fahs, Qian Shao, Anis Younes

RANDOM SAMPLING FROM JOINT PROBABILITY DISTRIBUTIONS DEFINED IN A BAYESIAN FRAMEWORK*

THIERRY A. MARA†‡, MARWAN FAHS§, QIAN SHAO¶, AND ANIS YOUNES§‖**

Abstract. Random variables characterized by a joint probability distribution function (jpdf) defined in a Bayesian framework are generally sampled with Markov chain Monte Carlo (MCMC). The latter can be computationally demanding when the number of variables is high. As an alternative, the maximal conditional probability distribution (MCPD) sampler was recently introduced by some of the authors of the present article to readily and efficiently draw values randomly sampled from the desired jpdf. The MCPD approach provides the probability distribution of a given variable under the condition that the other variables maximized the conditional jpdf. However, contrarily to MCMC, MCPD does not provide enough draws to allow posterior uncertainty and sensitivity analyses of the computer model responses. In the present work, we show how to draw random samples from the MCPD draws under the requirement that the target jpdf possesses a particular dependence structure. Several numerical tests are carried out to prove the efficiency of the new sampling method. The new approach is used to perform the predictive uncertainty and sensitivity analyses of numerical models posterior to their statistical calibration from experimental data.

Key words. Bayesian framework, model statistical calibration, maximal conditional posterior distribution, posterior uncertainty and sensitivity analyses, numerical drainage experiment

AMS subject classifications. 65C60, 62H20

1. Introduction. Sampling random variables from a given joint probability distribution function (jpdf) is a challenging issue if the latter is not defined in a closed-form. This is the case when modellers wish to statistically calibrate the input variables of their computer model in a Bayesian framework (e.g. [3, 1]). Sampling from the jpdf is necessary to perform, for instance, the predictive uncertainty and sensitivity analyses of the computer model responses (e.g. [36, 17, 24]. Many authors privilege sampling the input values from a given jpdf with Markov chain Monte Carlo (MCMC) [47, 33, 26, 38, 34, 18].

MCMC relies on a rejection/acceptance sampling to generate random draws sampled from the jpdf [28, 15]. It has been subject to several developments and improvements during the last two decades aiming at accelerating its convergence (e.g. [10, 12, 9, 11, 43, 5] among others). However, MCMC sampling is challenging in high-dimensions and remains computationally demanding.

Recently, a new approach has been introduced for the statistical calibration of computer models [22]. The authors named it the maximal conditional probability distribution sampler. The first step of this approach is to seek all the probable local optima of the jpdf (assuming the existence of a finite number of local optima). Then, several maximizations of the conditional jpdf are performed for different prescribed

*This work was supported by the French National Research Agency through the research project RESAIN n°ANR-12-BS06-0010-02.

†PIMENT, EA 4518, Université de La Réunion, FST, 15 Avenue René Cassin, 97715 Saint-Denis, Réunion (mara@univ-reunion.fr)

‡European Commission, Joint Research Centre, Directorate for Modelling, Indicators and Impact Evaluation, 21027 Ispra (VA), Italy (thierry.mara@ec.europa.eu)

§LHyGeS, UMR-CNRS 7517, Université de Strasbourg/EOST, 1 rue Blessig, 67084 Strasbourg, France

¶School of Civil Engineering, Wuhan University, 8 South Road of East Lake, Wuchang, 430072 Wuhan, PR China

‖UMR LISAH, INRA-IRD-SupAgro, 92761 Montpellier, France

**LMHE, Ecole Nationale d'Ingénieurs de Tunis, Tunisia

values of one selected variable. The values assigned to the selected variable are successively drawn around each local optimum while the values of the other variables are investigated by maximizing the conditional jpdf. This provides what is called the maximal conditional probability distribution (MCPD) of the selected variable.

The evaluation of the MCPDs is independent of each other. Consequently, they can be evaluated simultaneously by distributing the calculations over several computers (or a multi-core computer). This feature drastically decreases the computation time and makes the inversion of highly parameterized problems feasible. For instance, in [22], a flow model with $d = 104$ parameters has been calibrated with the MCPD approach by distributing the calculations over eight cores (each core assessing 13 MCPDs). An overall of 54 560 model calls was necessary to assess the MCPDs, but thanks to the parallelization the actual waiting time (also called computational time unit) was about 6 500 model calls. Several comparisons between MCPD and a MCMC sampler have been carried out in [21] which demonstrated the efficiency of the MCPD approach.

However, the MCPD draws represent a few probabilistic set of values sampled from the jpdf. If stochastic samples distributed over the target jpdf are desired (for the purpose of uncertainty and sensitivity analyses of computer models for instance), the MCPD draws are not sufficient. In the present work, we extend the approach in order to generate Monte Carlo samples from the MCPD draws. This extension is based on the orthogonalization procedure introduced in [23] which assumes a particular correlation structure between the variables. The new sampling method is tested on two numerical target distributions of increasing complexity before application to the identification of soil hydraulic parameters from a synthetic multi-step outflow experiment and to the calibration of a reactive transport model from field observations. In these last two applications, the use of Monte Carlo samples for uncertainty and sensitivity analyses is illustrated.

The paper is organized as follows: First, we briefly discuss the different methods to generate random samples from a given jpdf in Section 2. Then, we recall the concept of MCPD sampling and describe succinctly its assessment in Section 3. The approach for sampling random variables from the desired distribution is explained in Section 4. The new approach is evaluated through several numerical exercises of different dimensions and complexities in Section 5. In Section 6 we apply the new sampling approach to the calibration of a drainage model from synthetic multi-step outflow experiment. Finally, the statistical calibration of a reactive transport model from field data is undertaken in Section 7 before concluding in Section 8.

2. Sampling from joint distribution: Existing approaches. Let $\boldsymbol{x} = (x_1, \ldots, x_d)^T$ be a vector of random variables distributed over the joint probability density function $p(\boldsymbol{x})$, where the superscript T denotes the transpose operator. We are asked to generate $\mathbf{x}$, a sample of size N, of the random vector. The most straightforward approach to perform this task is to use the inverse Rosenblatt transformation [35],

that is,

$$\begin{cases} x_1 = F_1^{-1}(u_1) \\ x_2 = F_{2|1}^{-1}(u_2|x_1) \\ x_3 = F_{3|1,2}^{-1}(u_3|x_1, x_2) \\ \vdots \\ x_d = F_{d|1,\ldots,d-1}^{-1}(u_d|x_1, \ldots, x_{d-1}) \end{cases} \tag{2.1}$$

where $\boldsymbol{u}$ is a random vector uniformly and independently distributed within the unit hypercube $[0,1]^d$ and $F_{i|1,\ldots,i-1}^{-1}$ is the inverse cumulative distribution of x_i conditioned onto $(x_1, \ldots, x_{i-1})$. By sampling $\boldsymbol{u} \in [0,1]^d$ and knowing the set of inverse conditional cumulative distribution functions (CDF) $\{F_1^{-1}, F_{2|1}^{-1}, \ldots, F_{d|1,\ldots,d-1}^{-1}\}$, the desired sample of $\boldsymbol{x}$ is obtained by applying Eq. (2.1). Unfortunately, in many situations and especially in model statistical calibration, the conditional cumulative distribution functions are unknown.

More often, random variables are defined by their unconditional marginal CDFs $\{F_1(x_1), \ldots, F_d(x_d)\}$ and a copula density $c : [0,1]^d \to [0, +\infty[$. The copula density contains the dependence/correlation structure of the jpdf. It is convenient to define random variables in this way because many algorithms are proposed in the literature to generate random samples [31]. For instance, the inverse Nataf transform is designed to cope with Gaussian copulas [30]. We note that defining random variables from copula theory is tantamount to define the jpdf in a closed-form because of the following equality: $p(\boldsymbol{x}) = c(F_1(x_1), \ldots, F_d(x_d))p_1(x_1)\ldots p_d(x_d)$, where $p_i(x_i) = \mathrm{d}F_i(x_i)/\mathrm{d}x_i$ is the marginal pdf of x_i. However, there are many problems for which such an explicit definition is not possible as when, for model calibration purposes, the target jpdf is derived in a Bayesian framework. In such a situation, Markov chain Monte Carlo samplers are preferred.

MCMC samplers are based on acceptance-rejection algorithms. They employed the Metropolis-Hastings algorithm (or a variant of it) to sample $\mathbf{x}$ from $p(\boldsymbol{x})$ thanks to a given proposal distribution $q(\cdot|\cdot)$. There are several MCMC variants proposed in the literature. Amongst them, we can cite the Gibbs sampler [8] which requires the knowledge of the conditional distributions $p_{i|\sim i}(x_i|\boldsymbol{x}_{\sim i})$, $\forall i = 1, \ldots, d$ ($\boldsymbol{x}_{\sim i}$ stands for all the x-variables except x_i). When the conditional distributions are unknown, the Langevin [10] or Hamiltonian [16] MCMC sampler can be a good choice to generate $\mathbf{x}$. Notably, the efficiency of these samplers is enhanced if the Jacobian of $\log(p(\boldsymbol{x}))$ is provided. When the latter is not available or not differentiable, the DiffeRential Evolution Adaptive Metropolis sampler is a good alternative, in particular the version exploiting the archive of past states called DREAM$_{ZS}$ [19, 46]. DREAM$_{ZS}$ is particularly efficient, since instead of drawing new candidates from a prescribed proposal distribution q, new candidates are drawn thanks to an archive of past states.

Despite of the obvious improvements in MCMC sampling, the use of this method to generate a sample of $\boldsymbol{x} \sim p(\boldsymbol{x})$ can still be computationally demanding. This is particularly embarrassing when the computational time required to evaluate the target jpdf is high. In this case, the two stages surrogate-based MCMC approach of [4] seems a promising alternative.

The aim of the present work is to introduce a new fast approach to generate random draws from the desired jpdf based on the MCPD sample. This work capitalizes upon recent works of some of the authors of the present article [23, 22]. The approach,

described in the next sections, is particularly efficient if the Jacobian of $\log(p(\boldsymbol{x}))$ is provided. It also requires a particular correlation structure between the random variables. Although these constraints narrow its field of applications, it is believed that the proposed approach remains a valuable tool in many model calibration problems.

3. The maximal conditional posterior distribution. The maximal conditional posterior distribution of x_i writes

$$\mathcal{P}_i(x_i) = \max_{\boldsymbol{x}_{-i}} (p(\boldsymbol{x}_{-i}|x_i)) \times p(x_i). \tag{3.1}$$

$\mathcal{P}_i(x_i)$ is interpreted as the posterior probability function that maximizes the conditional posterior distribution $p(\boldsymbol{x}_{-i}|x_i)$. The MCPD of x_i is assessed in a discrete form by evaluating Eq. (3.1) at different values of x_i. In practice, the sampled values of x_i (denoted $\mathrm{x_i}^*$) are picked around each probable local optimum (estimated beforehand) within its prior uncertainty range. This gives,

$$\mathbf{x}^*_{-i} = \underset{\boldsymbol{x}_{-i}}{\operatorname{argmax}} \{p(\boldsymbol{x}_{-i}|x_i = \mathrm{x_i}^*)\} \tag{3.2}$$

$$\mathcal{P}_i(\mathrm{x_i}^*) = p(\mathbf{x}^*_{-i}|\mathrm{x_i}^*) \times p(\mathrm{x_i}^*) = p(\mathbf{x}^*) \tag{3.3}$$

The MCPD approach makes sense provided that the target jpdf $p(\boldsymbol{x})$ admits a finite number of modes. Notably, the success of MCPD sampling relies on the ability to retrieve all the probable local optima of $p(\boldsymbol{x})$. In the present work, all the local optima are searched by use of gradient-based methods with multiple starting points. Multiple tries ensure that the algorithm starts searching the local optima from different regions of the input space. Of course, the selection of the optimization technique is a matter of choice.

The algorithm used to compute the MCPDs is thoroughly explained in [22] and [21]. For convenience, it is succinctly recalled here. The algorithm is divided into three parts: in part 1, all the probable optima of $p(\boldsymbol{x})$ are investigated. In part 2, the posterior range of variation of each parameter around each probable optimum is roughly estimated by preliminarily evaluating its MCPD. In part 3, the evaluation of the MCPDs is refined. It is worth reminding that fast computation is possible because these three parts can take advantage of parallel computing.

Let us denote the MCPD draws in the vicinity of the local optimum $\mathbf{x}^{opt,m}$ by $\left\{\mathbf{x}^{k_i,m}\right\}_{i=1}^{d}$, $\forall k_i = 1, \ldots, N_i$. We set $p(\mathbf{x}^{opt,1}) \geq p(\mathbf{x}^{opt,2}) \cdots \geq p(\mathbf{x}^{opt,m})$ which means that $\mathbf{x}^{opt,1}$ is the global optimum. The subset $\left\{\left(\mathbf{x}^{k_i,m}, \mathcal{P}_i^{k_i,m} = p(\mathbf{x}^{k_i,m})/p(\mathbf{x}^{opt,1})\right)\right\}_{k_i=1}^{N_i}$ $\forall m = 1, \ldots, M$ represents the discretized MCPD of variable x_i. As defined, the discretized MCPD representation is scaled within $[0, 1]$. As explained above, the MCPD draws of x_i are estimated by setting $\mathrm{x_i}^* = \mathrm{x_i}^{k_i,m}$ and solving Eq. (3.2). The value $\mathrm{x_i}^{k_i,m}$ is successively drawn in the vicinity of the current optimal value of x_i. Figure 3.1 depicts an example of the MCPD draws for a unimodal target distribution. On the diagonal, the discretized MCPDs are plotted in blue crosses. On the lower off-diagonal, at row $\#j$ and column $\#i$, the pairs $(\mathrm{x_i}^{k_i,m}, \mathrm{x_j}^{k_i,m})$ and $(\mathrm{x_i}^{k_j,m}, \mathrm{x_j}^{k_j,m})$ are plotted. The first pair is obtained while $\mathcal{P}_i$ is assessed whereas the second one stems from the assessment of $\mathcal{P}_j$. This scatterplot shows possible correlations between the x_i and x_j draws. For instance, x_1 and x_2 show no correlation (row #2, column #1) because virtually two orthogonal curves are observed while the MCPD draws of x_2 and x_3 are correlated (row #3, column #2).

In [24], it was demonstrated that the MCPD draws were (by far) faster to sample than the MCMC draws. Nevertheless, the two samples are not comparable because the probabilistic MCPD draws are sampled with constraints (maximization of the conditional jpdf, see Eq. (3.2)) while the stochastic MCMC draws are obtained from the same jpdf but unconditionally. Therefore, posterior uncertainty and sensitivity analyses of computer models cannot be performed with the MCPD draws. For this purpose, a Monte Carlo (MC) sample that covers the posterior input space would be preferable. Hence, our aim in this work is to propose an approach to generate MC samples from the MCPD draws. To this end, an approach that can impose a desired correlation structure amongst independent MC samples is needed. This is described in the next section.

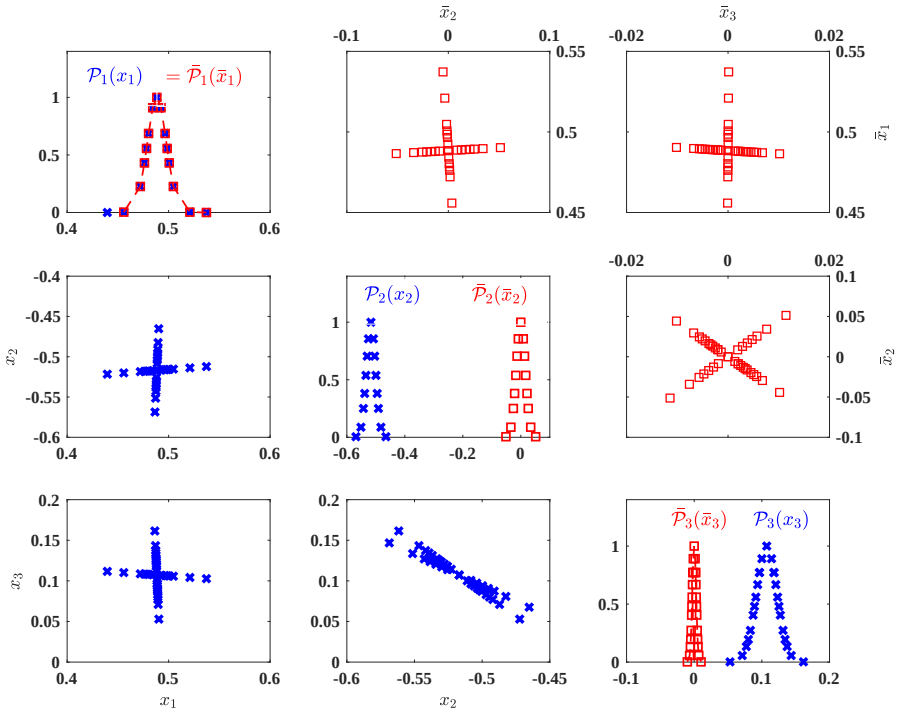

FIGURE 3.1. *Example of MCPD draws prior and posterior to the transformation procedure Eq. (4.2). Subplots on the diagonal display the MCPD draws of the original variable x_i and its transformed one $\bar{x}_i$ (blue crosses and red squares resp.). The lower off-diagonal plots show the pairwise correlations between the original variables. The draws decorrelated with the transformation Eq. (4.2) are depicted on the upper off-diagonal.*

4. The new sampling approach. In this section we describe the algorithm to generate stochastic Monte Carlo samples from a set of probabilistic MCPD sample. The success of the method heavily relies on the decorrelation procedure proposed in [23] . Here we recall the procedure in § 4.1. In the case of multimodal target jpdf, a balanced sampling scheme is necessary which is described in § 4.2. We discuss the validity of the approach and a possible alternative in case of failure in § 4.3. The algorithm of the proposed approach, called MCPD-MC, is given in § 4.4.

4.1. Decorrelation of the MCPD draws. Assume that the vector of dependent random variables $\boldsymbol{x} = (x_1, \ldots, x_d)^T$ is function of a set of independent latent random variables $\bar{\boldsymbol{x}} = (\bar{x}_1, \ldots, \bar{x}_d)^T$. Let us further assume that the dependence structure of the x-variables is strictly due to their relationships with the latent variables which are of the form,

$$\begin{cases} x_{i_1} = \bar{x}_1 \\ x_{i_2} = \bar{x}_2 + f_{i_2 1}(\bar{x}_1) \\ x_{i_3} = \bar{x}_3 + f_{i_3 1}(\bar{x}_1) + f_{i_3 2}(\bar{x}_2) \\ \vdots \\ x_{i_d} = \bar{x}_d + \sum_{j=1}^{d-1} f_{i_d j}(\bar{x}_j) \end{cases} \tag{4.1}$$

where, in the j-th equation, $\int_{\mathbb{R}} f_{i_j k}(\bar{x}_k) p(\bar{x}_k) \mathrm{d}\bar{x}_k = 0$, $\forall k = 1, \ldots, j-1$ which ensures that the set of functions $\{f_{i_j 1}(\bar{x}_1), \ldots, f_{i_j j-1}(\bar{x}_{j-1})\}$ is orthogonal (according to [41]). We note that each equation in Eq. (4.1) is related to the generalized additive model representation [14].

It is worth noticing that Eq. (4.1) may not always be possible for any set $\{i_1, \ldots, i_d\} = \{1, \ldots, d\}$, for instance, when the relationship between two variables is non-monotonic (see an example in § 5.2). However, we assume that it is at least possible for one of the sets (as for the non-monotonic case).

PROPOSITION 4.1. *Under assumption* (4.1)*, the transformation procedure proposed in [23] that writes*

$$\begin{cases} \bar{x}_1 = x_{i_1} \\ \bar{x}_2 = x_{i_2} - \mathbb{E}\left[(x_{i_2} - \mathbb{E}\left[x_{i_2}\right]) | \bar{x}_1\right] \\ \bar{x}_3 = x_{i_3} - \sum_{j=1}^{2} \mathbb{E}\left[(x_{i_3} - \mathbb{E}\left[x_{i_3}\right]) | \bar{x}_j\right] \\ \quad \vdots \\ \bar{x}_d = x_{i_d} - \sum_{j=1}^{d-1} \mathbb{E}\left[(x_{i_d} - \mathbb{E}\left[x_{i_d}\right]) | \bar{x}_j\right] \end{cases} \tag{4.2}$$

where $\mathbb{E}[\cdot]$ *is the mathematical expectation and* $\mathbb{E}[\cdot|\cdot]$ *is the conditional expectation, provides a set of independent variables* $\bar{\boldsymbol{x}}$*.*

Eq. (4.2) slightly differs from the original set of equations in [23] because in their paper it was assumed that $\boldsymbol{x}$ was a vector of standardized random variables (i.e. centered and reduced). The modified version here considers the conditional expectations, $\left(x_{i_j} - \mathbb{E}\left[x_{i_j}\right]\right)$ $j = 2, \ldots, d$, which are centered. Indeed, the reduction of the x-variables ($\mathbb{V}[x_i] = 1$) is not mandatory for the independence of the $\bar{x}$-variables. The proposed approach to generate random samples w.r.t. the target jpdf, first, requires to apply transformation (4.2) to the MCPD draws. Numerically this is achieved by approximating $\mathbb{E}[(x_i - \mathbb{E}[x_i])|\bar{x}_j] = f_{ij}(\bar{x}_j)$ with a regression method. In the present work, polynomial regressions are used in conjunction with Schwartz's criterion for the automatic selection of polynomial degree [39].

Applying the previous transformation to the set $\left\{\mathbf{x}^{k_i,m}\right\}_{k_i=1}^{N_i}$ and $\forall i = 1, \ldots, d$, which is the MCPD sample drawn in the vicinity of the local optimum $\mathbf{x}^{opt,m}$, yields the independent sample $\left\{\bar{\mathbf{x}}^{k_i,m}\right\}_{k_i=1}^{N_i}$, the pairs $\left(\bar{\mathrm{x}}_{\mathrm{i}}^{k_i,m}, \bar{\mathcal{P}}_i(\bar{\mathrm{x}}_{\mathrm{i}}^{k_i,m})\right)$ as well as the overall set of f_{ij} functions around the optimum. An example of such a transformation is

depicted in Figure 3.1 (red squares). One can note that the distribution of $\bar{x}_1$ is the same as the one of x_1 (row #1, column #1). The one of $\bar{x}_2$ is similar to the one of x_2 except that it is centered (row #2, column #2). $\bar{\mathcal{P}}_3(\bar{x}_3)$ is centered and narrower as compared to $\mathcal{P}_3(x_3)$ (row #3, column #3). The shrinkage of $\bar{x}_3$'s distribution is due to the correlation between x_3 and x_2. Pairwise analysis shows no correlation amongst the transformed draws (upper off-diagonal plots) contrarily to the MCPD draws depicted on the lower off-diagonal which shows a strong correlation between x_2 and x_3 (row #3, column #2).

4.2. Sampling from the target distribution. Given the estimated set of f_{ij} functions and the set of transformed MCPD $\bar{\mathcal{P}}_i$, we are now ready to propose a new approach to generate $\mathbf{x}$ from the joint probability distribution without further evaluating $p(\boldsymbol{x})$. First, given that $\bar{\boldsymbol{x}}$ is a vector of independent random variables in Eq. (4.1), we start by drawing the $\bar{x}$-variables from the $\bar{\mathcal{P}}_i(\bar{x}_i)$'s. For this purpose, we use the latin hypercube sampling [27] that requires the individual cumulative distribution functions (CDF) around each optimum. The CDF of $\bar{x}_i$ around the m-th optimum is defined as follows,

$$\begin{aligned} \mathcal{A}_m &= \int_{lo_i}^{up_i} \bar{\mathcal{P}}_i(\bar{x}_i)\mathrm{d}\bar{x}_i \\ \bar{F}_i^m(\bar{X}_i) &= \frac{1}{\mathcal{A}_m}\int_{lo_i}^{\bar{X}_i} \bar{\mathcal{P}}_i(\bar{x}_i)\mathrm{d}\bar{x}_i, \quad \forall i = 1,\ldots,d \end{aligned} \tag{4.3}$$

with $lo_i = \min\{\bar{\mathrm{x}}_\mathrm{i}^{1,m},\ldots,\bar{\mathrm{x}}_\mathrm{i}^{N_i,m}\}$, $up_i = \max\{\bar{\mathrm{x}}_\mathrm{i}^{1,m},\ldots,\bar{\mathrm{x}}_\mathrm{i}^{N_i,m}\}$ and $\bar{X}_i \in [lo_i, up_i]$. Numerically, $\bar{F}_i^m(\bar{X}_i)$ is estimated with the Simpson's quadrature rule for numerical integration. Let us denote $\{\bar{\mathbf{x}}^{k,m}\}_{k=1}^{n_m}$ the independent latin hypercube sample of size n_m. Then, the desired sample $\mathbf{x}$ of the original random variables are simply obtained by transforming the previous draws with Eq. (4.1). The f_{ij} functions are replaced by their polynomial approximations already computed in the decorrelation procedure. By repeating this procedure around all optima, one gets $N = \sum_{m=1}^{M} n_m$ draws of the x-variables sampled from the jpdf that was evaluated to get the MCPD draws. In the sequel, MCPD-MC refers to the MCPD-based Monte Carlo approach described in this section.

It has to be noted that, if a final sample of size N is sought, the sample size n_m for each optimum must respect the balance area of the modes, that is,

$$\frac{n_m}{N} = \frac{\mathcal{A}_m}{\sum_{k=1}^{M}\mathcal{A}_k}. \tag{4.4}$$

This ensures that each optimum is sampled in good proportion. For the sake of completeness, we show in Figure 4.1 the Monte Carlo draws obtained with the proposed approach applied to the MCPD draws depicted in Figure 3.1. Note that the MCPD-MC draws are randomly sampled with the desired correlation structure.

4.3. Discussion. Contrarily to MCMC, the MCPD sampler relies on the use of an optimizer. The computational effort of the algorithm resides in the optimization steps. In the numerical exercises below, the optimizations are performed with a gradient-based algorithm. The convergence of these algorithms is accelerated if the Jacobian matrix/vector of the target distribution is also provided. In our studies, the latter is systematically computed, unless the contrary is mentioned. As we underline above, the optimization processes can be distributed over several computers

(or parallel sessions). This advantage is also systematically exploited in the current studies.

In practice, it is not known beforehand whether the correlation structure of the input variables satisfies Eq. (4.1) (the strong assumption of the method). Thereby, it is recommended to evaluate the target distribution with the generated sample. Our experience suggests that, if for several candidates $\{\mathbf{x}^k\}_{k=1}^N$ of the MCPD-MC sample, the target distribution produces the value zero, then it is likely that the assumption is not valid.

Alternatively, if evidence yields to reject assumption Eq. (4.1), one can adopt a less restricted assumption by considering the following orthogonalization procedure,

$$\begin{cases} \bar{x}_1 &= x_{i_1} & & \\ \bar{x}_2 &= x_{i_2} & - & \mathbb{E}\left[(x_{i_2} - \mathbb{E}\left[x_{i_2}\right])\,|\bar{x}_1\right] \\ \bar{x}_3 &= x_{i_3} & - & \mathbb{E}\left[(x_{i_3} - \mathbb{E}\left[x_{i_3}\right])\,|\bar{x}_1, \bar{x}_2\right] \\ & \vdots & & \\ \bar{x}_d &= x_{i_d} & - & \mathbb{E}\left[(x_{i_d} - \mathbb{E}\left[x_{i_d}\right])\,|\bar{\boldsymbol{x}}_{\sim d}\right] \end{cases} \tag{4.5}$$

which relies on a dependence structure of the form,

$$\begin{cases} x_{i_1} = \bar{x}_1 \\ x_{i_2} = \bar{x}_2 + f_{i_2 1}(\bar{x}_1) \\ x_{i_3} = \bar{x}_3 + f_{i_3 1}(\bar{x}_1) + f_{i_3 2}(\bar{x}_2) + f_{i_3 12}(\bar{x}_1, \bar{x}_2) \\ \vdots \\ x_{i_d} = \bar{x}_d + \displaystyle\sum_{j_1=1}^{d-1} f_{i_d j_1}(\bar{x}_{j_1}) + \sum_{j_2>j_1}^{d-1} f_{i_d j_1 j_2}(\bar{x}_{j_1}, \bar{x}_{j_2}) + \cdots + f_{i_d 12\ldots d-1}(\bar{x}_1, \ldots, \bar{x}_{d-1}) \end{cases} \tag{4.6}$$

with still the orthogonality constraints on $f_{i_k j_1 j_2 \ldots j_s}(\bar{x}_{j_1}, \ldots, \bar{x}_{j_s})$ with $\{j_1, \ldots, j_s\} \subseteq \{1, \ldots, k-1\}$ in each equation to ensure its uniqueness. For the interested readers, the expansions in Eq. (4.6) are related to the analysis of variance decomposition [41]. The difficulty in this case is to approximate the multidimensional functions in the different equations. This can be achieved, for instance, with the non-parametric polynomial chaos expansion method of [40]. Adopting (4.6) has the drawback to slow down the MCPD-MC method as one has to build $(d-2)$ expansions (needless to perform such a complex expansion for x_{i_1} and x_{i_2}). This alternative approach is not considered in the numerical examples treated in § 5, § 6 and § 7 as assumption (4.1) is fulfilled.

4.4. The Algorithm. Given a MCPD sample $\left\{\mathbf{x}^{k_i,m}\right\}_{i=1}^d$, $\forall k_i = 1, \ldots, N_i$ and $m = 1, \ldots, M$, Monte Carlo samples of size N are obtained as follows,

1. Get the uncorrelated MCPD draws $\left\{\bar{\mathbf{x}}^{k_i,m}\right\}_{i=1}^d$ from Eq. (4.2) (or alternatively Eq. (4.6))
2. Generate independent random draws $\left\{\bar{\mathbf{x}}^{k,m}\right\}_{k=1}^{n_m}$ from Eq. (4.3) by paying attention to Eq. (4.4) in case $M > 1$
3. Get the desired Monte Carlo sample $\left\{\mathbf{x}^k\right\}_{k=1}^N$ by imposing the desired correlation structure Eq. (4.1) (or Eq. (4.6))
4. Evaluate the jpdf for each draws. If $p(\mathbf{x}^k) = 0$ for $k = 1, \ldots, n$ such that $n/N < 1\%$, then it is likely that Eq. (4.1) is not satisfied. In that case, repeat the algorithm by considering the orthogonalization procedure stemming from Eq. (4.6). Otherwise $\left\{\mathbf{x}^k\right\}_{k=1}^N$ is the desired Monte Carlo sample.

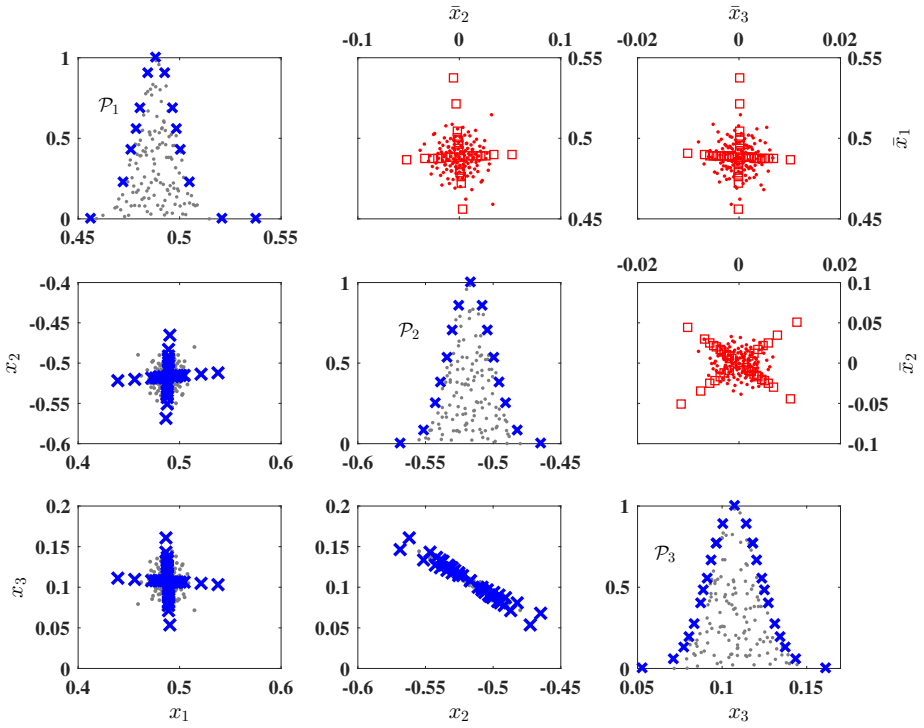

FIGURE 4.1. *Same as Figure 3.1 with the Monte Carlo draws obtained with the new approach. The upper off-diagonal plots show no correlation amongst the draws (red dots). The lower diagonal plots show the desired draws sampled from the target distributions (gray dots).*

Note that step 4 requires to evaluate the jpdf, and thus the likelihood function. This step might be computationally demanding. It is recommended to start with a small sample size N (say $N = 50$) and if the test is successful (i.e. $n/N > 1\%$), then increase the Monte Carlo sample size (which is computationally cheap). In all the following exercises, we find $n = 0$ and therefore this issue is not discussed much.

5. Numerical exercises.

5.1. Eleven-dimensional multimodal target distribution. We start with the following multimodal target distribution,

$$p(\boldsymbol{x}|\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_3, \mathbf{C}) = \frac{1}{6}\mathcal{N}(\boldsymbol{\mu}_1, 5\mathbf{C}) + \frac{2}{6}\mathcal{N}(\boldsymbol{\mu}_2, 5\mathbf{I}_d) + \frac{3}{6}\mathcal{N}(\boldsymbol{\mu}_3, 5\mathbf{I}_d) \tag{5.1}$$

where $\mathcal{N}(\boldsymbol{\mu}_i, 5\mathbf{I}_d)$ is the multiGaussian distribution of mean vector $\boldsymbol{\mu}_i$ and covariance $5\mathbf{I}_d$. $\mathbf{I}_d$ is the d-dimensional identity matrix which indicates that the parameters $(x_1, \ldots, x_d)$ are independent in the second and third Gaussian distributions in Eq. (5.1). In the present work, we consider the case $d = 11$ with the correlation matrix $\mathbf{C}$ having null off-diagonal elements except for $C_{1,2} = C_{2,1} = -0.5$ and $C_{1,3} = C_{3,1} = 0.8$. These non-null terms impose, for the first Gaussian distribution in Eq. (5.1), a negative correlation between x_1 and x_2 and a strong positive correlation between x_1 and x_3. The three modes of each variable are grouped in the vectors of means $\boldsymbol{\mu}_1 = (-5, -4, \ldots, 4, 5)^T$, $\boldsymbol{\mu}_2 = (1, 2, \ldots, 11)^T$ and $\boldsymbol{\mu}_3 = (11, 10, \ldots, 1)^T$.

The case $d = 11$ was studied in [21] in which the authors demonstrated that the MCPD draws were faster to generate than the MCMC draws (the case $d = 25$ was considered in [22]). Indeed, the MCPD sampler only required around one thousand evaluations of $p(\boldsymbol{x}|\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \mu_3, \boldsymbol{C})$ to obtain an accurate estimate of the variables' distributions. Thanks to the parallel computation, the computational time units (CTU) were about 190. This means that actually the waiting time corresponded to 190 model calls. This is very short compared to the CTU of the DREAM$_{ZS}$ sampler which was 10 000 (with 11 chains in parallel). The obtained MCPD sample is considered hereafter for generating the MCPD-MC sample.

Eq. (5.1) has been assessed with the MCPD-MC draws. The results are depicted in Figure 5.1 for a sample of size $N = 4\,096$. On the diagonal, the MCPD estimate of x_i (row #i, column # i) and the Monte Carlo draws $({\rm x_i}^{k,m}, p(\mathbf{x}^{k,m}))$, $\forall k = 1, \ldots, N$ and $m = 1, 2, 3$ are reported. Note that the MCPD-MC draws are located below the MCPD curves. This is because the MCPD-MC approach samples probable solutions while the MCPD sampler draws the most probable variable set for prescribed values of one of them.

The lower off-diagonal scatterplots show the pairwise correlations of MCPD-MC draws. The estimated Pearson correlation coefficients for the first mode are also reported. They are close to the analytical correlation coefficients, $r_{12}^{(1)} \approx C_{1,2} = -0.5$ and $r_{13}^{(1)} \approx C_{1,3} = 0.8$. The upper off-diagonal plots compare the empirical densities with the analytical marginal densities. It can be inferred that each mode has been sampled in good proportion.

5.2. Ten-dimensional twisted Gaussian target distribution. Let us now consider another challenging problem which was also analyzed in [21]. We target the twisted Gaussian density proposed in [13] and defined as follows,

$$p(\boldsymbol{x}) = p(x_1)p(x_2|x_1)\prod_{i=3}^{10} p(x_i)$$

with $p(x_1) = \mathcal{N}(0, 100)$, $p(x_2|x_1) = \mathcal{N}(-0.1x_1^2 + 10, 1)$ and $p(x_i) = \mathcal{N}(0, 1), \forall i = 3, \ldots, 10$. This jpdf is very challenging because it exhibits a non-monotonic relationship between x_2 and x_1. Indeed, the conditional expectation of x_2 on x_1 is $\mathbb{E}[x_2|x_1] = -0.1x_1^2 + 10$.

Generating the MCPD draws required approximately 1 900 evaluations of $p(\boldsymbol{x})$ while DREAM$_{ZS}$ required at least 20 000 evaluations to converge (see the aforementioned paper for more details). We can notice that the relationship between $\mathbb{E}[x_1|x_2]$ cannot be written in the form of $f_{12}(x_2)$ since for one value of x_2 two values of x_1 are possible. Hence, Eq. (4.1) is not valid for the variable set $(x_2, x_1, \ldots, x_{10})$ so ordered, but it is for the set $(x_1, x_2, x_3, \ldots, x_{10})$. Indeed, it is possible to write $\mathbb{E}[x_2|x_1] = f_{21}(x_1)$ with f_{21} the non-monotonic function mentioned above. Consequently, the transformation in Eq. (4.2) starts with $\bar{x}_1 = x_1$, then $\bar{x}_2 = x_2 - \mathbb{E}[(x_2 - \mathbb{E}[x_2])|\bar{x}_1]$, $\bar{x}_3 = x_3 - \mathbb{E}[(x_3 - \mathbb{E}[x_3])|\bar{x}_1] - \mathbb{E}[(x_3 - \mathbb{E}[x_3])|\bar{x}_2]$ and so on.

The results are depicted in Figure 5.2 for the first three variables. Note that, on the first diagonal plot (row #1, column #1) the MCPD and MCPD-MC draws of x_1 are depicted, while on row #2 and column #2 the draws of x_2 are represented. The MCPD-MC draws are located beneath the MCPD curves. We note that the uncertainty ranges of x_1 and x_2 are very large which is not an issue for the MCPD sampler. The upper off-diagonal plots prove that the MCPD draws have been successfully orthogonalized (they are actually independent). The lower off-diagonal scatterplot of

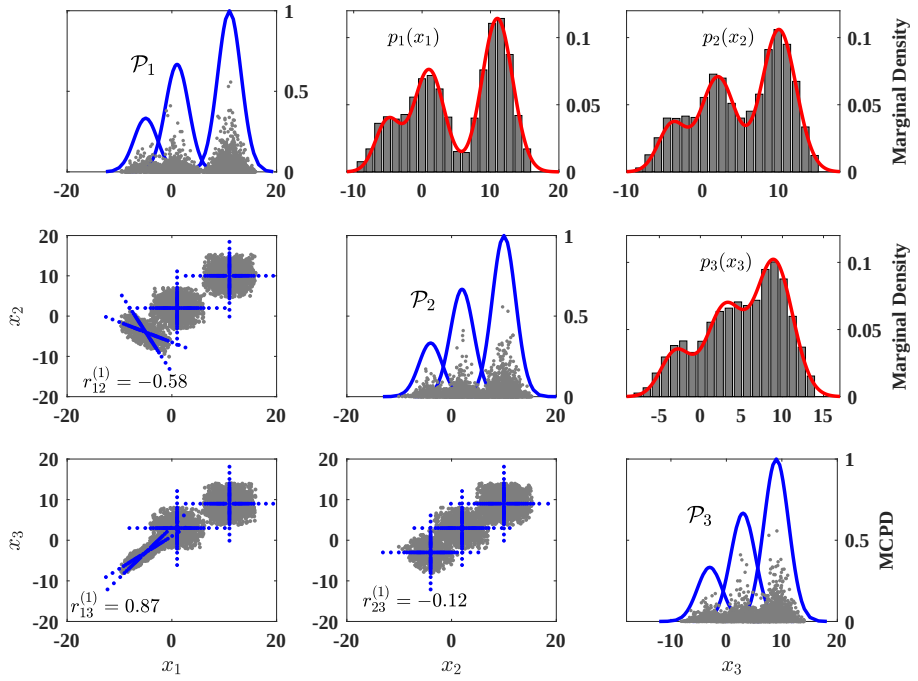

FIGURE 5.1. *Sampling from the multimodal target distribution. Results of first three variables for a sample of size $N = 4\,096$ are depicted. The on-diagonal plots represent the MCPDs estimate (blue line) as well as the MCPD-MC draws. The lower off-diagonal scatterplots represent the pairwise correlation of the MCPD draws (blue crosses) and the MCPD-MC draws (gray dots). $r_{ij}^{(1)}$ is the estimated pairwise correlation coefficient of x_i and x_j in the vicinity of the first mode. The upper off-diagonal confirms that the MCPD-MC sample of each variable fits its analytical marginal density (red curve).*

x_2 versus x_1 (row #2, column # 1) shows that, because of their strong dependence, the MCPD-MC draws (gray dots) are located close to the MCPD draws (blue dots).

6. Calibration of a drainage model.

6.1. The model and the dataset. The MCPD-MC approach proposed in the present work has been specifically developed to address the issue of model calibration in a Bayesian framework. In this section, we consider the calibration of a soil drainage model at the laboratory scale. Once again, this problem was also studied in [21] to compare the performance of the MCPD approach with the one of the MCMC sampler. Here, we repeat the same numerical experiment in order to illustrate the MCPD-MC approach.

We model a laboratory multistep outflow drainage experiment in which a column of length $L = 6$ cm and diameter $D = 8.5$ cm is filled with sand and initially saturated with water. The column is drained by imposing at the lower boundary of the column multistep prescribed negative pressure heads. Flow in the column is modelled by the

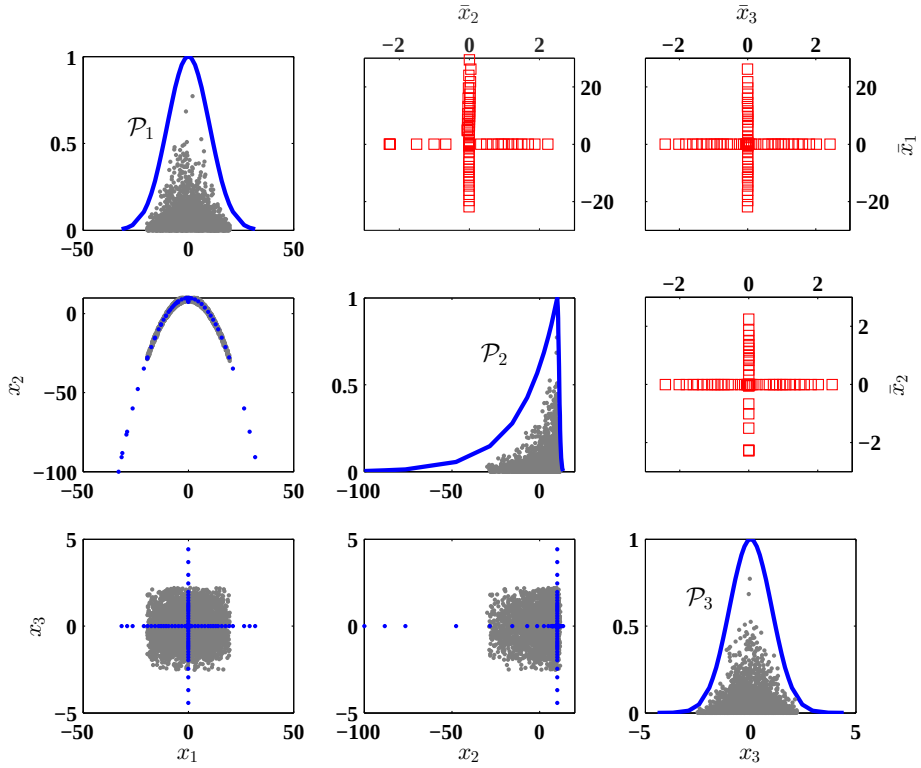

FIGURE 5.2. *Sampling from the banana-shaped distribution. Results of first three variables for a sample of size $N = 4\,096$ are depicted.*

non-linear one-dimensional Richard's equation,

$$\frac{\partial \theta}{\partial t} = \frac{\partial}{\partial z}\left[K(h)\left(\frac{\partial h}{\partial z} - 1\right)\right], \tag{6.1}$$

in conjunction with the Mualem-van Genuchten (MvG) retention curve [29, 45],

$$K(S_e) = k_s \cdot S_e^{\lambda}\left(1 - \left(1 - S_e^{1/m}\right)^m\right)^2, \tag{6.2}$$

where t (min) is time, z (cm) is the vertical coordinate (positive downward) and $m = 1 - 1/n$. The water content θ ($\mathrm{cm}^3.\mathrm{cm}^{-3}$) and the pressure head h (cm^{-1}) are the state variables. The effective saturation S_e is defined as follows,

$$S_e = \frac{\theta - \theta_r}{\theta_s - \theta_r} = \begin{cases} \dfrac{1}{|\alpha h|^n} & h < -1/\alpha \\ 1 & h \geq -1/\alpha \end{cases}. \tag{6.3}$$

and K ($\mathrm{cm}^3.\mathrm{min}^{-1}$) is the unsaturated hydraulic conductivity.

The soil hydraulic parameters to be estimated are: k_s ($\mathrm{cm}^3.\mathrm{min}^{-1}$) the saturated hydraulic conductivity, θ_s ($\mathrm{cm}^3.\mathrm{cm}^{-3}$) the saturated water content, θ_r ($\mathrm{cm}^3.\mathrm{cm}^{-3}$) the residual water content and the MvG fitting coefficients α (cm^{-1}), n ($-$) and λ ($-$).

Eqs.(6.1-6.3) are solved with a standard Galerkin finite element method in conjunction with the Newton linearization method. Synthetic data were obtained by running the numerical model for a given input parameter set and noising the model responses with independent Gaussian random noises. The responses of interest are

the pressure head h and the soil water content θ at the center of the column. The inverse modelling is performed from the *observed* pressure head and water content denoted by $\mathbf{y}_h$ and $\mathbf{y}_\theta$ respectively.

Setting the calibration problem in a Bayesian framework yields the following posterior jpdf,

$$p(\boldsymbol{x}, \sigma_h, \sigma_\theta | \mathbf{y}_h, \mathbf{y}_\theta) \propto \frac{1}{\sigma_h^{N_h} \sigma_\theta^{N_\theta}} \exp\left\{-\frac{1}{2}\left(\frac{SS_h(\boldsymbol{x})}{\sigma_h^2} + \frac{SS_\theta(\boldsymbol{x})}{\sigma_\theta^2}\right)\right\}, \tag{6.4}$$

where N_h and N_θ are the number of observed data of pressure and water content respectively and SS_h and SS_θ are the sum of square errors of pressure head and water content respectively. This posterior jpdf was derived by assuming normal errors for h and θ and independent uniform priors for the hydraulic parameters within large plausible ranges (see [21] for more details).

Eight unknowns were sought in the inverse problem, including the vector of hydraulic parameters $\boldsymbol{x} = (k_s, \theta_r, \theta_s, \alpha, n, \lambda)$ as well as the error variances σ_h^2 (cm^2), σ_θ^2 ($\mathrm{cm}^6.\mathrm{cm}^{-6}$). The different maximization processes were performed with the Levenberg-Marquardt algorithm [20, 25] as the partial derivatives of the model responses w.r.t. the hydraulic parameters were computed by the numerical model.

The performance of the MCPD sampler was discussed in [21] and a comparison with the DREAM$_{ZS}$ MCMC sampler of [18] was also carried out. We recall that the MCPD sampler required an overall of 7 500 model calls. But thanks to the parallelization, the computational time unit (related to the real waiting time) was about 2 000 model calls which corresponded to the assessment of θ_s's MCPD. An MCPD sample of size 185 was obtained. DREAM$_{ZS}$ required 64 000 model calls but with eight chains in parallel the computational time unit was 8 000. In the next subsections, we discuss the MCPD-MC sampling of the calibrated model parameters and the posterior predictive uncertainty of the model responses. Of particular interest is the posterior uncertainty of the predicted cumulative outflow at the end of the simulated experiment. In effect, the cumulative outflow was not considered as a measurement in the calibration problem.

6.2. Parameter uncertainty quantification. The MCPDs of the hydraulic parameters as well as those of the likelihood hyperparameters are depicted in Figure 6.1 (the on-diagonal plots). They show bell-shaped parameters' posterior marginal pdf with a clear optimal value. The support of the MCPDs are quite narrow except for the MvG parameter λ. This might be explained by a lack of sensitivity of $h(t)$ and $\theta(t)$ to this parameter. From the 185 MCPD draws, a MCPD-MC sample of size 512 was generated. They are plotted in Figure 6.1 (the scatterplots).

The MCPD-MC draws were then propagated through the model (Eq. (6.1)) in order to estimate the probability assigned to each draw (Eq. (6.4)). They are plotted with the MCPD curves (the on-diagonal scatterplots in Figure 6.1). Note that after evaluating Eq. (6.4), the probability value of each draw was scaled between [0,1] by dividing by the probability value of the MAP. The MCPD-MC draws are assigned high probability values, proving that they stem from the region of high probability of the posterior parameter space. Actually, none of the MCPD-MC draw was assigned a probability zero which is an indication of the reliability of the generated draws as explained in Step 4 of the algorithm defined in § 4.4.

Pairwise scatterplots of the MCPD-MC draws are depicted on the lower off-diagonal. They show either strong linear or strong non-linear correlations between

the hydraulic parameters, which is a sign of the overparameterization of the calibration problem. This was already noted in [21]. Indeed, k_s is positively correlated to $(\theta_s, \alpha, \lambda)$ and negatively correlated to (θ_r, n). This means that, for instance, an increase of k_s-value would cause a deviation of the predicted state variables from the observations that could be compensated by increasing $(\theta_s, \alpha, \lambda)$ and simultaneously decreasing (θ_r, n). The error variances are not correlated with the hydraulic parameters. Given that for this problem, Eq. (6.4) admits one single optimum, in the sequel the reference to the superscript m in the MCPD draws is dropped (see Section 3).

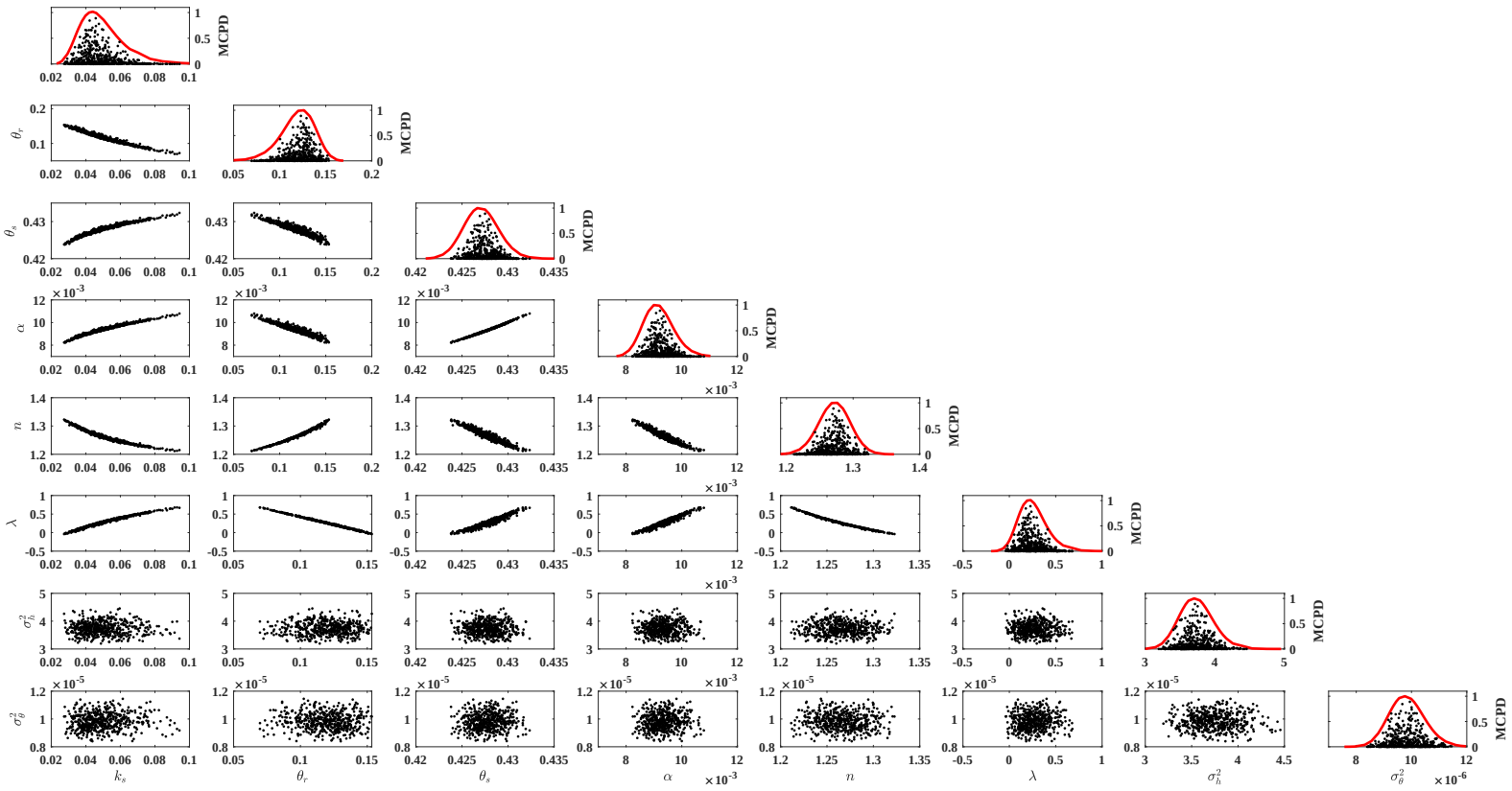

FIGURE 6.1. *The generated MCPD-MC draws of the numerical drainage experiment. On the diagonal, the red curves are the MCPDs, the black dots represent the MCPD-MC draws.*

6.3. Predictive uncertainty of the observed variables. Except the fact that the MCPD-MC draws are assigned high probabilities, so far, there is no clue that they are really sampled from the target posterior jpdf. One way to check this assumption is to assess the 95% credible intervals assigned to the *observed* state variables with the MCPD sample. In this way, one can compare the predictive uncertainty obtained with the 95% credible intervals evaluated with the few MCPD draws. Indeed, as explained in [21], it is possible to assess the posterior uncertainty of the *observed* model responses because, for the latter, the likelihood functions have been defined before calibration. Let us denote by y^*_{out} a new observation of either the pressure head ($out = h$) or the water content ($out = \theta$), then the posterior pdf of y^*_{out} reads,

$$\hat{p}(y^*_{out}|\mathbf{y}_h, \mathbf{y}_\theta) = \frac{\sum_{i=1}^{d}\sum_{k_i=1}^{N_i} \mathcal{P}_i(\mathbf{x}^{k_i}) p(y^*_{out}|\mathbf{y}_h, \mathbf{y}_\theta, \mathbf{x}^{k_i}, \sigma_h^{k_i}, \sigma_\theta^{k_i})}{\sum_{i=1}^{d}\sum_{k_i=1}^{N_i} \mathcal{P}_i(\mathbf{x}^{k_i})} \tag{6.5}$$

where one recognizes $\mathcal{P}_i(\mathbf{x}^{k_i})$ the probability assigned to the MCPD draw $\mathbf{x}^{k_i}$ (i.e. the posterior jpdf Eq. (6.4) evaluated at $\mathbf{x}^{k_i}$). The likelihood function at y^*_{out} evaluated at the k_i-th MCPD draw is denoted by $p(y^*_{out}|\mathbf{y}_h, \mathbf{y}_\theta, \mathbf{x}^{k_i}, \sigma_h^{k_i}, \sigma_\theta^{k_i})$. These two quantities are obtained by evaluating the model responses of interest at each MCPD draw $\mathbf{x}^{k_i}$, $k_i = 1, \ldots, N_i$, $i = 1, \ldots, d$.

The two predicted 95% credible intervals are represented on Figure 6.2. We can note that they match very well. Furthermore, they encompass most of the observations. In our viewpoint, this is a clear indication that the MCPD-MC sample was generated with respect to the posterior jpdf. We recall that in [21], MCPD and MCMC yielded similar predicted uncertainties for this calibration problem.

6.4. Posterior uncertainty analysis of the cumulative outflow. In subsurface hydrology, the cumulative outflow prediction is of high importance to predict water recharge of aquifers. While it is an issue to measure in-situ, this quantity is easily measured in laboratory experiment. In the present study, it corresponds to the amount of water exiting the column per unit surface in the end of the drainage experiment. The cumulative outflow is computed as follows,

$$C_s = \int_0^L (\theta(z, 0) - \theta(z, t_f))\, \mathrm{d}z,$$

where $t_f = 240$ min is the duration of the drainage experiment. Predicting the uncertainty of this model response directly with the 185 MCPD draws is not possible because it requires the knowledge of the likelihood of this variable. The uncertainty of cumulative outflow a posteriori (i.e. after calibration) can be assessed by propagating the MCPD-MC sample through this model response variable.

The predicted posterior density of the final cumulative outflow is depicted in Figure 6.3. It shows a very narrow support with $C_s \in [0.69, 0.75]$ cm with a mode around 0.71 cm. This indicates that, as far as the prediction of the cumulative outflow is concerned, the hydraulic parameters have been satisfactorily calibrated. On the opposite, a larger support of the predicted posterior density would indicate a poorly calibrated model. In that case, a sensitivity analysis would help identifying which input variables were responsible for the lack of accuracy [37]. It is worth mentioning that sensitivity analysis can be easily carried out with the MCPD-MC draws at hand. For this purpose, one can use the method developed in [23] which is valid under the assumption represented by Eq. (4.1). Although in the present case such an analysis is

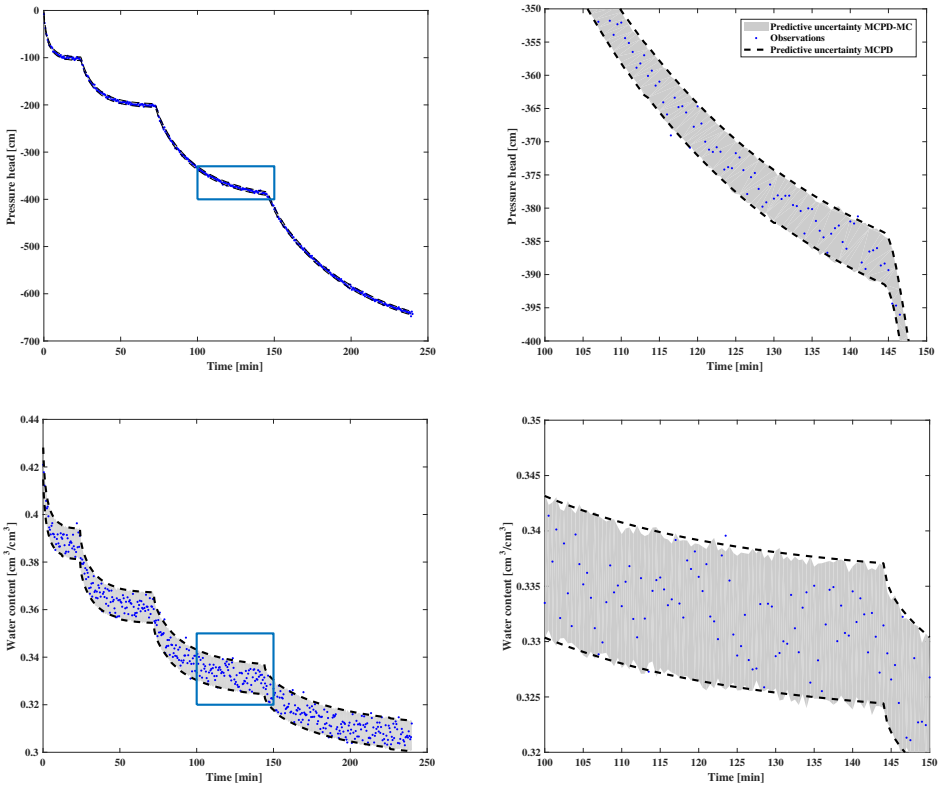

FIGURE 6.2. *The dots represent the data used for soil hydraulic parameter identification. The predictive uncertainty ranges (95% credible interval) estimated with the MCPD and MCPD-MC samples match satisfactorily well, confirming that the MCPD-MC draws are reliable.*

not necessary as the cumulative outflow is predicted accurately, we have undertaken a sensitivity analysis exercise for the sake of completeness.

The sensitivity analysis setting addressed in this exercise is the following: *What is the smallest subset of hydraulic parameters* $\boldsymbol{x_1} \subset \boldsymbol{x} = (k_s, \theta_r, \theta_s, \alpha, n, \lambda)$ *that mostly explained the predicted variance of* C_s*?* To answer this question, we consider the following Sobol' index [41]:

$$S_{\boldsymbol{x_1}}^{closed} = \frac{\mathbb{V}_{\boldsymbol{x_1}}\left[\mathbb{E}_{\boldsymbol{x_2}|\boldsymbol{x_1}}\left[C_s|\boldsymbol{x_1}\right]\right]}{\mathbb{V}_{\boldsymbol{x}}\left[C_s\right]} \tag{6.6}$$

where $\mathbb{V}$ is the variance operator and $\boldsymbol{x} = (\boldsymbol{x_1}, \boldsymbol{x_2})$. We investigate the smallest subset $\boldsymbol{x}_1$ with the highest Sobol' index and such that $S_{\boldsymbol{x_1}}^{closed} > 0.99$ which means that $\boldsymbol{x_1}$ explains more than 99% of the variance of C_s. From the MCPD-MC sample of size 512 already generated, the Sobol' indices of different subsets of various cardinality were computed with the method introduced in [23]. Obviously several subsets of equal cardinality that satisfied $S_{\boldsymbol{x_1}}^{closed} > 0.99$ were found. But the one with the highest Sobol' index was attributed to $\boldsymbol{x_1} = (k_s, \theta_r, n)$. This result indicate if narrower predictive uncertainty is required for the cumulative outflow, one should devote further effort to reduce the uncertainty in the input set $\boldsymbol{x_1} = (k_s, \theta_r, n)$.

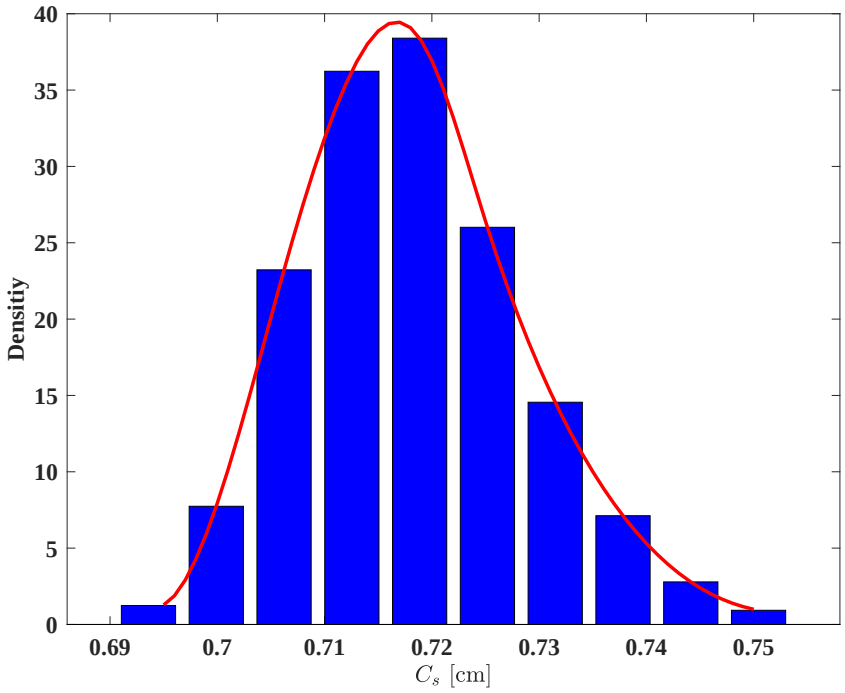

FIGURE 6.3. *Predictive uncertainty of the cumulative outflow in the end of the experiment.*

7. Calibration of a reactive transport model.

7.1. Problem setting. We consider the field observations of Valocchi et al. [44] that was used as a simulation benchmark by several authors ([48, 49, 6, 7]). The problem is a field experiment involving nonlinear ion exchange reactions in Palo Alto Baylands, California, where a multi-ion solution was injected into a well. For aqueous components, namely Mg^{2+}, Ca^{2+}, Na^{+} and Cl^{-}, were monitored at an observation well (named S23) located 16 m from the injection well. In this work, we only focus on the reactive species Mg^{2+} and Ca^{2+}. The experiment is modelled by a set of one-dimensional advective, dispersive and reactive transport equations,

$$\frac{\partial C_{sp}}{\partial t} = -U\frac{\partial C_{sp}}{\partial x} + D\frac{\partial^2 C_{sp}}{\partial x^2} + R_{sp}(x,t) \tag{7.1}$$

where C_{sp} (mmol.l^{-1}) is the concentration of cation $sp \in (Mg^{2+}, Ca^{2+}, Na^{+})$, U (m.h^{-1}) is the groundwater velocity, D (m^2.h^{-1}) is the dispersion coefficient of the homogeneous medium and R_{sp} (mmol.l^{-1}.h^{-1}) is a sink/source rate of cation due to chemical reactions. The different chemical reactions and associated log-constants of reactions are,

$Na^{+} + S^{-} \rightleftharpoons NaS, \quad \log(K_{Na})$
$Ca^{2+} + 2S^{-} \rightleftharpoons CaS_2, \quad \log(K_{Ca})$
$Mg^{2+} + 2S^{-} \rightleftharpoons MgS_2, \quad \log(K_{Mg})$

The chemical reactions are parameterized by two selectivity coefficients as follows, $\log(K_{CN}) = \log(K_{Ca}) - 2\log(K_{Na})$

$\log(K_{MN}) = \log(K_{Mg}) - 2\log(K_{Na})$
with $\log(K_{Na}) = 4.$ and S^- the surface charge of the soil. This set of equations is completed by initial conditions and is solved numerically with the sequential non-iterative approach [6]. The domain of length 16 m is discretized with a uniform mesh of 400 elements. The simulated time $T = 2000$ h is discretized with a uniform time step $\Delta t = 0.2$ h.

The model is parameterized by the vector of input parameters $\boldsymbol{x} = (\log(K_{CN}), \log(K_{MN}), T_f, U, D)$, where T_f (-) is the total fixed concentration. The observed data are the concentrations of Ca^{2+} and Mg^{2+} (mmol/l). We assume log-normal likelihood for each concentration data and independent uniform prior uncertainty ranges for $\boldsymbol{x}$. This leads to the following target jpdf,

$$p(\boldsymbol{x}, \sigma_{Ca}, \sigma_{Mg}|\mathbf{y}_{Ca}, \mathbf{y}_{Mg}) \propto \frac{1}{\sigma_{Ca}^{N_{Ca}} \sigma_{CMg}^{N_{Mg}}} \exp\left\{-\frac{1}{2}\left(\frac{SS_{Ca}(\boldsymbol{x})}{\sigma_{Ca}^2} + \frac{SS_{Mg}(\boldsymbol{x})}{\sigma_{Mg}^2}\right)\right\}, \quad (7.2)$$

which has the same form as Eq. (6.4) except that $\mathbf{y}_{Ca}$ and $\mathbf{y}_{Mg}$ are the vectors of observed log-concentrations while SS_{Ca} and SS_{Mg} are the sum of squares of the differences between predicted and observed log-concentration of cations.

7.2. Uncertainty quantification. The statistical calibration of the transport model involves seven variables, namely the model parameters $\boldsymbol{x} = (\log(K_{CN}), \log(K_{MN}), T_f, U, D)$, and the hyperparameters $(\sigma_{Ca}^2, \sigma_{Mg}^2)$ of the likelihood function. Their MCPD estimates are depicted in Figure 7.1. The optimization algorithm found a single set of optimum values which is, $\log(K_{CN}^{opt,1}) \simeq 8.69$, $\log(K_{MN}^{opt,1}) \simeq 8.43$, $T_f^{opt,1} \simeq 735.64$, $U^{opt,1} \simeq 0.74$, $D^{opt,1} \simeq 4.35$, $\sigma_{Ca}^{opt,1} \simeq 0.09$ and $\sigma_{Mg}^{opt,1} \simeq 0.14$. Indeed, the MCPD curves are unimodal (Figure 7.1, on the diagonal), rather Gaussian for the parameters in vector $(\log(K_{CN}), \log(K_{MN}), T_f)$ and skewed for those in $(U, D, \sigma_{Ca}^2, \sigma_{Mg}^2)$. The parameters in the former subset are well identified despite of the strong collinearity between $(\log(K_{CN}), \log(K_{MN}))$ (row #2, column #1). Those in the second subset have large posterior uncertainty ranges with a negative strong correlation between (U, D) (row #5, column #4). The hyperparameters are not correlated to the parameters (see the last two rows).

From the MCPD sample, a MCPD-MC sample of size $N = 512$ was generated. To check the validity of Eq. (4.1), for each MCPD-MC draw the predicted log-concentrations of Ca^{2+} and Mg^{2+} were collected after running the transport model. Then, with the collection of model responses the jpdf in Eq. (7.2) was evaluated. Therefore, each MCPD-MC draw is assigned a probability value which has been scaled properly within $[0, 1]$ and plotted on Figure 7.1 (dot plots). On the diagonal plots, we notice that the MCPD-MC draws are assigned high probability values. The dot plots filled the space between the MCPD curves and the x-axis. None of the MCPD-MC draw was assigned a probability zero. This is an indication that the MCPD-MC is successfully sampled from the desired jpdf. This is confirmed by the comparison of the predictive uncertainties obtained with the MCPD draws (by using Eq. (6.5)) and the Monte Carlo sample respectively (see Figure 7.2). Notably, the predictive uncertainties are narrow and encompass the observations. This indicates that the choice of the log-normal likelihood is reasonable.

In our study, cations Na^+ were not involved in the calibration process, only Ca^{2+} and Mg^{2+} were. Hence, it can be questioned whether the calibration of the model from the concentration measurements of Ca^{2+} and Mg^{2+} are sufficient to estimate the total mass of Na^+ (denoted M_{Na}) accurately. A Monte Carlo sample of size $N = 512$

was obtained by propagating the MCPD-MC draws through the model response M_{Na}. The estimated pdf is depicted in Figure 7.3. The support of the latter is large, ranging from 10 to 19 mol.m^{-2} with a mode around 16 mol.m^{-2}.

Identifying those uncertain input variables responsible for the large uncertainty in the model response is the role of sensitivity analysis. In the present analysis, we quantify the amount of the variance of M_{Na} explained by each individual input parameter. This statistic is called the Pearson correlation ratio or the first-order Sobol' index and is defined as follows [32, 42],

$$S_{x_i} = \frac{\mathbb{V}_{x_i}\left[\mathbb{E}_{\boldsymbol{x}_{\sim \boldsymbol{i}}|x_i}\left[M_{Na}|x_i\right]\right]}{\mathbb{V}_{\boldsymbol{x}}\left[M_{Na}\right]} \tag{7.3}$$

which is the same as Eq. (6.6) but for individual variable x_i such that $(x_i, \boldsymbol{x}_{\sim \boldsymbol{i}}) = \boldsymbol{x}$. Using the sensitivity analysis method of Mara & Tarantola [23] as previously, we find: $S_{K_{CN}} \simeq 0.74$, $S_{K_{MN}} \simeq 0.73$, $S_{T_f} \simeq 0.18$, $S_U \simeq 1$ and $S_D \simeq 0.99$. These results simply claim that the large uncertainty in M_{Na} prediction is due to the inaccurate assessment of the groundwater velocity U. The high sensitivity index of D is merely due to its strong correlation with U (see row #5, column #4 in Figure 7.1) which was poorly estimated from the measurements of Ca^{2+} and Mg^{2+} concentrations. Including measurements of chloride concentrations in the calibration dataset did not improve much the flow velocity estimate even though Cl^- is a tracer (not shown).

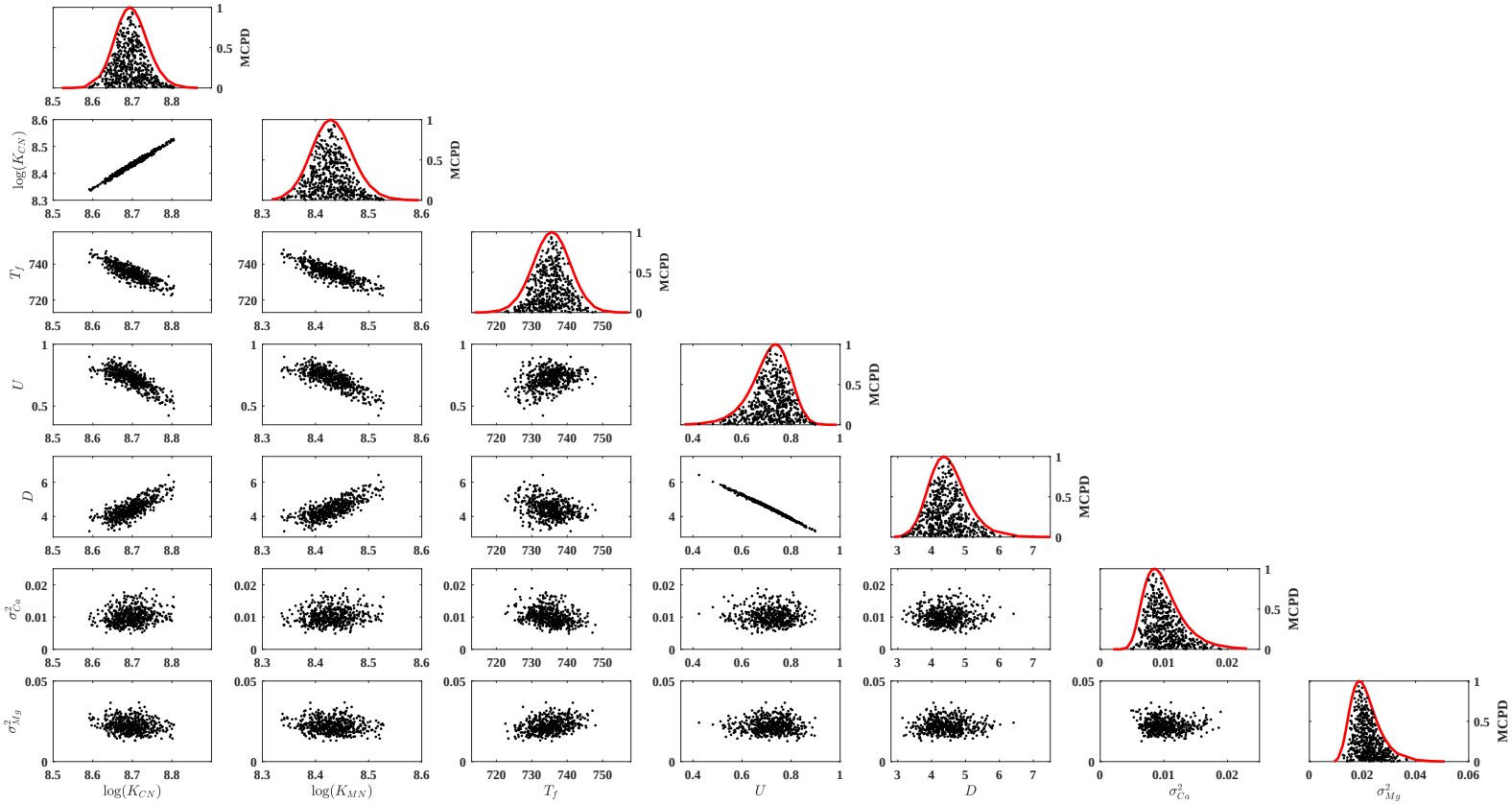

FIGURE 7.1. *The generated MCPD-MC draws of the reactive transport experiment. On the diagonal, the red curves are the MCPDs, the black dots represent the MCPD-MC draws.*

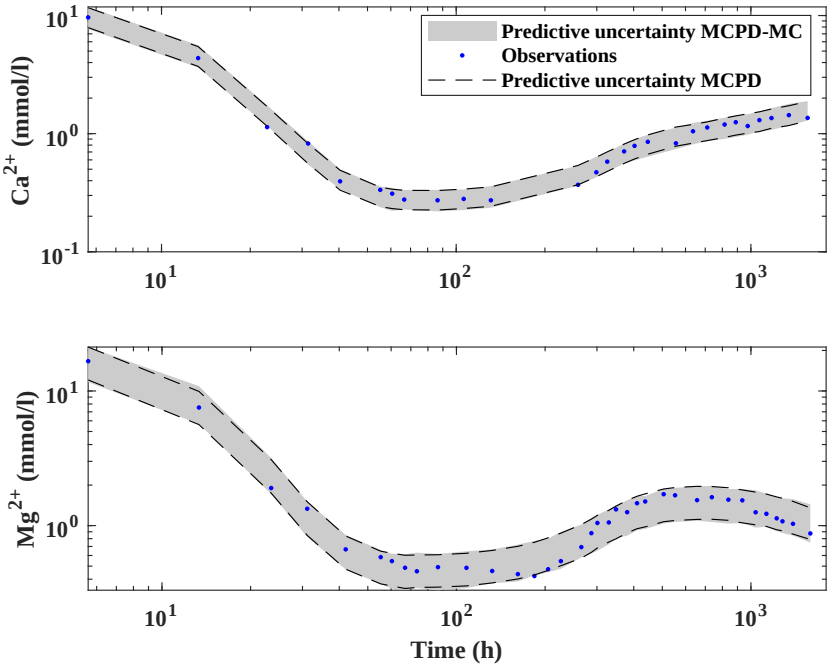

FIGURE 7.2. *Breakthrough curves of cations in the reactive transport experiment. The predictive uncertainty ranges (95% credible interval) estimated with the MCPD and MCPD-MC samples are comparable, confirming that the MCPD-MC draws are reliable.*

8. Conclusion. In this paper, we have extended the maximal conditional posterior distribution sampling to generate Monte Carlo draws from a joint probability distribution function. The proposed approach requires that the correlation structure amongst the random variables be such that Eq. (4.1) is valid at least for one possible ordering of the random variables in the set $(x_1, \ldots, x_d)$. Generating Monte Carlo samples is essential, for instance, to perform uncertainty and sensitivity analyses of model responses. This is illustrated in our work.

The numerical exercises have shown that the proposed approach, named MCPD-MC sampler, was able to sample from distribution functions with possibly multimodality and complex correlation structure. The MCPD-MC approach relies on the orthogonalization procedure introduced in [23] that produces independent samples under assumption (4.1). Fast generation of random samples from the MCPD draws is performed without evaluating the jpdf further. However, evaluation of the jpdf with each generated draw is necessary to verify that the latter is located in the region of high probability. This verification step is required to check the validity of Eq. (4.1).

Applications to model calibration have demonstrated the interest of the proposed approach. With the MCPD-MC sample, uncertainty and sensitivity analyses of the model responses that were not used in the calibration step could be performed. Furthermore, sensitivity analysis can be easily carried out with the method developed in [23] since the latter relies on the same assumption as the MCPD-MC sampler. However, it has to be underlined that Eq. (4.1) may not be satisfied in some situations. Therefore, great care must be taken when applying the MCPD-MC approach.

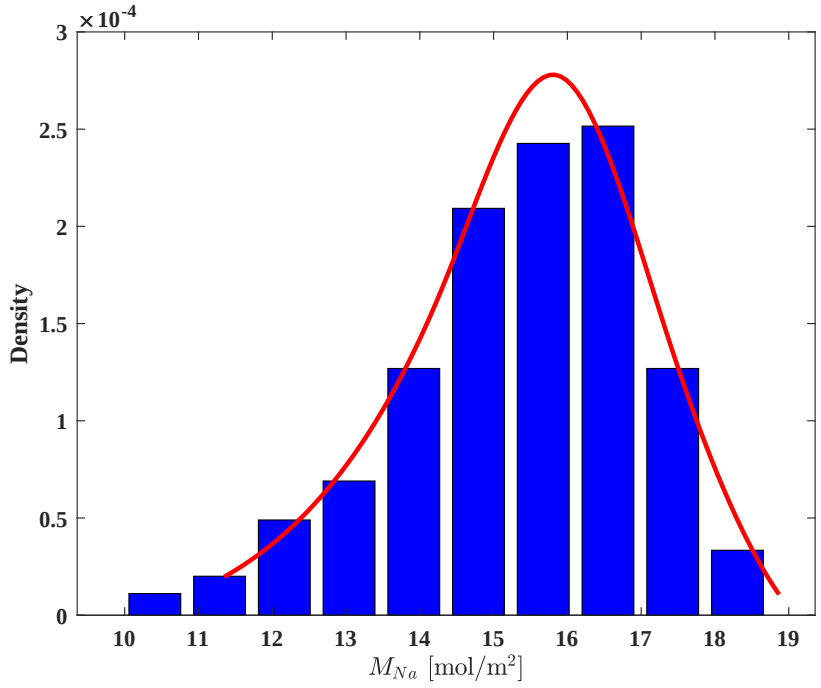

FIGURE 7.3. *Predictive uncertainty of the total mass of Na^+ involved in the chemical reactions during the flow transport experiment.*

Otherwise, another alternative (more expensive) relying on Eq. (4.6) can be employed.

It has to be underlined that the current version of the MCPD-MC sampler for statistical model calibration only handles Gaussian likelihood functions. Current development of the sampler aims at including a less restricted likelihood function by accounting for the Generalized Error Distribution (also called Exponential Power distribution [2]).

Acknowledgements. The MCPD-MC sampler developed in MATLAB is available upon request from the corresponding author (mara@univ-reunion.fr). The authors are grateful to the anonymous referees for their insightful comments that helped improving the manuscript. Qian SHAO acknowledges the support from the National Science Foundation of China (Grant No. 11702199).

References.

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 (2007), pp. 138–154.

[2] G. E. P. Box and G. C. Tiao, *Bayesian inference in statistical analysis*, Wiley and Sons, New York, 1992.

[3] K. Campbell, *Statistical calibration of computer simulations*, Reliability Engineering and System Safety, 91 (2006), pp. 1358–1363.

[4] T. Cui, C. Fox, and M. A. O'Sullivan, *Bayesian calibration of a large-scale geothermal reservoir model by a new adaptive delayed acceptance Metropolis-Hastings algorithm*, Water Resources Research, 47 (2011), p. W10521.

[5] P. Dostert, Y. Efendiev, and B. Mohanty, *Efficient uncertainty quantification techniques in inverse problems for Richards' equation using coarse-scale simulation models*, Advances in Water Resources, 32 (2009), pp. 329–339.

[6] M. Fahs, J. Carrayrou, A. Younes, and P. Ackerer, *On the efficiency of the direct substitution approach for reactive transport problems in porous media*, Water, Air and Soil Pollution, 193 (2008), pp. 299–308.

[7] M. Fahs, A. Younes, and P. Ackerer, *An efficient implementation of the Method of Lines for multicomponent reactive transport equations*, Water, Air and Soil Pollution, 215 (2011), pp. 273–283.

[8] A. E. Gelfand and A. F. Smith, *Sampling-based approaches to calculating marginal densities*, Journal of the American Statistical Association, 85 (1990), pp. 398–409.

[9] P. J. Green and A. Mira, *Delayed rejection in reversible jump Metropolis-Hastings*, Biometrika, 88 (2001), pp. 1035–1053.

[10] U. Grenander and M. Miller, *Representations of knowledge in complex system*, Journal of the Royal Statistical Society, Serie B, 56 (1994), pp. 549–603.

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

[12] H. Haario, Saksman, and J. Tamminen, *An adaptive Metropolis algorithm*, Bernouilli, 7 (2001), pp. 223–242.

[13] H. Haario, E. Saksman, and J. Tamminen, *Adaptive proposal distribution for random walk Metropolis algorithm*, Computational Statistics, 14 (1999), pp. 377–395.

[14] T. J. Hastie and R. J. Tibshirani, *Generalized additive models*, Chapman & Hall, London, 1990.

[15] H. Hastings, *Monte Carlo sampling methods using Markov chains and their applications*, Biometrika, 57 (1970), pp. 97–109.

[16] M. D. Hoffman and A. Gelman, *The No-U-Turn sampler: Adaptatively setting path lengths in Hamiltonian Monte Carlo*, Journal of Machine Learning Research, 15 (2014), pp. 1593–1623.

[17] S. Kucherenko, S. Tarantola, and P. Annoni, *Estimation of global sensitivity indices for models with dependent variables*, Computer Physics Communications, 183 (2012), pp. 937–946.

[18] E. Laloy, B. Rogiers, J. A. Vrugt, D. Mallants, and D. Jacques, *Efficient posterior exploration of a high-dimensional groundwater model from two-stage markov chain Monte Carlo simulation and polynomial chaos expansion*, Water Resources Research, 49 (2013), pp. 2664–2682.

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

[20] K. Levenberg, *A method for the solution of certain non-linear problems in least squares*, The Quarterly of Applied Mathematics, 2 (1944), pp. 164–168.

[21] T. A. Mara, F. Delay, F. Lehmann, and A. Younes, *A comparison of two Bayesian approaches for uncertainty quantification*, Environmental Modelling and Software, 82 (2016), pp. 21–30.

[22] 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 (2015), pp. 1–10.

[23] T. A. Mara and S. Tarantola, *Variance-based sensitivity indices for models with dependent inputs*, Reliability Engineering and System Safety, 107 (2012), pp. 115–121.

[24] T. A. Mara, S. Tarantola, and P. Annoni, *Non-parametric methods for global sensitivity*

analysis of model output with dependent inputs, Environmental Modelling and Software, 72 (2015), pp. 173–183.
[25] D. Marquardt, *An algorithm for least-squares estimation of nonlinear parameters*, SIAM Journal on Applied Mathematics, 11 (1963), pp. 431–441.
[26] Y. Marzouk and D. Xiu, *A stochastic collocation approach to Bayesian inference in inverse problems*, Communications in Computational Physics, 6 (2009), pp. 826–847.
[27] M.D. McKay, R.J. Beckman, and W.J. Conover, *A comparison of three methods for selecting values of input variables in the analysis of output from a computer code*, Technometrics, 21 (1979), pp. 239–245.
[28] 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 (1953), pp. 1087–1091.
[29] Y. Mualem, *A new model for predicting the hydraulic conductivity of unsaturated porous media*, Water Resources Research, 12 (1976), pp. 513–522.
[30] A. Nataf, *Détermination des distributions dont les marges sont données*, Comptes Rendus de l'Académie des Sciences, 225 (1962), pp. 42–43.
[31] R. B. Nelsen, *An introduction to copulas*, Springer series in statistics, second edition, 2006.
[32] K. Pearson, *On the general theory of skew correlation and non-linear regression*, in Mathematical contributions to the theory of evolution, vol. XIV, Drapers's Company Research Memoirs, 1905.
[33] E. S. Qian, C. A. Stow, and M. E. Borsuk, *On Monte Carlo methods for Bayesian inference*, J. Ecological Modelling, 159 (2003), pp. 269–277.
[34] B. Renard, D. Kavetsky, G. Kuczera, and M. Thyer, *Understanding predictive uncertainty in hydrology modeling: The challenge of identifying input and structural errors.*, Water Resources Research, 46 (2010), pp. 1–22.
[35] M. Rosenblatt, *Remarks on the multivariate transformation*, Annals of Mathematics and Statistics, 43 (1952), pp. 470–472.
[36] A. Saltelli, K. Chan, and E. M. Scott, *Sensitivity analysis*, John Wiley and Sons, Chichester, 2000.
[37] A. Saltelli and S. Tarantola, *On the relative importance of input factors in mathematical models: Safety assessment for nuclear waste disposal*, Journal of the American Statistical Association, 97 (2002), pp. 702–709.
[38] G. Schoups and J. Vrugt, *A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic and non-Gaussian errors*, Water Resources Research, 46 (2010), pp. 1–17.
[39] G. Schwartz, *Estimating the dimensions of a model*, Annal of Statistics, 6 (1978), pp. 461–464.
[40] Q. Shao, A. Younes, M. Fahs, and T. A. Mara, *Bayesian sparse polynomial chaos expansion for global sensitivity analysis*, Computer Methods in Applied Mechanics & Engineering, 318 (2017), pp. 474–496.
[41] I. M. Sobol', *Sensitivity estimates for nonlinear mathematical models*, Math. Mod. and Comput. Exp., 1 (1993), pp. 407–414.
[42] I. M. Sobol', S. Tarantola, D. Gatelli, S. Kucherenko, and W. Mauntz, *Estimating the approximation error when fixing unessential factors in global sensitivity analysis*, Reliability Engineering and System Safety, 92 (2007), pp. 957–960.
[43] C. J. F. ter Braak and J. Vrugt, *Differential evolution markov chain with snooker updater and fewer chains*, Statistics and Computing, 18 (2008), pp. 435–446.
[44] A. J. Valocchi, R. L. Street, and P. V. Roberts, *Transport of ion exchanging solutes in ground water: chromatographic theory field simulation*, Water Resources Research, 17 (1981), pp. 1517–1527.
[45] M. Th. van Genuchten, *A closed form equation for predicting the hydraulic properties of unsaturated soils*, Soil Science Society of America Journal, 44 (1980), pp. 892–898.
[46] J. A. Vrugt, *Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation*, Environmental Modelling and Software, 75 (2016), pp. 273–316.
[47] J. A. Vrugt and W. Bouten, *Validity of first-order approximations to describe parameter uncertainty in soil hydrologic models*, Soil Science Society of America Journal, 66 (2002), pp. 1740–1751.
[48] A. L. Walter, E. O. Frind, D. W. Blows, C. J. Ptacek, and J. W. Molson, *Modeling of multicomponent reactive transport in groundwater: 1. model development and evaluation*, Water Resources Research, 30 (1994), pp. 3137–3148.
[49] A. Zysset, F. Stauffer, and T. Dracos, *Modeling of chemically reactive groundwater transport*, Water Resources Research, 30 (1994), pp. 2217–2228.