Dimension reduction for uncertainty propagation and global sensitivity analyses of a cesium adsorption model

Pierre Sochala, Christophe Chiaberge, Francis Claret, Christophe Tournassat

Journal of Computational Science 75 (2024) 102197

Contents lists available at ScienceDirect

Journal of Computational Science

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

Dimension reduction for uncertainty propagation and global sensitivity analyses of a cesium adsorption model

Pierre Sochala [a,*], Christophe Chiaberge [b], Francis Claret [b], Christophe Tournassat [c,d]

[a] CEA, DAM, DIF, F-91297 Arpajon, France
[b] BRGM, 3 avenue Claude Guillemin, 45060 Orléans, France
[c] ISTO, Université d'Orléans-CNRS-BRGM, France
[d] Lawrence Berkeley National Laboratory, Berkeley, CA, USA

ARTICLE INFO

Keywords:
Radionuclide migration
System of solvers
Partial least squares
Arbitrary polynomial chaos
Closed and Cramer–Von Mises indices

ABSTRACT

This paper presents an efficient method to perform uncertainty and sensitivity analyses in a cesium adsorption model upstream chained with a pore water composition model. As the number of uncertain input parameters is about twenty for each of the two models, a dimension reduction technique is implemented to build a polynomial approximation of the cesium distribution coefficient in a reduced subspace. Two approaches are tested depending on the water composition and adsorption models are treated as a single block or two separate blocks. In view of assessing the robustness of the approaches, three initial cesium concentrations are considered to explore different regimes of the adsorption model. The interpretation of the linear transformations projecting the original inputs to the reduced coordinates is broadly consistent with the geochemical features of the model. Validation results show that the relative error levels of the surrogate models are around a few percent for both approaches with only one thousand realizations of the chained model. Global sensitivity analysis highlights that the variance of the cesium distribution coefficient is overwhelmingly governed by the adsorption model. Still, this conclusion is nuanced when considering the whole cumulative distribution function for which the interaction effects between the two models account for a fifth.

1. Introduction

Adsorption processes play a major role in the prediction of aqueous species migration in the geosphere. The safety arguments in support of many radioactive waste repository concepts are heavily relying on the existence of adsorption reactions in the geological formation and the multi-barriers systems [1–3]. Cation exchange and surface complexation models can be used to quantify the adsorption of radionuclides as a function of specific geochemical conditions that are considered to be representative of *in situ* conditions [4–6]. However, uncertainties in these model predictions must also be evaluated to inform performance assessment calculations. Uncertainty and sensitivity analyses make it also possible to identify the most important model parameters affecting uncertainty in radionuclide adsorption and help design new laboratory experiments [7]. In the present study, the emphasis is put on the use of surrogate model to underpin the influences of pore water chemistry on a cesium adsorption model outcome. Two difficulties must be addressed to estimate uncertainties of cesium adsorption models, namely the upstream chaining with a pore water composition model and the plentiful number of input parameters (≃40).

The use of surrogate models has developed significantly in the uncertainty quantification community and is now widely disseminated in many disciplines. In particular, some works have been devoted to the design of surrogate models for coupled problems described by systems of solvers (or codes). These works may be based on Gaussian processes [8,9] or polynomial approximations [10,11]. However, constructing and validating a surrogate model become intractable when the dimensionality increases because the computational complexity grows exponentially with the number of inputs as stated by the curse of dimensionality [12]. This issue has motivated the development of dimensionality reduction approaches with a mapping of the original high dimensional input parameters space to a suitable lower-dimensional subspace. A dimension reduction-based surrogate model construction relies on two key ingredients, which are an identification of the reduced subspace and an approximation in this subspace. The reduction step can be supervised or unsupervised whether the output is used to estimate the reduced space or not. The most widespread linear unsupervised technique is the principal component analysis but other possibly nonlinear methods have been developed using machine learning approaches

* Corresponding author.
E-mail address: pierre.sochala@cea.fr (P. Sochala).

https://doi.org/10.1016/j.jocs.2023.102197
Received 28 August 2023; Received in revised form 7 November 2023; Accepted 4 December 2023
Available online 6 December 2023

(see [13] for a review). Among the supervised techniques, the active subspace method [14,15] has received much attention to identify the reduced space using eigenpairs of a matrix derived from the output gradient with respect to the input parameters. In this study, we implemented a dimension reduction technique that linearly combines the input parameters to get a few reduced coordinates able to describe the main variations of the quantity of interest. We opted for the Partial Least Squares (PLS) method that determines the reduced variables without resorting to the model output gradient. Two strategies of surrogate model construction were tested, with the two chained models treated as a single block or two separate blocks. Then, we conducted a global sensitivity analysis to identify the respective contributions of each of the two models. This sensitivity study for groups of parameters was drawn on the variance-based indices as well as on indices defined with the conditional cumulative distribution function.

The paper is structured as follows. In Section 2, we present the general modeling framework including the geochemical model, the set of uncertain input parameters, and the different approaches implemented to build surrogate models. In Section 3, we detail the dimension reduction method combined with the polynomial expansion used to approximate the cesium distribution coefficient. Validation results are discussed in Section 4, and global sensitivity indices are analyzed in Section 5.

2. Framework

This section provides an overview of the present study with its issues related to the uncertainty propagation exercise. First, the geochemical system is introduced, it is a cesium adsorption model upstream chained with a pore water composition model. Second, the uncertain input parameters of each model and the linking parameters are listed with their ranges of variation. Third, the two approaches for building surrogate model are presented; one approach treats the different models as a single one whereas the other approach considers each model separately.

2.1. Cesium distribution estimation

We are interested in the cesium (Cs^+) distribution coefficient K_D [L kg^{-1}] which is representative of the cesium adsorbed on the solid normalized to the cesium remaining at equilibrium in solution. The cesium K_D is calculated with an adsorption model of which a part of the input parameters originates from an independent pore water chemical composition model. A complete description of the model used to calculate the pore water chemical composition in the Callovian-Oxfordian claystone can be found in [16], and for which we conducted an uncertainty propagation study in [17]. The input parameters of this model are the Cl^- and SO_4^{2-} total concentration obtained from core sample leaching measurements, the measured sodium Na^+, potassium K^+, calcium Ca^{2+}, magnesium Mg^{2+}, and strontium Sr^{2+} exchangeable concentrations, the related Na^+/K^+, Na^+/Ca^{2+}, Na^+/Mg^{2+}, Na^+/Sr^{2+} cation exchange selectivity coefficients, and the solubilities of Celestite, Calcite, Dolomite, Goethite, Quartz, Pyrite, Ripidolite, and Illite. The reference values of these $N_1 = 19$ parameters are reported in Table 1. The cesium adsorption model is based on a cation exchange model taking into account only two clay mineral phases illite and smectite (montmorillonite) [18], see [4] for a complete description. The model is briefly presented and made available in the form of a PHREEQC v3.5.0 [19] input file and its associated database (THERMOCHIMIE v9b [20]). The input parameters of the adsorption model contain the properties of the montmorillonite and illite planar sites, and the illite type II sites and Frayed Edge Sites (FES), which are relevant for the adsorption of Cs^+ present at trace concentration in the aqueous phase. The reference values of these $N_2 = 18$ parameters are reported in Table 2.

The computational software chain of the two previous models forms a directed system of solvers, meaning that the solvers can be ordered and the information can only be transferred forward in the system. This aspect is schematically illustrated in Fig. 1 where the outputs of the water composition model are a part of the inputs of the adsorption model. In the following, the global input parameters of the water composition model and the adsorption model are collected into N_1-dimensional and N_2-dimensional vectors denoted $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ respectively while the linking parameters are regrouped into a $N_{1,2}$-dimensional vector $\boldsymbol{\xi}_{1,2}$. Formally speaking, the global inputs are defined as the inputs of the models that are not an output of another model.

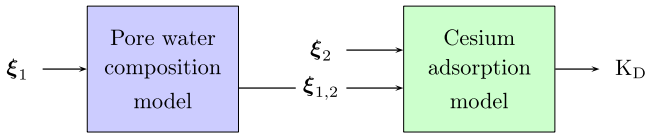


Fig. 1. Computational chain of the cesium distribution coefficient estimation.

Table 1
List of the N_1 uncertain input parameters of the pore water composition model with their mean, minimal and maximal values.

#	Type	Species	Unit	Mean	Min	Max
1	Leached parameter	Cl^-	mmol L^{-1}	41	37.4	44.6
2		SO_4^{2-}	mmol L^{-1}	66	60.3	71.7
3	Exchanged cation	Na^+	mol L^{-1}	1.0824	0.99	1.17
4		K^+	mol L^{-1}	0.417	0.38	0.45
5		Ca^{2+}	mol L^{-1}	1.549	1.41	1.69
6		Mg^{2+}	mol L^{-1}	0.602	0.55	0.65
7		Sr^{2+}	mol L^{-1}	0.0737	0.067	0.081
8	Selectivity coefficients ($\log K_{ex}$ value)	Na^+/K^+	–	1.2	1.03	1.37
9		Na^+/Ca^{2+}	–	0.7	0.53	0.87
10		Na^+/Mg^{2+}	–	0.7	0.53	0.87
11		Na^+/Sr^{2+}	–	0.6	0.43	0.77
12	Solubility ($\log K$ value)	Celestite	–	−6.62	−6.71	−6.53
13		Calcite	–	−8.48	−8.57	−8.39
14		Dolomite	–	−17.12	−17.5	−16.8
15		Goethite	–	0.39	0.044	0.74
16		Quartz	–	−3.74	−3.83	−3.65
17		Pyrithe	–	−58.78	−59.1	−58.4
18		Ripidolite	–	61.35	60.5	62.2
19		Illite	–	11.54	10.7	12.4

2.2. Input parameters perturbation

The N_1 inputs of the water composition model and the $N_2 + N_{1,2} = 18 + 5$ inputs of the adsorption model are listed in Tables 1 and 2 with their mean and extreme values. The ranges of variation have been chosen as a proportion of the mean values with the aim of generating plausible perturbations. In addition, the parameters are assumed to be uniformly distributed except the components of $\boldsymbol{\xi}_{1,2}$ whose probability distributions are unknown a priori since they are outputs of the pore water composition model. Under the assumption of independence, the parametric domain $\mathbb{X}_{\boldsymbol{\xi}} \subseteq \mathbb{R}^N$ and the probability distribution $p_{\boldsymbol{\xi}} : \mathbb{X}_{\boldsymbol{\xi}} \to \mathbb{R}^+$ of the random vector $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_N)$ have a product form,

$$\mathbb{X}_{\boldsymbol{\xi}} = \bigtimes_{i=1}^{N} I(\xi_i), \quad \text{and} \quad p_{\boldsymbol{\xi}}(\boldsymbol{\xi}) = \prod_{i=1}^{N} p_i(\xi_i),$$

where $p_i(\cdot)$ is the marginal distribution of the variable ξ_i with support $I(\xi_i) \subseteq \mathbb{R}$.

2.3. Surrogate modeling

Two approaches exist to build surrogate models of a system of solvers, namely the monolithic and the fragmented approaches [8]. On the one hand, the monolithic (or black-box) approach handles the computational software chain as a single block. The advantage of this approach is to rely on the global inputs, whose probability distributions

Table 2
List of the N_2 uncertain input parameters of the adsorption model with their mean, minimal and maximal values as well as their global numbering $\#_{gl}$. The $N_{1,2}$ input parameters stemming from the water composition model are also reported but their mean and extreme values are not available a priori.

#	$\#_{gl}$	Type\Clay	Species	Unit	Mean	Min	Max
1	20		$[\cdot]$	mol kg^{-1}	1.3 10^{-1}	9.9 10^{-2}	1.6 10^{-1}
2	21		K^+	–	1.1	0.8	1.4
3	22	Planar	Ca^{2+}	–	0.6	0.3	0.9
4	23	montmorillonite	Mg^{2+}	–	0.6	0.3	0.9
5	24		Sr^{2+}	–	0.3	0.0	0.6
6	25		Cs^+	–	1.7	1.4	2.0
7	26		$[\cdot]$	mol kg^{-1}	5.0 10^{-2}	3.0 10^{-2}	6.9 10^{-2}
8	27		K^+	–	1.2	0.9	1.5
9	28	Planar	Ca^{2+}	–	0.7	0.4	1.0
10	29	illite	Mg^{2+}	–	0.7	0.4	1.0
11	30		Sr^{2+}	–	0.7	0.4	1.0
12	31		Cs^+	–	1.6	1.3	1.9
13	32		$[\cdot]$	mol kg^{-1}	8.0 10^{-3}	4.0 10^{-3}	1.2 10^{-2}
14	33	Type II illite	K^+	–	2.1	1.8	2.4
15	34		Cs^+	–	3.6	3.3	3.9
16	35		$[\cdot]$	mol kg^{-1}	1.0 10^{-4}	5.0 10^{-5}	1.5 10^{-4}
17	36	"FES" illite	K^+	–	2.4	2.1	2.7
18	37		Cs^+	–	7.0	6.7	7.3
#		Type	Species	Unit	Mean	Min	Max
1			Na^+	mol L^{-1}	N/A	N/A	N/A
2			K^+	mol L^{-1}	N/A	N/A	N/A
3		Exchanged cations	Ca^{2+}	mol L^{-1}	N/A	N/A	N/A
4			Mg^{2+}	mol L^{-1}	N/A	N/A	N/A
5			Sr^{2+}	mol L^{-1}	N/A	N/A	N/A

can be chosen a priori. This approach, however, can be difficult to implement when the solvers are created and maintained by distinct teams. On the other hand, the fragmented approach deals with each solver of the computational chain separetely. This splitting is more attractive than the monolithic approach when the dimensionality of each solver is lower than the dimensionality of the global inputs because it moderates the curse of dimensionality. Nevertheless, this approach uses the linking parameters between the solvers, whose probability distributions are unknown a priori.

Both approaches are tested in this paper in order to evaluate the robustness of the dimension reduction method for approximating the model output $c = K_D$. The monolithic approach provides an approximation $\tilde{c}$ of c by solving a single problem in dimension $N_1 + N_2 = 37$,

$$c(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) \simeq \tilde{c}(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2), \quad \tilde{c} : \mathbb{X}_{\xi_1} \times \mathbb{X}_{\xi_2} \to \mathbb{R}.$$

The fragmented approach produces an approximation $\tilde{\tilde{c}}$ of c by solving $N_{1,2} = 5$ problems in dimension $N_1 = 19$,

$$\boldsymbol{\xi}_{1,2}(\boldsymbol{\xi}_1) \simeq \tilde{\boldsymbol{\xi}}_{1,2}(\boldsymbol{\xi}_1), \quad \tilde{\boldsymbol{\xi}}_{1,2} : \mathbb{X}_{\xi_1} \to \Xi_{\xi_{1,2}},$$

and one problem in dimension $N_2 + N_{1,2} = 23$,

$$c(\boldsymbol{\xi}_2, \boldsymbol{\xi}_{1,2}) \simeq \tilde{\tilde{c}}(\boldsymbol{\xi}_2, \tilde{\boldsymbol{\xi}}_{1,2}), \ \tilde{\tilde{c}} : \mathbb{X}_{\xi_2} \times \Xi_{\xi_{1,2}} \to \mathbb{R},$$

where the parametric domain $\Xi_{\xi_{1,2}} \subset \mathbb{R}^5$ of $\boldsymbol{\xi}_{1,2}$ is unknown a priori and does not necessarily have a product form. The advantage regarding the dimensionality decrease offered by the fragmented approach is seen in Table 3 where the number of terms in Polynomial Chaos (PC) expansions using a basis of total degree $d°$ is reported for both approaches. This table shows that it would require $\mathcal{O}(10^4)$ to $\mathcal{O}(10^6)$ model evaluations to fit the PC coefficients with the standard least squares method. Of course, there is extensive literature on sparse regression where most of the PC coefficients are equal to zero [21] but we decide to follow an alternative route using a dimensionality reduction approach.

Table 3
Number of terms $N_b = (N + d°)!/(N!d°!)$ in the PC basis of total degree $d°$ defined by the set of multi-indices $\mathcal{K}(d°) = \{\boldsymbol{k} \in \mathbb{N}^N, \|\boldsymbol{k}\|_1 \leq d°\}$.

$d°$	Monolithic	Fragmented	
	$N = 37$	$N = 19$	$N = 23$
1	38	20	24
2	741	210	300
3	9880	1540	2600
4	101 270	8855	17 550

3. Methodology

This section details the approach used to build a surrogate model of the cesium K_D in order to perform cheaply the uncertainty propagation in the cesium adsorption model upstream chained with a pore water composition model. The surrogate construction relies on a dimension reduction by means of partial least squares and a functional approximation in the reduced space with polynomial chaos expansion.

3.1. Partial least squares

PLS regression was developed by Wold and coworkers [22–24] to counteract the ill-posedness of the ordinary least squares solution when the predictors are highly collinear [25] and/or when the number of predictors exceeds the number of observations [26]. Initially introduced in econometrics and chemometrics, PLS has become popular in many scientific areas as a statistical tool of choice in multivariate analysis. The concept of PLS is to reduce the N original predictors to a smaller set of n uncorrelated latent components and to perform least squares regression on these components. The latent components express as linear combinations of the N original predictors and have been used recently in high-dimensional surrogate modeling for estimating the effective dimension of a problem [27]. Moreover, PLS can be applied to scalar (PLS1 version) or vectorial (PLS2 version) outputs [28].

3.1.1. Notation

Let us consider a system (or model) with N uncertain input parameters collected into a random vector $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_N) \in \mathbb{X} \subset \mathbb{R}^N$ and a single output quantity y. In PLS setting, the inputs are not necessarily independent and the output is assumed to be a function of $n \ll N$ underlying latent components. By sampling the parametric domain $\mathbb{X}$ with M realizations $\mathcal{X} = \{\boldsymbol{\xi}^{(m)}\}_{m=1}^M$ and computing the ensuing outputs $\mathcal{Y} = \{y^{(m)} = y(\boldsymbol{\xi}^{(m)})\}_{m=1}^M$, we can assemble the matrix of inputs $X = \left[\boldsymbol{\xi}^{(m)}\right] \in \mathbb{R}^{M,N}$ and the vector of output $\mathrm{y} = \left[y^{(m)}\right] \in \mathbb{R}^M$. Without loss of generality, the data X and y are centered.

3.1.2. Description

The principle of the PLS method is to identify a set of weight vectors $\{\boldsymbol{\omega}_k \in \mathbb{R}^N\}$ that maximize the covariance between the inputs and the output. The weights are calculated in a sequential manner by solving the following optimization problem at the current iteration k,

$$\begin{aligned}\boldsymbol{\omega}_k &= \underset{\boldsymbol{\omega},\ \|\boldsymbol{\omega}\|_2=1}{\arg\max} \left(\mathrm{Cov}\left(E_k\boldsymbol{\omega}, f_k\right)\right), \\ &= \underset{\boldsymbol{\omega},\ \|\boldsymbol{\omega}\|_2=1}{\arg\max} \left(\boldsymbol{\omega}^\top E_k^\top f_k\right), \\ &= E_k^\top f_k / \|E_k^\top f_k\|_2,\end{aligned}$$

where $E_k \in \mathbb{R}^{M,N}$ and $f_k \in \mathbb{R}^M$ denote the current residual of X and y, respectively (see Algorithm 1 for the definitions). The residual of X is first projected onto the weight vector to get the associated latent component (or score) $\tau_k \in \mathbb{R}^M$,

$$\tau_k = E_k\boldsymbol{\omega}_k / \|E_k\boldsymbol{\omega}_k\|_2,$$

and then projected onto the latent component to obtain the X-loading vector $p_k \in \mathbb{R}^N$,

$$p_k = E_k^\top \tau_k.$$

Both matrices of latent components $T = [\tau_k] \in \mathbb{R}^{M,n}$ and weights $W = [\omega_k] \in \mathbb{R}^{N,n}$ are orthonormal, that is $T^\top T = W^\top W = I_n$ (the identity matrix). When performing PLS regression, the latent components are used as regressors instead of the original ones (X is replaced by T in the normal equations). From a dimension reduction perspective, the latent components appear as a compressed version of the original inputs through the relation

$$T = XR,$$

where the reduction matrix $R \in \mathbb{R}^{N,n}$ is written as [28]

$$R = W\left(P^\top W\right)^{-1}, \tag{1}$$

with $P = [p_k] \in \mathbb{R}^{N,n}$ the X-loadings matrix.

3.1.3. Implementation

A detailed description of the PLS algorithm that we have implemented to compute the reduction matrix is given in Appendix A. The optimal number of latent components to retain in a PLS model or in a PLS-based surrogate can be selected by many techniques including the Jackknife method, cross-validation technique, or information criterion. The numerical experiments presented in Section 4 use the surrogate validation error and show that only two latent components are sufficient in our case. As a closing remark, a sparsity constraint can be added to the weight vectors in order to eliminate the input parameters with small contributions, thus improving the reduction dimension procedure. The sparse PLS [29] is not necessary in this study since it increases the computational cost (for tuning the penalty parameters controlling the sparsity) but without significant influence on the surrogate accuracy.

3.2. Polynomial chaos expansion

PC methods have been broadly used for uncertainty quantification [30,31] in many application domains including geosciences [32–34]. Given a model subject to uncertain input parameters, the principle of a PC expansion is to approximate the model output by using multivariate polynomial basis functions orthogonal with respect to the inputs' probability distributions. The main advantages of PC methods are the exponential convergence rate for smooth quantities of interest and the direct exact derivation of the moments and variance-based sensitivity indices. Although the standard framework assumes that the inputs are independent and have exact (or analytical) distributions, PC can be extended to arbitrary dependent distributions [35–37] only defined by their moments.

3.2.1. Generalized polynomial chaos

In classical generalized Polynomial Chaos (gPC), the inputs are assumed to be independent and the multivariate gPC basis functions $\{\phi_{\boldsymbol{k}}\}$ are explicitly defined as product of univariate orthonormal polynomials,

$$\phi_{\boldsymbol{k}}(\boldsymbol{\xi}) = \prod_{i=1}^{N} \varphi^i_{k_i}(\xi_i),$$

where $\boldsymbol{k} = (k_1, \dots, k_N) \in \mathbb{N}^N$ is a multi-index and $\{\varphi^i_{k_i}\}$ are univariate basis functions orthogonal with respect to the variable ξ_i. As an example, Legendre polynomials are orthogonal for uniform distributions. The orthogonality condition satisfied by the gPC reads

$$\langle \phi_{\boldsymbol{k}}, \phi_{\boldsymbol{l}} \rangle = \int_{\mathbb{X}_{\boldsymbol{\xi}}} \phi_{\boldsymbol{k}}(\boldsymbol{\xi})\phi_{\boldsymbol{l}}(\boldsymbol{\xi}) p_{\boldsymbol{\xi}}(\boldsymbol{\xi}) d\boldsymbol{\xi} = \|\phi_{\boldsymbol{k}}\|^2 \delta_{\boldsymbol{k},\boldsymbol{l}},$$

where $\langle \cdot, \cdot \rangle$ is the inner product in $L^2_{p_{\boldsymbol{\xi}}}(\mathbb{X}_{\boldsymbol{\xi}})$ and $\|\phi_{\boldsymbol{k}}\| = \langle \phi_{\boldsymbol{k}}, \phi_{\boldsymbol{k}} \rangle^{1/2}$ its induced norm.

3.2.2. Arbitrary polynomial chaos

When the inputs are dependent, arbitrary Polynomial Chaos (aPC) basis functions $\{\psi_{\boldsymbol{l}}\}$ can be constructed to satisfy a discrete orthogonality condition,

$$(\psi_{\boldsymbol{l}}, \psi_{\boldsymbol{m}}) = \frac{1}{|\mathcal{T}|} \sum_{\boldsymbol{\tau} \in \mathcal{T}} \psi_{\boldsymbol{l}}(\boldsymbol{\tau})\psi_{\boldsymbol{m}}(\boldsymbol{\tau}) = |||\psi_{\boldsymbol{l}}|||^2 \delta_{\boldsymbol{l},\boldsymbol{m}}, \tag{2}$$

where $(\cdot, \cdot)$ is the discrete weighted inner product and $|||\psi_{\boldsymbol{l}}||| = (\psi_{\boldsymbol{l}}, \psi_{\boldsymbol{l}})^{1/2}$ its induced norm. The random vector $\boldsymbol{\tau} = (\tau_1, \dots, \tau_n) \subset \mathbb{R}^n$ corresponds here to n reduced coordinates defined by the PLS reduction matrix R. A sampling $\mathcal{T}$ of $\boldsymbol{\tau}$ is obtained from a sampling $\mathcal{X}$ of $\boldsymbol{\xi}$ as follows

$$\mathcal{T} = \{\boldsymbol{\tau} = \boldsymbol{\xi} R,\ \boldsymbol{\xi} \in \mathcal{X}\}.$$

The aPCs are built empirically with the Gram–Schmidt orthogonalization procedure briefly summarized hereafter. Let $\{P_{\boldsymbol{l}}(\boldsymbol{\tau})\}$ be a set of N_{b} linearly independent polynomials of total degree d° defined by the set of multi-indices $\mathcal{L}(d^\circ) = \left\{\boldsymbol{l} \in \mathbb{N}^n, \|\boldsymbol{l}\|_1 \le d^\circ\right\}$. The orthonormal basis functions $\{\psi_{\boldsymbol{l}}(\boldsymbol{\tau})\}$ are constructed recursively from $\{P_{\boldsymbol{l}}(\boldsymbol{\tau})\}$. For all $\boldsymbol{l} \in \mathcal{L}$, we have

$$\begin{aligned} \widetilde{\psi}_{\boldsymbol{l}}(\boldsymbol{\tau}) &= P_{\boldsymbol{l}}(\boldsymbol{\tau}) - \sum_{\boldsymbol{m} \in \mathcal{M}(\boldsymbol{l})} r_{\boldsymbol{l},\boldsymbol{m}} \psi_{\boldsymbol{m}}(\boldsymbol{\tau}), \\ \psi_{\boldsymbol{l}}(\boldsymbol{\tau}) &= \widetilde{\psi}_{\boldsymbol{l}}(\boldsymbol{\tau}) / |||\widetilde{\psi}_{\boldsymbol{l}}(\boldsymbol{\tau})|||, \end{aligned} \tag{3}$$

where the set $\mathcal{M}(\boldsymbol{l})$ starts with $\mathcal{M}(\boldsymbol{0}) = \emptyset$ and then collects the multi-indices of the polynomials orthonormalized in the previous iterations. Normalization step (3) is optional but simplifies condition (2) into $(\psi_{\boldsymbol{l}}, \psi_{\boldsymbol{m}}) = \delta_{\boldsymbol{l},\boldsymbol{m}}$. A first way for computing the set of coefficients $\{r_{\boldsymbol{l},\boldsymbol{m}}\}$ is from a recursive algorithm [36] based on the definition $r_{\boldsymbol{l},\boldsymbol{m}} = (P_{\boldsymbol{l}}, \psi_{\boldsymbol{m}})$ and using multivariate monomials $P_{\boldsymbol{l}}(\boldsymbol{\tau}) = \boldsymbol{\tau}^{\boldsymbol{l}} = \prod_{j=1}^n \tau_j^{l_j}$ to simplify the inner products calculations. A more direct way followed here and described in Appendix B is through a QR factorization $\mathcal{P} = \mathcal{Q}\mathcal{R}$ of the matrix $\mathcal{P} = [T^{\boldsymbol{l}}] \in \mathbb{R}^{M,N_{\mathrm{b}}}$, $T \in \mathbb{R}^{M,n}$ being the PLS latent components matrix. Finally, the aPC basis functions appear as linear combinations of the $\boldsymbol{\tau}^{\boldsymbol{l}}$,

$$\forall \boldsymbol{l} \in \mathcal{L}, \quad \psi_{\boldsymbol{l}}(\boldsymbol{\tau}) = \sum_{\boldsymbol{m} \in \mathcal{L}} \alpha_{\boldsymbol{m},\boldsymbol{l}} \boldsymbol{\tau}^{\boldsymbol{m}}, \tag{4}$$

where $\{\alpha_{\boldsymbol{m},\boldsymbol{l}}\}$ are the coefficients of the inverse $\mathcal{R}^{-1}$ of the matrix $\mathcal{R}$.

3.2.3. Spectral coefficients computation

Once the aPC basis functions have been built, the second step is to estimate the set of spectral coefficients of the model output approximation. By exploiting the matrix $\mathcal{Q}$ of sampled aPC basis functions, we calculate the least squares solution,

$$\mathcal{Q}^\top \mathcal{Q} \boldsymbol{c} = \mathcal{Q}^\top \mathbb{c}, \tag{5}$$

where $\boldsymbol{c} \in \mathbb{R}^{N_{\mathrm{b}}}$ collects the spectral coefficients $c_{\boldsymbol{l}}$ and $\mathbb{c} \in \mathbb{R}^M$ is the vector of model output $c(\boldsymbol{\tau}^{(m)})$. Finally, the PLS-aPC approximation $\tilde{c}(\boldsymbol{\xi})$ of the model output $\mathrm{K_D}$ is expressed as

$$\tilde{c}(\boldsymbol{\xi}) = \sum_{\boldsymbol{l} \in \mathcal{L}} c_{\boldsymbol{l}} \psi_{\boldsymbol{l}}(\boldsymbol{\xi} R), \tag{6}$$

where R, $\{\psi_{\boldsymbol{\tau}}(\cdot)\}$ and $\{c_{\boldsymbol{l}}\}$ are defined by Eqs. (1), (4) and (5), respectively. The number of coefficients N_{c} in the PLS-aPC surrogate model (6) is

$$N_{\mathrm{c}}(n, d^\circ) = n \times N + N_{\mathrm{b}}(n, d^\circ),$$

where the first term corresponds to the number of coefficients of the reduction matrix R and the second term is the number of PC basis in the reduced subspace given a total degree d°. The impact of the number of reduced coordinates and polynomial basis functions on the approximation accuracy is investigated in the next section.

Table 4
Monolithic approach - RMSRE obtained using $M_* = 10^3$ realizations of the chained geochemical model for the three initial concentrations $[Cs]_0$ (mol L^{-1}).

$[Cs]_0$	d°	$n=1$	$n=2$	$n=3$
10^{-5}	2	$1.46 \cdot 10^{-1}$	$1.18 \cdot 10^{-1}$	$1.16 \cdot 10^{-1}$
	3	$1.16 \cdot 10^{-1}$	$6.59 \cdot 10^{-2}$	$6.41 \cdot 10^{-2}$
	4	$1.16 \cdot 10^{-1}$	$6.67 \cdot 10^{-2}$	$6.50 \cdot 10^{-2}$
10^{-3}	2	$7.92 \cdot 10^{-2}$	$5.31 \cdot 10^{-2}$	$5.22 \cdot 10^{-2}$
	3	$7.91 \cdot 10^{-2}$	$5.31 \cdot 10^{-2}$	$5.20 \cdot 10^{-2}$
	4	$7.87 \cdot 10^{-2}$	$5.32 \cdot 10^{-2}$	$5.23 \cdot 10^{-2}$
10^{-1}	2	$9.80 \cdot 10^{-2}$	$5.59 \cdot 10^{-2}$	$5.43 \cdot 10^{-2}$
	3	$9.77 \cdot 10^{-2}$	$5.49 \cdot 10^{-2}$	$5.24 \cdot 10^{-2}$
	4	$9.78 \cdot 10^{-2}$	$5.55 \cdot 10^{-2}$	$5.30 \cdot 10^{-2}$

Table 5
Monolithic approach - Number of coefficients $N_c(n, d^\circ)$ in the PLS-aPC surrogate model.

d°	$n=1$	$n=2$	$n=3$
2	40	80	121
3	41	84	131
4	42	89	146

4. Surrogate model validation and comparison

The adsorption model presented in Section 2.1 must be completed with an initial cesium concentration $[Cs]_0$. In what follows, we choose the three contrasting values 10^{-5}, 10^{-3}, and 10^{-1} mol L^{-1} to explore different regimes of the model. The influence of the sample size M on the cesium adsorption is shown on Fig. 2 where the empirical statistical moments are plotted for $M = 10^2$, 10^3, and 10^4 Monte-Carlo realizations. We observe on Fig. 3 that the probability density functions estimated for $M = 10^3$ are close to those obtained for $M = 10^4$. This aspect is consistent with the low bootstrap errors calculated for these values of M. On the basis of these observations, we use a training set $\mathcal{X}$ of $M = 10^3$ realizations to build the surrogate models in the monolithic and fragmented approaches.

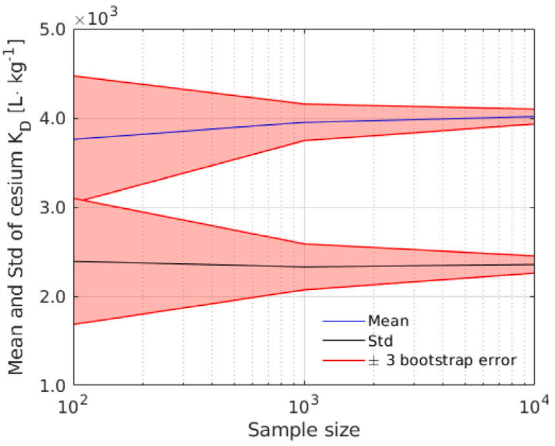
4.1. Monolithic approach

A linear transformation is applied to each component ξ_i of the global inputs vector $\boldsymbol{\xi} = (\xi_1, \xi_2)$ in order to manipulate, in the dimension reduction step, a set of random variables uniformly distributed over $[-1, 1]$. The surrogate model accuracy is measured with the root mean squared relative error (RMSRE) defined as

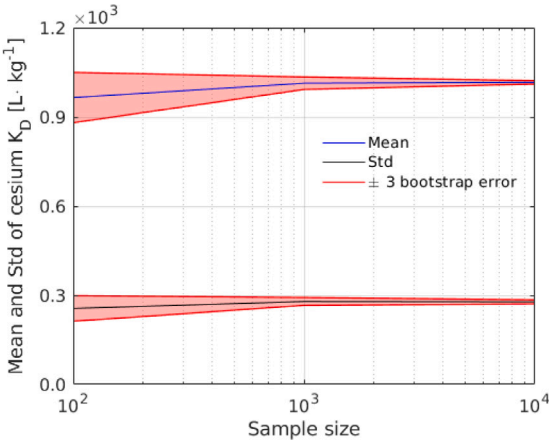
$$\text{RMSRE} = \sqrt{\frac{1}{M_*} \sum_{\boldsymbol{\xi} \in \mathcal{X}_*} \left(\frac{c(\boldsymbol{\xi}) - \tilde{c}(\boldsymbol{\xi})}{c(\boldsymbol{\xi})} \right)^2},$$

where $\mathcal{X}_*$ is a validation set of $M_* = 10^3$ realizations independent from the training set $\mathcal{X}$. The empirical errors of the surrogate models are reported in Table 4 where the number of reduced coordinates varies from one to three and the total polynomial degree varies from two to four. The corresponding number of coefficients are indicated in Table 5, they are much lower than the number of terms in high dimensional cases (see Table 3). The representation of cesium K_D by PLS-aPC surrogate models is very effective and robust since the error level is small for few reduced coordinates and low polynomial degree, whatever the initial condition. Indeed, a sampling of size 10^3 in dimension 37 yields a RMSRE around 5–6% when $n \geq 2$ and $d^\circ \geq 3$, for the three initial concentrations. Such a level of error can be appreciated from Fig. 4 where the surrogate model values and absolute errors are plotted versus the true values for $n = 2$ and $d^\circ = 3$.

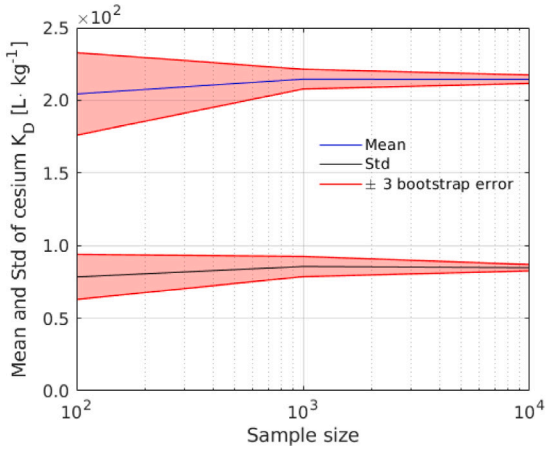
Let us now take a close look into the principle and effect of the dimension reduction. Fig. 5 depicts the first column coefficients of the reduction matrix for the three initial conditions. The first comment



(a) $[Cs]_0 = 10^{-5}$ mol · L^{-1}



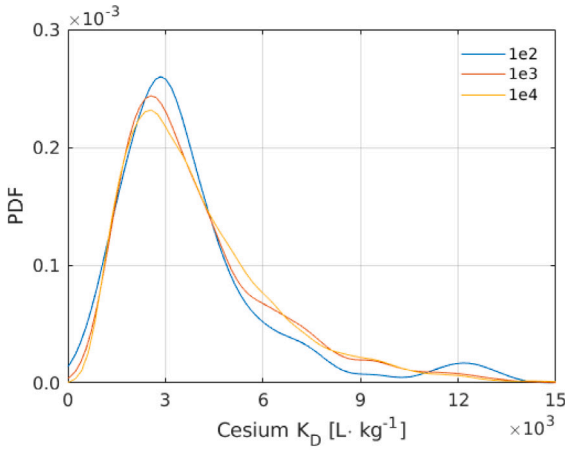
(b) $[Cs]_0 = 10^{-3}$ mol · L^{-1}



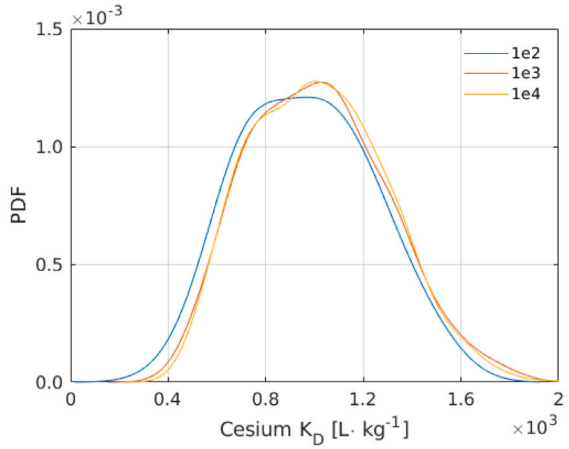
(c) $[Cs]_0 = 10^{-1}$ mol · L^{-1}

Fig. 2. Cesium K_D means and standard deviations as functions of the sample size M with the bootstrap errors [38] computed with 20 replicants from resampling with replacement.

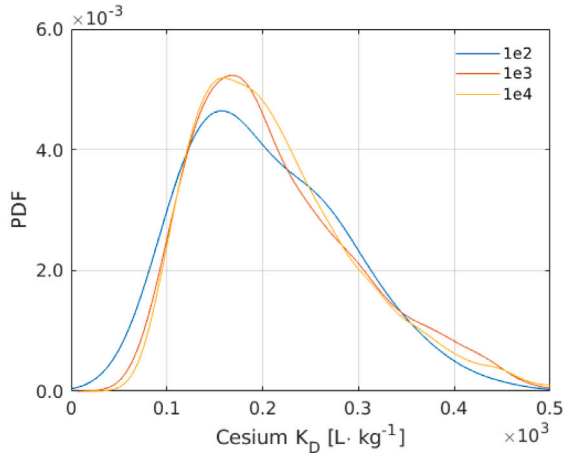
is that the Na^+/K^+ selectivity coefficient is the only input of the water composition model that stands out. This coefficient is indeed the main factor controlling the K^+ concentration in solution and K^+ is the main competitor for Cs^+ adsorption on the Type II sites and FES. The second comment is that the main contributors of the Cesium



(a) $[\mathrm{Cs}]_0 = 10^{-5}$ mol · L^{-1}

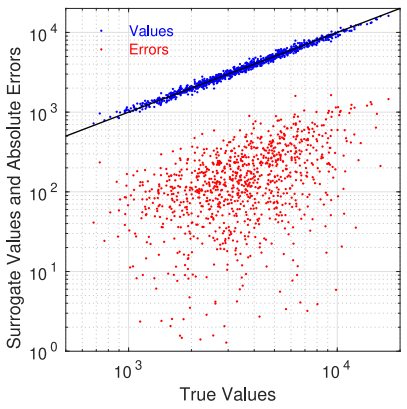


(b) $[\mathrm{Cs}]_0 = 10^{-3}$ mol · L^{-1}

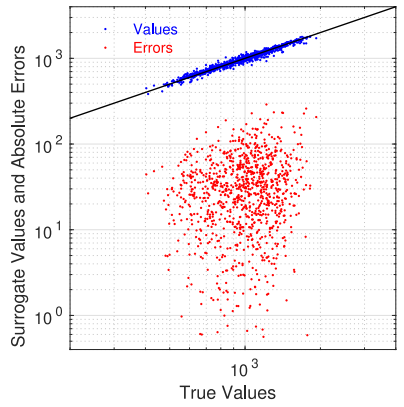


(c) $[\mathrm{Cs}]_0 = 10^{-1}$ mol · L^{-1}

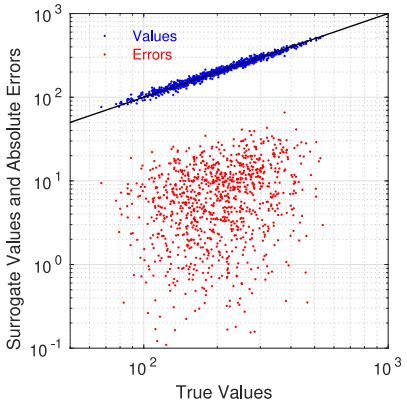
Fig. 3. Probability density functions of cesium K_D estimated using standard kernel density estimation for three sample sizes M.



(a) $[\mathrm{Cs}]_0 = 10^{-5}$ mol · L^{-1}



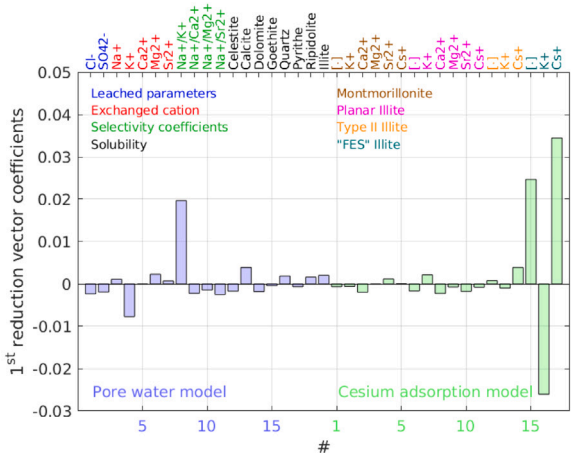
(b) $[\mathrm{Cs}]_0 = 10^{-3}$ mol · L^{-1}



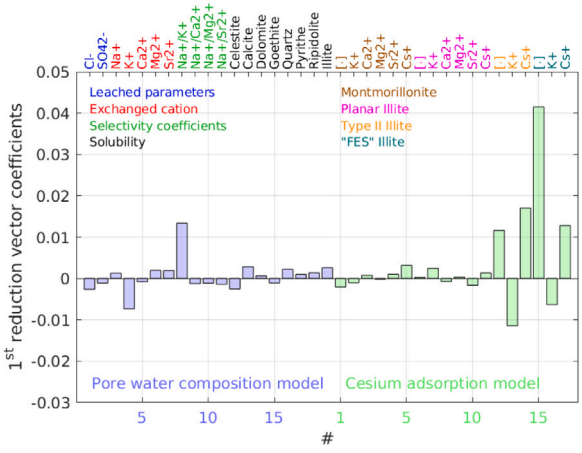
(c) $[\mathrm{Cs}]_0 = 10^{-1}$ mol · L^{-1}

Fig. 4. Monolithic approach - Surrogate model values $\tilde{c}(\boldsymbol{\xi})$ and absolute errors $|c(\boldsymbol{\xi}) - \tilde{c}(\boldsymbol{\xi})|$ versus true values $c(\boldsymbol{\xi})$ for the validation set $\mathcal{X}_*$ plotted in logarithmic scales. Results for $n = 2$ reduced coordinates and a total degree $d^\circ = 3$.

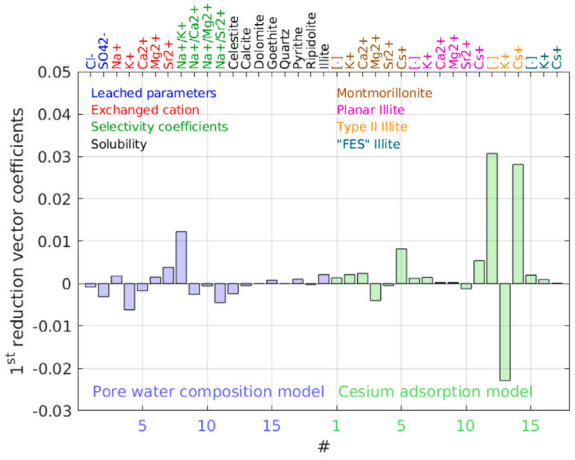
adsorption model depend on $[\mathrm{Cs}]_0$. For $[\mathrm{Cs}]_0 = 10^{-5}$ mol L^{-1}, the main contributors are the Na^+/Cs^+ and Na^+/K^+ selectivity coefficients for the exchange reactions on the illite FES and the FES site density on the illite surfaces. For $[\mathrm{Cs}]_0 = 10^{-3}$ mol L^{-1}, the main contributors are the Na^+/Cs^+ and Na^+/K^+ selectivity coefficients and the site densities for the exchange reactions on both FES and Type II sites on illite. For $[\mathrm{Cs}]_0 = 10^{-1}$ mol L^{-1}, the FES site parameters do not contribute significantly. This result can be understood with the gradual saturation by Cs^+ of the most reactive sites (FES, then Type II, then planar) with increasing Cs^+ aqueous concentration. Fig. 6 plots the cesium K_D with respect to the most important inputs of the adsorption model (in terms of amplitude in the first reduction vector) and the reduced coordinates. It is clear that the advantage conferred by the dimension reduction is to deliver a structuring of the data, which can be efficiently approximated with low-order polynomials.



(a) $[\mathrm{Cs}]_0 = 10^{-5}\ \mathrm{mol}\cdot\mathrm{L}^{-1}$



(b) $[\mathrm{Cs}]_0 = 10^{-3}\ \mathrm{mol}\cdot\mathrm{L}^{-1}$



(c) $[\mathrm{Cs}]_0 = 10^{-1}\ \mathrm{mol}\cdot\mathrm{L}^{-1}$

Fig. 5. Monolithic approach - Reduction matrix first column coefficients.

4.2. Fragmented approach

Contrary to the monolithic approach, the fragmented approach treats each solver separately by relying on the parameters connecting the different solvers. One complication encountered in this approach is that the linking parameters' probability distributions are unknown a priori since these parameters are outputs of solvers. In this section, we explain the procedure to deal with this aspect and present the analysis of cesium K_D in this context.

Table 6
Fragmented approach - RMSRE of the linking parameters obtained using $M_* = 10^3$ realizations of the pore water composition model.

$\xi_{1,2}$	d°	$n=1$	$n=2$	$n=3$
Na^+	2	$1.52\cdot10^{-2}$	$1.18\cdot10^{-2}$	$1.18\cdot10^{-2}$
	3	$1.52\cdot10^{-2}$	$1.18\cdot10^{-2}$	$1.19\cdot10^{-2}$
K^+	2	$4.20\cdot10^{-2}$	$1.68\cdot10^{-2}$	$1.56\cdot10^{-2}$
	3	$4.20\cdot10^{-2}$	$1.65\cdot10^{-2}$	$1.52\cdot10^{-2}$
Ca^{2+}	2	$6.11\cdot10^{-2}$	$4.29\cdot10^{-2}$	$4.28\cdot10^{-2}$
	3	$6.10\cdot10^{-2}$	$4.30\cdot10^{-2}$	$4.28\cdot10^{-2}$
Mg^{2+}	2	$6.24\cdot10^{-2}$	$4.45\cdot10^{-2}$	$4.49\cdot10^{-2}$
	3	$6.25\cdot10^{-2}$	$4.50\cdot10^{-2}$	$4.52\cdot10^{-2}$
Sr^{2+}	2	$2.60\cdot10^{-2}$	$1.87\cdot10^{-2}$	$1.84\cdot10^{-2}$
	3	$2.60\cdot10^{-2}$	$1.87\cdot10^{-2}$	$1.84\cdot10^{-2}$

Table 7
Fragmented approach - Fitted shape parameters of the beta distributions as well as Kolmogorov–Smirnov distance d_{KS}.

$\xi_{1,2}$	α	β	d_{KS}
Na^+	7.61	10.03	$6.0\cdot10^{-3}$
K^+	2.19	4.06	$2.9\cdot10^{-2}$
Ca^{2+}	4.36	9.05	$6.8\cdot10^{-3}$
Mg^{2+}	2.50	6.12	$8.9\cdot10^{-3}$
Sr^{2+}	4.14	7.57	$1.0\cdot10^{-2}$

4.2.1. Linking parameters

Our procedure to estimate the joint probability distribution of the linking parameters is to build surrogate models for these parameters and then to fit analytical probability distributions thanks to large samples of surrogate models. We build PLS-aPC surrogate models whose errors are reported in Table 6 for the five linking parameters. We see that the errors are a few percent and are converged when $n \geq 2$ and $d^\circ \geq 2$. The uncertainty propagation into the pore water composition model is out of the scope of the present paper but can be found in [17].

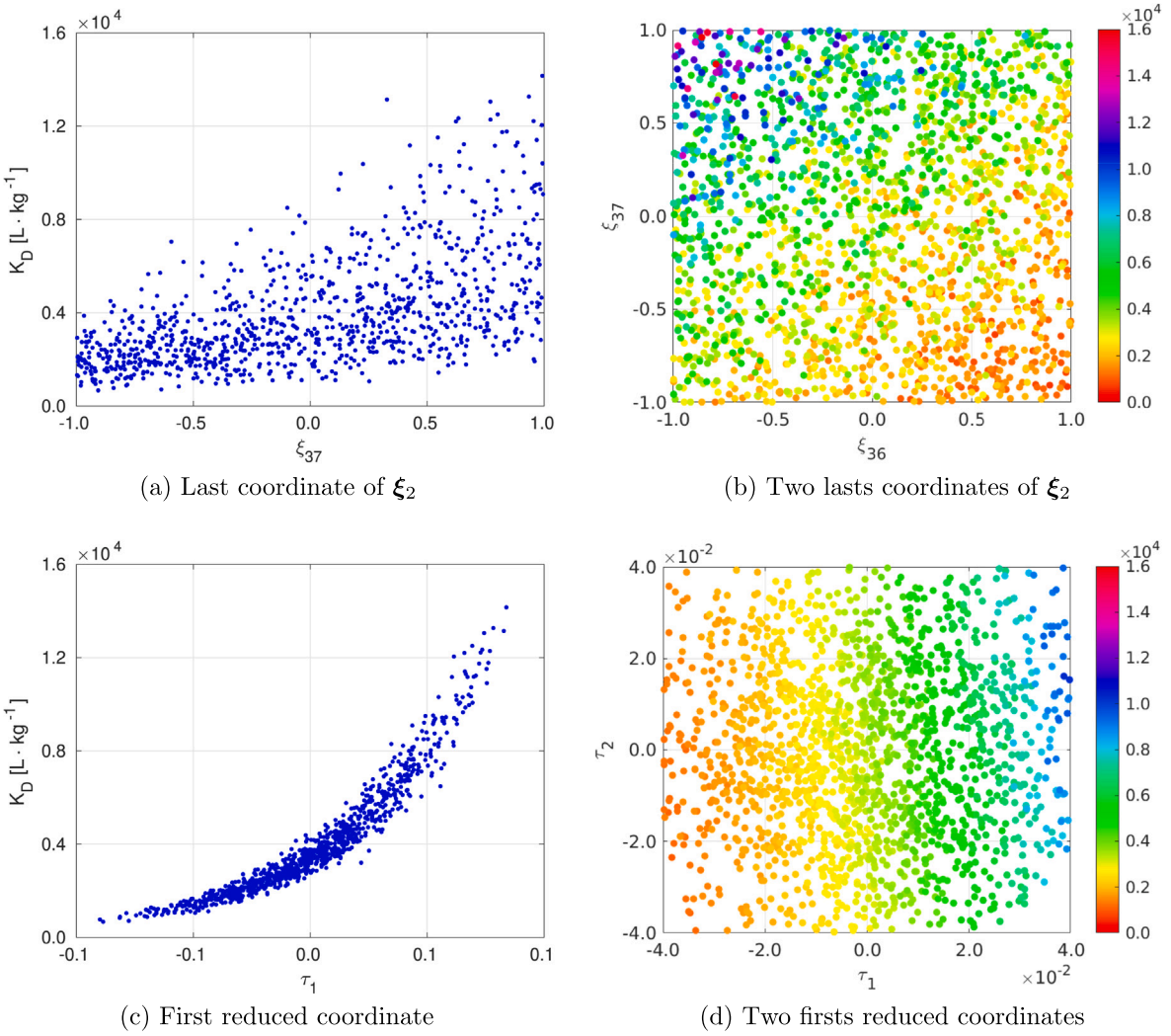
For the probability distributions fitting, we begin by measuring the dependence between the linking parameters with the Chatterjee coefficient [39] computed from a massive sampling $\mathcal{X}^\dagger$ of $M^\dagger = 10^6$ surrogate models realizations. The assumption of independence is valid here since this coefficient, not shown for brevity, is close to zero for all the pairs of distinct parameters. In case of dependency, a copula-based model [40] could be used to describe the relations between the linking parameters. For each of the five linking parameters, we then adjust beta distributions with the method of moments. The two shape parameters α and β of a beta distribution are directly obtained through the relations

$$\alpha = \frac{\hat{\mathbb{E}}\left(\hat{\mathbb{E}}(1-\hat{\mathbb{E}}) - \hat{\mathbb{V}}\right)}{\hat{\mathbb{V}}} \quad \text{and} \quad \beta = \frac{(\hat{\mathbb{E}}-1)\left(\hat{\mathbb{E}}(\hat{\mathbb{E}}-1) + \hat{\mathbb{V}}\right)}{\hat{\mathbb{V}}},$$

where $\hat{\mathbb{E}}$ and $\hat{\mathbb{V}}$ denote the empirical mean and variance computed from $\mathcal{X}^\dagger$. The adjusted shape parameters are indicated in Table 7, together with the Kolmogorov–Smirnov (KS) distance between the beta cumulative distributions and the empirical ones. The KS distances are relatively small, thereby validating the choice and fitting of beta distributions. For illustrative purposes, the adjusted and empirical distributions of the two linking parameters having the minimal and maximal KS distances are plotted in Fig. 7.

4.2.2. Global output

Linear transformations of the components of ξ_2 and inverse transformations of the linking parameters $\xi_{1,2}$, through their inverse cumulative distribution functions, are enforced to handle random variables uniformly distributed over $[-1, 1]$. The RMSRE listed in Table 8 and the



(a) Last coordinate of $\boldsymbol{\xi}_2$

(b) Two lasts coordinates of $\boldsymbol{\xi}_2$

(c) First reduced coordinate

(d) Two firsts reduced coordinates

Fig. 6. Monolithic approach - Cesium K_D scatter plot, top: as a function of the global inputs ξ_{36} and ξ_{37}, bottom: as a function of the reduced coordinates τ_1 and τ_2. Initial concentration $[Cs]_0 = 10^{-5}$ mol L^{-1}.

Table 8
Fragmented approach - RMSRE obtained using $M_* = 10^3$ realizations of the cesium adsorption model for the three initial concentrations $[Cs]_0$ (mol L^{-1}).

$[Cs]_0$	$d°$	$n = 1$	$n = 2$	$n = 3$
10^{-5}	2	$1.47 \cdot 10^{-1}$	$1.19 \cdot 10^{-1}$	$1.17 \cdot 10^{-1}$
	3	$1.08 \cdot 10^{-1}$	$6.22 \cdot 10^{-2}$	$6.20 \cdot 10^{-2}$
	4	$1.07 \cdot 10^{-1}$	$6.25 \cdot 10^{-2}$	$6.31 \cdot 10^{-2}$
10^{-3}	2	$6.68 \cdot 10^{-2}$	$5.23 \cdot 10^{-2}$	$5.20 \cdot 10^{-2}$
	3	$6.67 \cdot 10^{-2}$	$5.24 \cdot 10^{-2}$	$5.21 \cdot 10^{-2}$
	4	$6.65 \cdot 10^{-2}$	$5.23 \cdot 10^{-2}$	$5.22 \cdot 10^{-2}$
10^{-1}	2	$8.38 \cdot 10^{-2}$	$5.64 \cdot 10^{-2}$	$5.51 \cdot 10^{-2}$
	3	$8.31 \cdot 10^{-2}$	$5.40 \cdot 10^{-2}$	$5.20 \cdot 10^{-2}$
	4	$8.32 \cdot 10^{-2}$	$5.39 \cdot 10^{-2}$	$5.22 \cdot 10^{-2}$

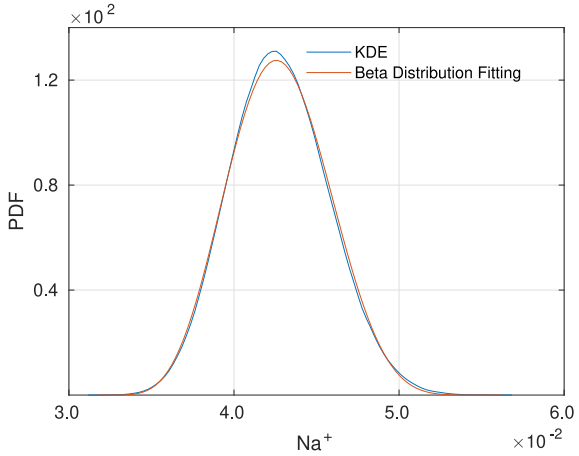
first column coefficients of the reduction matrix represented in Fig. 8 are strongly consistent with those obtained in the monolithic approach. One advantage offered by the fragmented approach is that the number of coefficients N_c in Table 9 is lower than in the monolithic approach. However, considering the similarity of the reduction matrix first column coefficients between the monolithic and fragmented approaches (see Figs. 5 and 8), more than likely that sparse PLS would have produced denoised reduction matrix with roughly the same number of non-zero coefficients.

Table 9
Fragmented approach - Number of coefficients $N_c(n, d°)$ in the PLS-aPC surrogate model.

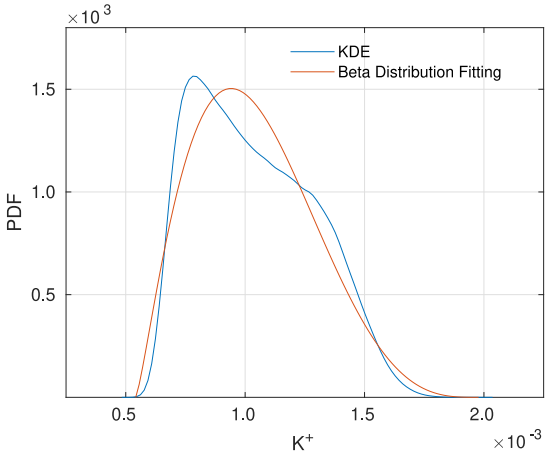
$d°$	$n = 1$	$n = 2$	$n = 3$
2	26	52	79
3	27	56	89
4	28	61	104

5. Global sensitivity analyses

A very useful information when examining a physical model with multiple inputs is the relative impact of each input and set of inputs on the model output. Of particular interest is the global sensitivity analysis which measures the contribution of inputs and their interactions over the whole parametric domain. Different measure criteria have been proposed in global sensitivity analysis such as the variance [41], higher-order moments [42] as well as probabilistic divergences and distances [43]. The most popular indices are the so-called Sobol indices that decompose the output variance into normalized elementary contributions. In this section, after recalling the definition of such variance-based indices in the case of a group of inputs, we specify the Cramér–Von Mises indices that are formulated on the whole distribution of the output.



(a) $d_{\mathrm{KS}} = 6.0 \cdot 10^{-3}$



(b) $d_{\mathrm{KS}} = 2.9 \cdot 10^{-2}$

Fig. 7. Fragmented approach - Empirical and fitted beta distributions with the method of moments using $M^{\dagger} = 10^6$ surrogate models evaluations.

5.1. Closed sensitivity indices

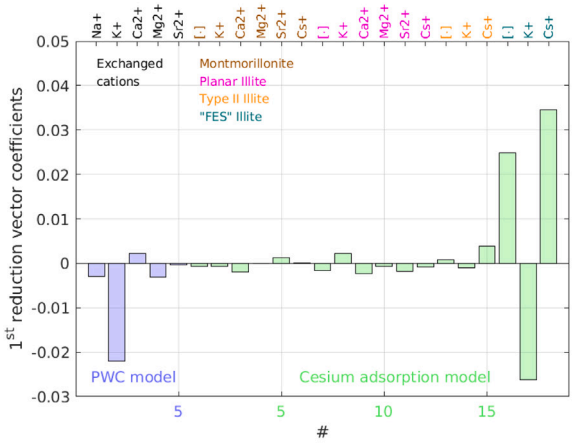
Variance-based global sensitivity analysis methods [41,44] determine the contribution of each input and set of inputs onto the variance of the model output. This analysis can be applied for groups of inputs [45,46] to quantify the influence of subsets of inputs. Specifically, when the vector of input parameters $\boldsymbol{\xi}$ is divided into two groups $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$, the variance decomposition of a quantity of interest Y is written as

$$\mathbb{V}(Y) = \mathbb{V}(\mathbb{E}(Y|\boldsymbol{\xi}_1)) + \mathbb{V}(\mathbb{E}(Y|\boldsymbol{\xi}_2)) + \mathbb{V}_{1,2}(Y), \tag{7}$$

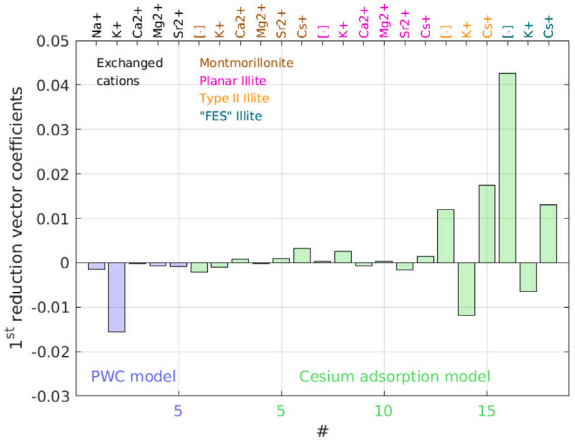
where $\mathbb{E}(Y|\boldsymbol{\xi})$ denotes the conditional expectation of Y given $\boldsymbol{\xi}$. The term $\mathbb{V}(\mathbb{E}(Y|\boldsymbol{\xi}_i))$ measures the part of the variance due to the group of inputs $\boldsymbol{\xi}_i$ while the term $\mathbb{V}_{1,2}(Y) = \mathbb{V}(Y) - \mathbb{V}(\mathbb{E}(Y|\boldsymbol{\xi}_1)) - \mathbb{V}(\mathbb{E}(Y|\boldsymbol{\xi}_2))$ quantifies the interaction effects between the two groups $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$. The influence of a group of parameters, that regroups the own effect of each parameter and all the interaction effects within the group, is measured by the closed sensitivity index C_i defined as

$$C_i = \frac{\mathbb{V}(\mathbb{E}(Y|\boldsymbol{\xi}_i))}{\mathbb{V}(Y)}.$$

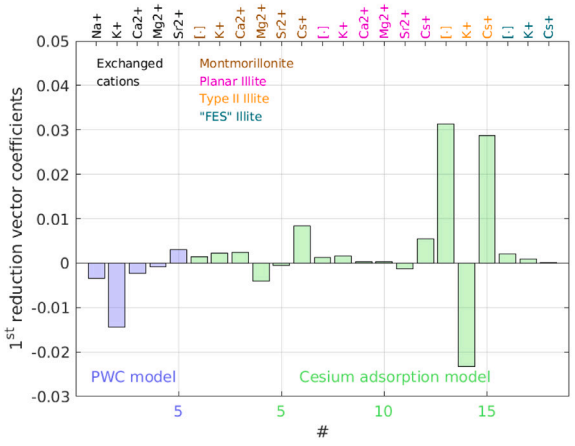
In other words, if the group contains n parameters, the closed index corresponds to the sum of the first-order, second-order, …, and nth-order Sobol indices. The closed sensitivity indices of $Y = \mathrm{K_D}$ are plotted



(a) $[\mathrm{Cs}]_0 = 10^{-5}\ \mathrm{mol \cdot L^{-1}}$



(b) $[\mathrm{Cs}]_0 = 10^{-3}\ \mathrm{mol \cdot L^{-1}}$



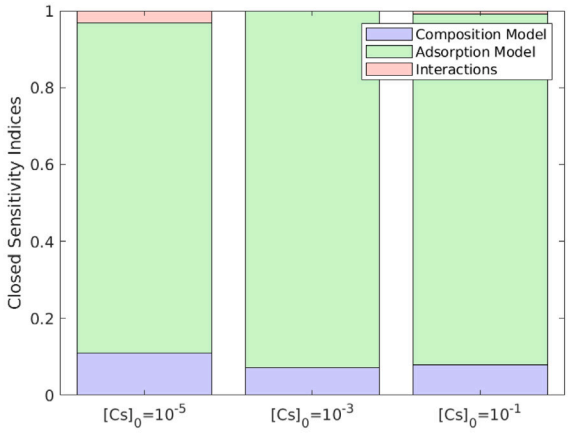
(c) $[\mathrm{Cs}]_0 = 10^{-1}\ \mathrm{mol \cdot L^{-1}}$

Fig. 8. Fragmented approach - Reduction matrix first column coefficients.

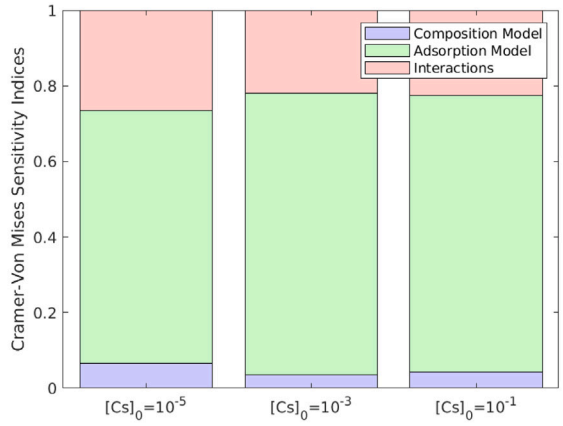
in Fig. 9(a) and show that the cesium variance is overwhelmingly governed by the adsorption model without interaction with the pore water composition model, regardless of the initial concentration.

5.2. Cramér–Von Mises indices

The variance-based indices are well adapted to describe the mean output behavior of a model [47] but, by nature, they do not quantify



(a) Closed sensitivity indices



(b) Cramér–Von Mises indices

Fig. 9. Global sensitivity indices obtained from the pick-freeze method applied with the monolithic surrogate model and samplings of 10^6 realizations.

the influence of an input over the whole distribution. For that purpose, more general indices exist in the available literature [47] including those based on the Cramér–Von Mises distance [48]. If $Y(x) = \mathbb{1}_{\{c\leq x\}}$, the variance decomposition (7) becomes

$$F_c(x)(1-F_c(x)) = \mathbb{V}(F_c(x|\boldsymbol{\xi}_1)) + \mathbb{V}(F_c(x|\boldsymbol{\xi}_2)) + \mathbb{V}_{1,2}(\mathbb{1}_{\{c\leq x\}}), \tag{8}$$

where $F_c(x) = \mathbb{P}(c \leq x) = \mathbb{E}\left(\mathbb{1}_{\{c\leq x\}}\right)$ is the cumulative distribution function of c and $F_c(x|\boldsymbol{\xi})$ the conditional cumulative distribution function given $\boldsymbol{\xi}$. The left-hand side of Eq. (8) is obtained through the relation $\mathbb{V}(\mathbb{1}_A) = \mathbb{P}(A)(1-\mathbb{P}(A))$ with $A = \{c \leq x\}$. The Cramér–Von Mises indices $\mathcal{W}_i$ defined as

$$\mathcal{W}_i = \frac{\int_{\mathbb{R}} \mathbb{V}(F_c(x|\boldsymbol{\xi}_i))dF_c(x)}{\int_{\mathbb{R}} F_c(x)(1-F_c(x))dF_c(x)}, \tag{9}$$

are obtained after integrating Eq. (8) in x with respect to the distribution of c and normalizing. The indices (9) have nice properties and advantages: they are based on the Hoeffding decomposition and sum to one (as Sobol indices), they always exist whatever the output distribution, and they can be simply obtained with a pick-freeze method (see Appendix C). The Cramer–Von Mises indices associated to cesium K_D are drawn on Fig. 9(b) and indicate that the interaction between the adsorption model and the pore water composition model represents about 20% of the uncertainty over the distribution of cesium K_D. It is interesting to note that these interaction indices strongly contrast with the Sobol ones, illustrating that a negligible Sobol index does not imply that the associated input (or group of inputs) has no effect over the whole distribution.

6. Conclusion

This work focused on investigating the use of dimension-reduction techniques to represent the output of a cesium adsorption model in various regimes. The specificity of the adsorption model is its upstream link with a pore water composition model as well as the plentiful number of input parameters. A linear supervised dimension reduction approach combined with a functional approximation has proved effective for constructing a surrogate model of the cesium distribution coefficient. Assessment of the surrogate models revealed that both monolithic and fragmented approaches have comparable mean squared relative errors. Direct exploitation of the surrogate models through global sensitivity analysis emphasized that the upstream pore water composition model has no effect on the cesium K_D variance. Also, we evidenced the benefit of going beyond the variance-based sensitivity analysis by estimating the impact of the inputs over the whole output distribution. The uncertainty propagation methodology presented in this paper could be repeated for other radionuclides and different models in order to investigate further the role of the pore water composition model into adsorption processes.

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.

Data availability

Data will be made available on request.

Acknowledgments

The European project DONUT has supported the work of P. Sochala, C. Chiaberge, and F. Claret. C. Tournassat acknowledges the funding support by a grant overseen by the French National Research Agency (ANR) as part of the "Investissements d'Avenir" Program LabEx VOLTAIRE, 10-LABX-0100.

All authors approved the version of the manuscript to be published.

Appendix A. NIPALS algorithm

The Nonlinear Iterative Partial Least Squares (NIPALS) algorithm with orthonormal latent components is detailed in Algorithm 1 for a single output (PLS1 model). Recall that the goal of the PLS method is to extract from the matrix of input data X, a set of n latent components that are more suitable to describe the output quantity y than the $N \gg n$ original predictors. Algorithm 1 starts with the initialization of the inputs and output residuals E_1 and f_1 to X and y, respectively. The first step of the iterative loop is to compute the current weight vector by maximizing the covariance between the X and y residuals. The second step is to calculate the current latent component (or score) by projecting the X residual onto the weight vector. The current X and y-loadings are then estimated by a projection of the residuals onto the current latent component. Finally, the residuals are updated by deflation and the procedure is started again. Once the loop is over, the different matrices are stored and the regression coefficients $\boldsymbol{\beta}$ as well as the PLS approximation $\hat{y}$ are computed. A number of variants of PLS exist for estimating the different matrices. For instance, the latent component matrix can be orthogonal or not, normalized or not, and we refer to [49,50] for a comparison of the different PLS variants.

Algorithm 1 NIPALS algorithm

Input: X, y, n

$E_1 = X$ and $f_1 = \mathrm{y}$, ▷ Residuals

for $k = 1, \dots, n$ **do**

$\boldsymbol{\omega}_k = E_k^\top f_k / \| E_k^\top f_k \|_2$, ▷ Weights

$\tau_k = E_k \boldsymbol{\omega}_k / \| E_k \boldsymbol{\omega}_k \|_2$, ▷ Scores

$p_k = E_k^\top \tau_k$, ▷ X-loadings

$q_k = f_k^\top \tau_k$, ▷ y-loading

$E_{k+1} = E_k - \tau_k p_k^\top$, ▷ X deflation

$f_{k+1} = f_k - \tau_k q_k$, ▷ y deflation

end for

$W = [\boldsymbol{\omega}_k]$, $T = [\tau_k]$, $P = [p_k]$, $\mathrm{q} = [q_k]$,

$R = W \left(P^\top W\right)^{-1}$, ▷ Reduction Matrix

$\boldsymbol{\beta} = R T^\top \mathrm{y}$, ▷ Regression coefficients

$\hat{\mathrm{y}} = X\boldsymbol{\beta}$, ▷ PLS1 Model

Output: $\hat{\mathrm{y}}, \boldsymbol{\beta}, W, T, P, q, R$

Appendix B. Weighted-QR factorization algorithm

The Gram–Schmidt (GS) procedure is a well-known method for orthogonalizing a set of vectors with respect to a given discrete (possibly weighted) inner product $(\cdot,\cdot)$. Starting from a finite set of m linearly independent vectors $\{\boldsymbol{p}_1, \dots, \boldsymbol{p}_m\}$ of $\mathbb{R}^{M \geq m}$, the GS procedure generates an orthogonal set of vectors $\{\boldsymbol{q}_1, \dots, \boldsymbol{q}_m\}$ that spans the same subspace of $\mathbb{R}^M$. The orthogonality condition of the $\boldsymbol{q}_k$ reads

$$(\boldsymbol{q}_k, \boldsymbol{q}_l) = \sum_{i=1}^{M} w_i q_{k,i} q_{l,i} = |||\boldsymbol{q}_k|||^2 \delta_{k,l},$$

where $|||\boldsymbol{q}_k||| = (\boldsymbol{q}_k, \boldsymbol{q}_k)^{1/2}$ and the weights w_i are here equal to $1/M$. The m vectors $\boldsymbol{q}_k$ express as

$$\boldsymbol{q}_k = \boldsymbol{p}_k - \sum_{j=1}^{k-1} \frac{(\boldsymbol{p}_k, \boldsymbol{q}_j)}{(\boldsymbol{q}_j, \boldsymbol{q}_j)} \boldsymbol{q}_j, \tag{B.1}$$

and can be normalized. In that case, the relations (B.1) can be rewritten as

$$\boldsymbol{p}_k = \sum_{j=1}^{k} r_{j,k} \boldsymbol{q}_j, \quad |||\boldsymbol{q}_k||| = 1, \tag{B.2}$$

where the coefficients $r_{j,k}$ can be obtained from a weighted-QR factorization of the matrix $\mathcal{P} = [\boldsymbol{p}_k] \in \mathbb{R}^{M,m}$,

$$\mathcal{P} = \mathcal{Q}\mathcal{R}, \quad \text{with} \quad \left(\sqrt{\mathcal{W}}\mathcal{Q}\right)^\top \sqrt{\mathcal{W}}\mathcal{Q} = I_m,$$

where $\mathcal{Q} = \left[\boldsymbol{q}_i\right] \in \mathbb{R}^{M,m}$, $\mathcal{R} = \left[r_{i,j}\right] \in \mathbb{R}^{m,m}$ is an upper triangular matrix and $\mathcal{W} = \mathrm{diag}(1/M) \in \mathbb{R}^{M,M}$ is the diagonal weight matrix coming from the definition of the inner product. Algorithm 2 details the two steps to get the factorization of $\mathcal{P}$, namely (i) the QR factorization of the matrix $\sqrt{\mathcal{W}}\mathcal{P}$ to calculate the matrix $\mathcal{R}$, and (ii) the computation of the matrix $\mathcal{Q} = \left(\sqrt{\mathcal{W}}\right)^{-1} \mathcal{U}$ that ensures the normalization with respect to the inner product. Note that the m relations (B.2) must be inverted to calculate the $\boldsymbol{q}_k$ from the $\boldsymbol{p}_k$ which is tantamount to inverse the matrix $\mathcal{R}$.

Algorithm 2 Weighted-QR factorization

Input: $\mathcal{P}$ and $\mathcal{W}$

$\sqrt{\mathcal{W}}\mathcal{P} = \mathcal{U}\mathcal{R}$, ▷ $\mathcal{R}$ computation

$\mathcal{Q} = \left(\sqrt{\mathcal{W}}\right)^{-1} \mathcal{U}$, ▷ $\mathcal{Q}$ computation

Output: $\mathcal{R}^{-1}$ and $\mathcal{Q}$

Appendix C. Pick-Freeze method

The Pick-Freeze (PF) procedure allows to compute several types of sensitivity indices by reformulating the variances of conditional expectations in terms of covariances which are estimated by empirical estimators. In this section, the model output $Y(\boldsymbol{\xi})$ is assumed to be a function of N independent input parameters ξ_i collected into the vector $\boldsymbol{\xi} = (\xi_1, \dots, \xi_N)$. Let $\boldsymbol{i}$ denote a non-empty subset of indices such that $\boldsymbol{i} \subseteq \mathcal{I} = \{1, \dots, N\}$ and let $\boldsymbol{i}^{\mathrm{c}} = \mathcal{I} \setminus \boldsymbol{i}$ be the complementary set of $\boldsymbol{i}$ in $\mathcal{I}$.

C.1. Closed indices

The numerator of the closed index with respect to $\boldsymbol{\xi}_{\boldsymbol{i}} = (\xi_i, i \in \boldsymbol{i})$ can be viewed as the covariance between the model output and its PF replication (see, e.g., [51]),

$$\mathbb{V}(\mathbb{E}(Y(\boldsymbol{\xi})|\boldsymbol{\xi}_{\boldsymbol{i}})) = \mathbb{C}\mathrm{ov}\left(Y(\boldsymbol{\xi}), Y(\boldsymbol{\xi}_{\boldsymbol{i}}, \boldsymbol{\xi}^*_{\boldsymbol{i}^{\mathrm{c}}})\right).$$

Using two independent M-samples of input variables $\mathcal{X} = \{\boldsymbol{\xi}^{(m)}\}$ and $\mathcal{X}^* = \{\boldsymbol{\xi}^{*(m)}\}$, the PF replication of Y is obtained by holding $\boldsymbol{\xi}_{\boldsymbol{i}}$ (frozen variable) and by replacing $\boldsymbol{\xi}_{\boldsymbol{i}^{\mathrm{c}}}$ with $\boldsymbol{\xi}^*_{\boldsymbol{i}^{\mathrm{c}}}$ (picked variables). Once the M replications $Y(\boldsymbol{\xi}_{\boldsymbol{i}}, \boldsymbol{\xi}^*_{\boldsymbol{i}^{\mathrm{c}}})$ have been computed, the PF empirical estimator of the index $C_{\boldsymbol{i}}$ reads

$$C_{\boldsymbol{i}} \simeq \frac{\widehat{\mathbb{E}}(YY_{\boldsymbol{i}}) - \widehat{\mathbb{E}}(Y)\widehat{\mathbb{E}}(Y_{\boldsymbol{i}})}{\widehat{\mathbb{V}}(Y)},$$

where $\widehat{\mathbb{E}}(\cdot)$ and $\widehat{\mathbb{V}}(\cdot)$ denotes the empirical mean and variance, $Y = [Y\left(\boldsymbol{\xi}^{(m)}\right)]$ and $Y_{\boldsymbol{i}} = [Y\left(\boldsymbol{\xi}^{(m)}_{\boldsymbol{i}}, \boldsymbol{\xi}^{*(m)}_{\boldsymbol{i}^{\mathrm{c}}}\right)]$.

C.2. Cramér–Von Mises indices

The numerator's integrand of the Cramér–Von Mises index with respect to $\boldsymbol{\xi}_{\boldsymbol{i}}$ can be rewritten as [52],

$$\mathbb{V}(F_Y(z|\boldsymbol{\xi}_{\boldsymbol{i}})) = \mathbb{V}(\mathbb{E}\left(\mathbb{1}_{\{Y \leq z\}}|\boldsymbol{\xi}_{\boldsymbol{i}}\right)) = \mathbb{C}\mathrm{ov}\left(\mathbb{1}_{\{Y \leq z\}}, \mathbb{1}_{\{Y_{\boldsymbol{i}} \leq z\}}\right).$$

The computation of the index $\mathcal{W}_{\boldsymbol{i}}$ requires a third independent sample $\mathcal{Z} = \{z^{(m)}\}$ of the output to calculate the integral with respect to dF_Y, leading to the following PF empirical estimator,

$$\mathcal{W}_{\boldsymbol{i}} \simeq \frac{\sum_{z \in \mathcal{Z}} \left[\widehat{\mathbb{E}}(\mathbb{1}_{\{Y \leq z\}} \mathbb{1}_{\{Y_{\boldsymbol{i}} \leq z\}}) - \widehat{F}_Y(z)\widehat{F}_{Y_{\boldsymbol{i}}}(z)\right]}{\sum_{z \in \mathcal{Z}} \left[\widehat{F}_Y(z)(1 - \widehat{F}_Y(z))\right]},$$

where $\widehat{F}_Y(z)$ denotes the empirical cumulative distribution function defined as

$$\widehat{F}_Y(z) = \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}_{\{Y^{(m)} \leq z\}}.$$

References

[1] F. Ewart, A. Haworth, S. Wisbey, On the derivation of a sorption database, Tech. Rep., AEA Technology, Harwell Laboratory, Oxfordshire, OX11 ORA, UK, 1992.

[2] B. Grambow, Geological disposal of radioactive waste in clay, Elements 12 (4) (2016) 239–245, http://dx.doi.org/10.2113/gselements.12.4.239.

[3] P. Toulhoat, Confinement and migration of radionuclides in a nuclear waste deep repository, C. R. Phys. 3 (7–8) (2002) 975–986, http://dx.doi.org/10.1016/S1631-0705(02)01381-6.

[4] Z. Chen, G. Montavon, S. Ribet, Z. Guo, J.C. Robinet, K. David, C. Tournassat, B. Grambow, C. Landesman, Key factors to understand *in-situ* behavior of Cs in Callovo–Oxfordian clay-rock (France), Chem. Geol. 387 (2014) 47–58, http://dx.doi.org/10.1016/j.chemgeo.2014.08.008.

[5] M.H. Bradbury, B. Baeyens, A generalised sorption model for the concentration dependent uptake of caesium by argillaceous rocks, J. Contam. Hydrol. 42 (2–4) (2000) 141–163, http://dx.doi.org/10.1016/S0169-7722(99)00094-7.

[6] S. Gaboreau, F. Claret, C. Crouzet, E. Giffaut, C. Tournassat, Caesium uptake by Callovian–Oxfordian clayrock under alkaline perturbation, Appl. Geochem. 27 (6) (2012) 1194–1201, http://dx.doi.org/10.1016/j.apgeochem.2012.02.002.

[7] A. Ayoub, W. Pfingsten, L. Podofillini, G. Sansavini, Uncertainty and sensitivity analysis of the chemistry of cesium sorption in deep geological repositories, Appl. Geochem. (ISSN: 0883-2927) 117 (2020) 104607, http://dx.doi.org/10.1016/j.apgeochem.2020.104607.

[8] F. Sanson, O. Le Maître, P.M. Congedo, Systems of Gaussian process models for directed chains of solvers, Comput. Methods Appl. Mech. Engrg. (ISSN: 0045-7825) 352 (2019) 32–55, http://dx.doi.org/10.1016/j.cma.2019.04.013.
[9] S. Marque-Pucheu, G. Perrin, J. Garnier, An efficient dimension reduction for the Gaussian process emulation of two nested codes with functional outputs, Comput. Statist. (ISSN: 1613-9658) 35 (2020) 1059–1099, http://dx.doi.org/10.1007/s00180-019-00926-7.
[10] P.G. Constantine, E.T. Phipps, A lanczos method for approximating composite functions, Appl. Math. Comput. (ISSN: 0096-3003) 218 (24) (2012) 11751–11762, http://dx.doi.org/10.1016/j.amc.2012.05.009.
[11] P.G. Constantine, E.T. Phipps, T. Wildey, Efficient uncertainty propagation for network multiphysics systems, Internat. J. Numer. Methods Engrg. 99 (2013) http://dx.doi.org/10.1002/nme.4667.
[12] R. Bellman, Adaptive Control Processes: A Guided Tour, Princeton University Press, 1961, URL https://books.google.fr/books?id=POAmAAAAMAAJ.
[13] L. Van Der Maaten, E. Postma, J. Van den Herik, Dimensionality reduction: a comparative review, J. Mach. Learn. Res. 10 (2009) 66–71.
[14] P.G. Constantine, Active Subspaces: Emerging Ideas for Dimension Reduction in Parameter Studies, SIAM, USA, ISBN: 1611973856, 2015.
[15] O. Zahm, P.G. Constantine, C. Prieur, Y.M. Marzouk, Gradient-based dimension reduction of multivariate vector-valued functions, SIAM J. Sci. Comput. 42 (1) (2020) A534–A558, http://dx.doi.org/10.1137/18M1221837.
[16] E.C. Gaucher, C. Tournassat, F.J. Pearson, P. Blanc, C. Crouzet, C. Lerouge, S. Altmann, A robust model for pore-water chemistry of clayrock, Geochim. Cosmochim. Acta 73 (21) (2009) 6470–6487, http://dx.doi.org/10.1016/j.gca.2009.07.021.
[17] P. Sochala, C. Chiaberge, F. Claret, C. Tournassat, Uncertainty propagation in pore water chemical composition calculation using surrogate models, Sci. Rep. 12 (2022) http://dx.doi.org/10.1038/s41598-022-18411-5.
[18] C. Tournassat, H. Gailhanou, C. Crouzet, G. Braibant, A. Gautier, E.C. Gaucher, Cation exchange selectivity coefficient values on smectite and mixed-layer illite/smectite minerals, Soil Sci. Soc. Am. J. 73 (3) (2009) 928–942, http://dx.doi.org/10.2136/sssaj2008.0285.
[19] D.L. Parkhurst, C. Appelo, et al., Description of input and examples for PHREEQC version 3—a computer program for speciation, batch-reaction, one-dimensional transport, and inverse geochemical calculations, US Geol. Surv. Tech. Methods 6 (A43) (2013) 497.
[20] E. Giffaut, M. Grivé, P. Blanc, P. Vieillard, E. Colàs, H. Gailhanou, S. Gaboreau, N. Marty, B. Made, L. Duro, Andra thermodynamic database for performance assessment: ThermoChimie, Appl. Geochem. 49 (2014) 225–236, http://dx.doi.org/10.1016/j.apgeochem.2014.05.007.
[21] N. Lüthen, S. Marelli, B. Sudret, Sparse polynomial chaos expansions: Literature survey and benchmark, SIAM-ASA J. Uncertain. Quantif. 9 (2) (2021) 593–649, http://dx.doi.org/10.1137/20M1315774.
[22] H. Wold, Path models with latent variables: The NIPALS approach, in: H.M. Blalock, A. Aganbegian, F.M. Borodkin, R. Boudon, V. Capecchi (Eds.), Quantitative Sociology, in: International Perspectives on Mathematical and Statistical Modeling, Academic Press, ISBN: 978-0-12-103950-9, 1975, pp. 307–357, http://dx.doi.org/10.1016/B978-0-12-103950-9.50017-4.
[23] H. Wold, Soft modeling: the basic design and some extensions, systems under indirect observation, structure, prediction. Part II, Contrib. Econ. Anal. 139 (1982) 1–54.
[24] H. Wold, Partial least squares, in: S. Kotz, N.L. Johnson (Eds.), Encyclopedia of Statistical Sciences, Vol. 6, Wiley, New York, 1985, pp. 581–591.
[25] S. Wold, A. Ruhe, H. Wold, W.J. Dunn III, The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses, SIAM J. Sci. Stat. Comput. (ISSN: 0196-5204) 5 (3) (1984) 735–743, http://dx.doi.org/10.1137/0905052.
[26] A.-L. Boulesteix, K. Strimmer, Partial least squares: a versatile tool for the analysis of high-dimensional genomic data, Brief. Bioinform. (ISSN: 1467-5463) 8 (1) (2006) 32–44, http://dx.doi.org/10.1093/bib/bbl016.
[27] Y. Zhou, Z. Lu, J. Hu, Y. Hu, Surrogate modeling of high-dimensional problems via data-driven polynomial chaos expansions and sparse partial least square, Comput. Methods Appl. Mech. Engrg. (ISSN: 0045-7825) 364 (2020) 112906, http://dx.doi.org/10.1016/j.cma.2020.112906.
[28] R. Manne, Analysis of two partial-least-squares algorithms for multivariate calibration, Chemom. Intell. Lab. Syst. (ISSN: 0169-7439) 2 (1) (1987) 187–197, http://dx.doi.org/10.1016/0169-7439(87)80096-5.
[29] K.-A. Lê Cao, D. Rossouw, C. Robert-Granié, P. Besse, A sparse PLS for variable selection when integrating omics data, Stat. Appl. Genet. Mol. Biol. 7 (1) (2008) http://dx.doi.org/10.2202/1544-6115.1390.
[30] R.G. Ghanem, S.D. Spanos, Stochastic Finite Elements: A Spectral Approach, Springer Verlag, 1991.
[31] O.P. Le Maître, O.M. Knio, Spectral Methods for Uncertainty Quantification, in: Scientific Computation, Springer, 2010.
[32] M. Botti, D.A. Di Pietro, O. Le Maître, P. Sochala, Numerical approximation of poroelasticity with random coefficients using Polynomial Chaos and Hybrid High-Order methods, Comput. Methods Appl. Mech. Engrg. (ISSN: 0045-7825) 361 (2020) 112736, http://dx.doi.org/10.1016/j.cma.2019.112736.
[33] P. Sochala, F. De Martin, O. Le Maître, Model reduction for large-scale earthquake simulation in an uncertain 3D medium, Int. J. Uncertain. Quantif. (ISSN: 2152-5080) 10 (2) (2020) 101–127, http://dx.doi.org/10.1615/Int.J.UncertaintyQuantification.2020031165.
[34] P. Sochala, C. Chen, C. Dawson, M. Iskandarani, A polynomial chaos framework for probabilistic predictions of storm surge events, Comput. Geosci. 24 (1) (2020) 109–128, http://dx.doi.org/10.1007/s10596-019-09898-5.
[35] S. Oladyshkin, W. Nowak, Data-driven uncertainty quantification using the arbitrary polynomial chaos expansion, Reliab. Eng. Syst. Saf. (ISSN: 0951-8320) 106 (2012) 179–190, http://dx.doi.org/10.1016/j.ress.2012.05.002.
[36] J.A. Paulson, E.A. Buehler, A. Mesbah, Arbitrary polynomial chaos for uncertainty propagation of correlated random variables in dynamic systems, IFAC-PapersOnLine (ISSN: 2405-8963) 50 (1) (2017) 3548–3553, http://dx.doi.org/10.1016/j.ifacol.2017.08.954.
[37] J.D. Jakeman, F. Franzelin, A. Narayan, M. Eldred, D. Plfüger, Polynomial chaos expansions for dependent random variables, Comput. Methods Appl. Mech. Engrg. (ISSN: 0045-7825) 351 (2019) 643–666, http://dx.doi.org/10.1016/j.cma.2019.03.049.
[38] B. Efron, R.J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall, New York, 1993.
[39] S. Chatterjee, A new coefficient of correlation, J. Amer. Statist. Assoc. 116 (536) (2021) 2009–2022, http://dx.doi.org/10.1080/01621459.2020.1758115.
[40] R.B. Nelsen, An Introduction to Copulas, Springer, New York, 2006.
[41] I.M. Sobol, Sensitivity estimates for nonlinear mathematical models, Math. Model. Comput. Exp. 1 (1993) 407–414.
[42] A.B. Owen, Variance components and generalized Sobol' Indices, SIAM-ASA J. Uncertain. Quantif. 1 (1) (2013) 19–41, http://dx.doi.org/10.1137/120876782.
[43] S. Da Veiga, Global sensitivity analysis with dependence measures, J. Stat. Comput. Simul. 85 (7) (2015) 1283–1305, http://dx.doi.org/10.1080/00949655.2014.945932.
[44] I.M. Sobol, Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates, Math. Comput. Simulation 55 (1–3) (2001) 271–280, http://dx.doi.org/10.1016/S0378-4754(00)00270-6.
[45] B. Broto, F. Bachoc, M. Depecker, J.-M. Martinez, Sensitivity indices for independent groups of variables, Math. Comput. Simulation (ISSN: 0378-4754) 163 (2019) 19–31, http://dx.doi.org/10.1016/j.matcom.2019.02.008.
[46] Z. Wang, G. Jia, Extended sample-based approach for efficient sensitivity analysis of group of random variables, Reliab. Eng. Syst. Saf. (ISSN: 0951-8320) 231 (2023) 108991, http://dx.doi.org/10.1016/j.ress.2022.108991.
[47] J.-C. Fort, T. Klein, N. Rachdi, New sensitivity analysis subordinated to a contrast, Comm. Statist. Theory Methods 45 (15) (2016) 4349–4364, http://dx.doi.org/10.1080/03610926.2014.901369.
[48] F. Gamboa, T. Klein, A. Lagnoux, Sensitivity analysis based on Cramér–von Mises distance, SIAM-ASA J. Uncertain. Quantif. 6 (2) (2018) 522–548, http://dx.doi.org/10.1137/15M1025621.
[49] M. Andersson, A comparison of nine PLS1 algorithms, J. Chemom. 23 (10) (2009) 518–529, http://dx.doi.org/10.1002/cem.1248.
[50] U.G. Indahl, The geometry of PLS1 explained properly: 10 key notes on mathematical properties of and some alternative algorithmic approaches to PLS1 modelling, J. Chemom. 28 (3) (2014) 168–180, http://dx.doi.org/10.1002/cem.2589.
[51] A. Janon, T. Klein, A. Lagnoux, M. Nodet, C. Prieur, Asymptotic normality and efficiency of two Sobol index estimators, ESAIM Probab. Stat. 18 (2014) 342–364, http://dx.doi.org/10.1051/ps/2013040.
[52] F. Gamboa, P. Gremaud, T. Klein, A. Lagnoux, Global sensitivity analysis: A novel generation of mighty estimators based on rank statistics, Bernoulli 28 (4) (2022) 2345–2374, http://dx.doi.org/10.3150/21-BEJ1421.

Pierre Sochala received his Ph.D. in applied mathematics from École des Ponts in 2008. He is currently researcher in scientific computing at French alternative energies and atomic energy commission (CEA). His work is focused on the development and implementation of uncertainty quantification and Bayesian inference methods to a number of geosciences applications including seismic and acoustic wave propagation, subsurface and oceanic flows, tomography, and geochemistry.