Polynomial chaos expansion for sensitivity analysis of model output with dependent inputs

Thierry A. Mara, William E Becker

Reliability Engineering and System Safety 214 (2021) 107795

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

journal homepage: www.elsevier.com/locate/ress

Polynomial chaos expansion for sensitivity analysis of model output with dependent inputs

Thierry A. Mara, William E. Becker [*]

European Commission, Joint Research Centre, 21027 Ispra (VA), Italy

ARTICLE INFO

Keywords:
Rosenblatt transformation
Nataf transformation
Mara–Tarantola transformation
Polynomial chaos expansion
Statistical calibration
Posterior sensitivity analysis

ABSTRACT

In this paper, we discuss the sensitivity analysis of model response when the uncertain model inputs are not independent of one other. In this case, two different kinds of sensitivity indices can be evaluated: (i) the sensitivity indices that account for the dependence/correlation of an input or group of inputs with the remainder and (ii) the sensitivity indices that do not account for this dependence. We argue that this distinction applies to any global sensitivity measure. In the present work, we focus on the estimation of variance-based sensitivity indices which are based on the second-order moment of the model response of interest. In particular, we derive new strategies and new computationally efficient methods to assess them, which rely on the polynomial chaos expansion. Several numerical exercises are carried out to demonstrate the performance of the new methods, including a sensitivity analysis of a drainage model posterior to its statistical calibration.

1. Introduction

Uncertainty and sensitivity analysis are essential ingredients of modelling [1]. In a decision-making process, uncertainty analysis aims to verify whether the decision made (according to the model responses) is robust to different sources of uncertainty. If the results are not robust, sensitivity analysis can be used to identify which sources of uncertainty are primarily responsible for the model output uncertainty [2]. Then, where feasible, further research can be conducted to improve knowledge (and reduce the uncertainty) of the most influential inputs. This refinement of knowledge is intended to yield a clearer decision. The importance of uncertainty and sensitivity analysis in the decision-making process is recognised by governmental organisations such as the European Commission, which has included them in their "Better Regulation Guidelines" [3]. These guidelines are the basis for all policy making at the European level.

Sensitivity analysis has received much attention during the two last decades in all disciplines that use and develop computer models, in particular since the seminal work of Ilya M. Sobol' [4] (who established the theoretical basis of the widely-used *variance-based sensitivity indices*, otherwise known as the *Sobol' indices*), and the promotion of variance-based sensitivity analysis by the European Commission's Joint Research Centre [1,5,6]. Sobol' proved that, provided that the model inputs (i.e. the sources of uncertainty in computer models) are independent of each other, there is a unique functional ANOVA (ANalysis Of VAriance) decomposition of the model response of interest with respect to the uncertain inputs. This functional representation decomposes the model response variance into a sum of partial variances which are contributions of each input individually, and further contributions due to interactions with other inputs. With such a decomposition, it is straightforward to measure the relative importance of the input variables in terms of their contributions to the model output variance.

Following the work of Sobol', many numerical methods have been proposed to estimate the Sobol' indices when model inputs are independent of each other. Monte Carlo methods are amongst the first estimators proposed in the literature [4,7,8]. They employ several (quasi) Monte Carlo samples to estimate Sobol' indices by following the *pick-freeze* strategy [9]. Another very popular class of methods are based on the so-called Fourier Amplitude Sensitivity Test [10–13]. Introduced in the 70s, they rely on Parseval–Plancherel theorem to estimate first- and total-order Sobol' indices. The most efficient methods (in most circumstances) are those based on *surrogate modelling* of the computer model responses [14–17]. A surrogate model (otherwise known as a 'emulator' or 'metamodel') is a statistical model that mimics the input–output relationship of the original model, but at a greatly reduced computational cost. It is constructed by running the original model a modest number of times for different input values and obtaining corresponding output values. The surrogate model is then constructed to fit this input–output data as well as possible by using a variety of possible approaches. Global sensitivity analysis (which typically requires a large number of model runs) can then be performed using Monte Carlo estimators on the emulator rather than the real

[*] Corresponding author.
E-mail addresses: thierry.mara@ec.europa.eu (T.A. Mara), william.becker@bluefoxdata.eu (W.E. Becker).

https://doi.org/10.1016/j.ress.2021.107795

Available online 26 May 2021

model, resulting in significant computational savings. In the present article, we consider the Polynomial Chaos Expansion (PCE) method. PCE is classified amongst the spectral approaches ([15,16,18] among others), and can be used as an approximation of the input–output relationship (inferred from a given input–output sample) by casting the model response onto orthogonal polynomials. The Parseval–Plancherel theorem is then invoked to compute Sobol' indices, which means that the variance decomposition and the Sobol' indices are analytically obtained from the PCE coefficients without running the PCE as a surrogate model.

The large majority of sensitivity analysis approaches and estimation procedures assume that model inputs are independent from one another. However in reality, input variables may be correlated with one another (i.e. linear dependence), or more generally *dependent* on each other, which can also imply nonlinear relationships. Performing global sensitivity analysis of computer model responses with dependent inputs is challenging. The difficulty stems from the non-uniqueness of the ANOVA decomposition and the definition of useful importance measures (see [19]). In the last decade, several mathematical frameworks and associated numerical methods have been proposed to address this issue [19–31].

Li et al. (2010) introduced the covariance-based sensitivity indices [21]. These are split into two kinds of sensitivity indices, namely, the *structural* and *correlative* sensitivity indices. They can be inferred after a functional input–output relationship (i.e. metamodel) is obtained. Caniou and Sudret (2013) use PCE as metamodel, identified by assuming first that the input variables are independent of each other [31–33]. Then, the covariance analysis (that the authors called ANCOVA) is performed by considering the possible correlation between the variables. Typically ANCOVA is a two-step approach. Chastaing et al. (2012) attempt to define an ANOVA representation (like the one of Sobol' in which the summands are orthogonal) of the functional input–output relationship in the case of correlated input variables [25]. Thus, with such a representation, it is straightforward to infer variance-based sensitivity indices. They find that this is possible only for a few joint probability distribution function. Therefore, the application of this approach is limited. In [22], a Monte Carlo estimation procedure was proposed for estimating sensitivity indices of correlated variables, although this requires knowledge of the joint distribution, and the computational cost is quite high. This was later extended to a metamodelling approach in [34], but in both cases there is no decomposition of sensitivity into contributions due to correlations, and independent contributions.

Owen and Prieur (2017) propose to use the Shapley measure [35] as a sensitivity index (called Shapley effect) when the input variables are dependent on each other [28] (see also [36]). The nice properties of the Shapley effect are that: (i) there is only one sensitivity index for each input, (ii) it takes into account the contribution of the input to the output variance in terms of correlation and interactions, (iii) mutual contributions are equally shared over the input variables that interact and that are correlated, and (iv) the overall Shapley effects sum to one. The main obstacles to the use of the Shapley effects for sensitivity analysis are the computational burden and their accurate estimation.

The authors in [20] observe that, when the input variables are correlated, the model response variance might be captured by only a few inputs; the remainder forming a subset of spurious inputs whose contribution to the response variance is embedded in the contribution of the former. This led the authors to introduce two types of variance-based sensitivity indices: one that accounts for the correlation of each input variable with the others (called *full* sensitivity indices in [23]) and one that does not (called *independent* sensitivity indices). The importance measures mentioned in previous paragraph do not decompose the contribution of each variable into correlated and non-correlated parts with the exception of the ANCOVA approach. Indeed, it is shown in [37] that it is possible to infer the correlated and non-correlated variance-based sensitivity indices with the ANCOVA approach. In our opinion, this decomposition is important in understanding whether an input is contributing to the model directly or through a correlation with another variable. Therefore, this paper will discuss approaches to computing the full and independent indices.

We note that this distinction is not completely new in statistics: for example, the *partial correlation coefficient* is essentially a version of the Pearson correlation coefficient which does not account for the mutual effect due to correlations with the other variables [38]. However, neither the partial correlation coefficient nor the sensitivity indices of [20] are "model free", because they assume linearity of the model response, and a linear dependence structure (i.e. correlated inputs). A first step beyond these limitations was made in [23], in which a polynomial chaos expansion method was used to extend the approach of [20] to nonlinear model responses, but still with limitations on the dependence structure between the input variables. Subsequently, a second step was made in [19] to handle any (given) dependence structure: sampling-based strategies (i.e. Monte Carlo estimators) were introduced to evaluate variance-based sensitivity index of any individual inputs or groups of inputs. In [30], the FAST method was adapted to compute the (*full* and *independent*) first-order and total sensitivity indices. The concept of *full* and *independent* sensitivity measures has also been extended to the elementary effects method of Morris in [29].

In the present paper, we make a further advance in global sensitivity analysis with dependent inputs, by extending the PCE-based method of [23] to a broader class of dependence structure (as in [19] and [30]). We focus on the estimation of the first- and total-order Sobol' indices, although any Sobol' index can be estimated in theory (at no extra cost). Incidentally, the number of model runs needed for accurate identification of sparse PCEs with the algorithm of [18] has been shown to be only weakly dependent on the dimensionality of the input space. This is not the case with the aforementioned Monte Carlo methods and the FAST method. Additionally, we show that the concept of distinguishing between *full* and *independent* sensitivity measures can be extended to any sensitivity index, although a full exploration of these new indices is left for future work.

The paper is organised as follows, in Section 2 we recall the definitions of the *full* and *independent* first-order and total sensitivity indices, and discuss the usage and implications of these indices in some detail. We also define the *full* and *independent* moment-independent sensitivity measures. Then, in Section 3 we recall three possible sampling techniques to generate dependent random samples from independent ones (and vice-versa). In Section 4, we describe the algorithm to compute the sensitivity indices of interest with PCE. Section 5 is devoted to numerical test cases and applications. Section 6 concludes.

2. Sensitivity indices for dependent inputs

2.1. Variance-based sensitivity indices

Let us consider a model response y, which is a function of a set of input variables $\boldsymbol{x} = (x_1, \dots, x_d) = (\boldsymbol{x}_1, \boldsymbol{x}_2)$ where $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are two complementary subsets. Let us assume that the inputs are random variables such that, $\boldsymbol{x} \sim p_{\boldsymbol{x}}(\boldsymbol{x})$, where $p_{\boldsymbol{x}}$ is a joint probability density function. If $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are independent of each other, then $p_{\boldsymbol{x}}(\boldsymbol{x}) = p_{\boldsymbol{x}_1}(\boldsymbol{x}_1)p_{\boldsymbol{x}_2}(\boldsymbol{x}_2)$. More generally (where $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ may be dependent on one another), we have the following relationships $p_{\boldsymbol{x}}(\boldsymbol{x}) = p_{\boldsymbol{x}_1}(\boldsymbol{x}_1)p_{\boldsymbol{x}_2|\boldsymbol{x}_1}(\boldsymbol{x}_2|\boldsymbol{x}_1)$ and $p_{\boldsymbol{x}}(\boldsymbol{x}) = p_{\boldsymbol{x}_2}(\boldsymbol{x}_2)p_{\boldsymbol{x}_1|\boldsymbol{x}_2}(\boldsymbol{x}_1|\boldsymbol{x}_2)$, where $p_{\boldsymbol{x}_i|\boldsymbol{x}_j}$ is a conditional pdf. If we now set $\bar{\boldsymbol{x}}_1 = \boldsymbol{x}_1|\boldsymbol{x}_2$ and $\bar{\boldsymbol{x}}_2 = \boldsymbol{x}_2|\boldsymbol{x}_1$, it can be inferred that the vectors $\bar{\boldsymbol{x}}_1$ and $\boldsymbol{x}_2$ (resp. $\boldsymbol{x}_1$ and $\bar{\boldsymbol{x}}_2$) are independent of each other since we have $p_{\boldsymbol{x}}(\boldsymbol{x}) = p_{\boldsymbol{x}_1}(\boldsymbol{x}_1)p_{\bar{\boldsymbol{x}}_2}(\bar{\boldsymbol{x}}_2) = p_{\bar{\boldsymbol{x}}_1}(\bar{\boldsymbol{x}}_1)p_{\boldsymbol{x}_2}(\boldsymbol{x}_2)$.

It is convenient to work with the pairs $(\bar{\boldsymbol{x}}_1, \boldsymbol{x}_2)$ and $(\boldsymbol{x}_1, \bar{\boldsymbol{x}}_2)$ instead of $(\boldsymbol{x}_1, \boldsymbol{x}_2)$, because the former contain independent subsets. Indeed, there is redundant information in the pair $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ due to the dependence of $\boldsymbol{x}_1$ on $\boldsymbol{x}_2$. We can now apply the well-established tools of sensitivity analysis with independent inputs, but instead of analysing the sensitivity of y with respect to $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ we can analyse its sensitivity with respect to $(\bar{\boldsymbol{x}}_1, \boldsymbol{x}_2)$ or $(\boldsymbol{x}_1, \bar{\boldsymbol{x}}_2)$. The notion of working with $(\bar{\boldsymbol{x}}_1, \boldsymbol{x}_2)$ and $(\boldsymbol{x}_1, \bar{\boldsymbol{x}}_2)$, rather than $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ directly, forms the basis of the sensitivity analysis approach in this work, although this does imply that *ordering* of variables (i.e. whether to use $(\bar{\boldsymbol{x}}_1, \boldsymbol{x}_2)$ or $(\boldsymbol{x}_1, \bar{\boldsymbol{x}}_2)$) becomes important. This is discussed further in Section 4.2.

Following this idea, the resulting sensitivity indices can be interpreted as follows: the sensitivity index of $\boldsymbol{x}_1$ is the *full* sensitivity index of $\boldsymbol{x}_1$ (taking into account dependence with the other input variables within $\boldsymbol{x}_2$) whereas the sensitivity index of $\bar{\boldsymbol{x}}_2$ is the *independent* sensitivity index of $\boldsymbol{x}_2$, which does not account for dependence with other input variables (see [19,23,30]).

Within the context of variance-based sensitivity analysis, the following sensitivity indices can be defined,

$$S_{\boldsymbol{x}_1} = \frac{\mathbb{V}\left[\mathbb{E}\left[y|\boldsymbol{x}_1\right]\right]}{\mathbb{V}\left[y\right]}, \tag{1}$$

$$ST^{ind}_{\boldsymbol{x}_1} = \frac{\mathbb{E}\left[\mathbb{V}\left[y|\boldsymbol{x}_2\right]\right]}{\mathbb{V}\left[y\right]}, \tag{2}$$

$$S^{ind}_{\boldsymbol{x}_1} = \frac{\mathbb{V}\left[\mathbb{E}\left[y|\bar{\boldsymbol{x}}_1\right]\right]}{\mathbb{V}\left[y\right]}, \tag{3}$$

$$ST_{\boldsymbol{x}_1} = \frac{\mathbb{E}\left[\mathbb{V}\left[y|\bar{\boldsymbol{x}}_2\right]\right]}{\mathbb{V}\left[y\right]}, \tag{4}$$

where $\mathbb{V}[\cdot]$ is the variance operator, $\mathbb{E}[\cdot]$ is the mathematical expectation, and $\mathbb{V}[\cdot|\cdot]$ and $\mathbb{E}[\cdot|\cdot]$ are the conditional variance and expectation respectively. $S_{\boldsymbol{x}_1}$ and $ST_{\boldsymbol{x}_1}$ are respectively the full first- and total-order sensitivity indices while $S^{ind}_{\boldsymbol{x}_1}$ and $ST^{ind}_{\boldsymbol{x}_1}$ are respectively the independent first- and total-order sensitivity indices.

These formulas can be interpreted as follows, $\mathbb{V}[y] = \mathbb{V}_{\boldsymbol{x}}[y]$ is the total variance of y when $\boldsymbol{x} \sim p_{\boldsymbol{x}}$. On the one hand, $\mathbb{V}\left[\mathbb{E}\left[y|\boldsymbol{x}_1\right]\right] = \mathbb{V}_{\boldsymbol{x}_1}\left[\mathbb{E}_{\boldsymbol{x}_2|\boldsymbol{x}_1}\left[y|\boldsymbol{x}_1\right]\right]$ in Eq. (1) (resp. $\mathbb{E}_{\boldsymbol{x}_2|\bar{\boldsymbol{x}}_2}\left[\mathbb{V}_{\boldsymbol{x}_1}\left[y|\boldsymbol{x}_1\right]\right]$ in Eq. (4)) is the partial variance of y due to $\boldsymbol{x}_1$ alone (resp. alone and by interaction with $\bar{\boldsymbol{x}}_2$) when the latter is left free to vary w.r.t. its marginal density, that is $\boldsymbol{x}_1 \sim p_{\boldsymbol{x}_1}$, while $\boldsymbol{x}_2$ is constrained to vary conditionally onto $\boldsymbol{x}_1$, i.e. $\bar{\boldsymbol{x}}_2 \sim p_{\boldsymbol{x}_2|\boldsymbol{x}_1}$. Therefore, $\mathbb{V}_{\boldsymbol{x}_1}\left[\mathbb{E}_{\boldsymbol{x}_2|\boldsymbol{x}_1}\left[y|\boldsymbol{x}_1\right]\right]$ (resp. $\mathbb{E}_{\boldsymbol{x}_2|\boldsymbol{x}_1}\left[\mathbb{V}_{\boldsymbol{x}_1}\left[y|\boldsymbol{x}_1\right]\right]$) carries away the mutual contribution to the total variance due to the dependence of $\boldsymbol{x}_1$ on $\boldsymbol{x}_2$. On the other hand, $\mathbb{V}\left[\mathbb{E}\left[y|\bar{\boldsymbol{x}}_1\right]\right] = \mathbb{V}_{\boldsymbol{x}_1|\boldsymbol{x}_2}\left[\mathbb{E}_{\boldsymbol{x}_2}\left[y|\boldsymbol{x}_1\right]\right]$ in Eq. (3) (resp. the numerator in Eq. (2)) is the partial variance of y due to $\bar{\boldsymbol{x}}_1$ alone (resp. alone and by interaction with $\boldsymbol{x}_2$) when this time $\boldsymbol{x}_1$ is constrained to vary conditionally onto $\boldsymbol{x}_2$, i.e. $\bar{\boldsymbol{x}}_1 \sim p_{\boldsymbol{x}_1|\boldsymbol{x}_2}$ while $\boldsymbol{x}_2 \sim p_{\boldsymbol{x}_2}$. Thus, these partial variances do not contain the mutual contribution due to the dependence of $\boldsymbol{x}_1$ on $\boldsymbol{x}_2$.

The first two sensitivity indices (Eqs. (1)–(2)) are the classical definitions of the Sobol' indices as highlighted in [22], while the last two are only defined for dependent input variables. This implies that in the "classical" Sobol' indices, the first-order sensitivity index accounts for correlations with other variables, while the total-order index does not (see also [37]). The overall indices (1)–(4) are scaled within $[0,1]$ and we have $S_{\boldsymbol{x}_1} \leq ST_{\boldsymbol{x}_1}$, $S^{ind}_{\boldsymbol{x}_1} \leq ST^{ind}_{\boldsymbol{x}_1}$ (see for instance [30]).

The full first-order sensitivity index $S_{\boldsymbol{x}_1}$ measures the amount of variance of y due to $\boldsymbol{x}_1$ and its dependence with $\boldsymbol{x}_2$ but does not include the interactions of $\boldsymbol{x}_1$ with $\bar{\boldsymbol{x}}_2$. The full total sensitivity index $ST_{\boldsymbol{x}_1}$ takes into account both dependence and interaction. The independent first-order sensitivity index $S^{ind}_{\boldsymbol{x}_1}$ measures the contribution of $\boldsymbol{x}_1$ by ignoring its correlations with $\boldsymbol{x}_2$ and interactions with $\bar{\boldsymbol{x}}_2$ while $ST^{ind}_{\boldsymbol{x}_1}$ accounts for interactions and ignores correlations. Consider an input x_i, which contributes to the model response variance only because of its strong dependence to the other inputs. In that case, we shall have $ST_{x_i} \geq 0$ and $ST^{ind}_{x_i} = 0$.

It is worth underlining here that *interactions* in the case of dependent inputs (following the partitioning of variables used in this paper) do not have the same interpretation as in the case of independent ones. While in the dependent case interactions concern $(\boldsymbol{x}_1, \bar{\boldsymbol{x}}_2)$ and $(\bar{\boldsymbol{x}}_1, \boldsymbol{x}_2)$, in the independent case they are related to the original variables $(\boldsymbol{x}_1, \boldsymbol{x}_2)$. This has the following consequence: a non-interacting input/output relationship with respect to $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ can turn out to have an interacting relationship with respect to $(\boldsymbol{x}_1, \bar{\boldsymbol{x}}_2)$ or $(\bar{\boldsymbol{x}}_1, \boldsymbol{x}_2)$. This complicates somewhat the analysis, but recall that we use the pairs $(\boldsymbol{x}_1, \bar{\boldsymbol{x}}_2)$ and $(\bar{\boldsymbol{x}}_1, \boldsymbol{x}_2)$ because they are *independent* variables, which allows the use of the variance decomposition of y in the Sobol' sense [4]. However, in the dependent case, note that such a decomposition is not unique as one can derive it with respect to either $(\boldsymbol{x}_1, \bar{\boldsymbol{x}}_2)$ or $(\bar{\boldsymbol{x}}_1, \boldsymbol{x}_2)$.

To illustrate, let us consider a model response which is supposedly a function of two dependent random vectors $(\boldsymbol{x}_1, \boldsymbol{x}_2)$, but actually only a function of $\boldsymbol{x}_1$, say, $y = f(\boldsymbol{x}_1)$. By computing the sensitivity indices w.r.t. $(\boldsymbol{x}_1, \bar{\boldsymbol{x}}_2)$, one will find $ST_{\boldsymbol{x}_1} = S_{\boldsymbol{x}_1} = 1$ and $ST^{ind}_{\boldsymbol{x}_2} = S^{ind}_{\boldsymbol{x}_2} = 0$. This leads to the conclusion that, because of the dependence structure between $(\boldsymbol{x}_1, \boldsymbol{x}_2)$, all the information (i.e. variance) of y is explained by the subset $\boldsymbol{x}_1$. By further computing the sensitivity indices w.r.t. $(\bar{\boldsymbol{x}}_1, \boldsymbol{x}_2)$, one will get $ST^{ind}_{\boldsymbol{x}_1} \geq S^{ind}_{\boldsymbol{x}_1} > 0$ and $ST_{\boldsymbol{x}_2} \geq S_{\boldsymbol{x}_2} > 0$ which indicates that because of its dependence with $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ also has an influence on the model response. It can be inferred that the value of $\boldsymbol{x}_2$ cannot be fixed at an arbitrary value because this has an impact on the conditional pdf of $\boldsymbol{x}_1$ (to do so, one should have $ST_{\boldsymbol{x}_2} = ST^{ind}_{\boldsymbol{x}_2} = 0$, according to [29]). Finally, it can be concluded that $\boldsymbol{x}_2$ is a spurious vector of inputs only relevant because of its dependence with $\boldsymbol{x}_1$ (see [20]). However, it cannot be concluded that the original model structure (w.r.t. $(\boldsymbol{x}_1, \boldsymbol{x}_2)$) is of the form $y = f(\boldsymbol{x}_1)$. Indeed, it is only when $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ is a vector of independent random variables that the original model structure can be revealed because the variance decomposition in the Sobol' sense w.r.t. $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ is unique. Nevertheless, with this result one can conclude that a surrogate model with only $\boldsymbol{x}_1 \sim p_{\boldsymbol{x}_1}$ can be built that could be very accurate as compared to the original model, thus contributing to a *model reduction* objective.

The authors of [23] propose to estimate Eqs. (1)–(4) with polynomial chaos expansions after decorrelating the input sample. The advantage of the proposed approach is its computational efficiency. Indeed, only one input sample of a relatively modest size (albeit a function of the model complexity) is sufficient to estimate all the first-order and total sensitivity indices of each individual input variable. The problem is that decorrelation does not necessarily mean independence. Actually, with the decorrelation procedure of [23] recalled in Section 3.3, independence is ensured only for a particular form of the dependence structure, which is explained later in Section 3.3. In Section 3.2 we extend the method to correlation structures represented by a Gaussian copula [22,39] and in Section 3.1 to the general dependence structure represented by the Rosenblatt transform. In Section 4, we then extend the PCE method to the case of dependent variables represented by one of these dependence structures.

2.2. Moment-independent sensitivity indices

In the following, we show how the concept of full and independent sensitivity indices can also be applied to moment-independent measures. At the end of this section, we also comment on how this can apply in the more general case.

In the case of independent inputs, variance-based sensitivity indices indicate whether a model output y is a function of an input (or subset of inputs) in the form of a functional relationship like $y = f(\boldsymbol{x}_1, \boldsymbol{x}_2)$. Specifically, they examine the contribution to the variance of the model output distribution p_y. If a relationship of the form $y = f(\boldsymbol{x}_2)$ is discovered, then $\boldsymbol{x}_1$ is deemed irrelevant. Arguably a more natural way to infer that y does not depend on $\boldsymbol{x}_1$ is to prove that $p_{y|\boldsymbol{x}_1} = p_y$. This has led to the definition of the so-called moment-independent sensitivity measures defined as [40],

$$\delta_{\boldsymbol{x}_1} = \frac{1}{2}\mathbb{E}_{\boldsymbol{x}_1}\left[\int_{\mathbb{R}} \left|p_y - p_{y|\boldsymbol{x}_1}\right| \mathrm{d}y\right] \in [0,1] \tag{5}$$

This measure is equal to half of the average of the differences between the unconditional pdf of y, denoted p_y, and the conditional pdf of y given $\boldsymbol{x}_1$, denoted $p_{y|\boldsymbol{x}_1}$. Finding $\delta_{\boldsymbol{x}_1} = 0$ allows us to conclude that y is not dependent on $\boldsymbol{x}_1$, therefore $\boldsymbol{x}_1$ can be regarded as non-influential. We note that to evaluate y, $\boldsymbol{x}_1$ is drawn from $p_{\boldsymbol{x}_1}$ while the complementary input subset $\boldsymbol{x}_2$ is sampled from $p_{\boldsymbol{x}_2|\boldsymbol{x}_1}$. In the independent case discussed here we naturally have $p_{\boldsymbol{x}_2|\boldsymbol{x}_1} = p_{\boldsymbol{x}_2}$.

In the case of dependent input variables, the functional relationship between y and $\boldsymbol{x}$ can take one of the following forms, $y = f(\boldsymbol{x}_1, \bar{\boldsymbol{x}}_2)$ or $y = f(\bar{\boldsymbol{x}}_1, \boldsymbol{x}_2)$. If one shows that $y = f(\boldsymbol{x}_1)$ (i.e. $S_{\boldsymbol{x}_1} = 1 \Leftrightarrow ST^{ind}_{\boldsymbol{x}_2} = 0$), then it can be concluded that $\boldsymbol{x}_2$ is a subset of spurious inputs. Additionally, if $ST_{\boldsymbol{x}_2} = 0$ then it can be inferred that $\boldsymbol{x}_2$ can be fixed without any impact on the model response. However, such a distinction cannot be made with the δ-importance measure as defined in Eq. (5) because the latter actually assesses the importance of $\boldsymbol{x}_1$ by accounting for its dependence with $\boldsymbol{x}_2$. This remark leads us to introduce the following complementary δ-importance measure that does not account for the dependence of $\boldsymbol{x}_1$ to $\boldsymbol{x}_2$,

$$\delta^{ind}_{\boldsymbol{x}_1} = \frac{1}{2}\mathbb{E}_{\bar{\boldsymbol{x}}_1}\left[\int_{\mathbb{R}} \left| p_y - p_{y|\bar{\boldsymbol{x}}_1} \right| \mathrm{d}y\right] \tag{6}$$

where $\bar{\boldsymbol{x}}_1 \sim p_{\boldsymbol{x}_1|\boldsymbol{x}_2}$ while $\boldsymbol{x}_2 \sim p_{\boldsymbol{x}_2}$. Finding $\delta^{ind}_{\boldsymbol{x}_1} = 0$ indicates that $\boldsymbol{x}_1$ is a spurious subset of inputs which impacts the model response only via its dependence with $\boldsymbol{x}_2$ (if any exists) while if $\delta_{\boldsymbol{x}_1} = \delta^{ind}_{\boldsymbol{x}_1} = 0$, it can be additionally concluded that $\boldsymbol{x}_1$ can be fixed without affecting the model response.

The brute-force estimates of $\delta_{\boldsymbol{x}_1}$ and $\delta^{ind}_{\boldsymbol{x}_1}$ subtly differ. In both cases, p_y is obtained after sampling $\boldsymbol{x} \sim p_{\boldsymbol{x}}$ with a (quasi) Monte Carlo technique, running the model for each draw and collecting the model response $y = f(\boldsymbol{x})$. Subsequently, p_y can be estimated with an appropriate method (such as the kernel density estimator of [41]). To estimate $p_{y|\boldsymbol{x}_1}$ one has to draw one random value of $\boldsymbol{x}_1 \sim p_{\boldsymbol{x}_1}$, then generate a (quasi) Monte Carlo sample of $\bar{\boldsymbol{x}}_2 \sim p_{\boldsymbol{x}_2|\boldsymbol{x}_1}$ and infer from $\bar{\boldsymbol{x}}_2$ the sample of $\boldsymbol{x}_2$, run the model and get the model responses before finally estimating $p_{y|\boldsymbol{x}_1}$ for the drawn value of $\boldsymbol{x}_1$. This process is repeated several times for different values of $\boldsymbol{x}_1 \sim p_{\boldsymbol{x}_1}$. In the end, one obtains one single estimate of p_y and several estimates of $p_{y|\boldsymbol{x}_1}$ for different values of $\boldsymbol{x}_1$ sampled from $p_{\boldsymbol{x}_1}$. These can be used to compute $\delta_{\boldsymbol{x}_1}$.

On the contrary, to compute $\delta^{ind}_{\boldsymbol{x}_1}$, one has to get several estimates of $p_{y|\bar{\boldsymbol{x}}_1}$ for different values of $\bar{\boldsymbol{x}}_1$ sampled from $p_{\boldsymbol{x}_1|\boldsymbol{x}_2}$. We proceed by first generating a (quasi) Monte Carlo sample of $\boldsymbol{x}_2 \sim p_{\boldsymbol{x}_2}$, then we deduce the draws of $\bar{\boldsymbol{x}}_1$ sampled from $p_{\boldsymbol{x}_1|\boldsymbol{x}_2}$, and therefore the sample of $\boldsymbol{x}_1$. Finally $p_{y|\bar{\boldsymbol{x}}_1}$ is deduced after running the model and getting the response vector. This process is repeated several times for different values of $\bar{\boldsymbol{x}}_1$. This tricky sampling procedure is further explained in Section 3.

The present extension of Borgonovo's moment-independent measure is given here to show how the underlying approaches in this paper can be generalised outside of variance-based measures. Indeed, we could conceivably apply the same principles to any global sensitivity measures defined in the following general way in [42],

$$\gamma_{\boldsymbol{x}_1} = \mathbb{E}_{\boldsymbol{x}_1}\left[\Delta(y, y|\boldsymbol{x}_1)\right] \tag{7}$$

where $\Delta(\cdot,\cdot)$ is a dissimilarity measure between two random variables [43]. For example, in the case of dependent inputs one can introduce the complementary sensitivity index that does not account for the effect of $\boldsymbol{x}_1$ due to its dependence on $\boldsymbol{x}_2$, that is,

$$\gamma^{ind}_{\boldsymbol{x}_1} = \mathbb{E}_{\boldsymbol{x}_1|\boldsymbol{x}_2}\left[\Delta(y, y|\boldsymbol{x}_1)\right] \tag{8}$$

In the remainder of this paper, however, we focus on variance-based sensitivity analysis and leave investigation of other measures to future work.

3. Generating dependent and independent samples

In the previous section it was explained that a sensitivity analysis of the system of dependent variables $f(\boldsymbol{x}_1, \boldsymbol{x}_2)$ can be achieved by instead considering the independent pair $\boldsymbol{x}_1, \bar{\boldsymbol{x}}_2$. Modellers need samples of $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ to run their model but they need samples of $\bar{\boldsymbol{x}}_1$ to compute $S^{ind}_{\boldsymbol{x}_1}$ and $ST_{\boldsymbol{x}_1}$ (see Eqs. (3)–(4). The key challenge in the present approach to estimate the set of variance-based sensitivity indices is to obtain samples of the conditional and unconditional variables (e.g. $\bar{\boldsymbol{x}}_1$ and $\boldsymbol{x}_1$) respectively. In this section, we present several possibilities to do so, with different properties and requirements.

3.1. The Rosenblatt transformation

Let $\boldsymbol{u} = (\boldsymbol{u}_1, \boldsymbol{u}_2)$ be the vector of independent random variables stemming from the isoprobabilistic transformation,

$$\begin{cases} \boldsymbol{u}_1 = F_{\boldsymbol{x}_1}(\boldsymbol{x}_1) \\ \boldsymbol{u}_2 = F_{\boldsymbol{x}_2|\boldsymbol{x}_1}(\boldsymbol{x}_2|\boldsymbol{x}_1) \end{cases} \tag{9}$$

where $F_{\boldsymbol{x}_1}$ is the cumulative density function (cdf) of $\boldsymbol{x}_1$ and $F_{\boldsymbol{x}_2|\boldsymbol{x}_1}$ is the conditional cdf of $\boldsymbol{x}_2$. We note that $\boldsymbol{u}$ is uniformly distributed over the unit hypercube $[0,1]^d$. We also notice that Eq. (9) is not unique as one can swap the subscripts 1 and 2. Eq. (9) is called the Rosenblatt transform (RT, [44]) of $\boldsymbol{x}$ and allows independent random variables to be obtained from dependent ones provided that the conditional cdfs are known.

Propagating the input uncertainty through the model response requires the generation of (quasi or pseudo) random draws of $\boldsymbol{x}$. This can be achieved with one of the inverse RT as follows,

$$\begin{cases} \boldsymbol{x}_1 = F^{-1}_{\boldsymbol{x}_1}(\boldsymbol{u}_1) \\ \boldsymbol{x}_2 = F^{-1}_{\boldsymbol{x}_2|\boldsymbol{x}_1}(\boldsymbol{u}_2|\boldsymbol{u}_1) \end{cases} \tag{10}$$

which requires that the cdfs be invertible. This is performed in practice by first generating a sample of $\boldsymbol{u}$ with a (quasi) random generator such as the LP$_\tau$ sequences of [45]. Then, this sample is transformed into a sample of $\boldsymbol{x}$ using Eq. (10). From Eq. (9), $\bar{\boldsymbol{x}}_2 = F^{-1}_{\bar{\boldsymbol{x}}_2}(\boldsymbol{u}_2)$, therefore it is straightforward to generate $\bar{\boldsymbol{x}}_2$ independently of $\boldsymbol{x}_1$ as the former only depends on $\boldsymbol{u}_2$. Therefore, fixing the value of $\boldsymbol{u}_2$ is equivalent to fixing $\bar{\boldsymbol{x}}_2$.

3.2. The Nataf transformation

The Rosenblatt transformation requires knowledge of the conditional cdfs, which might not always be the case in practice. When the uncertainty of $\boldsymbol{x}$ is characterised by a set of marginal cdfs $\{F_{x_1}, \dots, F_{x_d}\}$ and a product–moment correlation matrix $\boldsymbol{R}_{\boldsymbol{x}}$, Nataf transformation is more suitable to generate random samples of $\boldsymbol{x}$ (NT, [46]). We note however that the Nataf transformation is equivalent to the Rosenblatt transform with a Gaussian copula [47]. Given $\boldsymbol{R}_{\boldsymbol{z}}$ and $\boldsymbol{z} \sim \mathcal{N}(0, \boldsymbol{I}_d)$ the steps of the NT are,

$$\begin{cases} \text{Find the upper Cholesky matrix } \boldsymbol{U} : \boldsymbol{R}_{\boldsymbol{z}} = \boldsymbol{U}^T\boldsymbol{U} \\ \boldsymbol{z}^c = \boldsymbol{z}\boldsymbol{U} \\ x_j = F^{-1}_{x_j}\left(\Phi\left(z^c_j\right)\right), \forall j = 1, \dots, d \end{cases} \tag{11}$$

The crux of this algorithm is to choose $\boldsymbol{R}_{\boldsymbol{z}}$ so that the product–moment correlation matrix of $\boldsymbol{x}$, *in fine*, is the one desired (i.e. $\boldsymbol{R}_{\boldsymbol{x}}$). For some special distributions, the link between $\boldsymbol{R}_{\boldsymbol{z}}$ and $\boldsymbol{R}_{\boldsymbol{x}}$ are well-known (see [48]). Generally, an iterative procedure can be required to tune $\boldsymbol{R}_{\boldsymbol{z}}$. We note that Eq. (11) is equivalent to the procedure of Iman & Conover [49] when $\boldsymbol{R}_{\boldsymbol{x}}$ is a rank correlation matrix. In that case, there is no need to tune $\boldsymbol{R}_{\boldsymbol{z}}$ as the latter exactly equals $\boldsymbol{R}_{\boldsymbol{x}}$.

In the NT transformation, two complementary independent subsets $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ such that $\boldsymbol{z} = (\boldsymbol{z}_1, \boldsymbol{z}_2)$, produce two correlated subsets $(\boldsymbol{x}_1, \boldsymbol{x}_2)$. It is worth noticing that because $\boldsymbol{U}$ is a upper triangular matrix, the information in $\boldsymbol{z}_2$ propagates only through $\boldsymbol{x}_2$. Therefore, $\bar{\boldsymbol{x}}_2$ is related to $\boldsymbol{z}_2$ which means that to fix $\bar{\boldsymbol{x}}_2$ one simply needs to fix $\boldsymbol{z}_2$. As for RT, NT is not unique as the elements in the subsets $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ can be chosen arbitrarily. Freedom in the definition of the subsets implies columns and rows permutation of $\boldsymbol{R}_{\boldsymbol{x}}$. Of particular interest are subsets of the form $(x_j, \boldsymbol{x}_{-j})$ in order to compute individual sensitivity indices.

3.3. The transformation of Mara and Tarantola

In [23], a non-linear Gram–Schmidt-like transformation is introduced to obtain independent variables $v_1, v_2, \dots, v_d$ from dependent ones. This transformation (hereafter referred to as "MT") reads as follows,

$$\begin{cases} v_1 &= x_{i_1} \\ v_2 &= x_{i_2} - \mathbb{E}\left[\left(x_{i_2} - \mathbb{E}\left[x_{i_2}\right]\right) | v_1\right] \\ v_3 &= x_{i_3} - \sum_{j=1}^{2} \mathbb{E}\left[\left(x_{i_3} - \mathbb{E}\left[x_{i_3}\right]\right) | v_j\right] \\ & \vdots \\ v_d &= x_{i_d} - \sum_{j=1}^{d-1} \mathbb{E}\left[\left(x_{i_d} - \mathbb{E}\left[x_{i_d}\right]\right) | v_j\right] \end{cases} \tag{12}$$

and assumes an *additive* dependence structure between the x-variables of the form,

$$\begin{cases} x_{i_1} = v_1 \\ x_{i_2} = v_2 + f_{i_2,1}(v_1) \\ x_{i_3} = v_3 + f_{i_3,1}(v_1) + f_{i_3,2}(v_2) \\ \vdots \\ x_{i_d} = v_d + \sum_{j=1}^{d-1} f_{i_d,j}(v_j) \end{cases} \tag{13}$$

where $\int_{\mathbb{R}} f_{i_j,k}(v_k) p_{v_k} \mathrm{d}v_k = 0$, which ensures that the set of functions $\{f_{i_j,1}, \dots, f_{i_j,j-1}\}$ in the jth row of Eq. (13) are orthogonal (according to [4]). We note that each equation in Eq. (13) is related to the generalised additive model representation [50]. We also note that $\boldsymbol{v} = (\boldsymbol{v}_1, \boldsymbol{v}_2)$ produces $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2)$ with the information in $\boldsymbol{v}_2$ only found in $\boldsymbol{x}_2$ and not in $\boldsymbol{x}_1$. Therefore, fixing $\boldsymbol{v}_2$ implies fixing $\bar{\boldsymbol{x}}_2$.

This transformation was introduced to overcome NT that only deals with linear pairwise dependencies between variables. In effect, MT tackles nonlinear pairwise dependencies. The transformation of [23] in Eq. (12) is particularly suited to dealing with given data. In that case, the orthogonal functions in Eq. (13) can be estimated with an iterative univariate regression procedure [50–53], among others). Note that a less restrictive assumption than (13) can be adopted (see [23]), but then a more sophisticated approach is required to infer the independent vector $\boldsymbol{v}$. Importantly, the MT transformation requires no knowledge of cdfs, either marginal or conditional. In that sense, it is the most generally-applicable transformation of the three described here (RT being the most generally-theoretical), albeit containing the strong assumption in (13). The transformation is also sensitive to the ordering of the variables — for this reason, variables are circularly reordered in calculating sensitivity indices (see Section 4.2). Specifically, circular permutation is chosen to make sure that one obtains a single estimate of first- and total-order sensitivity indices of each variable, both accounting for and ignoring the mutual dependent variance contributions. Finally, we note that as with any estimation procedure, its accuracy will be sensitive to the sample size — this may be checked by monitoring convergence as the sample size is incrementally increased, and/or by bootstrapping over the full sample.

4. Polynomial chaos expansion

4.1. The case of independent variables

In [15], the author demonstrated that it was straightforward to estimate the variance-based sensitivity indices from a polynomial chaos expansion. Since then, PCE for global sensitivity analysis has received much attention in the community of modellers [16,54–58]. PCE is a spectral representation of any square-integrable function $f(\boldsymbol{x})$, when the inputs are independent random variables. A PCE representation is written,

$$f(\boldsymbol{x}) = \sum_{\boldsymbol{\alpha} \subset \mathbb{N}^d} a_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}(\boldsymbol{x}) \tag{14}$$

where $\boldsymbol{\alpha} = \alpha_1 \alpha_2 \dots \alpha_d \in \mathbb{N}^d$ is a d-dimensional index and the $a_{\boldsymbol{\alpha}}$ are the PCE coefficients. The $\Psi_{\boldsymbol{\alpha}}$ are the multidimensional polynomial basis elements of degree $|\boldsymbol{\alpha}| = \sum_{i=1}^{d} \alpha_i$. Because $\boldsymbol{x}$ is a vector of independent random variables, $\Psi_{\boldsymbol{\alpha}}$ is obtained by tensor product of univariate polynomials $\psi_{\alpha_i}(x_i)$. Depending on the marginal probability distribution function of x_i, ψ_{α_i} belongs to different families of orthonormal polynomials (such as Legendre polynomials for uniform distribution, Hermite polynomials for normal distribution, Laguerre polynomials for exponential distributions, and so on — see [59]).

Given the PCE coefficients, and knowing that the inputs are independent, the variance-based sensitivity indices can be computed as follows [15],

$$\begin{cases} S_{x_j} = \dfrac{\sum_{\alpha_j>0} a^2_{0\dots\alpha_j\dots0}}{\sum_{\boldsymbol{\alpha}\subset\mathbb{N}^d} a^2_{\boldsymbol{\alpha}} - a^2_{0\dots0}} \\ \\ ST_{x_j} = \dfrac{\sum_{\boldsymbol{\alpha}\subset\mathbb{N}^d:\alpha_j>0} a^2_{\boldsymbol{\alpha}}}{\sum_{\boldsymbol{\alpha}\subset\mathbb{N}^d} a^2_{\boldsymbol{\alpha}} - a^2_{0\dots0}} \end{cases} \quad \forall i \in (1, \dots, d) \tag{15}$$

To estimate the coefficients $a_{\boldsymbol{\alpha}}$, we use the Bayesian sparse polynomial chaos expansion approach of [18], which builds sparse PCEs in a Bayesian framework with the help of the Kashyap information criterion for model selection [60]. This means that any statistic computed with this approach is assigned a credible interval, which is crucial to assess its significance level. There are other alternatives proposed in the literature to build sparse PCEs, notably [16].

4.2. The case of dependent variables

In the case of dependent inputs, PCEs must be constructed with respect to one of the vectors of independent variables discussed in Section 3. If RT is used, then we consider $\boldsymbol{u}$ in conjunction with the shifted-Legendre polynomials, if NT is employed then we consider the vector $\boldsymbol{z}$ and Hermite polynomials while with MT, $\boldsymbol{v}$ is considered in association with the orthogonal polynomials derived with the Gram–Schmidt procedure. In the sequel, we denote by $\bar{\boldsymbol{x}}$ the vector of independent variables obtained by one the aforementioned transformations.

Now, suppose that $\bar{\boldsymbol{x}}$ is obtained by transforming $(x_1, \dots, x_d)$. Then, the sensitivity indices of $\bar{x}_1$ represent the full sensitivity indices of x_1 which account for its dependencies with the other variables (see Section 2). The sensitivity indices of $\bar{x}_2$ represent the sensitivity indices of x_2 which account for its dependencies with the other variables except x_1 (see [23]). The sensitivity indices of $\bar{x}_d$ are interpreted as the independent sensitivity indices of x_d, and the full sensitivity indices of (x_1, x_2) are those of $(\bar{x}_1, \bar{x}_2)$. In the same way, the independent sensitivity indices of (x_{d-1}, x_d) are those of $(\bar{x}_{d-1}, \bar{x}_d)$.

However, as underlined previously, the different transformations to generate $\bar{\boldsymbol{x}}$ are not unique. The order of the variables in the vector to be transformed determines which sensitivity indices can be computed with the identified PCE. To compute the overall set of full sensitivity indices (S_{x_j}, ST_{x_j}) and independent sensitivity indices $(S^{ind}_{x_j}, ST^{ind}_{x_j})$, for all $j = 1, \dots, d$, all the transformations obtained by circularly reordering the input vector $(x_1, \dots, x_d)$ must be considered. This implies d transformations and subsequently the identification of d sets of PCE coefficients. Nonetheless, with PCE, only one Monte Carlo sample of the input variables $\boldsymbol{x}$ is needed to assess any variance-based sensitivity indices. Hence, the number of model calls N can still be reasonable. In many cases, $N < 1\ 000$ is sufficient to obtain an accurate estimation of the sensitivity indices but this depends on the effective dimension of the model and the degree of smoothness of $f(\boldsymbol{x})$. This makes PCE-based sensitivity analysis a compelling approach compared with existing methods. The algorithm to compute the overall $(S_{x_j}, ST_{x_j}, S^{ind}_{x_j}, ST^{ind}_{x_j})$, for all $j = 1, \dots, d$, is given in Appendix.

We also note that if the objective is to estimate higher-order interaction effects, circularly reordering the variables is not sufficient, and other permutations would have to be considered. However, for many cases the estimation of first and total-order sensitivity indices is sufficient, and higher-order effects are left to future work.

Depending on the transformation used, PCEs are built upon $\bar{\boldsymbol{x}}$ which is either $\boldsymbol{u}$, or $\boldsymbol{z}$ or $\boldsymbol{v}$. If $\boldsymbol{u}$ is chosen, then PCEs are built with the shifted-Legendre polynomials, if it is $\boldsymbol{z}$ then Hermite polynomials will be used, while if the random variables are arbitrarily distributed (like in the case of $\boldsymbol{v}$), then a Gram–Schmidt orthogonalisation procedure is used to obtain the subset of orthogonal polynomials. But as highlighted in [61], working with the transformed independent variables might pose problems of convergence of PCEs. In that case, subsequent individual transformation of the $\bar{x}$-variables can help to overcome this issue. In the numerical exercises Section 5, we will indicate which further transformation has been used if so.

4.3. Computational issues

Eq. (15) indicates that the Sobol' indices are computed directly from the PCE coefficients. Thus, there is no need to run the PCE as a surrogate model to estimate variance-based sensitivity indices. This is the reason why variance-based sensitivity analysis with PCE can be classified among the spectral methods along with the Fourier amplitude sensitivity test [10,13]. Therefore, the challenge with PCE-based sensitivity analysis is the PCE coefficient estimation. In practice, it is expected that a sparse PCE often suffices to be a good proxy of the input–output relationship, that is,

$$f(\boldsymbol{x}) = \sum_{\boldsymbol{\alpha}\in\mathcal{A}} a_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}(\bar{\boldsymbol{x}}) + \epsilon \tag{16}$$

where $\mathcal{A}$ is a subset of multi-indexes $\boldsymbol{\alpha} \subset \mathbb{N}^d$ of maximal polynomial degree $p_\mathcal{A} = |\boldsymbol{\alpha}|$ and level of interaction $q_\mathcal{A}$, $\bar{\boldsymbol{x}}$ the vector of independent variables and ϵ the approximation error.

We employ the stepwise regression algorithm of [18], derived in a Bayesian framework, to identify the best PCE. Let $\boldsymbol{X}$ be an input sample of $\boldsymbol{x} \sim p_{\boldsymbol{x}}$ of size $N \times d$ and $\boldsymbol{Y}$ the associated vector of model responses, Bayesian sparse PCE starts by obtaining the standardised vector $\boldsymbol{y}$ from the original vector $\boldsymbol{Y}$. We denote by $\bar{\boldsymbol{X}}$ one of the possible independent input samples obtained from the transformation of $\boldsymbol{X}$ with either RT, NT or MT. From the dataset $\left(\bar{\boldsymbol{X}}, \boldsymbol{y}\right)$, the best sparse PCE is obtained as follows,

1. *Initialisation*: Set initial polynomial degree $p = 4$ and interaction level $q = 2$ (or $p = 2$, $q = 1$ if d high). Create the initial subset of candidates $\mathcal{A}_c = \{\boldsymbol{\alpha} \in \mathbb{N}^d : p_\mathcal{A} \le p, q_\mathcal{A} \le q\}$.
2. Define $P = \operatorname{Card}\left(\mathcal{A}_c\right)$, denote by $(\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_P)$ the subset of multi-indexes, set $m = 1$, $\mathcal{A} = \boldsymbol{\alpha}_1$, and $KIC_m^{MLE} = +\infty$.
3. *Model selection*: Set $m = m+1$, $\mathcal{A} = (\mathcal{A}, \boldsymbol{\alpha}_m)$. Compute the current PCE coefficients and KIC_m^{MLE}. If $KIC_m^{MLE} > KIC_{m-1}^{MLE}$ remove $\boldsymbol{\alpha}_m$ from $\mathcal{A}$. Resume until $m = P$.
4. *Enrichment or Stop*: Update $p_\mathcal{A}$, $q_\mathcal{A}$. If $p_\mathcal{A} < (p-1)$ and $q_\mathcal{A} < q$ Stop. Otherwise, set $p = p+2$ and/or $q = q+1$, create a new subset $\mathcal{A}_c$ with elements of degree within $[p_\mathcal{A}, p]$ and level of interaction within $[q_\mathcal{A}, q]$ and resume from 2.

The PCE coefficients are estimated at step 3 by assuming $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$, yielding

$$\hat{\boldsymbol{a}}_\mathcal{A} \sim \mathcal{N}\left(\hat{\boldsymbol{a}}_\mathcal{A}^{MLE}, \hat{\boldsymbol{C}}_{aa}\right) \tag{17}$$

with,

$$\hat{\boldsymbol{a}}_\mathcal{A}^{MLE} = \left(\boldsymbol{\Psi}_\mathcal{A}^T \boldsymbol{\Psi}_\mathcal{A}\right)^{-1} \boldsymbol{\Psi}_\mathcal{A}^T \boldsymbol{y} \tag{18}$$

$$\hat{\boldsymbol{C}}_{aa} = \left(\hat{\sigma}_\epsilon^{MLE}\right)^2 \left(\boldsymbol{\Psi}_\mathcal{A}^T \boldsymbol{\Psi}_\mathcal{A}\right)^{-1} \tag{19}$$

and the posterior pdf of the error variance is the following inverse Gamma distribution,

$$\hat{\sigma}_\epsilon^2 \sim \mathrm{IG}\left(\frac{N+2}{2}, \frac{\left(\boldsymbol{y} - \boldsymbol{\Psi}_\mathcal{A}\hat{\boldsymbol{a}}_\mathcal{A}^{MLE}\right)^T \left(\boldsymbol{y} - \boldsymbol{\Psi}_\mathcal{A}\hat{\boldsymbol{a}}_\mathcal{A}^{MLE}\right)}{2}\right) \tag{20}$$

whose mode is,

$$\left(\hat{\sigma}_\epsilon^{MLE}\right)^2 = \frac{\left(\boldsymbol{y} - \boldsymbol{\Psi}_\mathcal{A}\hat{\boldsymbol{a}}_\mathcal{A}^{MLE}\right)^T \left(\boldsymbol{y} - \boldsymbol{\Psi}_\mathcal{A}\hat{\boldsymbol{a}}_\mathcal{A}^{MLE}\right)}{N}. \tag{21}$$

The estimated vector of coefficients $\hat{\boldsymbol{a}}_\mathcal{A}^{MLE}$ is known as the maximum likelihood estimate (MLE) which differs from [18] who identified the maximum a posteriori estimate since Gaussian prior was assigned to the PCE coefficients. The error vector $\boldsymbol{\varepsilon} = \boldsymbol{y} - \boldsymbol{\Psi}_\mathcal{A}\hat{\boldsymbol{a}}_\mathcal{A}^{MLE}$ can be compared a posteriori with $\mathcal{N}(0, \left(\hat{\sigma}_\epsilon^{MLE}\right)^2)$ to ensure that the initial assumption (Gaussian likelihood) is reasonable. It is worth mentioning here that assuming Gaussian likelihood might provide non-robust estimates for some problems (e.g. when $\boldsymbol{y}$ contains outliers).

In step 3, the current sparse PCE (associated with $\mathcal{A}$) is judged better than the previous one provided that its model selection criterion (namely, KIC_m^{MLE}) is smaller than the previous one (i.e. KIC_{m-1}^{MLE}). The Kashyap information criterion (KIC) at the mth iteration is defined as follows

$$KIC_m^{MLE} = N \ln\left(\hat{\sigma}_\epsilon^{MLE}\right)^2 - P_m \ln(2\pi) - \ln\left|\left(\hat{\sigma}_\epsilon^{MLE}\right)^2 \left(\boldsymbol{\Psi}_\mathcal{A}^T \boldsymbol{\Psi}_\mathcal{A}\right)^{-1}\right| \tag{22}$$

where $P_m = \operatorname{Card}(\mathcal{A})$, $|\cdot|$ stands for determinant.

With the previous algorithm, the final maximal polynomial degree $p_\mathcal{A}$ and level of interaction $q_\mathcal{A}$ are automatically identified. Suppose that, at iteration $m-1$, elements of polynomial degree within $[p, p+2]$ and level of interaction within $[q, q+1]$ has been added to the subset of multi-indexes (step 4) but at iteration m none of these elements has been kept during the stepwise regression. Then, at the end of the mth iteration, one comes up with a subset of maximal polynomial degree $p_\mathcal{A} = p$ and level of interaction $q_\mathcal{A} = q$. In this case, the iteration stop because it is unlikely that higher-order elements would provide better sparse PCE. Otherwise, the enrichment procedure should go further.

As an illustration, suppose that at step 4, the current subset of multi-indexes for a problem with $d = 4$ variables is the subset $\mathcal{A}$ below. We note that $\mathcal{A}$ is of polynomial degree $p_\mathcal{A} = 4$ and level interaction $q_\mathcal{A} = 3$. Hence, at step 4, $\mathcal{A}$ is enriched with elements of degrees 5 and 6 (see $\mathcal{A}_c$).

$$\mathcal{A} = \left\{\begin{matrix} 0000 \\ 1000 \\ 0010 \\ 2000 \\ 0020 \\ 1010 \\ 1110 \\ 2010 \\ 1020 \\ 2020 \end{matrix}\right\},\ \mathcal{A}_c = \left\{\begin{matrix} 3110 \\ 1310 \\ 1130 \\ 4010 \\ 2210 \\ 2030 \\ 3020 \\ 1220 \\ 1040 \\ 4020 \\ 2220 \\ 2040 \end{matrix}\right\}$$

To circumvent the curse of dimensionality, we proceed as follows

- Find in $\mathcal{A}$ all elements of degree within $[p_\mathcal{A} - 1, p_\mathcal{A}]$ or level interaction $q_\mathcal{A}$ (e.g. 1110)
- Make copies of each selected element by increasing by 2 the value (i.e. polynomial degree) of each index except if the index of one variable in $\mathcal{A}$ is always 0 (e.g. 1110 has provided $3110; 1310; 1130$ and not 1112 as the fourth index is always zero).

Another feature of Bayesian sparse PCE is that the joint probability density function of all quantities are estimated (see Eqs. (17)–(21)). Therefore, the sensitivity indices (15) can be computed with uncertainty range. This is useful to assess significant differences between PCE-based computed statistics.

5. Numerical examples

5.1. PCE-RT on non-rectangular domain

We illustrate the use of polynomial chaos expansion in conjunction with the Rosenblatt transformation (method denoted PCE-RT) by considering a ten dimensional problem with an inequality constraint. The input vector $\boldsymbol{x} = (x_1, \dots, x_d) \in [0,1]^d$ is uniformly distributed within the non-rectangular domain defined by the constraint $\sum_{j=1}^{d} x_j < 1$. Sampling a joint pdf of this nature can be a challenging issue if one uses the basic acceptance–rejection sampling proposed in [62]. This is because the joint pdf of $\boldsymbol{x}$ is: $p_{\boldsymbol{x}}(\boldsymbol{x}) = d!$ if $\sum_{j=1}^{d} x_j < 1$, otherwise $p_{\boldsymbol{x}}(\boldsymbol{x}) = 0$. This means that the probability of drawing a good candidate (that satisfies the constraint) by randomly drawing from the unit hypercube is $\frac{1}{d!}$ which is very low for $d = 10$ as discussed in the present example. On the contrary, knowing the conditional cdfs one can efficiently generate random samples w.r.t. the joint pdf by using Eq. (10) (inverse RT). It can be shown that the inverse conditional cdfs in this case are,

$$\begin{cases} x_{i_1} = 1 - (1 - u_1)^{\frac{1}{d}} \\ x_{i_k} = \left(1 - (1 - u_k)^{\frac{1}{d-k+1}}\right) \prod_{j=1}^{k-1} \left(1 - u_j\right)^{\frac{1}{d-j+1}} \end{cases} \tag{23}$$

In the present work, this sample is generated with the LP$_\tau$ sequences of [45]. A sample size of $N = 1\ 024$ is chosen.

Now, let us consider the following model response: $y = \sum_{j=1}^{d} c_j \log(x_j)$. We set $d = 10$, and $c_j = 0$ for all $j < 7$, and for all $j \geq 7$ $c_j = 1$. This implies that the model structurally only depends on $(x_7, \dots, x_{10})$. It is straightforward to see that, as a function of the u-variables, the model response is of the form,

$$y = \sum_{j=1}^{d} f_j(u_j).$$

Indeed, by replacing Eq. (23) in the model response y, it can be inferred that,

$$f_j(u_j) = \log\left[\left(1 - \left(1 - u_j\right)^{\frac{1}{d-j+1}}\right)^{c_j} \cdot \left(1 - u_j\right)^{\sum_{k=j+1}^{d} \frac{c_k}{d-j+1}}\right]$$

Because the u-variables are independent of each other it can be concluded that the model is additive w.r.t. to $\boldsymbol{u}$. To verify this, the Sobol' indices of the u-variables are estimated with the PCE approach. PCEs are built upon the u-variables with the shifted-Legendre orthogonal polynomials. The results are gathered in Table 1 for two different RT transformations. They have been estimated with fair accuracy. Examining Table 1, despite the estimation error, one can easily verify that $\sum_{j=1}^{10} \hat{S}_{u_j} = 1$, whatever the transformation ordering.

Recall that the interpretation of the sensitivity indices of the u-variables as those of the x-variables depends on how the former have been obtained from the RT (Eq. (9)). In Table 1, we have reported the first-order sensitivity indices of the u-variables for two different transformations. In the first one (row #2), the sample of $\boldsymbol{u}$ has been computed after transforming the sample of $\boldsymbol{x}$ set in the canonical order $(x_1, \dots, x_{10})$. Consequently, $\hat{S}_{u_1} = \hat{S}_{x_1}$, which is the full first-order effect of x_1 that accounts for its dependence with all other variables. We note that $\hat{S}_{x_1}$ only represents 3% of the total variance. The sensitivity index $\hat{S}_{u_2}$ is of $x_2|x_1$, which is the first-order effect of x_1 which accounts for its dependence with all other variables except x_1 (in [23] the authors denote it $\hat{S}_{2-1}$). The interpretation of the other indices follows the same reasoning until the interpretation of $\hat{S}_{u_{10}}$, where the latter is simply the independent first-order effect of x_{10}, namely $\hat{S}^{ind}_{x_{10}}$. We note that the independent contribution of x_{10} to the total variance is much higher than the full contribution of x_1. This is explained by the structural independence of y to x_1. Although the model does not structurally depend on x_1, because of its dependence with the other variables, x_1 still has a substantial impact on the model response. By fixing the value of x_1, the uncertainty of the other variables would be impacted, which would marginally affect the variance of the model response (a small reduction of 3%). However, because the model is additive, fixing $(x_1, \dots, x_6)$ would lead to a significant reduction of the response variance (i.e. $\sum_{j=1}^{6} \hat{S}_{u_j} = 40\%$).

The second RT is applied to $(x_7, \dots, x_{10}, x_1, \dots, x_6)$ (Table 1, row #3). Because y has a structural dependence only on the first four variables $(x_7, x_8,\ x_9, x_{10})$, we should find $S_{(u_1,\dots,u_4)} = S_{(x_7,\dots,x_{10})} = 1$. This is confirmed by our results. Indeed, we first note that $\hat{S}_{u_j} = 0$ for $j > 4$. This is explained by the fact that, with the current RT, the sensitivity indices of these u-variables (for $j > 4$) are those of $(x_1, \dots, x_6)$ without accounting for their dependencies with $(x_7, \dots, x_{10})$. Since the former only impact y via their dependencies with the latter, their independent contributions are null. We also note that the first-order sensitivity indices of $(x_7, \dots, x_{10})$ sum-up to one as expected.

Table 1
Estimated first-order sensitivity indices of the RT-variables with the PCE-RT method.

RT ordering	$\hat{S}_{u_1}$	$\hat{S}_{u_2}$	$\hat{S}_{u_3}$	$\hat{S}_{u_4}$	$\hat{S}_{u_5}$	$\hat{S}_{u_6}$	$\hat{S}_{u_7}$	$\hat{S}_{u_8}$	$\hat{S}_{u_9}$	$\hat{S}_{u_{10}}$
$(x_1, \dots, x_{10})$	0.03	0.04	0.05	0.06	0.09	0.13	0.13	0.13	0.14	0.20
(95% CI×e^{-3})	(2)	(2)	(3)	(3)	(3)	(4)	(4)	(4)	(4)	(5)
$(x_7, \dots, x_{10}, x_1, \dots, x_6)$	0.21	0.23	0.26	0.30	0.00	0.00	0.00	0.00	0.00	0.00
(95% CI×e^{-3})	(1)	(1)	(1)	(2)	[-]	[-]	[-]	[-]	[-]	[-]

Finally, the fact that we have found that $S_{(x_7,\dots,x_{10})} = 1$ and $S^{ind}_{(x_1,\dots,x_6)} = 0$, does not allow us to infer that the model response is structurally independent of $(x_1, \dots, x_6)$. As previously discussed in Section 2, the non-uniqueness of the ANOVA decomposition (in the Sobol' sense) precludes us making this conclusion. However, this finding is still useful in that it gives information relevant to a *model reduction* setting. In effect, we can conclude that all the information in the model response is contained in the input subset (x_7, x_8, x_9, x_{10}). Therefore, prioritisation should be given to the accurate assessment of these variables in order to better estimate the model response.

5.2. PCE-NT on the Ishigami function

As a second example and to demonstrate the PCE-NT approach, let us consider the Ishigami function, which is defined as follows,

$$f(\boldsymbol{x}) = \sin x_1 + 7 \sin^2 x_2 + 0.1 x_3^4 \sin x_1 \tag{24}$$

where the input variables are uniformly distributed over $[-\pi, \pi]$, with a linear correlation between x_1 and x_3. We denote ρ_{13} the moment–product correlation coefficient. The Sobol' indices of this problem for linear correlations varying over $(-1, 1)$ were investigated in [22] and [30]. In [22] only S_{x_j} and $ST^{ind}_{x_j}$ were estimated with a Monte Carlo approach. The overall first-order and total-order sensitivity indices were assessed in [30] with the Fourier amplitude sensitivity test (the approach was called EFAST-INT). For the sake of comparison, PCE-NT is employed to estimate $(S_{x_j}, ST^{ind}_{x_j}, S^{ind}_{x_j}, ST_{x_j})$, for all $j = 1, 2, 3$. PCE-NT relies on transformation Eq. (11) which requires a vector of independent standard normal variables $\boldsymbol{z}$, due to the choice of a Gaussian copula. The latter is obtained from the isoprobabilistic transformation of the vector $\boldsymbol{u}$ uniformly distributed over $[0,1]^d$, i.e. $z_j = \Phi^{-1}(u_j)$.

A sample of $\boldsymbol{u}$ is drawn with the LP$_\tau$ sequences. PCEs are built upon the u-variables with shifted Legendre polynomials because PCEs built upon the z-variables with Hermite polynomials faced convergence issues. This is due to the fact that $\boldsymbol{u}$ is uniformly distributed like the original model input $\boldsymbol{x}$ (albeit the correlation structure) while the z-variables are non-linear transformations of the u-variables. Such a non-linear transformation can complicate the relationship between y and $\boldsymbol{z}$ as compared with $\boldsymbol{u}$. This issue has been highlighted by other authors [63].

The estimated sensitivity indices are depicted in Fig. 1. Note that we do not report the estimation errors as the indices are estimated accurately. They are in good accordance with those obtained in [30] with the EFAST-INT method. Nevertheless, while with EFAST-INT an overall number of 3 × 8 192 function calls were necessary to obtain accurate estimates of the Sobol' indices, with PCE-NT only $N = 1\ 024$ function calls were performed. This demonstrates the efficiency of the Bayesian sparse polynomial chaos expansion proposed by [18].

The results indicate that x_2 is not correlated to the other two inputs as its full and independent sensitivity indices are equal. We can also

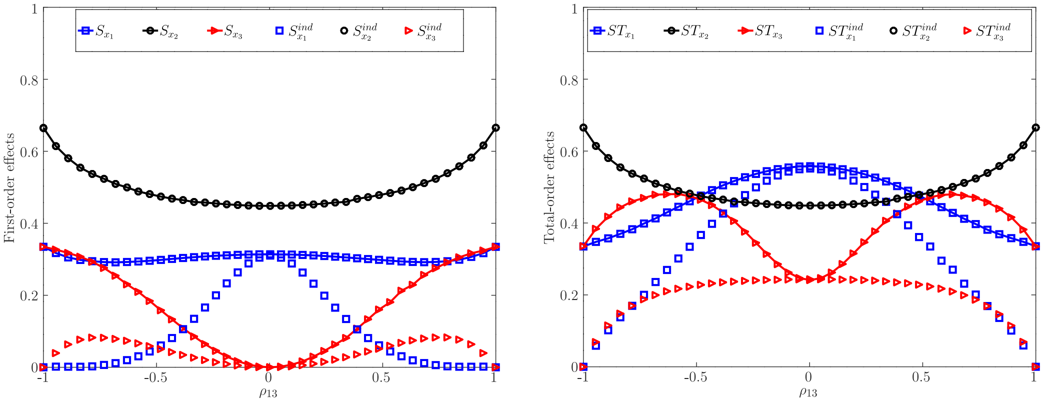

Fig. 1. First-order (left) and total-order (right) sensitivity indices versus the correlation coefficient (ρ_{13}) between x_1 and x_3 in the Ishigami function. See text for comments and explanations.

guess that x_2 does not interact with the other two inputs, because its total and first-order indices are also the same. As far as x_1 and x_3 are concerned we note that, when the correlation coefficient is zero, the full and independent sensitivity indices are equal. When the correlation coefficient is close to ±1, the independent sensitivity indices (both first and total order) of x_1 and x_3 are null because all the information in $f(\boldsymbol{x})$ is captured by only one of the pairs (x_1, x_2) or (x_2, x_3). This finding informs the analyst that the output uncertainty is explained by one of these two pairs only, thus, allowing some kind of dimensionality reduction.

5.3. PCE-MT and model calibration

5.3.1. Motivation

We consider now a case in which model input distributions are inferred from statistical calibration. Statistical calibration of computer models enhances their reliability by demonstrating that they can match observations. If the model at hand is not able to match the observations, statistical calibration can help in identifying the sources of discrepancy, if done properly. Model calibration is usually undertaken in a Bayesian framework which requires the definition of a likelihood function (characterising discrepancies between model predictions and observations) and prior uncertainties of the model inputs (independence is usually assumed). Model inputs might also include hyperparameters of discrepancy models [14,64–66]. This yields the so-called joint posterior distribution of the model inputs. Sampling random draws from the posterior distribution is not an easy job, and Markov Chain Monte Carlo (MCMC) methods are typically used for such a task [67–70].

We consider the sensitivity analysis of a drainage experiment model posterior to its calibration. This model was studied in different papers. First, [71] used this model to compare the performances of two Bayesian approaches for statistical calibration. Then, [72] performed the global sensitivity analysis of the model responses prior to its calibration. Recently, [73] used this model to illustrate the extension of the calibration method developed in [74] to the generation of stochastic random samples from a joint pdf defined in a Bayesian framework. In the present exercise, we consider the data obtained in [73]. We stress on the fact that this is a purely numerical exercise meaning that the *observations* were generated by simulations corrupted with random noise. Hereafter, we recall the setting of the statistical calibration problem.

5.3.2. The case study

In the modelled experiment, a vertical column filled with water-saturated soil is drained by imposing multistep pressure heads on either side of the column. This is achieved by moving vertically (downward), at different timesteps, a reservoir tank connected to the bottom of the column. The soil drainage experiment is modelled by several equations and initial/boundary conditions. The flow through the porous medium is governed by the following partial differential equation:

$$\frac{\partial \omega}{\partial t} = \frac{\partial}{\partial z}\left[K(h)\left(\frac{\partial h}{\partial z} - 1\right)\right] \tag{25}$$

where t [T] is time, z [L] is the vertical coordinate (positive downward), and K [L][T]$^{-1}$ is the unsaturated hydraulic conductivity. The water content ω [−] and the pressure head h [L] are the state variables. The unsaturated hydraulic conductivity $K(h)$ is modelled by the Mualem–van Genuchten (MvG) retention curve [75,76],

$$K(S_e) = k_s \cdot S_e^{\lambda}\left(1 - \left(1 - S_e^{1/m}\right)^m\right)^2 \tag{26}$$

where k_s [L][T]$^{-1}$ is the saturated hydraulic conductivity, and S_e (−) is the effective saturation defined as follows:

$$S_e = \frac{\omega - \omega_r}{\omega_s - \omega_r} = \begin{cases} \dfrac{1}{\left(1 + |\alpha h|^n\right)^m} & h < 0 \\ 1 & h \geq 0 \end{cases} \tag{27}$$

with $m = 1 - 1/n$. The soil hydraulic parameters are the saturated hydraulic conductivity k_s, the saturated water content ω_s [−], the residual water content ω_r [−] and the MvG fitting coefficients α [L]$^{-1}$, n [−] and λ [−].

The calibration exercise of this model was performed in [71]. For sake of completeness, we recall here the target joint posterior distribution which is the likelihood function (using independent uniform priors for the calibrated parameters),

$$p(\boldsymbol{x}|\boldsymbol{y}_h, \boldsymbol{y}_\omega, \sigma_h, \sigma_\omega) \propto \frac{1}{(\sigma_h \sigma_\omega)^{N_o}} \exp\left\{-\frac{1}{2}\left(\frac{SS_h(\boldsymbol{x})}{\sigma_h^2} + \frac{SS_\omega(\boldsymbol{x})}{\sigma_\omega^2}\right)\right\} \tag{28}$$

where SS_h and SS_ω are the sum of square errors of the pressure head and water content respectively while σ_h^2 and σ_ω^2 are their error variances. We recall that the data sets are simulated model outputs corrupted with Gaussian noise. With true experimental data sets, the choice of the target joint posterior distribution must be chosen with care. The random vector $\boldsymbol{x} = (k_s, \omega_r, \omega_s, \alpha, n, \lambda)$ contains the soil hydraulic parameters to be calibrated. Independent uniform priors were assumed within ranges as shown in Table 2. These uncertainty ranges were considered in the sensitivity analysis of the observed model responses prior to the calibration exercise [72]. The observations are $\boldsymbol{y}_h$, the hydraulic heads at $z = 3$ cm, and the average water content $\boldsymbol{y}_\omega$ at the same location. For both variables, $N_o = 481$ observations at different timesteps were considered.

A random sample of size $N = 512$, drawn from the joint posterior distribution Eq. (28), was generated with the numerical approach described in [73]. This allows sampling of the posterior without an explicit analytical representation of the distribution. It is worth mentioning that a MCMC sample could have been used. However, the

Table 2
Prior uncertainty ranges of hydraulic parameters; cm and minutes are used following the convention in hydrology.

Parameter	k_s (cm/min)	ω_r (–)	ω_s (–)	α (cm^{-1})	n (–)	λ (–)
Range	[0.01,0.5]	[0.01,0.20]	[0.40,0.45]	[0.005,0.02]	[1.0,1.4]	[−0.5,1.0]

sample obtained in [73] is preferred because the method used to generate it relies on the assumption defined by Eq. (13). Therefore, this sample fits the requirements of the PCE-MT approach described in Section 3.3. The random sample is depicted in Fig. 2.

The main diagonal of Fig. 2 shows the posterior marginal distributions of the inputs. We note that, while the distributions of (ω_s, α, n) appear to be close to normal, the distributions of (k_s, ω_r, λ) are skewed. The posterior ranges of (ω_s, α, n) are remarkably narrow, proving that they are satisfactorily identified. The plots below the main diagonal show the pairwise scatter plots of the generated random sample (see [73] for more details). It is clear that there are strong correlations in the sample, with some parameters being positively correlated (e.g. k_s and ω_s) and others negatively correlated (e.g. k_s and ω_r). This correlation structure is dictated by the model which is based on the physics of the system. This suggests that to fit the observed data, one must account for the complex interplay between the hydraulic parameters.

The plots above the main diagonal depict the sample where variables have been decorrelated using the approach in Eq. (12), and considering the canonical order $\boldsymbol{x} = (k_s, \omega_r, \omega_s, \alpha, n, \lambda)$. This transformation provides the decorrelated sample $\bar{\boldsymbol{x}} = (k_s, \bar{\omega}_r, \bar{\omega}_s, \bar{\alpha}, \bar{n}, \bar{\lambda})$. For instance, on row #1 column #2 k_s is plotted against $\bar{\omega}_r$, which corresponds to ω_r decorrelated from k_s. On row #2 column #3 $\bar{\omega}_r$ is plotted against $\bar{\omega}_s$, which is ω_s decorrelated from both k_s and $\bar{\omega}_r$, and so forth. We notice that the ranges of variation of $(\bar{\omega}_r, \bar{\omega}_s, \bar{\alpha}, \bar{n}, \bar{\lambda})$ have been shrunk significantly compared with their original ranges (i.e. of $(\omega_r, \omega_s, \alpha, n, \lambda)$). Notably, $\bar{\alpha}$ varies within $[-2, 1]e^{-5}$ which is virtually zero (column # 4, row # 1,2,3). Therefore, it is expected that its independent Sobol' indices would be equal to zero. The interpretation of the independent Sobol' indices Eqs. (2)–(3) holds provided that $\bar{\boldsymbol{x}}$ are effectively independent of each other. Visual inspection of the scatter plots of the decorrelated draws (upper diagonal) suggests that this is likely the case, except notably for $\bar{k}_s$ and $\bar{\alpha}$ (row #1, column #4). This implies that, the sensitivity indices of $\bar{\alpha}$ (resp $\bar{k}_s$) also contains some contribution of $\bar{k}_s$ (resp. $\bar{\alpha}$).

5.3.3. Posterior sensitivity analysis

The quantity of interest in the present study is the predicted cumulative outflow at $t = 240$ min, which is related to the quantity of water removed from the soil. It is defined as follows,

$$C_s = \int_0^L (\omega(z,0) - \omega(z,t))\,\mathrm{d}z. \tag{29}$$

To perform the posterior uncertainty and sensitivity analysis of C_s, we propagate the random sample previously discussed in Eq. (29). Because the joint posterior pdf in Eq. (28) from which the input sample has been drawn is not tractable, one cannot infer the conditional cdfs and get sets of independent samples with RT. Hence, PCEs are built upon the υ-variables obtained after orthogonalisation of the MCMC sample from Eq. (12). The associated orthogonal polynomials are inferred from Gram–Schmidt orthogonalisation. We recall that the υ-variables are not completely independent (see previous discussion about Fig. 2). Therefore, the sensitivity analysis has to be conducted with care: a visual inspection can help to gauge whether any residual dependence still exists. This limitation is however expected with any decorrelation procedure. An important remark is that PCE identification with the method of [18] is computationally cheap because of the strong correlation between the input variables that yields very sparse PCEs.

The total-order and first-order variance-based sensitivity indices are gathered in Table 3. It results that the first-order effects are equal to the total-order effects for all input variables. The full sensitivity indices are all greater than 0.81, meaning that none of them can be fixed to their best estimate without impacting the prediction of the cumulative flow. We also observe that all the independent sensitivity indices are close to zero (the highest one is $S_n^{ind} \simeq 0.04$). This means that all the hydraulic parameters mainly contribute to the model response variance through their mutual correlations. These results are related to the strong correlations observed in Fig. 2 (lower diagonal). It is then expected that only few parameters suffice to explain the predicted flow variance.

The MvG parameter α is the one which has the highest full first-order sensitivity index ($S_\alpha \simeq 0.94$, Table 3 in bold font), although k_s has a full first-order effect with almost the same value ($S_{k_s} \simeq 0.93$). The difference between S_{k_s} and S_α is likely to be insignificant if one recalls that the samples of k_s and $\bar{\alpha}$ are not completely independent (read the discussion in the previous subsection). Anyway, if we assume this difference does not matter, we can conclude that the independent mutual contribution of the other parameters to the model response variance is $S^{ind}_{(k_s,\omega_r,\omega_s,n,\lambda)} = 1 - S_\alpha \simeq 0.06$. This implies that if one were able to find the true value of α, and sample the remainder conditionally on that value, then propagating the sample through Eq. (29) would yield a reduction of 94% of the variance of the cumulative flow (because there are no interactions in the model, $S_{x_j} = ST_{x_j}$). However, given the high accuracy with which α has been estimated during the calibration process ($\alpha \in [0.008, 0.011]$ cm^{-1}, see Fig. 2 row # 4, column # 4), it is unlikely that its uncertainty can be further reduced.

What is the smallest subset of input parameters that explains the uncertainty in the predicted flow C_s? The answer to this question requires the calculation of the first-order sensitivity indices of groups of parameters until one of them is equal to one. These indices are called "closed" sensitivity indices because they are closed within a subset of inputs [77]. They are usually denoted by $S^c_{(\dots)}$ with a superscript c. Here, they are simply denoted $S_{(\dots)}$ to be consistent with our notation in Eq. (1). The sensitivity to pairs of variables (Table 4) indicates that the couplet which has the highest contribution to the response variance is (k_s, n) (with $S_{(k_s,n)} \simeq 0.97$, in bold font), although there are several other pairs with similarly high values (e.g. $S_{(\omega_s,\alpha)} \simeq 0.96$). By looking at triplets of parameters, it is found that the total response variance is almost entirely explained by (k_s, ω_r, n), since the effect of this group of parameters is $S_{(k_s,\omega_r,n)} \simeq 1$ (bold font in Table 5). We note that $(k_s, \omega_r, \omega_s)$, (k_s, ω_s, α), (k_s, α, n), (k_s, n, λ) and (ω_r, α, n) also capture most of the variance of the predicted cumulative flow as their sensitivity index is higher than 0.99. These results confirm our preliminary guess that because of the strong correlation between the input parameters, only a few explain the total variance of the predicted flow.

Table 3
Sensitivity analysis of the model of drainage experiment. Estimated individual first-order sensitivity indices of the hydraulic parameters for predictive cumulative flow. In parenthesis, the 95% credible intervals. The highest value is highlighted in bold font.

$\hat{S}_{k_s} = \hat{ST}_{k_s}$	$\hat{S}_{\omega_r} = \hat{ST}_{\omega_r}$	$\hat{S}_{\omega_s} = \hat{ST}_{\omega_s}$	$\hat{S}_\alpha = \hat{ST}_\alpha$	$\hat{S}_n = \hat{ST}_n$	$\hat{S}_\lambda = \hat{ST}_\lambda$
0.929 ($\pm 2e^{-3}$)	0.881 ($\pm 3e^{-3}$)	0.926 ($\pm 4e^{-3}$)	**0.936** ($\pm 3e^{-3}$)	0.874 ($\pm 2e^{-3}$)	0.896 ($\pm 4e^{-3}$)
$\hat{S}^{ind}_{k_s} = \hat{ST}^{ind}_{k_s}$	$\hat{S}^{ind}_{\omega_r} = \hat{ST}^{ind}_{\omega_r}$	$\hat{S}^{ind}_{\omega_s} = \hat{ST}^{ind}_{\omega_s}$	$\hat{S}^{ind}_\alpha = \hat{ST}^{ind}_\alpha$	$\hat{S}^{ind}_n = \hat{ST}^{ind}_n$	$\hat{S}_\lambda = \hat{ST}_\lambda$
$1.1e^{-3}$ ($\pm 5e^{-4}$)	$1.9e^{-3}$ ($\pm 6e^{-4}$)	0.00 ($\pm$0.00)	0.00 ($\pm$0.00)	$4.3e^{-2}$ ($\pm 3e^{-3}$)	0.00 ($\pm$0.00)

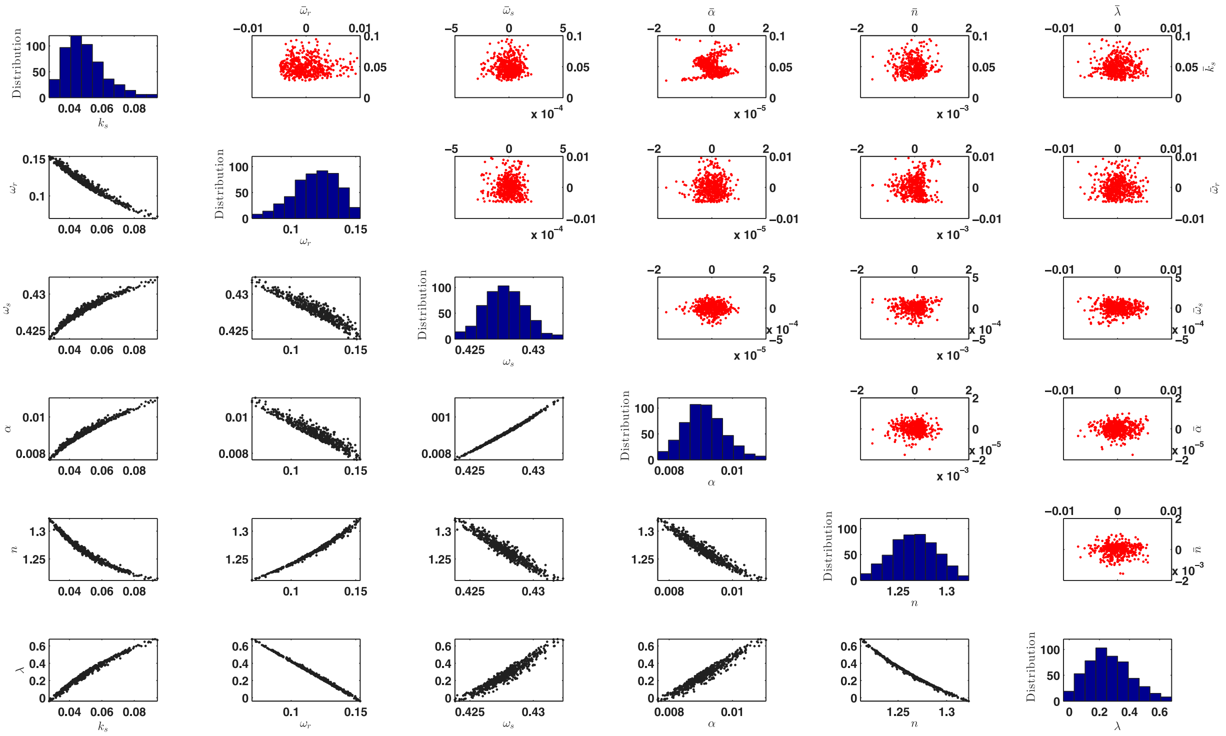

Fig. 2. Draws of the MvG hydraulic parameters generated by statistical calibration. On the diagonal, are depicted the posterior marginal distributions. On the lower diagonal are the scatter plots of the draws generated in a Bayesian framework, showing correlations in the sample. On the upper diagonal, the draws transformed with Eq. (12) are depicted. The latter are not correlated. See text for explanation.

Table 4
Sensitivity analysis of the model of drainage experiment. Estimated first-order sensitivity indices of couplets of hydraulic parameters (95% credible intervals). The highest value is highlighted in bold font.

$\hat{S}_{(k_s,\omega_r)}$	$\hat{S}_{(k_s,\omega_s)}$	$\hat{S}_{(k_s,\alpha)}$	$\hat{S}_{(k_s,n)}$	$\hat{S}_{(k_s,\lambda)}$
0.942 (±0.002)	0.958 (±0.001)	0.942 (±0.002)	**0.970** (±0.001)	0.938 (±0.002)
$\hat{S}_{(\omega_r,\omega_s)}$	$\hat{S}_{(\omega_r,\alpha)}$	$\hat{S}_{(\omega_r,n)}$	$\hat{S}_{(\omega_r,\lambda)}$	$\hat{S}_{(\omega_s,\alpha)}$
0.949 (±0.002)	0.933 (±0.002)	0.881 (±0.003)	0.929 (±0.004)	0.959 (±0.003)
$\hat{S}_{(\omega_s,n)}$	$\hat{S}_{(\omega_s,\lambda)}$	$\hat{S}_{(\alpha,n)}$	$\hat{S}_{(\alpha,\lambda)}$	$\hat{S}_{(n,\lambda)}$
0.959 (±0.003)	0.943 (±0.004)	0.947 (±0.002)	0.944 (±0.002)	0.950 (±0.002)

Table 5
Sensitivity analysis of the model of drainage experiment. Estimated first-order sensitivity indices of triplets of hydraulic parameters. The highest value is highlighted in bold font.

$\hat{S}_{(k_s,\omega_r,\omega_s)}$	$\hat{S}_{(k_s,\omega_r,\alpha)}$	$\hat{S}_{(k_s,\omega_r,n)}$	$\hat{S}_{(k_s,\omega_r,\lambda)}$	$\hat{S}_{(k_s,\omega_s,\alpha)}$
0.990 $\pm 7e^{-4}$	0.942 $\pm 2e^{-3}$	**0.999** $\pm 1e^{-4}$	0.944 $\pm 2e^{-3}$	0.998 $\pm 8e^{-4}$
$\hat{S}_{(k_s,\omega_s,n)}$	$\hat{S}_{(k_s,\omega_s,\lambda)}$	$\hat{S}_{(k_s,\alpha,n)}$	$\hat{S}_{(k_s,\alpha,\lambda)}$	$\hat{S}_{(k_s,n,\lambda)}$
0.979 $\pm 1e^{-3}$	0.984 $\pm 1e^{-3}$	0.997 $\pm 4e^{-4}$	0.942 $\pm 2e^{-3}$	0.995 $\pm 4e^{-4}$
$\hat{S}_{(\omega_r,\omega_s,\alpha)}$	$\hat{S}_{(\omega_r,\omega_s,n)}$	$\hat{S}_{(\omega_r,\omega_s,\lambda)}$	$\hat{S}_{(\omega_r,\alpha,n)}$	$\hat{S}_{(\omega_r,\alpha,\lambda)}$
0.959 $\pm 2e^{-3}$	0.995 $\pm 7e^{-4}$	0.940 $\pm 3e^{-3}$	0.995 $\pm 5e^{-4}$	0.935 $\pm 2e^{-3}$
$\hat{S}_{(\omega_r,n,\lambda)}$	$\hat{S}_{(\omega_s,\alpha,n)}$	$\hat{S}_{(\omega_s,\alpha,\lambda)}$	$\hat{S}_{(\omega_s,n,\lambda)}$	$\hat{S}_{(\alpha,n,\lambda)}$
0.978 $\pm 2e^{-3}$	0.977 $\pm 2e^{-3}$	0.984 $\pm 2e^{-3}$	0.981 $\pm 2e^{-3}$	0.976 $\pm 2e^{-3}$

6. Conclusion

Global sensitivity analysis of models with dependent inputs is a challenging issue. The reason is that even if a model response is structurally independent of a given input (by “structurally” we mean in its mathematical definition), it can appear to be sensitive to that input if it is strongly correlated to others. Computing the independent sensitivity indices helps to identify the input variables that are structurally linked to the model response of interest. This is the case when the independent total-order sensitivity index of an input is not null. However, if its independent total-order sensitivity index is null, it cannot be inferred that the model response does not structurally depend on that input. To conclude this, one has to perform the ANOVA decomposition by assuming independence, because in that case the ANOVA decomposition is unique.

In the present work, we have used the Polynomial Chaos Expansion (PCE) approach to estimate variance-based sensitivity indices. For this purpose, three transformations have been used: namely, the Rosenblatt transformation [44], the Nataf transformation [46] and [23] transformation. The decision of which transformation to use is problem-dependent. For example, the Rosenblatt transformation should be preferred if the conditional distributions are known. The Nataf transformation is adequate when the input uncertainty is defined by their marginal distributions and their moment–product correlation matrix. With a given input sample, the [23] transformation is appropriate, provided that Eq. (13) holds.

The case studies here have demonstrated the efficiency of the proposed methodology, allowing the estimation of any variance-based sensitivity indices from one single sample. Finally, the sensitivity analysis of a drainage model posterior to its calibration highlights the importance of defining the objective of the sensitivity analysis, and thereby computing the sensitivity indices that are appropriate for the task, as recommended in [77]. For the calibrated drainage model, the focus was on identifying the smallest subset of inputs that mostly explain the total variance of the predicted cumulative flow. It was found that the solution to this question is not unique. Several subsets of inputs of identical cardinality can explain the same amount of the output variance.

CRediT authorship contribution statement

Thierry A. Mara: Conceptualization, Methodology, Software, Writing (original and revisions). **William E. Becker:** Writing (original and revisions).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix. Algorithm of PCE with dependent inputs

Let $\mathbf{x}$ be an $N \times d$ dependent sample and let $\mathbf{y}$ be the associated $N \times 1$ vector of model responses. The input sample may have been generated from either the Rosenblatt transformation Eq. (10), the Nataf transformation (11) or the transformation of Mara & Tarantola (13). The generation of $\mathbf{x}$ is obtained from an independent sample denoted $\bar{\mathbf{x}}$ of size $N \times d$. Depending on the method chosen to sample $\boldsymbol{x}$, we have $\bar{\mathbf{x}} \in (\mathbf{u}, \mathbf{z}, \mathbf{v})$ (see Section 4.2 for more details). The algorithm to compute the (full and independent) first-order and total-order sensitivity indices is as follows,

1. Set $i = 1$ and $\mathbf{x} = [\mathbf{x}_1, \dots, \mathbf{x}_d]$, where $\mathbf{x}_j$ is the sample of x_j
2. Depending on the method chosen, find $\bar{\mathbf{x}}$ from either Eq. (9), Eq. (11) or Eq. (12)
3. Identify the Bayesian sparse PCE associated with $(\bar{\mathbf{x}}, \mathbf{y})$ and compute the associated first-order and total sensitivity indices [18]. Let us denote them $\bar{S}_j$ and $\bar{ST}_j$, $j = 1, \dots, d$
4. Set $S_i = \bar{S}_1$, $ST_i = \bar{ST}_1$, $S^{ind}_{i-1} = \bar{S}_d$ and $ST^{ind}_{i-1} = \bar{ST}_d$ with the following convention, $S^{ind}_d = S^{ind}_0$ and $ST^{ind}_d = ST^{ind}_0$.
5. Set $i = i + 1$. If $i = d + 1$ stop. Otherwise set $\mathbf{x} = [\mathbf{x}_i, \dots, \mathbf{x}_d, \mathbf{x}_1, \dots, \mathbf{x}_{i-1}]$ and resume from 2

References

[1] Saltelli A, Ratto M, Andres T, Campolongo F, Cariboni J, Gatelli D, et al. Global sensitivity analysis: the primer. Probability and statistics, Chichester: John Wiley and Sons; 2008.

[2] Strong M, Oakley JE. When is a model good enough? Deriving the expected value of model improvement via specifying internal model discrepancies. SIAM/ASA J Uncertain Quantif 2014;2:106–25.

[3] European Commission. Better regulation guidelines. 2017.

[4] Sobol' IM. Sensitivity estimates for nonlinear mathematical models. Math Mod Comput Exp 1993;1:407–14.

[5] Saltelli A, Chan K, Scott EM. Sensitivity analysis. Chichester: John Wiley and Sons; 2000.

[6] Saltelli A, Tarantola S, Campolongo F, Ratto M. Sensitivity analysis in practice. Probability and statistics, Chichester: John Wiley and Sons; 2004.

[7] Archer G, Saltelli A, Sobol' IM. Sensitivity measures ANOVA like techniques and use of bootstrap. J Stat Comput Simul 1997;58:99–120.

[8] Saltelli A. Making best use of model evaluations to compute sensitivity indices. Comput Phys Commun 2002;145:280–97.

[9] Gamboa F, Janon A, Klein T, Lagnoux A, Prieur C. Statistical inference for Sobol pick-freeze Monte Carlo method. Statistics 2016;50:881–902.

[10] Cukier RI, Fortuin CM, Shuler KE, Petschek AG, Schaibly JH. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. I. theory. J Chem Phys 1973;59:3873–8.

[11] Cukier RI, Schaibly JH, Shuler KE. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. iii. analysis of the approximations. J Chem Phys 1975;63:1140–9.

[12] Cukier RI, Levine RI, Shuler KE. Nonlinear sensitivity analysis of multiparameter model systems. J Comput Phys 1978;26:1–42.

[13] Saltelli A, Tarantola S, Chan K. A quantitative model independent method for global sensitivity analysis of model output. Technometrics 1999;41:39–56.

[14] Oakley JE, O'Hagan A. Probabilistic sensitivity analysis of complex models: a Bayesian approach. J R Stat Soc B 2004;66:751–69.

[15] Sudret B. Global sensitivity analysis using polynomial chaos expansions. Reliab Eng Syst Saf 2008;93:964–79.

[16] Blatman G, Sudret B. Adaptive sparse polynomial chaos expansion based on least angle regression. J Comput Phys 2011;230:2345–67.

[17] Buzzard GT, Xiu D. Variance-based global sensitivity analysis via sparse-grid interpolation and cubature. Commun Comput Phys 2011;9:542–67.

[18] Shao Q, Younes A, Fahs M, Mara TA. Bayesian sparse polynomial chaos expansion for global sensitivity analysis. Comput Methods Appl Mech Eng 2017;318:474–96.

[19] Mara TA, Tarantola S, Annoni P. Non-parametric methods for global sensitivity analysis of model output with dependent inputs. Environ Modell Softw 2015;72:173–83.

[20] Xu C, Gertner GZ. Uncertainty and sensitivity analysis for models with correlated parameters. Reliab Eng Syst Saf 2008;93:1563–73.

[21] Li G, Rabitz H, Yelvington PE, Oluwole OO, Bacon F, Kolb CE, et al. Global sensitivity analysis for systems with independent and/or correlated inputs. J Phys Chem 2010;114:6022–32.

[22] Kucherenko S, Tarantola S, Annoni P. Estimation of global sensitivity indices for models with dependent variables. Comput Phys Comm 2012;183:937–46.

[23] Mara TA, Tarantola S. Variance-based sensitivity indices for models with dependent inputs. Reliab Eng Syst Saf 2012;107:115–21.

[24] Most T. Variance-based sensitivity analysis in the presence of correlated input variables, lecture. Dynamic software and engineering, 2012, http://www.dynardo.de.

[25] Chastaing G, Gamboa F, Prieur C. Generalized Hoeffding–Sobol decomposition for dependent variables – application to sensitivity analysis. Electron J Stat 2012;6:2420–48.

[26] Zhou C, Lu Z, Zhang L, Hu J. Moment independent sensitivity analysis with correlations. Appl Math Model 2014;38:4885–96.

[27] Song E, Nelson B, Staum J. Shapley effects for global sensitivity analysis: Theory and computation. SIAM/ASA J Uncertain Quantif 2016;4:1060–83.

[28] Owen A, Prieur C. On Shapley value for measuring the importance of dependent inputs. SIAM J Uncertain Quantif 2017;5:986–1002.

[29] Ge Q, Menendez M. Extending Morris method for qualitative global sensitivity analysis of models with dependent inputs. Reliab Eng Syst Saf 2017;162:28–39.

[30] Tarantola S, Mara TA. Variance-based sensitivity indices of computer models with dependent inputs: the Fourier amplitude sensitivity test. Int J Uncertain Quantif 2017;7:511–23.

[31] Caniou Y. Analyse de sensibilité globale pour les modèles imbriqués et multiéchelles [Ph.D. thesis], Clermont-Ferrand: University Blaise Pascal; 2012.

[32] Sudret B, Caniou Yves. Analysis of covariance (ANCOVA) using polynomial chaos expansions. In: Deodatis G, editor. Proc. 11th international conference on structural safety and reliability. New York, USA: 2013.

[33] Caniou Y, Sudret B. Covariance-based sensitivity indices based on polynomial chaos functional decomposition. In: 7th international conference on sensitivity analysis of model output. Nice, France: 2013.

[34] Zuniga M, Kucherenko S, Shah N. Metamodelling with independent and dependent inputs. Comput Phys Comm 2013;184(6):1570–80.

[35] Shapley Lloyd S. A value for n-person game. In: Khun HW, Tucker AW, editors. Contributions to the theory of games II. Anals of mathematics studies, vol. 28, Princeton Univ. Press NJ; 1953, p. 307–17.

[36] Iooss B, Prieur C. Shapley effects for sensitivity analysis with correlated inputs: comparisons with Sobol' indices, numerical estimation and applications. Int J Uncertain Quantif 2019;9:493–514.

[37] Li G, Rabitz H. Relationship between sensitivity indices defined by variance- and covariance-based methods. Reliab Eng Syst Saf 2017;167:136–57.

[38] Helton JC, Johnson JD, Sallaberry CJ, Storlie CB. Survey of sampling-based methods for uncertainty and sensitivity analysis. Reliab Eng Syst Saf 2006;91:1175–209.

[39] Nelsen RB. An introduction to copulas. Springer series in statistics, 2nd ed. 2006.

[40] Borgonovo E. Measuring uncertainty importance: Investigation and comparison of alternative approaches. Risk Anal 2006;26:1349–61.

[41] Botev ZI, Grotowski JF, Kroese DP. Kernel density estimation via diffusion. Ann Statist 2010;38:2916–57.

[42] Baucells M, Borgonovo E. Invariant probabilistic sensitivity analysis. Manage Sci 2013;59:2536–49.

[43] Da Veiga S. Global sensitivity analysis with dependence measures. J Stat Comput Simul 2015;85:1283–305.

[44] Rosenblatt M. Remarks on the multivariate transformation. Ann Math Stat 1952;43:470–2.

[45] Sobol' IM, Turchaninov VI, Levitan YuL, Shukman BV. Quasi-random sequence generator. Keldysh Institute of Applied Mathematics, Russian Academy of Sciences; 1992.

[46] Nataf A. Détermination des distributions dont les marges sont données. C R Acad Sci 1962;225:42–3.

[47] Lebrun R, Dutfoy A. Do Rosenblatt and Nataf isoprobabilistic transformations really differ? Probab Eng Mech 2009;24:577–84.

[48] Liu PL, Der Kiureghian A. Multivariate distribution models with prescribed marginals and covariances. Probab Eng Mech 1986;1:105–12.

[49] Iman RI, Conover WJ. A distribution-free approach to inducing rank correlation among input variables. Commun Statist Simulation Comput 1982;11:311–34.

[50] Hastie TJ, Tibshirani RJ. Generalized additive models. London: Chapman & Hall; 1990.

[51] Lewandowski D, Cooke R, Tebbens RJD. Sample-based estimation of correlation ratio with polynomial approximation. ACM Trans Model Comput Simul 2007;18:1–16.

[52] Ratto M, Pagano A, Young P. State dependent parameter metamodelling and sensitivity analysis. Comput Phys Comm 2007;117:863–76.

[53] Mara TA, Rakoto Joseph O. Comparison of some efficient methods to evaluate the main effect of computer model factors. J Stat Comput Simul 2008;78:167–78.

[54] Fajraoui N, Ramasomanana F, Younes A, Mara TA, Ackerer P, Guadagnini A. Use of global sensitivity analysis and polynomial chaos expansion for interpretation of nonreactive transport experiments in laboratory-scale porous media. Water Resour Res 2011;47:W02521. http://dx.doi.org/10.1029/2010WR009639.

[55] Ciriello V, Di Federico V, Riva M, Cadini F, De Sanctis J, Zio E, et al. Polynomial chaos expansion for global sensitivity analysis applied to a model of radionuclide migration in a randomly heterogeneous aquifer. Stoch Environ Res Risk Assess 2012.

[56] Sochala P, Le Maître OP. Polynomial chaos expansion for subsurface flows with uncertain soil parameters. Adv Water Resour 2013;62:139–54.

[57] Riva M, Guadagnini A, Dell'Oca A. Probabilistic assessment of seawater intrusion under multiple sources of uncertainty. Adv Water Resour 2015;75:93–104.

[58] Rajabi MM, Ataie-Ashtaiani B, Simmons CT. Polynomial chaos expansions for uncertainty propagation and moment independent sensitivity analysis of seawater intrusion simulations. J Hydrol 2015;520:101–22.

[59] Xiu D, Karniadakis GE. The Wiener–Askey polynomial chaos for stochastic differential equations. SIAM J Sci Comput 2002;24:619–44.

[60] Kashyap RL. Optimal choice of AR and MA parts in autoregressive moving average models. IEEE Trans Pattern Anal Mach Intell 1982;4:99–104.

[61] Rahman S. Extended polynomial dimensional decomposition for arbitrary probability distributions. J Eng Mech 2009;135:1439–51.

[62] Kucherenko S, Klymenko OV, Shah N. Sobol' indices for problems defined in non-rectangular domains. Reliab Eng Syst Saf 2017;167:218–31.

[63] Torre E, Marelli S, Embrechts P, Sudret B. Data-driven polynomial chaos expansion for machine learning regression. J Comput Phys 2019;388:601–23.

[64] Bayarri MJ, Berger JO, Paulo R, Sacks J, Cafeo JA, Cavendish J, et al. A framework for validation of computer models. Technometrics 2007;49:138–54.

[65] Renard B, Kavetsky D, Kuczera G, Thyer M. Understanding predictive uncertainty in hydrology modeling: The challenge of identifying input and structural errors. Water Resour Res 2010;46:1–22.

[66] Xu T, Valocchi AJ. A Bayesian approach to improved calibration and prediction of groundwater models with structural error. Water Resour Res 2015;51:9290–311.

[67] Metropolis N-A, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equations of state calculations by fast computing machines. J Chem Phys 1953;21:1087–91.

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

[69] Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. Statist Sci 1992;7:457–511.

[70] Robert CP, Casella G. Monte Carlo statistical method. 2nd ed. Springer series in statistics. New York: 2004.

[71] Mara TA, Delay F, Lehmann F, Younes A. A comparison of two Bayesian approaches for uncertainty quantification. Environ Modell Softw 2016;82:21–30.

[72] Mara TA, Belfort B, Fontaine V, Younes A. Addressing factors fixing setting from given data: A comparison of different methods. Environ Modell Softw 2017;87:29–38.

[73] Mara TA, Fahs M, Shao Q, Younes A. Random sampling from joint probability distributions defined in a bayesian framework. SIAM J Sci Comput 2019;41:A316–38.

[74] Mara TA, Fajraoui N, Younes A, Delay F. Inversion and uncertainty of highly parameterized models in a Bayesian framework by sampling the maximal conditional posterior distribution of parameters. Adv Water Resour 2015;76:1–10.

[75] Mualem Y. A new model for predicting the hydraulic conductivity of unsaturated porous media. Water Resour Res 1976;12:513–22.

[76] van Genuchten MTh. A closed form equation for predicting the hydraulic properties of unsaturated soils. Soil Sci Am J 1980;44:892–8.

[77] Saltelli A, Tarantola S. On the relative importance of input factors in mathematical models: Safety assessment for nuclear waste disposal. J Amer Statist Assoc 2002;97:702–9.