

Uncertainty propagation in pore water chemical composition calculation using surrogate models

Pierre Sochala, Christophe Chiaberge, Francis Claret, Christophe Tournassat

▶ **To cite this version:**

Pierre Sochala, Christophe Chiaberge, Francis Claret, Christophe Tournassat. Uncertainty propagation in pore water chemical composition calculation using surrogate models. Scientific Reports, 2022, 12 (1), pp.15077. 10.1038/s41598-022-18411-5 . hal-03779374

HAL Id: hal-03779374
https://brgm.hal.science/hal-03779374

Submitted on 16 Sep 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

scientific reports

Check for updates

OPEN

Uncertainty propagation in pore water chemical composition calculation using surrogate models

Pierre Sochala[1,2]✉, Christophe Chiaberge[2], Francis Claret[2] & Christophe Tournassat[3,4]

Performance assessment in deep geological nuclear waste repository systems necessitates an extended knowledge of the pore water chemical conditions prevailing in host-rock formations. In the last two decades, important progress has been made in the experimental characterization and thermodynamic modeling of pore water speciation, but the influence of experimental artifacts and uncertainties of thermodynamic input parameters are seldom evaluated. In this respect, we conducted an uncertainty propagation study in a reference geochemical model describing the pore water chemistry of the Callovian-Oxfordian clay formation. Nineteen model input parameters were perturbed, including those associated to experimental characterization (leached anions, exchanged cations, cation exchange selectivity coefficients) and those associated to generic thermodynamic databases (solubilities). A set of 13 quantities of interest were studied by the use of polynomial chaos expansions built non-intrusively with a least-squares forward stepwise regression approach. Training and validation sets of simulations were carried out using the geochemical speciation code PHREEQC. The statistical results explored the marginal distribution of each quantity of interest, their bivariate correlations as well as their global sensitivity indices. The influence of the assumed distributions for input parameters uncertainties was evaluated by considering two parametric domain sizes.

Knowledge of pore water chemical composition is crucial for the building of nuclear waste repository performance assessments[1]. First, pore water chemical composition controls radionuclides solubility and adsorption properties on geological and engineered materials. Second, pore water chemical composition influences the transport and mechanical properties of clayey materials, which are essential constituents of existing multi-barrier concepts[2]. Third, pore water chemical composition dictates the nature and kinetics of chemical alteration processes of repository exogenous materials, such as concrete and nuclear glass[3]. But pore water chemical composition models contain a significant number of input parameters, exhibit strong nonlinearities, and have tightly coupled output results. For these reasons, it is difficult to estimate uncertainties on each of the input parameters and to evaluate the uncertainties of the model outputs, using e.g. error propagation methods. A direct sampling of clay pore water that retains the main characteristics representative of in situ conditions is particularly complex because of a range of side reactions taking place during sampling procedures[4]. Consequently, confidence in the knowledge of pore water chemical composition must be built on a consistent combination of several factors, which includes in situ seepage water collection and characterization, experimental water rock interactions results, and geochemical modeling results linking observations about solid material composition and reactivity with quantitative thermo-kinetic concepts[5]. While considerable effort with this experimental and modeling coupled approach enabled to produce predictive models for claystone pore water chemical composition that are consistent with experimental characterization, little attention has been directed to evaluate the model output uncertainties induced by input parameters uncertainties. Indeed, although the treatment of uncertainties in the performance assessment of geologic high-level radioactive waste repositories is recognized as an important topic for more than three decades[6], most of the studies focus on uncertainties related to retention processes[7]. Using Monte Carlo methods, the effects of database parameter uncertainty have been evidenced on geochemical equilibrium calculations but limited to very simple (as stated by the authors) modeling scenarios[8]. The goal of the present study is the implementation of a methodology based on surrogate models designed to propagate parametric uncertainties into a pore water chemical composition model with a moderate number of input parameters (around twenty).

Propagation of uncertainty gained wide popularity in many geosciences disciplines[9–12]. Its principle consists in perturbing a set of input parameters and then estimating the ensuing effects on the output quantities. The

[1]CEA, DAM, DIF, 91297 Arpajon, France. [2]BRGM, 3 avenue Claude Guillemin, 45060 Orléans, France. [3]ISTO, Université d'Orléans-CNRS-BRGM, Orléans, France. [4]Lawrence Berkeley National Laboratory, Berkeley CA, USA. ✉email: pierre.sochala@cea.fr

#	Type	Species	Unit	μ
1	Leached parameter	Cl^-	mmol L^{-1}	41
2		SO_4^{2-}	mmol L^{-1}	66
3	Exchanged cation	Na^+	mol L^{-1}	1.0824
4		K^+	mol L^{-1}	0.417
5		Ca^{2+}	mol L^{-1}	1.549
6		Mg^{2+}	mol L^{-1}	0.602
7		Sr^{2+}	mol L^{-1}	0.0737
8	Selectivity coefficients (log K_{ex} value)	Na^+/K^+	–	1.2
9		Na^+/Ca^{2+}	–	0.7
10		Na^+/Mg^{2+}	–	0.7
11		Na^+/Sr^{2+}	–	0.6
12	Solubility (log K value)	Celestite	–	−6.62
13		Calcite	–	−8.48
14		Dolomite	–	−17.12
15		Goethite	–	0.39
16		Quartz	–	−3.74
17		Pyrithe	–	−58.78
18		Ripidolite	–	61.35
19		Illite	–	11.54

Table 1. List of the 19 uncertain input parameters with their reference values μ (unperturbed state of the geochemical model).

interest of such statistical framework is to produce richer and more useful information than a single deterministic simulation can deliver. Parametric uncertainty analyses in geochemistry are motivated by different sources of uncertainty such as reaction kinetic rate constants, thermodynamic constants (e.g. solubility and aqueous complex formation constants), initial and boundary conditions, and transport properties[13–16]. Among available approaches, surrogate models have the advantage of providing a fast approximation everywhere in the parametric domain from a small ensemble of simulations, whereas Monte-Carlo techniques evaluate the direct model for a finite number of samples and require a large ensemble to achieve the convergence of the statistical estimators.

In this study, we are interested in using a surrogate model approach to propagate uncertainty into a pore water chemical composition model of the Callovian-Oxfordian (COx) clay formation in the Paris Basin (France), which has been the target of many studies investigating the feasibility of deep nuclear waste repository[17]. First, we briefly summarize the geochemical model and the parametric domain on which statistical approximations of the different quantities of interest (QoI) were built. Second, we describe the construction and validation of the surrogate models, with a Polynomial Chaos (PC) method and an orthogonal matching pursuit procedure, which are particularly efficient if the QoI exhibit smooth variations when the uncertain inputs vary. At last, we focus discussion on moments, marginal distributions, correlations and joint distributions as well as on global sensitivity indices, which quantify the influence of the input parameter distributions onto the variance of the QoIs.

Framework

Pore water composition model. The estimation of pore water chemical composition in the COx claystone relies on a geochemical model, of which complete description can be found in[4]. The model is briefly presented and made available in the form of a PHREEQC[18] input file and its associated database (THERMOCHIMIE v9b[19]) in the supplementary information file. The complete list of pore water chemical composition model input parameters are: Cl^- and SO_4^{2-} total concentration obtained from core sample leaching measurements; measured sodium Na^+, potassium K^+, calcium Ca^{2+}, magnesium Mg^{2+}, and strontium Sr^{2+} exchangeable concentrations; related Na^+/K^+, Na^+/Ca^{2+}, Na^+/Mg^{2+}, Na^+/Sr^{2+} cation exchange selectivity coefficients; and solubilities of Celestite, Calcite, Dolomite, Goethite, Quartz, Pyrite, Ripidolite, and Illite (corresponding to illite_Imt-2 of the database). The reference values of these $N = 19$ parameters are reported in Table 1.

Uncertainty model. Once the uncertain input parameters have been identified, the next step is to determine their statistical distributions. For a scalar parameter, it consists of specifying a range (or support) and an associated probability density function. The N uncertain inputs were collected into a random vector $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_N) \in \boldsymbol{\Xi} \subset \mathbb{R}^N$ whose components ξ_i were assumed to be independent and uniformly distributed over the range $[\xi_i^-, \xi_i^+]$, namely

$$\xi_i \sim \mathcal{U}\left(\xi_i^-, \xi_i^+\right), \quad \xi_i \perp \xi_j \quad \text{if} \quad i \neq j. \tag{1}$$

Species	Small range case			Large range case		
	σ	ξ^-	ξ^+	σ	ξ^-	ξ^+
Cl^-	2.05 (5%)	37.4	44.6	4.10 (10%)	33.9	48.1
SO_4^{2-}	3.30 (5%)	60.3	71.7	6.60 (10%)	54.6	77.4
Na^+	0.05 (5%)	0.99	1.17	0.1 (10%)	0.91	1.25
K^+	0.02 (5%)	0.38	0.45	0.04 (10%)	0.35	0.49
Ca^{2+}	0.08 (5%)	1.41	1.69	0.16 (10%)	1.27	1.83
Mg^{2+}	0.03 (5%)	0.55	0.65	0.06 (10%)	0.50	0.70
Sr^{2+}	0.04 (5%)	0.067	0.081	0.08 (10%)	0.06	0.088
Na^+/K^+	0.1 (8%)	1.03	1.37	0.2 (16%)	0.85	1.55
Na^+/Ca^{2+}	0.1 (14%)	0.53	0.87	0.2 (28%)	0.35	1.05
Na^+/Mg^{2+}	0.1 (14%)	0.53	0.87	0.2 (28%)	0.35	1.05
Na^+/Sr^{2+}	0.1 (17%)	0.43	0.77	0.2 (34%)	0.25	0.95
Celestite	0.05	−6.71	−6.53	0.1	−6.79	−6.45
Calcite	0.05	−8.57	−8.39	0.1	−8.65	−8.31
Dolomite	0.2	−17.5	−16.8	0.4	−17.8	−16.4
Goethite	0.2	0.044	0.74	0.4	−0.30	1.08
Quartz	0.05	−3.83	−3.65	0.1	−3.91	−3.57
Pyrite	0.2	−59.1	−58.4	0.4	−59.5	−58.1
Ripidolite	0.5	60.5	62.2	1	59.6	63.1
Illite	0.5	10.7	12.4	1	9.81	13.3

Table 2. Standard deviation σ, minimal value ξ^- and maximal value ξ^+ of the 19 uncertain input parameters for the small range case and the large range case.

The assumption of independence implies that the joint distribution $p_{\boldsymbol{\xi}}$ of the vector $\boldsymbol{\xi}$ and therefore its range $\boldsymbol{\Xi}$ factorizes to

$$p_{\boldsymbol{\xi}}(\boldsymbol{\xi}) = \prod_{i=1}^{N} p_{\xi_i}(\xi_i; \xi_i^-, \xi_i^+) \quad \text{and} \quad \boldsymbol{\Xi} = \prod_{i=1}^{N} \left[\xi_i^-, \xi_i^+\right]. \tag{2}$$

In case of a uniform distribution, the probability density function of each parameter ξ_i is defined as

$$p_{\xi_i}(\xi_i; \xi_i^-, \xi_i^+) := \begin{cases} 1/(\xi_i^+ - \xi_i^-), & \xi_i \in \left[\xi_i^-, \xi_i^+\right], \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

The extreme values ξ_i^- and ξ_i^+ of the ith parameter are defined as

$$\xi_i^- = \mu_i - \sqrt{3}\sigma_i \quad \text{and} \quad \xi_i^+ = \mu_i + \sqrt{3}\sigma_i, \tag{4}$$

where the mean $\mu_i = \mathbb{E}(\xi_i)$ corresponds here to the reference value indicated in Table 1 and the standard deviation $\sigma_i = \sqrt{\mathbb{V}(\xi_i)}$ is reported in Table 2. Recall that the mean $\mathbb{E}(\cdot)$ and the variance $\mathbb{V}(\cdot)$ of a random variable u are defined as

$$\mathbb{E}(u) := \int_{\boldsymbol{\Xi}} u(\boldsymbol{\xi}) p_{\boldsymbol{\xi}}(\boldsymbol{\xi}) d\boldsymbol{\xi}, \tag{5}$$

$$\mathbb{V}(u) := \mathbb{E}\left[(u - \mathbb{E}(u))^2\right]. \tag{6}$$

We have chosen the uniform distribution since it is the maximum entropy distribution[20,21] among all continuous distributions which are supported in a given finite range. The maximum entropy distribution is often preferred because it represents the least informative distribution but other types of distributions can be adopted. Two cases were considered in order to investigate the effect of the amplitude of perturbations around the reference values onto the uncertainty of the QoIs. Hereafter, these cases are referred to as the "small range case" and "large range case", respectively. Each parameter range of the latter case is twice the range of the former case, implying from Eq. (2) that the measure (or area) of the parametric domain $\boldsymbol{\Xi}$ in the large range case is $2^N \simeq 5 \cdot 10^5$ times higher than in the small range case.

Quantities of interest. We are interested in pH, pe + pH where pe is the redox potential, total aqueous concentrations of sodium (Na^+), potassium (K^+), calcium (Ca^{2+}), magnesium (Mg^{2+}), strontium (Sr^{2+}), iron

(Fe), silicium (Si), aluminum (Al), sulphate (S(VI)) and sulfur (S(−II)), as well as the $\log_{10}$ of CO_2 partial pressure $\log_{10} p_{CO_2}$. The set of these $\mathcal{N} = 13$ QoI is denoted $\mathbb{U}$,

$$\mathbb{U} := \left\{\text{pH}, \text{pe} + \text{pH}, \text{Na}^+, \text{K}^+, \text{Ca}^{2+}, \text{Mg}^{2+}, \text{Sr}^{2+}, \text{Fe}, \text{Si}, \text{Al}, \text{S(VI)}, \text{S}(-\text{II}), \log_{10} p_{CO_2}\right\}. \tag{7}$$

Surrogate model

The non-intrusive construction of a surrogate model relies on a training set $\mathcal{X} := \{\boldsymbol{\xi}^{(m)}\}$ that samples the parametric domain. The corresponding outputs were computed using PHREEQC, and we obtained $\mathcal{U} := \{u^{(m)} := u(\boldsymbol{\xi}^{(m)})\}$ for each $u \in \mathbb{U}$. The input-output relations $\boldsymbol{\xi}^{(m)} \to u^{(m)}$ were then exploited to build an approximation of u over the whole parametric domain. Several families of methods have been developed over the past decades to construct surrogate models including Gaussian processes[22] and (possibly deep) neural networks[23]. In this study, in which 19 input parameters were perturbed, we chose polynomial chaos surrogates[24,25] for their relatively low computational costs of construction in moderate dimensional case. In this section, after a brief description of PC expansions and a short reminder on the least squares method, we present the orthogonal matching pursuit procedure as well as the validation of the surrogate models.

Polynomial chaos. Any random variable u with finite variance can be approximated by a spectral expansion[26,27] of the form

$$u^{\mathscr{K}}(\boldsymbol{\xi}) = \sum_{\boldsymbol{k}\in\mathscr{K}} u_{\boldsymbol{k}}\phi_{\boldsymbol{k}}(\boldsymbol{\xi}), \tag{8}$$

where $\{u_{\boldsymbol{k}}\}$ is the set of spectral coefficients of $u^{\mathscr{K}}$ and $\{\phi_{\boldsymbol{k}}(\boldsymbol{\xi})\}$ is a complete orthogonal set constituting a basis of $L_2(\boldsymbol{\Xi}, p_{\boldsymbol{\xi}})$. The $\phi_{\boldsymbol{k}}(\boldsymbol{\xi})$ are N-variate Legendre polynomials for uniform distributions as is the case here. Each multivariate polynomial is defined by an integer-valued multi-index $\boldsymbol{k} = (k_1, \ldots, k_N) \in \mathbb{N}^N$ where k_i is the polynomial degree associated to the ith variable ξ_i. The truncated PC expansion (8) is then defined using a finite set $\mathscr{K}$ of multi-indices and we denote $N_\text{b} := \left|\mathscr{K}\right|$ the PC basis dimension. Sets of multi-indices are often chosen by prescribing a maximal degree d° leading to

$$\mathscr{K}(d^\circ) = \left\{\boldsymbol{k} \in \mathbb{N}^N, \|\boldsymbol{k}\|_1 \leq d^\circ\right\} \quad \text{and} \quad N_\text{b}(d^\circ) = \frac{(N+d^\circ)!}{N!d^\circ!}. \tag{9}$$

Least squares method is an efficient approach to estimate the spectral coefficients but cannot be applied if the sample size M is much lower than the PC basis dimension N_b. In this case, more advanced methods are used to produce sparse PC.

Ordinary least squares. A first way of estimating the spectral coefficients of a PC expansion is to use the Ordinary Least Squares (OLS) method that consists of minimizing the squared norm of the residual,

$$\min_{\boldsymbol{u}} \|A\boldsymbol{u} - \mathbb{u}\|_2^2, \tag{10}$$

where $A \in \mathbb{R}^{M,N_\text{b}}$ is the matrix of basis functions $\phi_{\boldsymbol{k}}(\boldsymbol{\xi}^{(m)})$, $\boldsymbol{u} \in \mathbb{R}^{N_\text{b}}$ collects the spectral coefficients $u_{\boldsymbol{k}}$ and $\mathbb{u} \in \mathbb{R}^M$ is the vector of model output $u(\boldsymbol{\xi}^{(m)})$. The solution of the minimization problem (10) satisfies the system of normal equations

$$A^\top A\boldsymbol{u} = A^\top \mathbb{u}, \tag{11}$$

provided that the matrix $A^\top A$ is invertible.

Orthogonal matching pursuit. When dealing with high dimensional case, sparse approximation theory has been developed for finding solutions to underdetermined linear systems under sparsity constraint. Such parsimonious solutions can be justified by the sparsity-of-effects principle stating that most models are usually dominated by main effects and low-order interactions[28]. This principle is illustrated in PC by sparse expansions in which most of the coefficients are zeroes.

Numerous algorithms have been developed recently for the computation of sparse PC expansions (see[29] for a review of the existing methods). We relied here on the Orthogonal Matching Pursuit (OMP) method that is a classical greedy algorithm to select a set of active basis functions among a large set (or dictionary) of functions. Initially developed in signal processing[30], the matching pursuit algorithm starts with an empty approximation and adds sequentially the most correlated basis function to the current residual. The index γ^k of the new basis function satisfies (for $k \geq 1$),

$$\gamma^k = \arg\max_j \left(\left|\mathbb{d}_j^\top \mathbb{r}^{k-1}\right|\right), \tag{12}$$

where $\mathbb{d}_j \in \mathbb{R}^M$ is the j-th column of the dictionary D and $\mathbb{r}^{k-1} \in \mathbb{R}^M$ the current residual. The orthogonal version of the method[31] computes the coefficients of the approximation to ensure that the residual is orthogonal to the span of the active functions,

QoI #	1	2	3	4	5	6	7	8	9	10	11	12	13
Small range case	320	530	820	860	380	460	1130	420	1210	640	1170	960	190
Large range case	320	490	640	790	530	460	820	370	780	820	890	900	180

Table 3. Number of terms retained in the OMP method for each QoI.

$$\left(A^k\right)^\top A^k \boldsymbol{u}^k = \left(A^k\right)^\top \mathfrak{u}, \tag{13}$$

where $A^k \in \mathbb{R}^{M,k}$ is the matrix of the active basis functions at iteration k. The OMP method is a least-squares forward stepwise regression approach that can be easily implemented (see Supplementary material for the detailed algorithm). Several criteria are possible to stop the iterations, such as the residual norm or cross-validation errors. Here, we compute every ten iterations (until 1500) the Mean Squared Error (MSE) using a validation set $\mathscr{X}_*$ of M_* realizations, $\text{MSE} := \sum_{\boldsymbol{\xi}\in\mathscr{X}_*} \left(u(\boldsymbol{\xi}) - u^{\mathscr{K}}(\boldsymbol{\xi})\right)^2 / M_*$, and then select the number of active functions that minimizes the MSE.

Validation. We assessed and compared the PC expansions computed using either the OLS or OMP methods. Except for pH, pe + pH, S(VI) and $\log_{10} p_{CO_2}$, a logarithmic transformation improved the surrogate approximations. Indeed, the log variables exhibited smoother dependences with respect to the uncertain input parameters than the original ones, and their use reduced the approximation errors of the original variables. In practice, the change of variable is trivial and consists of (i) building a PC expansion $v^{\mathscr{K}}(\boldsymbol{\xi})$ of $v(\boldsymbol{\xi}) := \log(u(\boldsymbol{\xi}))$ using the set $\mathscr{V}$ of logarithmically transformed outputs

$$\mathscr{V} := \left\{v^{(m)} := \log\left(u^{(m)}\right)\right\}, \tag{14}$$

and (ii) applying the backward transformation to retrieve the original variables

$$u^{\mathscr{K}}(\boldsymbol{\xi}) := \exp\left(v^{\mathscr{K}}(\boldsymbol{\xi})\right). \tag{15}$$

The PC expansions were built with a training set $\mathscr{X}$ of $M = 10^4$ Monte-Carlo realizations and their errors were estimated using an independent validation set $\mathscr{X}_*$ of $M_* = 10^4$ Monte-Carlo realizations. The accuracy of four PC expansions were compared with three obtained with the OLS method in which different maximal degrees were used $d^\circ = 1, 2, 3$, and one obtained with the OMP using a dictionary of $N_b(5) = 42504$ functions. The number of PC basis functions for the OLS method is $N_b(1) = 21$, $N_b(2) = 210$, $N_b(3) = 1540$ while the number of active functions retained in the OMP method depends on the QoI and is reported in Table 3.

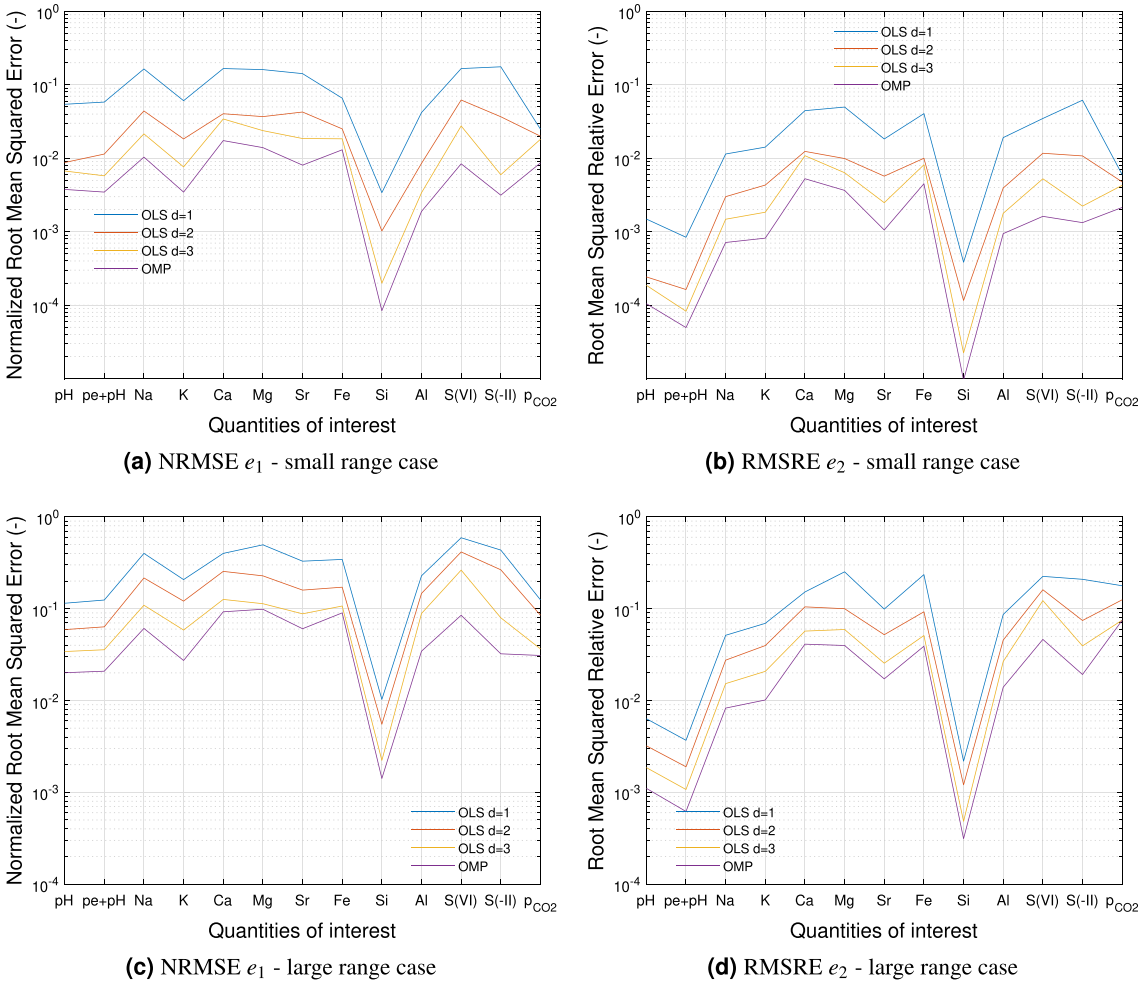
Two error metrics were used to estimate the accuracy of the approximations (Fig. 1): the root mean squared error normalized by the empirical variance $\widehat{\mathbb{V}}_{\mathscr{X}_*}(\cdot)$ of the QoI,

$$e_1 := \left[\frac{1}{M_*}\sum_{\boldsymbol{\xi}\in\mathscr{X}_*} \frac{\left(u(\boldsymbol{\xi}) - u^{\mathscr{K}}(\boldsymbol{\xi})\right)^2}{\widehat{\mathbb{V}}_{\mathscr{X}_*}(u)}\right]^{1/2}, \tag{16}$$

and the root mean squared relative error,

$$e_2 := \left[\frac{1}{M_*}\sum_{\boldsymbol{\xi}\in\mathscr{X}_*} \left(\frac{u(\boldsymbol{\xi}) - u^{\mathscr{K}}(\boldsymbol{\xi})}{u(\boldsymbol{\xi})}\right)^2\right]^{1/2}. \tag{17}$$

The global normalization of error e_1 allows to express the approximation error of the QoI in comparison with its uncertainty level whereas the local normalization of error e_2 is suitable when the approximation error and/or the QoI have different magnitudes across the parametric domain. The error levels obtained for the large range case were higher than for the small range case (roughly one order of magnitude) because large variations of input parameters induced more complex dependencies in geochemical reactions. As expected, the errors associated with the OLS method decreased when the maximal degree increased since the addition of higher order terms improved the approximations of the stochastic nonlinearities. A further increase of the maximal degree was not an option to reduce the error because the number of basis functions $N_b(4) = 8855$ was too close to the sample size $M = 10^4$, thereby producing an ill-conditioned matrix $A^\top A$ in (11). On the contrary, the PC expansions obtained by the OMP method exhibited a higher accuracy and a lower number of terms (Table 3). Therefore, in subsequent analyses, we used the OMP surrogate models for which the error level was at most 1% for e_1 and 0.5% for e_2 in the small range case and 10% for e_1 and 7% for e_2 in the large range case. Lastly, we note that the input parameters distributions can be changed retroactively on a subset of the parametric domain provided that the surrogate model error is sufficiently low over this subset.



(a) NRMSE e_1 - small range case

(b) RMSRE e_2 - small range case

(c) NRMSE e_1 - large range case

(d) RMSRE e_2 - large range case

Figure 1. Validation errors for different PC expansions.

Results and discussion

A direct exploitation of the PC coefficients was not feasible because of the logarithmic transformation. Statistical information were then derived promptly from extensive samplings of the surrogate models. We processed each QoI individually by studying their moments and marginal distributions. We then computed the correlations and plotted the joint distributions of the most correlated pairs of QoI. In closing, a global sensitivity analysis was carried out in order to rank the contribution of the uncertain input parameters onto the variance of each QoI.

Moments. The empirical estimators of the mean μ, the standard deviation σ and the coefficient of variation $c_v = \sigma/\mu$ of each QoI (Table 4) were obtained from a set $\mathcal{Y}$ of $N = 10^6$ Monte-Carlo realizations of the surrogate models,

$$\widehat{\mu} = \widehat{\mathbb{E}}(u^{\mathscr{K}}) := \frac{1}{N}\sum_{\boldsymbol{\xi}\in\mathcal{Y}} u^{\mathscr{K}}(\boldsymbol{\xi}) \quad \text{and} \quad \widehat{\sigma} := \left[\frac{1}{N-1}\sum_{\boldsymbol{\xi}\in\mathcal{Y}}\left(u^{\mathscr{K}}(\boldsymbol{\xi}) - \widehat{\mu}\right)^2\right]^{1/2}. \tag{18}$$

For most QoI, mean values and standard deviations had the same characteristics as the uncertain input parameters for the large and small range cases, i.e. the means were roughly identical while the standard deviations were multiplied by a factor 2. Iron and Aluminum concentrations were an exception because their mean were respectively 4 (Fe) and 1.7 (Al) times higher for the large range case than for the small range case. The ratio is 13 (Fe) and 3.5 (Al) for their standard deviation. Except for Al in the small range case, their standard deviations were larger than their mean values, pointing out a high dependence of Al and Fe concentrations to input model parameters variations. Low concentration of these two elements and a tight coupling of solubility controls exerted by two mineral phases, Ripidolite and Illite, for which the chosen uncertainty on solubility products were the highest, explained these findings (Table 2). On the contrary, pH values were remarkably stable despite the complex coupled control on this parameter exerted by many phases in the system[5], thus showing a strong thermodynamic buffering of this parameter by the mineralogical assemblage.

QoI	Small range case					Large range case				
	$\widehat{\mu}$	$\widehat{\sigma}$	$\widehat{s}$	$\widehat{k}$	$\widehat{c_v}$ [%]	$\widehat{\mu}$	$\widehat{\sigma}$	$\widehat{s}$	$\widehat{k}$	$\widehat{c_v}$ [%]
pH	7.17	0.20	−0.004	2.66	3	7.19	0.39	−0.03	2.68	5
pe + pH	4.20	$6.13 \cdot 10^{-2}$	0.043	2.79	1.5	4.20	$1.26 \cdot 10^{-1}$	0.17	2.97	3
Na^+	$4.29 \cdot 10^{-2}$	$3.05 \cdot 10^{-3}$	0.2	2.93	7	$4.43 \cdot 10^{-2}$	$7.07 \cdot 10^{-3}$	0.94	5.85	16
K^+	$1.04 \cdot 10^{-3}$	$2.50 \cdot 10^{-4}$	0.38	2.27	24	$1.17 \cdot 10^{-3}$	$5.75 \cdot 10^{-4}$	0.92	3.85	49
Ca^{2+}	$8.46 \cdot 10^{-3}$	$2.21 \cdot 10^{-3}$	0.52	3.16	26	$9.33 \cdot 10^{-3}$	$4.80 \cdot 10^{-3}$	1.48	7.89	51
Mg^{2+}	$5.91 \cdot 10^{-3}$	$2.14 \cdot 10^{-3}$	0.62	3.13	36	$6.68 \cdot 10^{-3}$	$5.06 \cdot 10^{-3}$	2.61	21.0	76
Sr^{2+}	$2.10 \cdot 10^{-4}$	$2.87 \cdot 10^{-5}$	0.45	3.08	14	$2.17 \cdot 10^{-4}$	$6.42 \cdot 10^{-5}$	1.05	4.95	30
Fe	$4.59 \cdot 10^{-5}$	$5.48 \cdot 10^{-5}$	3.33	20.0	119	$1.85 \cdot 10^{-4}$	$6.99 \cdot 10^{-4}$	13.1	319.0	378
Si	$1.83 \cdot 10^{-4}$	$2.11 \cdot 10^{-5}$	0.14	1.83	12	$1.87 \cdot 10^{-4}$	$4.29 \cdot 10^{-5}$	0.28	1.89	23
Al	$9.34 \cdot 10^{-8}$	$5.25 \cdot 10^{-8}$	1.05	3.83	56	$1.55 \cdot 10^{-7}$	$1.83 \cdot 10^{-7}$	2.38	10.9	117
S(VI)	$1.46 \cdot 10^{-2}$	$3.00 \cdot 10^{-3}$	0.32	2.84	21	$1.66 \cdot 10^{-2}$	$8.63 \cdot 10^{-3}$	2.43	18.3	52
S(−II)	$7.50 \cdot 10^{-10}$	$3.93 \cdot 10^{-10}$	1.80	8.15	52	$1.31 \cdot 10^{-9}$	$1.84 \cdot 10^{-9}$	5.65	62.2	141
$\log_{10} p_{CO_2}$	−2.07	0.41	−0.014	2.66	−20	−2.11	0.79	−0.058	2.62	−38

Table 4. Empirical mean $\widehat{\mu}$, standard deviation $\widehat{\sigma}$, skewness $\widehat{s}$, kurtosis $\widehat{k}$ and coefficient of variation $\widehat{c_v}$ estimated with 10^6 realizations of the surrogate models.

Marginal distributions. Small range case results exhibited three types of empirical marginal distribution profiles (Fig. 2): bell-shaped distributions for pH, pe + pH (not shown), Na^+, K^+, Ca^{2+}, Mg^{2+}, Sr^{2+}, S(VI), and $\log_{10} p_{CO_2}$; right-skewed distributions for Al, S(−II), and Fe; and a piecewise linear distribution for Si. Large range case results led to a flattening of the distributions (except for Fe), which was coherent with variances increase. The shape of a distribution can be described by skewness s and kurtosis k that are defined as the third and fourth standardized moments, respectively. The empirical estimators of s and k, indicated in Table 4, are

$$\widehat{s} := \frac{\widehat{\mathbb{E}}\left[\left(u^{\mathscr{K}}(\boldsymbol{\xi}) - \widehat{\mu}\right)^3\right]}{\widehat{\sigma}^3} \quad \text{and} \quad \widehat{k} := \frac{\widehat{\mathbb{E}}\left[\left(u^{\mathscr{K}}(\boldsymbol{\xi}) - \widehat{\mu}\right)^4\right]}{\widehat{\sigma}^4}. \tag{19}$$

The skewness of a distribution measures its asymmetry and a distribution is commonly said to be fairly symmetrical if $|s| \leq 1/2$, moderately skewed if $1/2 \leq |s| \leq 1$ and highly skewed if $|s| \geq 1$. In the small range case, we observed a slight asymmetry for Ca^{2+} and Mg^{2+} and a high asymmetry for Al, S(−II), and Fe. In the large range case, the asymmetry became important for all the QoIs except for pH, pH + pe, Si and $\log_{10} p_{CO_2}$. The kurtosis of a distribution measures the combined weight of the tails relative to the rest of the distribution. It is common to compare the kurtosis to 3 which is the kurtosis of a normal distribution; a high kurtosis ($k > 3$) indicates heavy tails while low kurtosis ($k < 3$) denotes light tails. In the small range case, the kurtosis is between 2 and 4 except for Fe and S(−II) which have strong heavy-tailed distributions and Si due to its piecewise linear distribution. In the large range case, we observed that the kurtosis of each distribution increases substantially (except for pH, pH + pe, Si and $\log_{10} p_{CO_2}$), meaning that the heaviness of the tails grows in importance. We noted that Fe was the only quantity of which the distribution was more peaked for the larger parametric domain; the mean values obtained with each of these cases were significantly different but the medians were very close (Fig. 2).

Linear correlations. Linear correlation between two random variables u and v were measured with the Pearson's correlation coefficient $r(u, v) \in [-1, 1]$ defined as follows

$$r(u, v) := \frac{\mathbb{C}\mathrm{ov}(u, v)}{\sigma(u)\sigma(v)}, \tag{20}$$

where $\mathbb{C}\mathrm{ov}(u, v) := \mathbb{E}[(u - \mathbb{E}(u))(v - \mathbb{E}(v))]$ is the covariance between u and v. The square of Pearson's coefficient is the coefficient of determination $R^2(u, v) := r(u, v)^2 \in [0, 1]$, which represents the percentage of variation of u due to a linear variation of v.

Empirical estimates of $r(u, v)$ and $R^2(u, v)$ are plotted on Fig. 3 in which the lower (resp. upper) parts of the matrices correspond to the small (resp. large) range case. Three pairs presented a particularly strong correlation regardless of the parametric domain size: the pairs (pH, pe + pH) and (pH, $\log_{10} p_{CO_2}$) were negatively correlated with $R^2 = 88\%$ and $R^2 = 97\%$ respectively, whereas the pair (pe + pH, $\log_{10} p_{CO_2}$) was positively correlated with $R^2 = 80\%$. The (pH, $\log_{10} p_{CO_2}$) pair correlation can be understood by noting that the standard deviation of Ca^{2+} concentration (Table 2) was small compared to its mean value (Table 1) and that the $\log_{10} p_{CO_2}$ value is directly related to pH by the Calcite equilibrium reaction. The correlation in the pair (pH, pe + pH) cannot be explained by the known negative correlation of the pair (pe, pH) at constant dioxygen or dihydrogen fugacity through corresponding Nernst's equation, which results in a −1 slope in the pe − pH diagram representation: the QoI transformation from pe to pe + pH was indeed meant to suppress this correlation. Consequently, the observed negative correlation must be attributed to particular equilibrium reactions. Goethite equilibrium, the reaction of which results in a −3 slope in a pe − pH diagram, may explain the observed correlation.

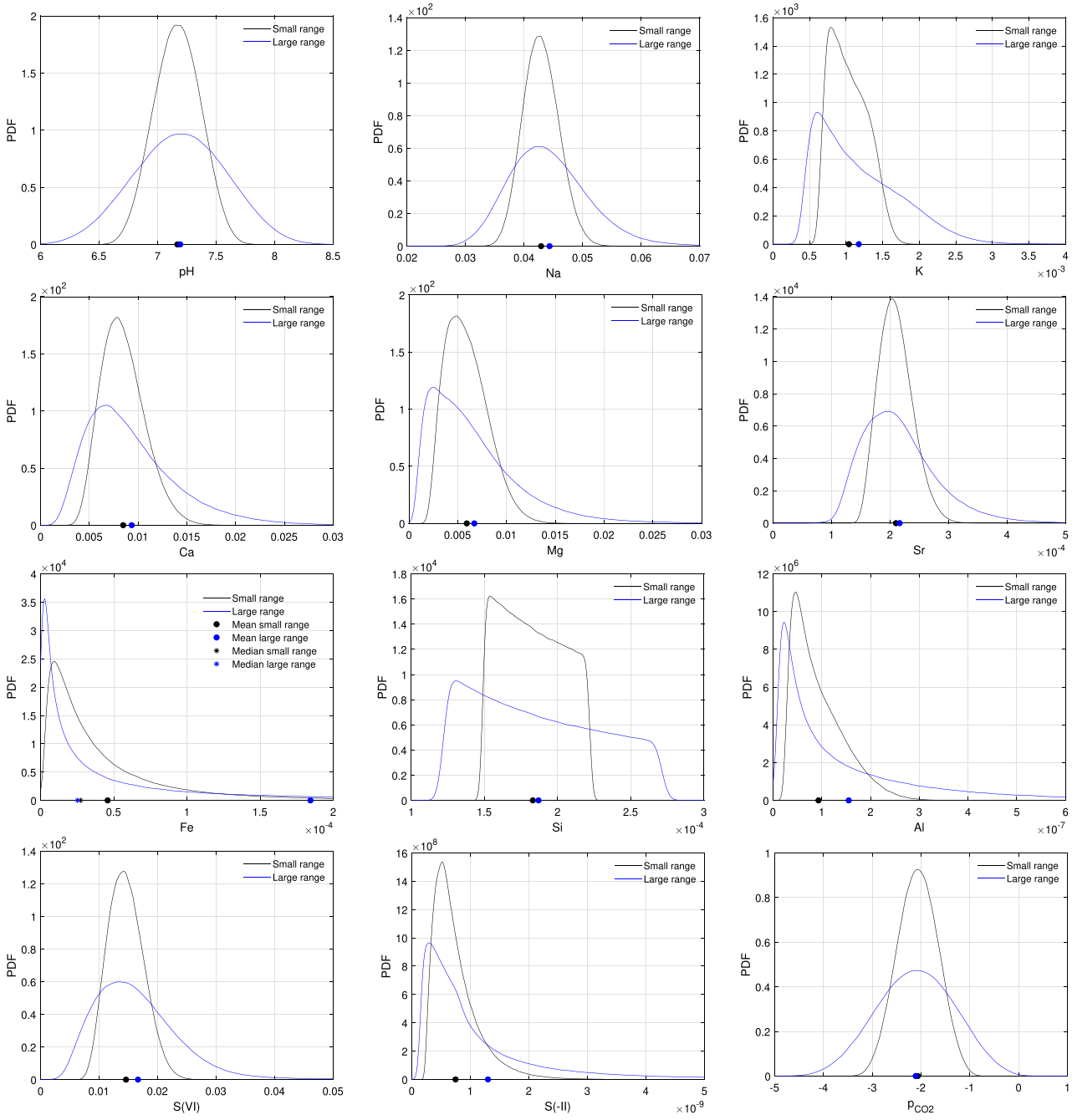
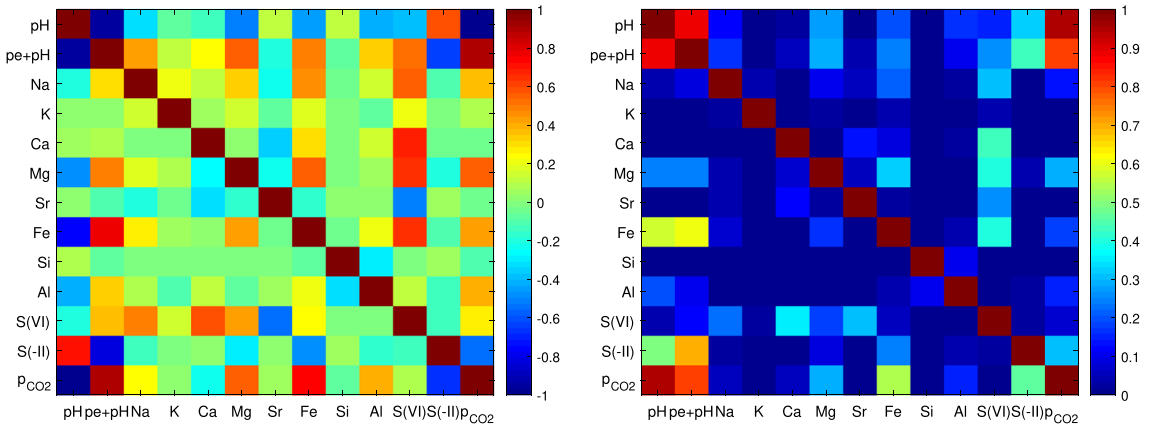


Figure 2. Marginal distributions and means of the QoI for the small range case (black curves) and the large range case (blue curves) estimated with 10^6 realizations of the surrogate models and the kernel density estimation method[32].

Four pairs had a moderate correlation for the small range case that significantly decreased for the large range case: the pairs (pH, Fe) and (pe + pH, S(−II)) were negatively correlated with $R^2 = 57\%$ and $R^2 = 70\%$. The correlation is well explained by the sensitivity of S(−II) and Fe concentration to redox conditions. The pairs (Fe, $\log_{10} p_{CO_2}$) and (pe + pH, Fe) were positively correlated with $R^2 = 53\%$ and $R^2 = 60\%$. Inversely, the correlation of some pairs involving S(VI) was higher for the large range case: (Fe, S(VI)), (Ca^{2+}, S(VI)), and (Mg^{2+}, S(VI)) with $R^2 = 46\%$, $R^2 = 42\%$, and $R^2 = 40\%$, respectively (instead of 0.5%, 35%, and 19% for the small range case). These observations can be related to the charge balance requirement in aqueous solution during the calculation. In the model, Na^+ and Cl^- total concentrations (aqueous + exchange) are stabilized at their final values before the reaction step with minerals. Mineral phases exert no further control on their concentrations. Hence, a variation of S(VI) concentration, which is the second major anion in solution, must be compensated by an equivalent variation of cations concentrations to fulfill solution electroneutrality. This compensation is mostly achieved by Ca^{2+}, Mg^{2+}, and Fe because Na^+ total concentration is fixed by the amount available on the



(a) $r(u,v)$ **(b)** $R^2(u,v)$

Figure 3. Matrices of correlation coefficient $r(u, v)$ and coefficient of determination $R^2(u, v)$ for the small range case (lower part of matrices) and the large range case (upper part of matrices) estimated with 10^6 realizations of the surrogate models.

cation exchanger, and because Sr^{2+} concentration is controlled by Celestite solubility, which is itself linked to S(VI) concentration.

Bivariate distributions. The shapes of the isolines contours of the most correlated pairs of QoI ($R^2 > 50\%$) were clearly consistent with the sign of the correlation coefficient (Fig. 4), namely negative for the pairs $(\mathrm{pH}, \log_{10} p_{CO_2})$, (pH, Fe), (pe + pH, S(−II)) and positive for the pairs $(\mathrm{pe} + \mathrm{pH}, \log_{10} p_{CO_2})$, $(\mathrm{Fe}, \log_{10} p_{CO_2})$, (pe + pH, Fe). Also, the pairs $(\mathrm{pH}, \log_{10} p_{CO_2})$ and $(\mathrm{pe} + \mathrm{pH}, \log_{10} p_{CO_2})$ followed a bivariate normal distribution whereas the other pairs exhibited more complex asymmetrical distributions. Isolines of the pair (pH, pe + pH) had the same pattern as those of the pair $(\mathrm{pH}, \log_{10} p_{CO_2})$ (not shown).

Global sensitivity analysis. An essential aspect of uncertainty propagation is the global sensitivity analysis[33,34], which quantifies the relative contribution of each uncertain input parameter (or group of input parameters) to the variance of the QoI. This analysis across the whole parametric domain should not be confused with local sensitivity analysis[35], which estimates the effect of small perturbations around specific input values by means of the partial derivatives of the model. The global sensitivity analysis was based on the decomposition of the total variance[36] into $2^N - 1$ terms ($N = 19$ in this study), as follows

$$\mathbb{V}(u) = \sum_{i=1}^{N} \mathbb{V}_i + \sum_{i<j} \mathbb{V}_{ij} + \cdots + \mathbb{V}_{1\ldots N}, \tag{21}$$

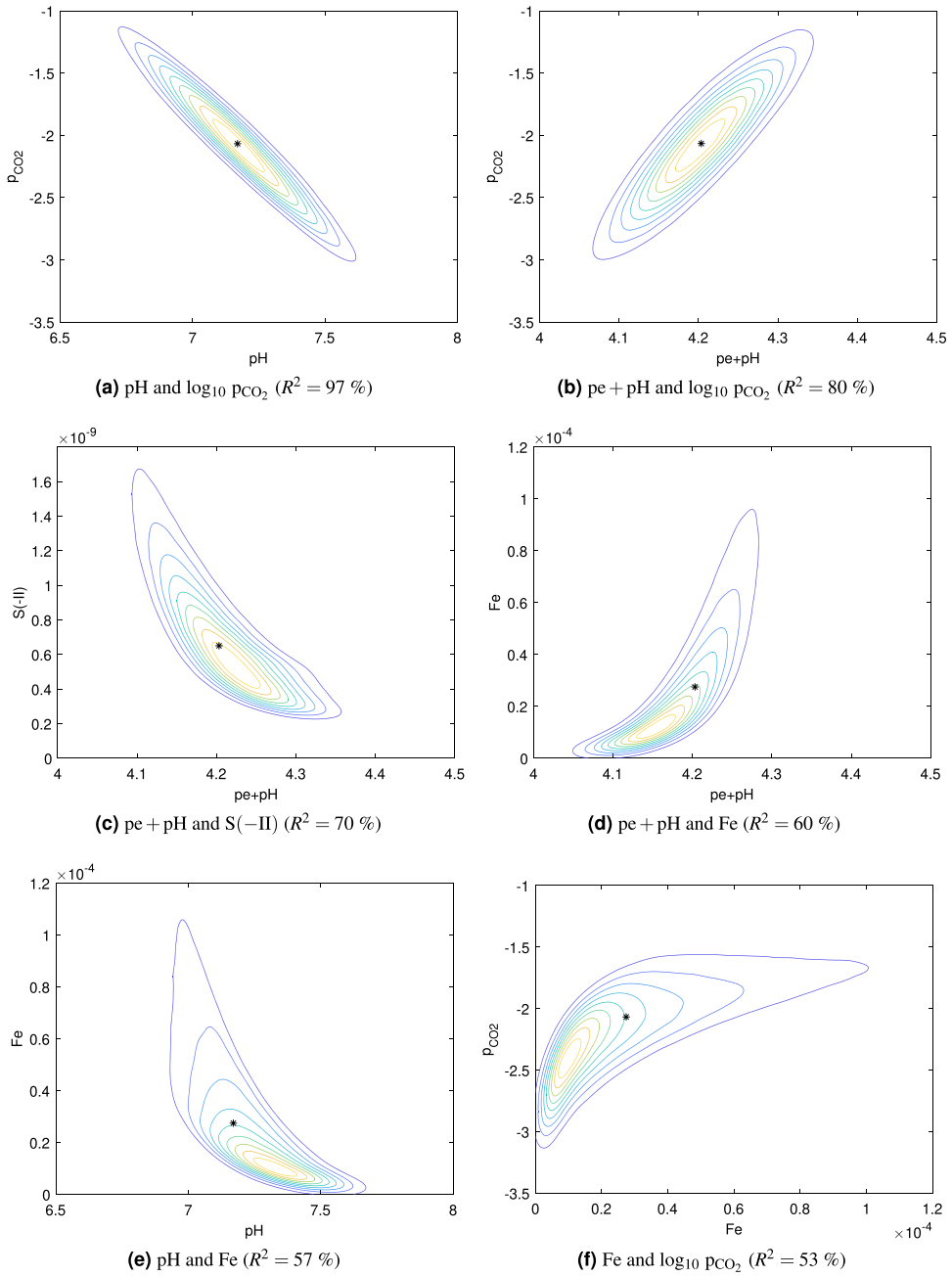
where $\{\mathbb{V}_i\}$ are the first-order interaction terms, $\{\mathbb{V}_{ij}\}$ the second order terms, and so on. Of particular interest are the $\mathbb{V}_i$ which measure the own effects of the input parameter ξ_i on the output variance. Typically, these effects are normalized by the total variance defining the first-order sensitivity indices S_i by

$$S_i := \frac{\mathbb{V}_i}{\mathbb{V}}. \tag{22}$$

The first-order sensitivity indices were estimated from the Monte-Carlo pick-freeze algorithm[33,37], which requires a sample of size M of the input variables (Fig. 5). For a given case, the number of surrogate model evaluations was $M(N + 1) = 2 \cdot 10^7$ for each QoI. A sum of the first-order indices close to one is representative of low interactions between parameters and of an essentially additive model. Interaction effects were minor for the small range case (except for Fe and S(−II)), but increased significantly for the large range case. The first-order sensitivity indices of eight quantities, pH, pe + pH, Mg^{2+}, Fe, Si, Al, S(−II), $\log_{10} p_{CO_2}$ were mainly governed by solubilities, while four other quantities, Na^+, Ca^{2+}, Sr^{2+}, S(VI), depended on the four input categories. In addition, three QoIs were strongly dependent on a single parameter: $\log K_{ex}^{Na^+/K^+}$ for K^+ consistently with the known control of K^+ concentration by cation exchange reactions in clay minerals rich systems[38]; quartz for Si consistently with the negligible variation of quartz solubility product and with Si aqueous speciation in the explored range of pH variations; and Illite for Al consistently with the fact that only Illite and Ripidolite react with Al.

Conclusion

Our uncertainty propagation study using surrogate models proved to be successful in analyzing the sensitivity of a reference pore water geochemical model to its various input parameters. The results, and validation with direct Monte-Carlo simulations, show that sparse polynomial chaos are well-adapted to approximate the quantities of interest. Most significant correlations and anti-correlations were tractable from geochemical constraints, giving



(a) pH and $\log_{10} p_{CO_2}$ ($R^2 = 97\ \%$)

(b) pe + pH and $\log_{10} p_{CO_2}$ ($R^2 = 80\ \%$)

(c) pe + pH and S(−II) ($R^2 = 70\ \%$)

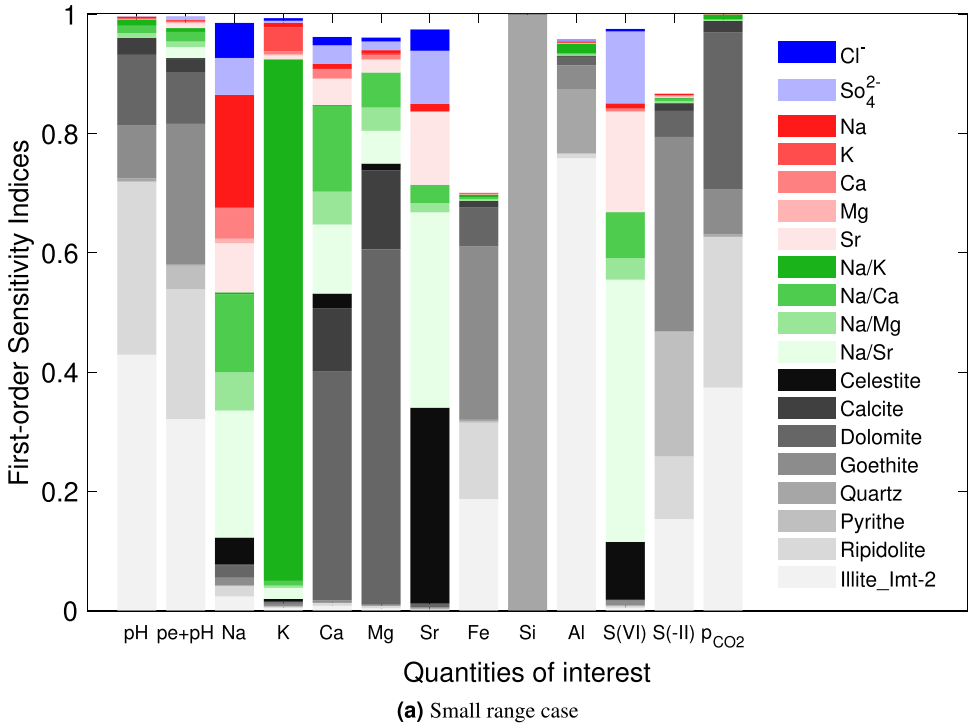
(d) pe + pH and Fe ($R^2 = 60\ \%$)

(e) pH and Fe ($R^2 = 57\ \%$)

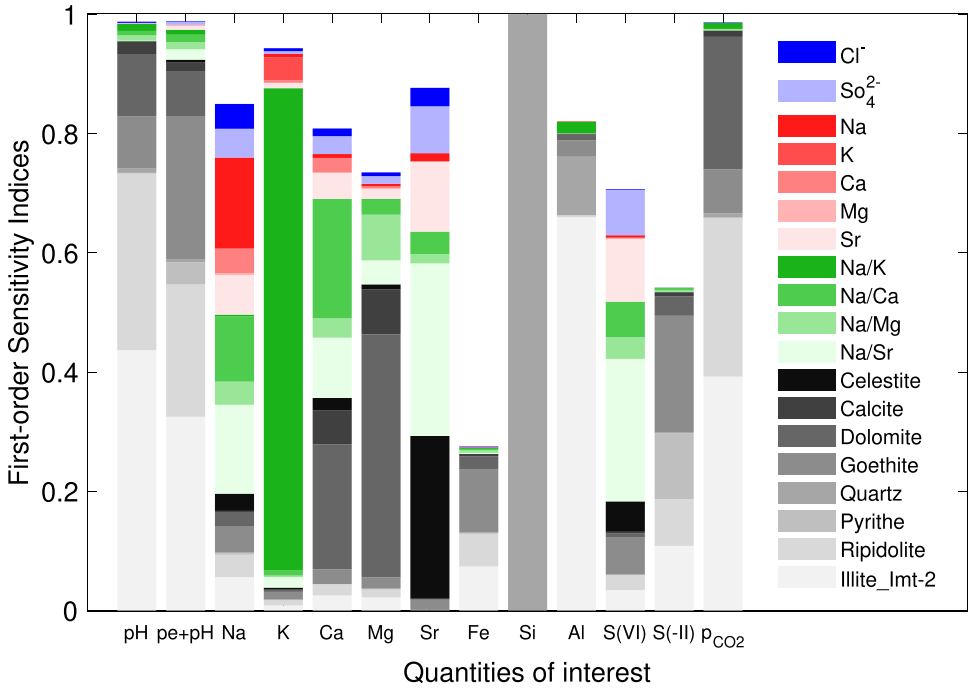
(f) Fe and $\log_{10} p_{CO_2}$ ($R^2 = 53\ \%$)

Figure 4. Isolines of bivariate distributions and medians (asterisks) estimated with 10^6 realizations of the surrogate models and the kernel density estimation method for the most correlated pairs of QoI (small range case). The plots are ordered according to the value of R^2.

confidence in the overall analysis. The method makes it possible not only to quantify the uncertainties of the quantities of interest for future performance evaluation calculations, but also to identify the main influential input parameters. This latter information is particularly valuable to guide further research efforts in view of reducing uncertainties on specific aspects of performance assessment analyses. Because pore water chemistry influences many important parameters such as radionuclides transport and retardation by adsorption and precipitation, uncertainty analyses of reactive transport modeling outcomes would certainly benefit from a coupling with our surrogates models to decipher uncertainties in adsorption models predictions, and to speed up calculations in fully coupled approaches.



(a) Small range case



(b) Large range case

Figure 5. First-order sensitivity indices S_i (own effects). The color depends on the input parameter category: blue for the leached parameters, red for the exchanged cations, green for the selectivity coefficients, and gray for the solubilities. The color shade varies within each category.

Data availability

All data generated or analysed during this study are included in this published article and its supplementary information files.

Received: 16 May 2022; Accepted: 10 August 2022
Published online: 05 September 2022

References

1. Altmann, S. 'Geo'chemical research: A key building block for nuclear waste disposal safety cases. *J. Contam. Hydrol* **102**, 174–179 (2008).
2. Tournassat, C. & Steefel, C. I. Reactive transport modeling of coupled processes in nanoporous media. *Rev. Mineral Geochem.* **85**, 75–110 (2019).
3. Claret, F., Marty, N. & Tournassat, C. *Modeling the Long-term Stability of Multi-barrier Systems for Nuclear Waste Disposal in Geological Clay Formations*, chap. 8, 395–451 (Wiley, 2018). https://doi.org/10.1002/9781119060031.ch8.
4. Tournassat, C., Vinsot, A., Gaucher, E. C. & Altmann, S. Chapter 3—Chemical conditions in clay-rocks. In Tournassat, C., Steefel, C. I., Bourg, I. C. & Bergaya, F. (eds.) *Natural and Engineered Clay Barriers*, vol. 6 of *Developments in Clay Science*, 71 – 100 (Elsevier, 2015).
5. Gaucher, E. *et al.* A robust model for pore-water chemistry of clayrock. *Geochim. Cosmochim. Acta.* **73**, 6470–6487 (2009).
6. Bonano, E. J. & Cranwell, R. M. Treatment of uncertainties in the performance assessment of geologic high-level radioactive waste repositories. *Math. Geol.* **20**, 543–565. https://doi.org/10.1007/BF00890336 (1988).
7. Ayoub, A., Pfingsten, W., Podofillini, L. & Sansavini, G. Uncertainty and sensitivity analysis of the chemistry of cesium sorption in deep geological repositories. *Appl. Geochem.* **117**, 104607 https://doi.org/10.1016/j.apgeochem.2020.104607 (2020).
8. Denison, F. H. & Garnier-Laplace, J. The effects of database parameter uncertainty on uranium (vi) equilibrium calculations. *Geochim. Cosmochim. Acta.* **69**, 2183–2191. https://doi.org/10.1016/j.gca.2004.09.033 (2005).
9. Sochala, P. & Le Maître, O. Polynomial Chaos expansion for subsurface flows with uncertain soil parameters. *Adv. Water. Resour.* **62**, 139–154. https://doi.org/10.1016/j.advwatres.2013.10.003 (2013).
10. Li, G. *et al.* Quantifying initial and wind forcing uncertainties in the Gulf of Mexico. *Comput. Geosci.* **20**, 1133–1153. https://doi.org/10.1007/s10596-016-9581-4 (2016).
11. Sochala, P., De Martin, F. & Le Maître, O. Model reduction for large-scale earthquake simulation in an uncertain 3d medium. *Int. J. Uncertain. Quantif.* **10**, 101–127 (2020).
12. Snelling, B., Neethling, S., Horsburgh, K., Collins, G. & Piggott, M. Uncertainty quantification of landslide generated waves using Gaussian process emulation and variance-based sensitivity analysis. *Water* **12**, 416 (2020).
13. Phenix, B. D. *et al.* Incorporation of parametric uncertainty into complex kinetic mechanisms: Application to hydrogen oxidation in supercritical water. *Combust. Flame* **112**, 132–146 (1998).
14. Reagan, M. T., Najm, H. M., Ghanem, R. G. & Knio, O. M. Uncertainty quantification in reacting-flow simulations through non-intrusive spectral projection. *Combust. Flame* **132**, 545–555. https://doi.org/10.1016/S0010-2180(02)00503-5 (2003).
15. Alexanderian, A., Le Maître, O., Najm, H., Iskandarani, M. & Knio, O. Multiscale stochastic preconditioners in non-intrusive spectral projection. *SIAM J. Sci. Comp.* **50**, 306–340. https://doi.org/10.1007/s10915-011-9486-2 (2012).
16. Srinivasan, G., Tartakovsky, D. M., Robinson, B. A. & Aceves, A. B. Quantification of uncertainty in geochemical reactions. *Water Resour. Res.* https://doi.org/10.1029/2007WR006003 (2007).
17. Delay, J. *et al.* Three decades of underground research laboratories: What have we learned?. *Geol. Soc. Spec. Publ.* **400**, SP400-1 (2014).
18. Parkhurst, D. L. & Appelo, C. A. J. *Description of input and examples for PHREEQC Version 3—a computer program for speciation,batch-reaction, one-dimensional transport, and inverse geochemical calculations*, U.S. Geological Survey Techniques and Methods, book 6, chap. A43, http://pubs.usgs.gov/tm/06/a43/, (2013).
19. Giffaut, E. *et al.* Andra thermodynamic database for performance assessment: ThermoChimie. *Appl. Geochem.* **49**, 225–236 (2014).
20. Shannon, C. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x (1948).
21. Jaynes, E. Information theory and statistical mechanics. *Phys. Rev.* **106**, 620–630. https://doi.org/10.1103/PhysRev.106.620 (1957).
22. Rasmussen, C. E. & Williams, C. K. I. *Gaussian Processes for Machine Learning* (The MIT Press, 2005).
23. Aleksander, I. & Morton, H. *An Introduction to Neural Computing* (Chapman and Hall, 1990).
24. Ghanem, R. G. & Spanos, S. D. *Stochastic Finite Elements: A Spectral Approach* (Springer, 1991).
25. Le Maître, O. P. & Knio, O. M. *Spectral Methods for Uncertainty Quantification. Scientific Computation* (Springer, 2010).
26. Cameron, R. & Martin, W. The orthogonal development of nonlinear functionals in series of Fourier-Hermite functionals. *Ann. Math.* **48**, 385–392 (1947).
27. Ernst, O. G., Mugler, A., Starkloff, H.-J. & Ullmann, E. On the convergence of generalized polynomial chaos expansions. *Esaim Math. Model. Numer. Anal.* **46**, 317–339. https://doi.org/10.1051/m2an/2011045 (2012).
28. Montgomery, D. *Design and Analysis of Experiments. Student Solutions Manual* (Wiley, 2004).
29. Lüthen, N., Marelli, S. & Sudret, B. Sparse polynomial chaos expansions: Literature survey and benchmark. *SIAM-ASA J. Uncertain.* **9**, 593–649. https://doi.org/10.1137/20M1315774 (2021).
30. Mallat, S. & Zhang, Z. Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal Process.* **41**, 3397–3415 (1993).
31. Pati, Y., Rezaiifar, R. & Krishnaprasad, P. Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. In *Proceedings of 27th Asilomar Conference on Signals, Systems and Computers*, Vol. 1, 40–44, https://doi.org/10.1109/ACSSC.1993.342465 (1993).
32. Parzen, E. On estimation of a probability density function and mode. *Ann. Math. Stat.* **33**, 1065–1076. https://doi.org/10.1214/aoms/1177704472 (1962).
33. Sobol, I. M. Sensitivity estimates for nonlinear mathematical models. *Math. Model. Comput. Exp.* **1**, 407–414 (1993).
34. Homma, T. & Saltelli, A. Importance measures in global sensitivity analysis of nonlinear models. *Reliab. Eng. Syst. Saf.* **52**, 1–17. https://doi.org/10.1016/0951-8320(96)00002-6 (1996).
35. Cacuci, D. G. Sensitivity theory for nonlinear systems. I. Nonlinear functional analysis approach. *J. Math. Phys.* **22**, 2794–2802. https://doi.org/10.1063/1.525186 (1981).
36. Hoeffding, W. A class of statistics with asymptotically normal distribution. *Ann. Math. Stat.* **19**, 293–325 (1948).
37. Sobol, I. M. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Math. Comput. Simul.* **55**, 271–280 (2001).
38. Tremosa, J. *et al.* Geochemical characterization and modelling of the Toarcian/Domerian porewater at the Tournemire underground research laboratory. *Appl. Geochem.* **27**, 1417–1431 (2012).

Acknowledgements

The work of P. S., C. C. and F. C. has been supported by the European project DONUT. C. T. would like to acknowledge 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. The authors would like to thank the two anonymous reviewers for taking the time necessary to assess the manuscript and for their thoughtful comments and constructive suggestions.

Author contributions

C.C., F.C. and C.T. conceived the uncertainty models, C.T. launched the PHREEQC ensemble simulations, P.S. built the surrogate models and analysed the results together with F.C. and C.T. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-18411-5.

Correspondence and requests for materials should be addressed to P.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022