Journal of Computational Physics 399 (2019) 108909

Contents lists available at ScienceDirect

Journal of Computational Physics

www.elsevier.com/locate/jcp

Regression-based sparse polynomial chaos for uncertainty quantification of subsurface flow models

Tarakanov Alexander*, Elsheikh Ahmed H.

ARTICLE INFO

Article history:
Received 28 February 2019
Received in revised form 19 August 2019
Accepted 22 August 2019
Available online 27 August 2019

Keywords:
Polynomial chaos expansion
Surrogate modeling
CO2 sequestration
Multi-phase flow

ABSTRACT

Surrogate-modeling techniques including Polynomial Chaos Expansion (PCE) is commonly used for statistical estimation (aka. Uncertainty Quantification) of quantities of interests obtained from expensive computational models. PCE is a data-driven regression-based technique that relies on spectral polynomials as basis-functions. In this technique, the outputs of few numerical simulations are used to estimate the PCE coefficients within a regression framework combined with regularization techniques where the regularization parameters are estimated using standard cross-validation as applied in supervised machine learning methods.

In the present work, we introduce an efficient method for estimating the PCE coefficients combining Elastic Net regularization with a data-driven feature ranking approach. Our goal is to increase the probability of identifying the most significant PCE components by assigning each of the PCE coefficients a numerical value reflecting the magnitude of the coefficient and its stability with respect to perturbations in the input data. In our evaluations, the proposed approach has shown high convergence rate for high-dimensional problems, where standard feature ranking might be challenging due to the curse of dimensionality.

The presented method is implemented within a standard machine learning library (scikit-learn [1]) allowing for easy experimentation with various solvers and regularization techniques (e.g. Tikhonov, LASSO, LARS, Elastic Net) and enabling automatic cross-validation techniques using a widely used and well tested implementation. We present a set of numerical tests on standard analytical functions, a two-phase subsurface flow model and a simulation dataset for CO2 sequestration in a saline aquifer. For all test cases, the proposed approach resulted in a significant increase in PCE convergence rates.

1. Introduction

Uncertainty Quantification (UQ) and Uncertainty Propagation (UP) in the subsurface flow related problems have been the subject of intensive research activities over the last decades [2–6]. For instance, UQ of oil production forecasts from a given reservoir has far-reaching economical consequences [7]. Also, the accurate risk assessment of CO2 trapping in an underground reservoir [8] is of high importance from ecological and social perspectives [9].

The main challenge for UQ in subsurface flow related tasks is the complexity of the modeled physical systems [10] and the lack of information about the rock properties that determine underground flow [11]. Therefore, UQ for a Quantity of

* Corresponding author.
E-mail address: a.tarakanov@hw.ac.uk (A. Tarakanov).

https://doi.org/10.1016/j.jcp.2019.108909

Interest (QoI) is usually performed numerically through multiple evaluations of expensive reservoir simulations [12]. This corresponds to significant computational resources especially when dealing with a high number of uncertain parameters [13] and for the cases where model high resolution is requirement [14]. Several lines of research have been pursued to address this challenge. For instance, models of multiple continuum media [15–17], dual mesh approaches [18–20], upscaling [21–24] and model reduction [25–28] techniques have been developed to decrease the run-time of a single simulation. Also, various surrogate modeling techniques [29–31] emerged in order to reduce the cost of evaluating a large number of expensive numerical simulations.

In the current manuscript, we focus on surrogate modeling approaches using PCE-based response surfaces. There are several advantages of using PCE as a proxy model. First of all, surrogate models based on sparse PCE do not require significant computational resources to compute a value at any given point within the interpolation domain as it is simply a direct polynomial function evaluation. Secondly, important statistical properties such as moments and sensitivities can be computed directly from the PCE coefficients without the need for a Monte-Carlo simulations [32]. This is attributed to a special design properties of PCE that links the probability distribution of random variables with the orthogonality of polynomial basis functions [32].

Generally, two techniques could be utilized to estimate the PCE coefficients: collocation-based and regression-based methods. For collocation approaches, the QoI values are evaluated at pre-specified set of points called collocation nodes [33]. These specific points are designed in such a way that the PCE coefficients can be expressed as linear combination of the QoI values, allowing for direct computation of the PCE coefficients. The optimal choice of the collocation points especially for high-dimensional problems is a subject of extensive research activities [34–36]. In regression-based approaches, the PCE coefficients correspond to the solution of an error-minimization problem [37]. It is simple to show that the mean-square error minimization can be reduced to a linear regression problem to estimate the PCE expansion coefficients. Designing fast and accurate solution techniques to this minimization problem including various preconditioning methods is also a subject of intensive research activities [38–40]. Hampton and Doostan [41] developed a hybrid collocation and regression technique, where the training points for the surrogate model are generated with collocation techniques while the PCE coefficients are estimated by solving an error-minimization problem. One of the advantages of this approach is the better conditioning of the regression problem when compared to training using random samples [41].

In generic cases, sparse collocation techniques and hybrid approaches provide accurate response surfaces using reasonable computational resources [42–44]. However, these methods rely on evaluating the QoI at specific set of points. This strategy can be successfully adopted for UQ of oil production and CO2 storage capacity [45–47]. However, computation of QoI values in the case of subsurface flow problems can be challenging if the collocation points correspond to extreme values of parameters that significantly affect convergence properties of the numerical scheme. Therefore, such collocation nodes can either increase computational costs of the response surface construction or reduce the overall accuracy of the surrogate model if significant numerical error is introduced to the QoI values at collocation nodes. Additionally, for many practical problems sampling of data points can not be controlled. For instance, samples could be generated randomly (e.g. Latin hypercube sampling), or in accordance with a prescribed probability distribution [48], or based on another meta-modeling technique used in combination with PCE for model stacking [49]. Under these conditions, collocation techniques cannot be directly applied. For this reason, regression based PCE (utilized in this manuscript) have wider applicability for any set of training samples where optimal response surfaces could be built.

In regression methods, PCE coefficients are computed through the minimization of mean-square error over the training data. Therefore, for low-dimensional problems, a direct approach could be adopted. In generic setting, the number of PCE coefficients for a problem with n variables can be expressed as follows:

$$D = D(n, d) = \binom{n+d}{n} \tag{1}$$

where D is the number of PCE coefficients and d is the degree of polynomials used. It is simple to observe the fast growth of D with both d and n. This exponential growth of PCE coefficients imposes significant constraints on building PCE-based response surfaces. First of all, solving the error-minimization regression problem in high dimensions is a challenging task, because of the high number of numerical operations needed till convergence. Secondly, the number of QoI values (i.e. training samples) needed for accurate estimation of the PCE coefficients increases with D, which corresponds to additional runs of an expensive numerical simulator. In other words, the curse of dimensionality makes it impractical to solve for PCE coefficients directly. However, for a large class of problems it was observed that PCE coefficients are sparse [50,51]. Therefore, various techniques for sparse regression can be adopted. For example, ℓ^1 regularization techniques [52] can be considered as a first step towards enforcing sparsity on the PCE regression coefficients. This approach is widely adopted and will be referred to as standard PCE [53] in the rest of this manuscript. Further dimension reduction could be achieved through fitting both the data and the QoI derivatives at the training points [54]. The additional information from the gradients increases the quality of PCE response surface [55]. Unfortunately, for many problems it is not possible to obtain the gradient information at the training points. Another line of research focuses on reducing the problem dimension by using advanced methods for solving nonlinear regression problems. For instance, sparse PCE coefficients can be computed efficiently through the application of support vector regression [56] or preconditioned conjugate gradient [57] techniques. Another direction of development relies on coupling the iterative solvers with algorithms for ranking the importance of the basis polynomials (e.g. orthogonal matching pursuit [37]) or ranking based on the impact on the residual [58]. Further reduction of dimension could also be achieved by adaptive truncation of the spectrum of the expansion. For instance, it has been observed empirically for a broad class of problems, that higher order interactions between the polynomial basis from different dimension have less impact on the quality of the response surface when compared to the one dimensional low order polynomials. This empirical observation is the foundation for hyperbolic truncation techniques [59]. Moreover, the performance of all regression-based approaches could be improved by transformation of the input variables (e.g. scaling, normalization). For example, variable rotations [60] or generic linear transformations [61] could significantly reduce the complexity of the error minimization problem corresponding to finding the PCE coefficients.

In the current paper, we focus on further improvement of dimension reduction techniques for regression based PCE. We present a novel iterative approach for solving the error minimization problem. We introduce a new data-driven ranking procedure for sequential identification of the most significant PCE basis functions with the closest relation to the interpolated QoI values. The ranking procedure is based on the correlation between the basis functions and the QoI values penalized by factors that measure the sensitivity of the corresponding coefficient to the noise in the input data. The aim of the introduction of correction/penalty factors is to avoid overestimation of the significance of a given polynomial basis function due to occasional location of data-points. The introduced ranking approach enables us to determine the most significant PCE terms and subsequently solving a reduced regression problem at each iteration. The new method could be easily combined with various regularization techniques.

The proposed approach has been integrated in scikit-learn [1], a widely used machine learning library. This integration enables uniform testing of a huge variety of techniques such as Lasso, Lars and Elastic Net [62] in order to formulate and solve the regularized regression problem. We implement PCE as an input feature transformation using machine learning terminology. Therefore, PCE can be naturally included in any machine-learning pipeline allowing one to combine different methods for variable transformation with advanced cross-validation techniques. Moreover, this implementation allows for an easy comparisons to alternative machine-learning techniques (e.g. Random Forests, Support Vector Machines). In the numerical evaluation section, we compare the proposed approach to classical methods for sparse PCE namely, the Orthogonal Matching Pursuit (OMP) and Least Angular Regression (LARS). We consider four data-sets for evaluation. The first two data-sets are generated using analytical functions and the last two data-sets are based on subsurface simulations of fluid flow in porous media. In all the test cases, extensive comparisons are performed in terms of Mean-Square Error (MSE) using a hold-out (aka. validation) set of points following the best practices in the machine learning literature.

The rest of this manuscript is organized as follows: In the following section, a general introduction to PC is presented followed by the proposed ranking procedure. In section 3 we present a set of numerical examples. Finally, the conclusion of our work is presented in section 4.

2. Methodology

Polynomial chaos expansion PCE is a meta-modeling technique that relies on orthogonal polynomials. One of the main advantages of PCE when compared to alternative surrogate modeling techniques is the ability to estimate the QoI sensitivity to given combination of variables through simple analytical formulae. This is only possible due to the close relation between the orthogonality of basis polynomials and the probability distribution of the input variables. This relation is explained in subsection 2.1, along with an overview of basic ideas of PCE. The proposed reordering of PCE basis is then introduced in subsection 2.2.

2.1. Basics of polynomial chaos

The essence of PCE is the relation between the statistics of input data and orthogonality of the utilized basis polynomials. The relation concerned gives a powerful tool for calculating the PCE coefficients and for further statistical analysis of the data. We first explain this relation for the single-variate case and then extend this formulation to multi-variate cases. Additionally, examples of applying this concept to study the statistical properties of PCE are presented.

For single-variate function $f(x)$, a PCE is defined as series of orthogonal polynomials:

$$f(x) = \sum_{\alpha} c_\alpha p_\alpha(x) \tag{2}$$

where $p_\alpha(x)$ is an orthogonal single-variate polynomial with the index α and c_α is the corresponding PCE coefficient. The specific type of utilized orthogonal polynomials is not of a principal importance in the definition introduced in Eq. (2). Therefore, PCE can be naturally formulated for all well-known families of orthogonal polynomials. For example Hermite, Legendre and Chebyshev polynomials [63].

The analysis of the PCE relies on the orthogonality of the basis polynomials, which is introduced through the notion of an inner product defined as follows:

$$\langle g_1, g_2 \rangle = \int_{-\infty}^{+\infty} \mathcal{K}(x) g_1(x) g_2(x) \mathrm{d}x \tag{3}$$

where $g_1(x)$ and $g_2(x)$ are certain square-integrable functions and $\mathcal{K}(x)$ is a non-negative function referred to as the kernel function or simply the kernel. Classical families of orthogonal polynomials are related to a specific form of the kernel function. For instance, Hermite polynomials correspond to a kernel function identical to Gaussian distribution function with zero mean and unit variance [64]:

$$\mathcal{K}(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \tag{4}$$

For a generic case, the PCE basis functions are constructed by applying Gram-Schmidt orthogonalization to the set of monomial functions (e.g. $1,\ x,\ x^2,\ \ldots$) [65]. Therefore, PCE techniques could be naturally extended to any arbitrary kernel functions $\mathcal{K}(x)$.

A central idea of PCE is the statistical interpretation of $\mathcal{K}(x)$ as probability density function for a given random variable [66]. This interpretation allows one to reformulate the inner product defined in Eq. (3) in terms of expectations:

$$\langle g_1, g_2 \rangle = \int_{-\infty}^{+\infty} \mathcal{K}(x) g_1(x) g_2(x) \mathrm{d}x = \mathbb{E}[g_1, g_2] \tag{5}$$

In the setting, the orthogonality of polynomials $p_\alpha(x)$ with respect to the inner product can be reformulated as:

$$\langle p_\alpha, p_\beta \rangle = \mathbb{E}[p_\alpha, p_\beta] = \|p_\alpha\|^2 \delta_{\alpha\beta} \tag{6}$$

where $\delta_{\alpha\beta}$ is a Kronecker symbol. In the present work, we consider orthonormal polynomials with $\|p_\alpha\|^2 = 1$ in order to simplify the numerical analysis of PCE. Therefore, Eq. (7) can be transformed as follows:

$$\langle p_\alpha, p_\beta \rangle = \mathbb{E}[p_\alpha, p_\beta] = \delta_{\alpha\beta} \tag{7}$$

Moreover, the basis polynomials orthogonality can be used to estimate the PCE coefficients:

$$c_\alpha = \langle f, p_\alpha \rangle = \mathbb{E}[f, p_\alpha] \tag{8}$$

For multi-variate functions, similar analysis could be performed through the introduction of the tensor-product concept where the set of multivariate basis functions is formed as products of single-variate polynomials:

$$p_A(\mathbf{x}) = p^{(1)}_{\alpha_1}(x_1) p^{(2)}_{\alpha_2}(x_2) \ldots p^{(n)}_{\alpha_n}(x_n) \tag{9}$$

where α_k is the degree of single-variate polynomial, $p^{(k)}_{\alpha_k}(x_k)$ is a uni-variate polynomial that depends only on the k-th coordinate of the vector $\mathbf{x}$. The degree of polynomial $p_A(\mathbf{x})$ is defined as:

$$\deg(p_A(\mathbf{x})) = \sum_k \deg(p^{(k)}_{\alpha_k}(x_k)) = \sum_k \alpha_k \tag{10}$$

Similar to Eq. (2), the PCE of multivariate function $f(\mathbf{x})$ is defined as:

$$f(\mathbf{x}) = f(x_1, \ldots, x_n) = \sum_A c_A p_A(\mathbf{x}) \tag{11}$$

where c_A is the PCE coefficient corresponding to polynomial basis function with multi-index A. The inner product in multi-dimensional case is defined as:

$$\langle g_1, g_2 \rangle = \int \mathcal{K}(\mathbf{x}) g_1(\mathbf{x}) g_2(\mathbf{x}) \mathrm{d}\mathbf{x} \tag{12}$$

where $\mathcal{K}(\mathbf{x})$ is a multi-variate kernel function and $g_1(\mathbf{x})$, $g_2(\mathbf{x})$ are certain square-integrable functions. It is important to note that the polynomial basis functions obtained by tensor multiplications inherit the orthogonality and orthonormality from single-variate PCE if the multi-variate kernel function $\mathcal{K}(\mathbf{x})$ equals the product of single-variate kernel functions:

$$\mathcal{K}(\mathbf{x}) = \mathcal{K}_1(x_1) \cdots \mathcal{K}_n(x_n) \tag{13}$$

where $\mathcal{K}_1(x_1), \cdots, \mathcal{K}_n(x_n)$ are single-variate kernel functions. From a probabilistic point of view, this is equivalent to the mutual independence of the coordinates of the vector $\mathbf{x}$.

The inner product defined in Eq. (12) can be utilized to derive an expression for the PCE coefficients similar to Eq. (8):

$$c_A = \langle f, p_A \rangle = \mathbb{E}[p_A, f] \tag{14}$$

The relation between the input data statistics and the polynomial basis orthogonality can be used to derive analytical expressions for the mean, variance and Sobol' indices of the function $f(\mathbf{x})$. For example, the mean can be estimated by:

$$\mathbb{E}[f] = \langle 1, f\rangle = \sum_A c_A \langle 1, p_A\rangle = c_{0,\dots,0} = c_0 \tag{15}$$

Where $c_{0,\dots,0}$ is the constant polynomial coefficient. In the present work, we simplify the notations and use c_0 instead of $c_{0,\dots,0}$. The mean-square deviation can be calculated in the similar fashion:

$$\sigma^2 = \mathbb{E}[(f - c_0), (f - c_0)] = \sum_{A_1, A_2 \neq 0} c_{A_1} c_{A_2} \delta_{A_1 A_2} = \sum_{A>0} c_A^2 \tag{16}$$

Calculation of other quantities for sensitivity analysis and UQ such as partial variances and Sobol' indices could be performed naturally with PCE. A partial standard deviation represents the sensitivity to a given combination of variables. It is defined as the standard deviation of the function $f(\mathbf{x})$ averaged with respect to certain collection of variables [36]:

$$\sigma^2_{r_1,\dots,r_k}(f) = \sigma^2(\mathbb{E}_{t_1,\dots,t_{n-k}}[f]) \tag{17}$$

where $\sigma_{r_1,\dots,r_k}$ is the standard deviations with respect to the components of the vector $\mathbf{x}$ with indices $r_1, \cdots, r_k$ and $\mathbb{E}_{t_1,\dots,t_{n-k}}[f]$ is the average with respect to components of the vector $\mathbf{x}$ with indices $t_1, \cdots, t_{n-k}$ that form a complement to $r_1, \cdots, r_k$ [32], n is the dimension of $\mathbf{x}$ and k is a certain integer number from 1 to n. Sobol' indices are commonly used as a measure for sensitivity and are defined as the normalized partial standard deviations:

$$S_{r_1,\dots,r_k}(f) = \frac{\sigma^2_{r_1,\dots,r_k}(f)}{\sigma^2(f)} \tag{18}$$

For the response function f with a PCE representation, the partial standard deviations can be calculated in a similar fashion as the normal standard deviation defined in Eq. (16) following [36]:

$$\sigma^2_{r_1,\dots,r_k}(f) = \sum_{\alpha_{r_l}>0, \alpha_{t_j}=0} c_A^2 \tag{19}$$

The relation between orthogonality of polynomial basis functions and the probability distribution of input data has an important consequence on the numerical calculation of the PCE coefficients. In practice, for regression based response surfaces, the PCE coefficients for a given function $f(x)$ can be computed for given input data through the minimization of the mean-square error (MSE) functional:

$$\mathcal{F}(\mathbf{c}) = \sum_i \frac{(y_i - \sum_A c_A p_A(\mathbf{x}_i))^2}{N} \tag{20}$$

where $\mathbf{x}_i$ is the i^{th} vector of input variables, N is the number of data points and y_i is the value of the function f at the point $\mathbf{x}_i$. In the present work the spectrum of PCE is truncated to a certain polynomial degree d. Therefore, the dimension of $\mathbf{c}$ is given by Eq. (1).

It is simple to see that minimizing the functional defined in Eq. (20) is equivalent to solving a system of linear equations:

$$\mathbf{M}_{AB} c_B = \mathbf{V}_A \tag{21}$$

where the square matrix $\mathbf{M}$ and vector $\mathbf{V}$ are defined as:

$$\mathbf{M}_{AB} = \sum_i \frac{p_A(\mathbf{x}_i) p_B(\mathbf{x}_i)}{N}, \qquad \mathbf{V}_A = \sum_i \frac{y_i p_A(\mathbf{x}_i)}{N} \tag{22}$$

The relation between the basis orthogonality and the statistical distribution of the input data imposes several constraints on the value of the matrix $\mathbf{M}$ and the vector $\mathbf{V}$. If the data is sampled in agreement with the probability distribution determined by the kernel function defined in Eq. (13), then the matrix $\mathbf{M}$ should converge to a unit matrix:

$$\mathbf{M}_{AB} = \sum_i \frac{p_A(\mathbf{x}_i) p_B(\mathbf{x}_i)}{N} = \mathbb{E}[p_A p_B] + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right) = \delta_{AB} + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right) \tag{23}$$

where the term $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$ represents the convergence in accordance with the law of large numbers [67]. Similar reasoning could be applied to the vector $\mathbf{V}$ showing the close correlation between the data and the basis functions:

$$\mathbf{V}_A = \sum_i \frac{p_A(\mathbf{x}_i) y_i}{N} = \mathbb{E}[y(x) p_A(x)] + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right) = c_A + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right) \tag{24}$$

Eq. (23) and Eq. (24), simply means that the coefficients $\mathbf{c}$ minimizing the MSE functional defined in Eq. (20) is close to the correlation vector $\mathbf{V}$ if a sufficient number N of training data points is available. Moreover, the difference between $\mathbf{V}$ and $\mathbf{c}$ can be estimated as follows:

$$|V_A - c_A| \leq \frac{k_A}{\sqrt{N}} \tag{25}$$

Where k_A is a positive number. In other words, $\mathbf{V}$ provides a reasonable approximation for $\mathbf{c}$ if a sufficient number of data-points is available. We utilize this observation to introduce a novel ranking-based approach to approximate the PCE coefficients as described in the next subsection.

2.2. Ranking based sparse PCE

In the present work, we estimate the PCE coefficients by minimizing the mean-square error functional defined in Eq. (20). It is well-known that a straight-forward minimization of mean square errors could provide an unstable solution or a response surface that is not quite accurate at points that are not included in the training data-set. Therefore, we utilize a mixed ℓ_1 and ℓ_2 regularization technique known as Elastic Net model [68] (i.e., combined Tikhonov and Lasso regularization). This results in a regularized functional for error minimization of the following form:

$$\mathbf{c} = \arg\min_{\mathbf{c}} \mathcal{L}(\mathbf{c}) = \arg\min_{\mathbf{c}} \left(\mathcal{F}(\mathbf{c}) + \lambda_1 \sum_A |c_A| + \lambda_2 \sum_A c_A^2 \right) \tag{26}$$

where $\mathcal{L}(\mathbf{c})$ is a functional for minimization and λ_1, λ_2 are hyperparameters that could be tuned in order to maximize the quality of the surrogate model. In the present work, λ_1 and λ_2 are determined through cross-validation.

We utilize a coordinate descent algorithm [62] in order to find the solution for the minimization problem defined in Eq. (26). This is an iterative algorithm that sequentially updates the solution vector $\mathbf{c}$ by minimizing the functional $\mathcal{L}(\mathbf{c})$ with respect to one of the coordinates at each step as summarized in Algorithm 1.

Algorithm 1 Coordinate descent.

```
$\mathbf{c} = 0$                                          ▷ Set vector of parameters to zero
while $\Delta\mathcal{L} > \varepsilon$ do                          ▷ Iterate while change in $\mathcal{L}(c)$ is significant
    Select a value $k$ from 1 to dim($\mathbf{c}$)                ▷ Select one of the coordinates
    $c_k = \arg\min_{c_k} \mathcal{L}(\mathbf{c})$                   ▷ Minimize with respect to single parameter
    Update $\Delta\mathcal{L}$
return $\mathbf{c}$
```

One of the essential parts in Algorithm 1 is the selection of the next component for update. Classical approaches include: random selection or selection based on the cyclic order on the set of components [62]. In the present work, we introduce a novel scheme for reordering the polynomial basis functions that increases the algorithm convergence rate and increases the response surface quality when utilizing small number of training samples. The aim of the reordering procedure is to identify the polynomial basis functions with the highest PCE coefficients in order to determine its values first. It should be noted that the assumption about the agreement between sampling of training data and orthogonality of basis polynomial functions Eq. (5) is of principal importance for the proposed reordering procedure. For the cases where this assumption is violated, data transformation techniques should be applied before using the proposed reordering approach. For instance, the desired distribution of input variables can be achieved through quantile transformation [69] or Rosenblatt transformation [70].

The reordering technique utilizes a ranking of PCE coefficients inspired by Eq. (24), which states that the vector of moments is close to the actual PCE coefficients given a sufficient number of training points. However, for certain polynomial basis functions the difference between c_A and V_A can be significant leading to an overestimation of the importance of those components due to the lack of the available data, which can be considered as a noise. In order to address this issue, we introduce a ranking of polynomial basis functions in a form of the signal-to-noise ratio which is a correlation coefficient divided by a correction factor that quantifies the sensitivity of a given PCE coefficient to the data noise.

Two sources of noise are considered in the current work: noise in the values of QoI and noise in the deviation of matrix $\mathbf{M}$ due to the random sampling of the training data. In order to quantify both sources of noise, we perform two series of Monte-Carlo simulations. In the first series of Monte-Carlo simulations, the sensitivity of the correlation vector $\mathbf{V}$ to the QoI values is estimated. For that purpose, we introduce random perturbations θ_i to each of the training data-points. In the present study, the noise part is sampled from a normal distribution with zero mean and unit variance. The correlation of the basis polynomial $p_A(\mathbf{x})$ with perturbed data to the QoI is estimated using:

$$\mathbf{U}_A = \sum_i \frac{(y_i + \theta_i) p_A(\mathbf{x}_i)}{N} \tag{27}$$

The mean-square deviation $\sigma_{Y,A}$ of $\mathbf{U}_A$ from $\mathbf{V_A}$ is considered as a measure for stability:

$$\sigma_{Y,A} = \sqrt{\mathbb{E}_\theta[(\mathbf{U}_A - \mathbf{V}_A)^2]} \tag{28}$$

where the mean $\mathbb{E}_\theta$ is taken over several realization of θ.

The second series of Monte-Carlo simulations is performed to quantify the stability with respect to the location of training points. As long as the location of training data points $\tilde{\mathbf{x}}$ is considered as a random parameter, a set of N points is generated at each Monte-Carlo simulation. Then the mean-square deviation $\sigma_{X,A}$ of $\mathbf{M}_{AA}$ from the unit matrix can be computed numerically as follows:

$$\sigma_{X,A} = \sqrt{E_{\tilde{\mathbf{x}}}[(\mathbf{M}_{AA} - \mathbf{I})^2]} \tag{29}$$

where the mean $\mathbb{E}_{\tilde{\mathbf{x}}}$ is taken over a number of realizations of $\tilde{\mathbf{x}}$. Finally, the ranking coefficient for the basis polynomial $p_A(\mathbf{x})$ is defined as:

$$r_A = \frac{1}{\sqrt{\sigma^2_{Y,A} + \sigma^2_{X,A}}} \frac{|\mathbf{V}_A|}{H} \tag{30}$$

where parameter H is introduced for normalization purposes. The value of H is given by the expression:

$$H = \frac{1}{2} \min_A |\mathbf{V}_A| + \frac{1}{2} \max_A |\mathbf{V}_A| \tag{31}$$

where minimum and maximum are taken over all values of multi-index A. In this work we consider high values of r_A as an indicator of a high value of the corresponding PCE coefficient.

In the present work, the ranking parameter r_A is used within the coordinate descent Algorithm 1 to select the next PCE coefficient for updates. Therefore, we solve iteratively for PCE coefficients by performing the following steps sequentially: ranking of basis functions based on the residual $\eta^{(k)}$ at the step k, select the first N_B basis functions and solve for the corresponding PCE coefficients using coordinate descent method. These steps are combined in Algorithm 2. In our numerical testing, we set $N_B = 5$ based on some initial testing. However, a more rigorous approach could utilize cross validation to select the optimal number of N_B.

Algorithm 2 Ranking based sparse PCE solver.

```
k=0                                   ▷ Set iteration counter to zero
$\eta^{(0)}$ = y                      ▷ Residual equals to initial data
$\mathbf{c} = 0$                      ▷ Set vector of parameters to zero
while $\Delta\mathcal{L} > \varepsilon$ do    ▷ Iterate while change in $\mathcal{L}(c)$ is significant
    $\mathbf{r} = \mathbf{r}(\eta^{(k)})$    ▷ Compute ranking
    Select $A_1, \cdots, A_{N_B}$     ▷ indices of first $N_B$ components with highest rank
    Solve $\Delta\mathbf{c} = \arg\min_{\Delta\mathbf{c}}(\mathcal{L}(\Delta c_{A_1}, \cdots, \Delta c_{A_{N_B}}))$ with respect to selected components    ▷ Use Algorithm 1
    Update coefficients: $\mathbf{c} = \mathbf{c} + \Delta\mathbf{c}$
    Update residual: $\eta^{(k+1)} = y - \sum_A c_A p_A(\mathbf{x})$
    $k = k + 1$
return $c$
```

3. Numerical examples

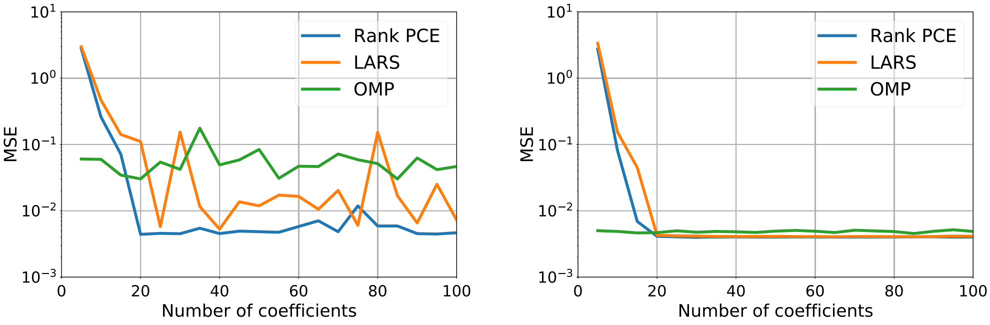
In this section, the proposed ranking based sparse PCE is evaluated on four test cases. The first test case is the Ishigami function [71], the second test case is a ten-dimensional Ackley function, the third case is a waterflooding problem with uncertain permeability field and the forth test case utilizes a data-set from simulations of CO2 injection [72]. In all test cases, the proposed PCE approach is compared to two standard techniques for sparse regression-based PCE: Least Angular Regression [73] and the Orthogonal Matching Pursuit (OMP) algorithm [74]. The numerical implementations are all based on scikit-learn, a machine library including with standard implementation of the LARS, OMP and coordinate descent algorithm. Moreover, cross-validation tools within this library are used to select the optimal regularization parameters for the Elastic Net functional defined in Eq. (26).

Test case 1: Ishigami function

Ishigami function is one of the standard benchmarks [71]:

$$y = 1 + \frac{1 + \pi^4/10 + \sin(\pi x_1) + 7\sin^2(\pi x_2) + 0.1(\pi x_3)^4 \sin(\pi x_1)}{9 + \pi^4/5} \tag{32}$$

This three dimensional function shows a strong nonlinear behavior and is commonly used as a test function for evaluating different response surface techniques. Typically, the evaluation domain is $[-\pi, \pi]^3$. In the present work, PCE with Legendre



(a) 200 training points

(b) 2000 training points

Fig. 1. Mean-square error on the test data set versus the number of free coefficients for the Ishigami function.

polynomials is used because of the finite length of the interval concerned. We have rescaled the input parameters linearly to the interval $[-1, 1]$, given that the Legendre polynomials are defined over the interval $[-1, 1]$. For each of the rescaled variables, a uniform distribution over the interval $[-1, 1]$ is assumed. We consider two training sets of 200 samples and 2000 samples. Another set of 2000 points uniformly sampled over the cube $[-1, 1]^3$ is utilized for out-of-sample MSE calculations (aka. test set). We construct a PCE of polynomial functions up to degree $d = 10$. The aim of the example is to compare the convergence rates of the proposed ranking based PCE approach against the standard sparse regularization techniques (i.e. LARS and OMP), while increasing the number of free coefficients N_D available for fitting by these iterative techniques. In the case of LARS and OMP, the value of N_D is well-defined. For the proposed ranking based approach, N_D is defined as:

$$N_D = N_I N_B \tag{33}$$

where N_I is the number of iterations and N_B is the number of PCE coefficients that can be modified by coordinate descent solver after each ranking update in Algorithm 2. It should be emphasized that PCE coefficients are selected solely based on ranking. In other words, the overlapping with previously selected PCE coefficients can occur. Therefore, the value of N_D given by Eq. (33) is a conservative upper bound for the total number of polynomial basis functions involved in the response surface construction (i.e. with non-zero coefficient).

The numerical level of tolerance has been set to 10^{-6} in all numerical schemes. Fig. 1a and Fig. 1b shows the MSE for each method versus the number of free coefficients N_D for 200 and 2000 training points, respectively.

The results presented in Fig. 1, shows that response surface built using the ranking based sparse PCE is of higher quality when compared to those obtained by the standard LARS or OMP algorithm, especially when the size of training data is limited. However, all three techniques perform similarly in the case with higher number of training data-points as shown in Fig. 1b. This is a major advantage when collecting training samples corresponds to running computationally expensive simulations.

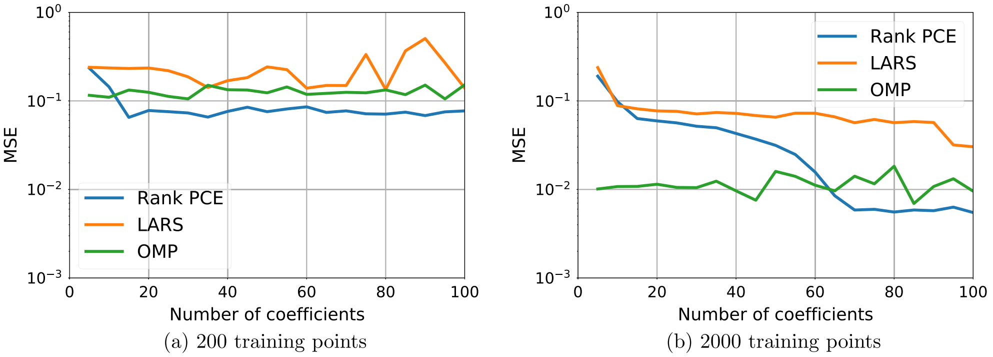
Test case 2: Ackley function

In this example, we build a response surface for a 10-dimensional Ackley function [75] of the form:

$$y = -20\exp\left(-0.2\left(\frac{1}{n}\sum_{k=1}^{n} x_k^2\right)^{1/2}\right) - \exp\left(\frac{1}{n}\sum_{k=1}^{n}\cos 2\pi x_k\right) + 20 + \exp(1) \tag{34}$$

where n is the dimension of the input vector. This function shows a strong nonlinear behavior with plenty of local minimums and is frequently used as a benchmark for optimization algorithm. The setup of the current numerical example is similar to the first test case. However, we assume that each of the input variables is uniformly distributed in the interval $[-5, 5]$. Legendre polynomials are utilized as basis function for the PCE. Therefore, input rescaling is applied to map all input variables to the interval $[-1, 1]$. In other words, we consider data to be uniformly distributed in the cube $[-1, 1]^{10}$. Similar to the first test case, two training sets sizes are considered (200, 2000 samples) and 2000 samples points uniformly distributed in the cube $[-1, 1]^{10}$ are set aside as a test set for calculating the out-of-sample MSE. We truncate the PCE spectrum to polynomial functions up to degree $d = 8$.

Fig. 2 shows the MSE convergence for the ranking based sparse PCE versus LARS and OMP algorithms. Similar to the first test case, the proposed approach produces a response surface of higher quality than LARS or OMP if the size of training data is limited as shown in Fig. 2a, while all techniques perform similarly a higher number of training data-points is used and a high number of polynomial basis functions is utilized as shown in Fig. 2b.



(a) 200 training points

(b) 2000 training points

Fig. 2. Mean-square error on the test data set versus the number of free coefficients for a ten dimensional Ackley function.

Test case 3: waterflooding problem

In the present test case, we evaluate the developed PCE approach on an uncertainty propagation for a waterflooding problem with uncertain permeability field. Dimension reduction using PCA technique is applied to the spatial field as an effective parametrization techniques [76]. Waterflooding is a commonly used process within the petroleum industry for achieving higher hydrocarbon recovery rates. The essence of this approach is to inject water through a number of wells in a given reservoir in order to displace the existing oil and increase the oil production from another set of wells, which are commonly spatially scattered to surround the injection wells. The increased productivity is observed until the injected water starts to appear at the production wells. Thus estimating when water will appear at the production wells (commonly known as the water breakthrough time [77]) is of significant practical importance. We note, that this time is commonly measured in terms of volume of water injected relative to the total reservoir pore volume (PVI). Prediction of the water breakthrough time t_b is of high importance for hydrocarbon field development because of the economical effects associated with it. In addition, t_b is highly sensitive to the spatial distribution of reservoir properties [78] (e.g. porosity, permeability) which are highly uncertain because of lack of observations. Moreover, reliable forecast for hydrocarbon production rate after the water breakthrough is significant for economical decisions. Therefore, in the present test case we develop a surrogate model for the production rate at late stages of the well life. In particular, we focus on the oil production rate q_{oil} when 60% of PVI has been injected.

The waterflooding system is modeled via mass and momentum conservation laws coupled with Darcy's law. A simplified model for waterflooding is utilized where flow of two incompressible fluids (water and oil) is considered. In this setting, we are interested in predicting the spatial distribution of volumetric fractions s_a (saturation) of each of the fluids. The index a could be replaced by either w or o for water and oil, respectively. The evolution of saturations is governed by mass and momentum conservation laws expressed through the following partial differential equation (PDE):

$$\frac{\partial \phi s_a \rho_a}{\partial t} - \sum_{\gamma=1}^{3} \frac{\partial}{\partial x^\gamma}\left(\frac{\rho_a k k_a}{\mu_a}\frac{\partial P}{\partial x^\gamma}\right) = Q_a \tag{35}$$

where $\gamma = 1, 2, 3$ is a spatial index of the coordinate vector $\mathbf{x}$, $\rho_a = \rho_a(\mathbf{x})$ and $\mu_a = \mu_a(\mathbf{x})$ are the density and viscosity of fluid a at the point $\mathbf{x}$ respectively, $k = k(\mathbf{x})$ is the permeability, $\phi = \phi(\mathbf{x})$ is the porosity at a given point, $P(\mathbf{x})$ is a pressure at point $\mathbf{x}$, $k_a(s)$ is a relative phase permeability of fluid a, $s = s(\mathbf{x})$ saturations of fluids at the point $\mathbf{x}$, $Q_a = Q_a(\mathbf{x})$ is a source term for fluid a at the point $\mathbf{x}$. Generally, the permeability k is a tensor. In the present example, we assume k to be a spherical tensor which can vary in space. Therefore, it is fully described by a single spatial function $k = k(\mathbf{x})$. In the present work we neglect capillary pressure effects. Therefore, both fluids are subjected to the same pressure at any given point. The source terms Q_a are considered to be non-zero only for cells with injection and production wells. Finally, incompressible fluids and rocks (solid matrix) are considered. Therefore, Eq. (35) could be simplified:

$$\phi\frac{\partial s_a}{\partial t} - \sum_{\gamma=1}^{3} \frac{\partial}{\partial x^\gamma}\left(\frac{k k_a}{\mu_a}\frac{\partial P}{\partial x^\gamma}\right) = q_a \tag{36}$$

where $q_a = Q_a/\rho_a$ is the source term for fluid a normalized to the density of corresponding fluid. For calculation of relative phase permeabilities, Brooks-Corey model [79] is used:

$$\begin{aligned} k_w(S_{\text{wn}}) &= k_w^{(0)} S_{\text{wn}}^{p_w} \\ k_w(S_{\text{wn}}) &= k_o^{(0)}(1 - S_{\text{wn}})^{p_o} \end{aligned} \tag{37}$$

Table 1
Fluid properties and parameters of the model for relative-phase permeability.

μ_o, cP	μ_w	p_o	p_w	$k_o^{(0)}$	$k_w^{(0)}$
10.0	1.0	2.0	2.0	1.0	1.0

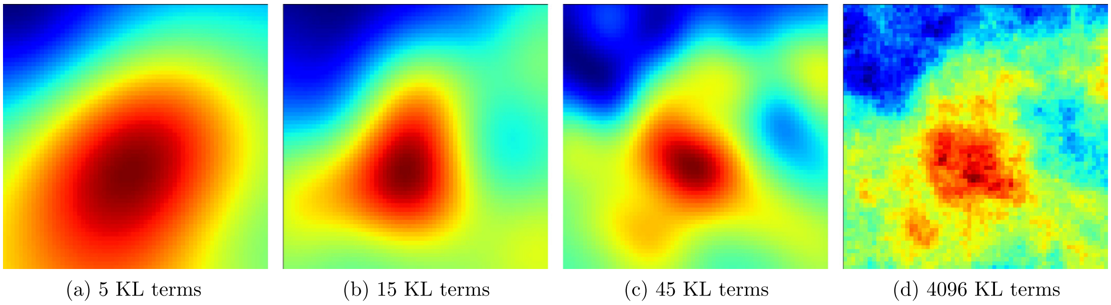

(a) 5 KL terms (b) 15 KL terms (c) 45 KL terms (d) 4096 KL terms

Fig. 3. Permeability realizations projected to a different number of KL-terms.

where k_w and k_o are the values of relative phase permeability for water and oil, respectively and $k_w^{(0)}$ and $k_o^{(0)}$ are maximum the values of relative phase permeability for water and oil, respectively. The values p_w and p_o are dimensionless parameters of the model and S_{wn} is the normalized water saturation defined as:

$$S_{\text{wn}} = \frac{S - S_{\text{wir}}}{1 - S_{\text{wir}} - S_{\text{owr}}} \tag{38}$$

where S_{wir} and S_{owr} are irreducible water and oil saturations, respectively.

In this test case, we consider a five-spot injection pattern where an injection well is located in the center of a square surrounded by four production wells. Given the symmetry of this pattern, only one quarter of the domain is modeled with one producer and one injector located at the opposite corners of a square domain. The length of the edge of that square is set to $L = 640$ m. The thickness of the reservoir is $h = 10$ m. We do not consider discretization along the vertical direction and we only consider a two-dimensional flow problem. For the purposes of simplicity, incompressible immiscible fluids is considered while neglecting gravity effects. A uniform square grid is used for simulations and the dimensions of each grid-block is 10 m by 10 m by 10 m. In other words, a 64 by 64 by 1 mesh is used for discretization. The porosity of the reservoir is assumed to be constant and equal to 0.2. Both injection and production rates are considered to be constant and equal to 10 m^3/day. The fluid properties and parameters of Corey model are presented in the Table 1.

In the present work, the reservoir permeability $k(\mathbf{x})$ is assumed to be a random field with a predefined distribution given the correlation between values at different points within the domain. In reservoir modeling, it is natural to assume that the values of logarithm of permeability $\log(k(\mathbf{r}))$ at different points $\mathbf{r}_1$ and $\mathbf{r}_2$ are exponentially correlated [76]:

$$\langle \log(k(\mathbf{r}_1)), \log(k(\mathbf{r}_2)) \rangle = \exp\left(- \frac{|\mathbf{r}_1 - \mathbf{r}_2|}{L_c} \right) \tag{39}$$

where L_c is a correlation length. In the present example, the correlation length is set to $L_c = 1/4L = 160m$. The utilized distribution of log-permeability allows one to implement Karhunen-Loeve expansion and express $\log(k(\mathbf{r}))$ as a linear combination of mutually independent random variables:

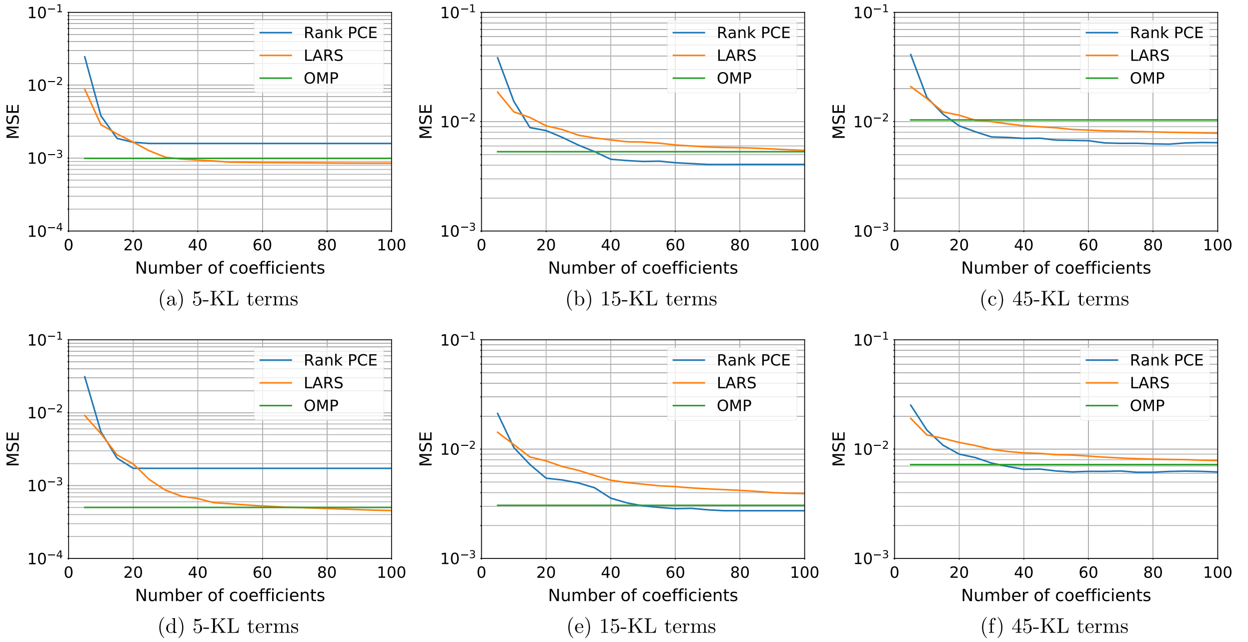
$$\log(k(\mathbf{r})) = \sum_{\alpha} \theta_\alpha \lambda_\alpha \xi_\alpha(\mathbf{r}) \tag{40}$$

where $\lambda_\alpha, \xi_\alpha(\mathbf{r})$ are the eigen-values and eigen-functions of the KL expansion, respectively. The θ_α are random mutually independent coefficients. In the present example θ_α are considered as input random variables for the PCE response surface. The permeability field is normalized such that a zero value of $\log(k(\mathbf{r}))$ corresponds to a permeability of 1 mD.

We truncate the KL expansion spectrum by taking the first 5, 15 or 45 KL components. Because of the long correlation length with respect to the size of the domain, a significant part of the energy of the spectrum is captured in all truncation scenarios. In this work, the fraction of the energy of the spectrum is defined as follows:

$$H(n) = \frac{\sum_{\alpha=1}^{n} \lambda_\alpha^2}{\sum_{\alpha=1}^{\infty} \lambda_\alpha^2} \tag{41}$$

where n is the number of components in the truncated KL expansion. In the present example, $H(5) = 0.9898$, $H(15) = 0.9948$ and $H(45) = 0.9972$. It is important to notice that despite the fact that KL expansion captures significant portion of the energy spectrum, it provides smooth reconstruction of the permeability field as shown in Fig. 3.



(a) 5-KL terms (b) 15-KL terms (c) 45-KL terms

(d) 5-KL terms (e) 15-KL terms (f) 45-KL terms

Fig. 4. Mean square error (MSE) over the test data for breakthrough time (a, b, c) and for oil production rate (d, c, e) for different PCE algorithms versus the response surface free parameters.

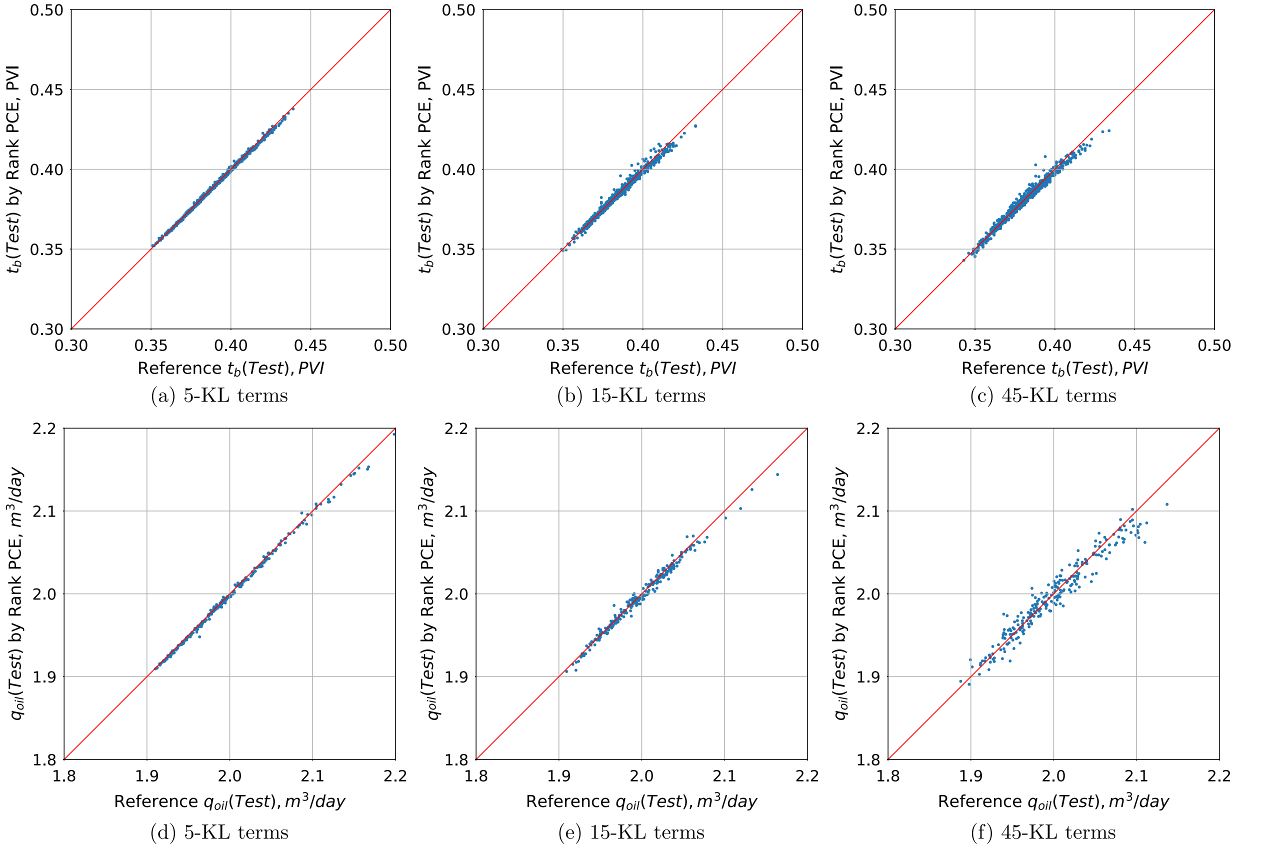
Three different PCE response surfaces are built corresponding to the 5, 15, 45-KL terms where 1000 samples (i.e. reconstructed permeability realizations) are evaluated. Each of these samples has been generated from a normal distribution of the coordinates θ_α corresponding to the truncated eigen-vectors of the KL expansion. Water breakthrough times are estimated through numerical simulations for each of the permeability realization using a forward simulation run. A training set of 750 samples is used for building the PCE response surface and the remaining 250 samples are used for testing. Legendre polynomials with degree $d \leq 5$ are considered as basis functions. The tolerance in all numerical schemes used to estimate the PCE coefficients has been set to 10^{-6}.

Fig. 4 shows the MSE for various KL truncation levels. The results presented in this figure, demonstrate that the proposed rank based PCE technique has similar accuracy when compared to the OMP algorithm for low-dimensional problems. However, the rank based PCE has clear advantages in higher dimensions. Also, both the Rank-PCE and OMP methods, perform slightly better than LARS. However, all three techniques do not perform perfectly, because the MSE is around 5% of the mean-value of the QoI. The cross-plot shown in Fig. 5 demonstrates how the quality of prediction is affected by the dimension of the problem or truncation scheme for KL expansion. The best accuracy of the response surface has been achieved for the problem with the lowest dimension corresponding to 5-KL truncation level (left). The numerical error is the highest for 45-KL truncation level (right). The reason for such behavior is that the training set of the same cardinality has been used for all truncation schemes. It should be noted that the accuracy of the permeability representation increases with the increase of parameter space dimension. However, capabilities of the response surface to reproduce the simulation results for a fixed number of direct simulations drop with dimension. In other words, more training data is required to build a high quality response surface for the high-dimensional case when compared with problems of lower dimensionality.

Test case 4: data from CO2 injection simulations

In this test case, PCE-based response surface is used as a fast emulator for CO2 injection process. The QoI is the mass of CO2 in a gas phase after given time period from the end of CO2 injection [72]. The data is based on the simulations results developed by Manceau and Rohmer [72]. The key uncertain parameters in these simulations are: average field porosity, average field permeability, regional hydraulic gradient relative phase permeability, capillary pressure and the permeability anisotropy k_v/k_h ratio. More detailed description of this problem can be found in [72]. The average field porosity ϕ and permeability k are considered as independent continuous variables with a uniform probability distribution via density-function variable transformation [70]:

$$\begin{cases} \phi = f_\phi(x_1) \\ k = f_k(x_2) \end{cases} \tag{42}$$



(a) 5-KL terms (b) 15-KL terms (c) 45-KL terms

(d) 5-KL terms (e) 15-KL terms (f) 45-KL terms

Fig. 5. Cross-plots of breakthrough time and oil production rate in PVI units for Rank PCE algorithm. Top row (a, b, c) shows the results for test samples for breakthrough time and the bottom row (d, e, f) shows the results for oil production rate.

Table 2
Summary of variables notations and types.

Variable	Notation	Type
Porosity	x_1	Continuous, $\mathcal{U}[-1; 1]$
Permeability	x_2	Continuous, $\mathcal{U}[-1; 1]$
Relative phase permeability	x_3	Discrete, 10 different models
Regional hydraulic gradient	x_4	Discrete, 2 different values
Capillary pressure	x_5	Discrete, 2 different models
Permeability anisotropy	x_6	Discrete, 3 different values

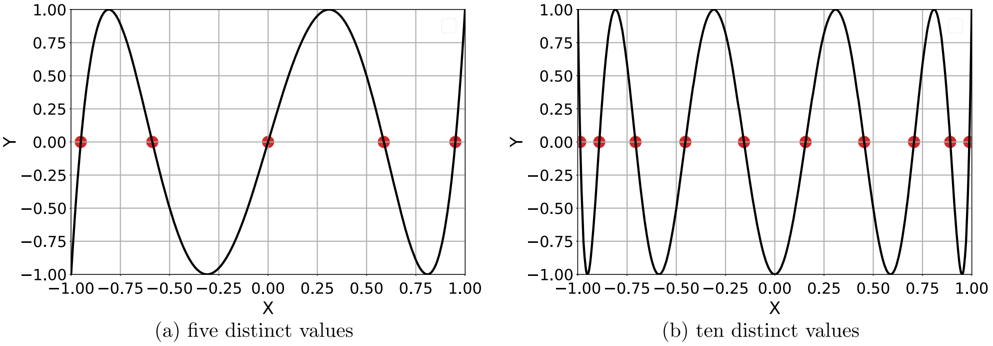
where x_1 and x_2 are independent random variables uniformly distributed in the interval $[-1; 1]$, f_ϕ and f_k are functions for transformation of variables. All other variables are considered as discrete variables with equal probabilities over all discretized values. Table 2 summarizes the variable names and types used in this test case.

We note that this test case includes categorical variables in the input space. In order to handle this type of data, we utilize Chebyshev polynomials for categorical data. Additionally, we establish a one-to-one correspondence between the values of a given categorical variable and Chebyshev nodes:

$$t_m \rightarrow \cos\left(\frac{2m-1}{2M}\pi\right) \tag{43}$$

where M is the total number of possible values for a given categorical variable. The mapping given by this equation is illustrated in Fig. 6.

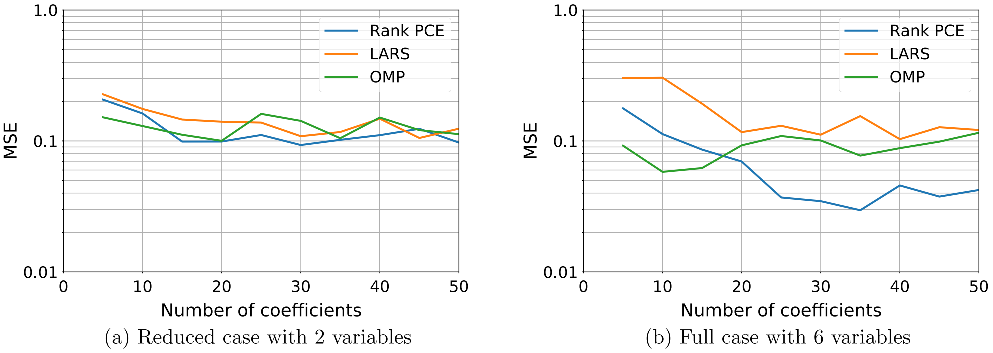
We note that Gauss-quadrature rules for Chebyshev polynomials have the same weight [65] for each of the nodes. This justifies using Chebyshev polynomials for categorical data and the corresponding mapping to the Chebyshev nodes presented in Eq. (43) especially when training samples are uniformly distributed over the distinct categories. Therefore, the polynomials orthogonality and the distribution of categorical variables are consistent with each other.



(a) five distinct values

(b) ten distinct values

Fig. 6. Location of Chebyshev nodes corresponding to roots of the polynomials with the same degrees as the number of distinct values present in the categorical variable.



(a) Reduced case with 2 variables

(b) Full case with 6 variables

Fig. 7. Mean square error (test data) for the different PCE algorithm versus the response surface free parameters.

$$\int_{-1}^{+1} \frac{p_\alpha(t)p_\beta(t)}{\sqrt{1-t^2}}\mathrm{d}t = \sum_m \frac{\pi}{M} p_\alpha(t_m)p_\beta(t_m) = \pi\ \mathbb{E}[p_\alpha p_\beta] \tag{44}$$

In the present work we utilize normalized polynomials $p_\alpha(t)$ to $q_\alpha(t)$:

$$\mathbb{E}[q_\alpha q_\beta] = \delta_{\alpha\beta} \tag{45}$$

Using Chebyshev polynomials provides a natural extension of standard PCE to problems with categorical variables while preserving the fundamental relation between the orthogonality of basis functions and probability distribution as defined in Eq. (5).

In the current example, sampling of the data is performed using uniform distributions over the parameter ranges. A total of 998 data points are generated in accordance with the proposed probability distributions of variables and we used 250 data-points for training (i.e. constructing the PCE) and the remaining data points are used for testing. The mass of CO2 injection is computed via detailed numerical simulations (see [72] for more details). We normalized the QoI such that following equality holds for the training data:

$$\sum_i \frac{y_i^2}{N} = 1 \tag{46}$$

We observed empirically that the QoI is highly sensitive to the permeability and relative phase permeability. Therefore, we constructed two evaluation cases with the same data set. For the first case which we refer to as the reduced case, we built a two-dimensional response surface using the permeability and relative phase permeability only as an input. The second case, which we denote as the full case, we utilize all the six uncertain variables in the response surface. In both cases, we evaluate the proposed ranking based sparse PCE against standard sparse regression PCE algorithms (i.e. LARS and OMP methods) for different numbers of expansion coefficients N_D. For both the reduced and full problems, PCE is performed with polynomials of degree $d \leq 10$. The number of terms in PCE varies from 5 to 50 and the tolerance has been set similar to all other test cases to 10^{-6}. Legendre polynomials were used for continuous variables x_1, x_2 and Chebyshev polynomials were used for the discrete/categorical variables.

Fig. 7, shows the mean square error over the test data set for both the reduced and full cases in Fig. 7a and Fig. 7b, respectively. The introduced ranking based approach shows better convergence rates for both problems. Moreover, the results in Fig. 7b demonstrate that advantages of the proposed Rank-PCE are more pronounced for higher dimensional problems, where the search space inside the iterative solver is large. For this case, the introduced ranking step allows for an efficient identification of the most significant components of PCE resulting in a higher quality response surfaces.

4. Concluding remarks

In the current manuscript, we introduced a ranking based sparse PCE technique (Rank-PCE). The core idea of the proposed approach is to rank the PCE features in accordance with the magnitude of a given PCE coefficient based on the correlation with data while estimating for the accuracy of computed correlations. We demonstrated, via a set of numerical examples, the superior performance of Rank-PCE when compared to standard sparse regularization techniques. Rank-PCE resulted in an increase in convergence rates for generative function with sparse spectrum. We also noticed that the improvements in convergence is more pronounced for high-dimensional problems, enabling the application of PCE to problems with significant number of independent variables. Moreover, the advantages of Rank-PCE are also evident for problems with limited number of training samples as demonstrated in the analytical test cases.

In addition to novel ranking procedure, we presented an extension of PCE response surfaces to problems with both continuous and categorical data through the utilization of Chebyshev polynomials to represent the discrete variables. The proposed technique might be not optimal for general cases, however under the uniform sampling conditions, it provides a simple approach to handle categorical data in PCE that is consistent with the statistical properties of PCE for sensitivity analysis and UQ. In other words, the proposed approach maintains the relation between basis orthogonality and statistics of the input variables, which is fundamental for UQ with PCE. This technique is also easy to implement given the availability of Chebyshev polynomials in most scientific computing libraries.

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.

Acknowledgements

This work is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 653718. We thank J.C. Manceau and J. Rohmer (BRGM, France) for providing the CO2 injection dataset used in test cases. We also thank Dr. Shing Chan (formally at HWU and currently at the Oxford Big Data Institute) for developing the reservoir simulation codes used in the third example.

References

[1] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[2] Murat Fatih Tugan, Caglar Sinayuc, A new fully probabilistic methodology and a software for assessing uncertainties and managing risks in shale gas projects at any maturity stage, J. Pet. Sci. Eng. (ISSN 0920-4105) 168 (2018) 107–118, https://doi.org/10.1016/j.petrol.2018.05.001, http://www.sciencedirect.com/science/article/pii/S0920410518303905.

[3] Qianlin Wang, Laibin Zhang, Jinqiu Hu, An integrated method of human error likelihood assessment for shale-gas fracturing operations based on SPA and UAHP, Process Saf. Environ. Prot. (ISSN 0957-5820) 123 (2019) 105–115, https://doi.org/10.1016/j.psep.2019.01.003, http://www.sciencedirect.com/science/article/pii/S0957582018308449.

[4] Bailian Chen, Jincong He, Xian-Huan Wen, Wen Chen, Albert C. Reynolds, Uncertainty quantification and value of information assessment using proxies and Markov chain Monte Carlo method for a pilot project, J. Pet. Sci. Eng. (ISSN 0920-4105) 157 (2017) 328–339, https://doi.org/10.1016/j.petrol.2017.07.039, http://www.sciencedirect.com/science/article/pii/S0920410517305909.

[5] Xiang Ma, Nicholas Zabaras, A stochastic mixed finite element heterogeneous multiscale method for flow in porous media, J. Comput. Phys. (ISSN 0021-9991) 230 (12) (2011) 4696–4722, https://doi.org/10.1016/j.jcp.2011.03.001, http://www.sciencedirect.com/science/article/pii/S0021999111001318.

[6] Yingqi Zhang, Curtis M. Oldenburg, Stefan Finsterle, Preston Jordan, Keni Zhang, Probability estimation of CO2 leakage through faults at geologic carbon sequestration sites, in: Greenhouse Gas Control Technologies 9, Energy Proc. (ISSN 1876-6102) 1 (1) (2009) 41–46, https://doi.org/10.1016/j.egypro.2009.01.008, http://www.sciencedirect.com/science/article/pii/S1876610209000095.

[7] Steinar M. Elgsaeter, Olav Slupphaug, Tor Arne Johansen, Oil and gas production optimization; lost potential due to uncertainty, in: 17th IFAC World Congress, IFAC Proc. Vol. (ISSN 1474-6670) 41 (2) (2008) 4540–4547, https://doi.org/10.3182/20080706-5-KR-1001.00764, http://www.sciencedirect.com/science/article/pii/S1474667016396598.

[8] Wei Jia, Brian McPherson, Feng Pan, Zhenxue Dai, Ting Xiao, Uncertainty quantification of CO2 storage using Bayesian model averaging and polynomial chaos expansion, Int. J. Greenh. Gas Control (ISSN 1750-5836) 71 (2018) 104–115, https://doi.org/10.1016/j.ijggc.2018.02.015, http://www.sciencedirect.com/science/article/pii/S175058361730419X.

[9] Klaus S. Lackner, A guide to CO2 sequestration, Science (ISSN 0036-8075) 300 (5626) (2003) 1677–1678, https://doi.org/10.1126/science.1079033, http://science.sciencemag.org/content/300/5626/1677.

[10] Stefan Bachu, Didier Bonijoly, John Bradshaw, Robert Burruss, Sam Holloway, Niels Peter Christensen, Odd Magne Mathiassen, CO2 storage capacity estimation: methodology and gaps, Int. J. Greenh. Gas Control (ISSN 1750-5836) 1 (4) (2007) 430–443, https://doi.org/10.1016/S1750-5836(07)00086-2, http://www.sciencedirect.com/science/article/pii/S1750583607000862.

[11] Richard Webster , Margaret A. Oliver, Basic Steps in Geostatistics: The Variogram and Kriging, Springer, Cham, 2015.

[12] Jingwen Zheng, Juliana Y. Leung, Ronald P. Sawatzky, Jose M. Alvarez, A cluster-based approach for visualizing and quantifying the uncertainty in the impacts of uncertain shale barrier configurations on SAGD production, in: SPE Canada Heavy Oil Technical Conference, Calgary, Alberta, Canada, 2014.

[13] T. Dodwell, C. Ketelsen, R. Scheichl, A. Teckentrup, A hierarchical multilevel Markov Chain Monte Carlo algorithm with applications to uncertainty quantification in subsurface flow, SIAM/ASA J. Uncertain. Quantificat. 3 (1) (2015) 1075–1108, https://doi.org/10.1137/130915005.

[14] M. Khasanov, V. Babin, O. Melchaeva, O. Ushmaev, D. Echeverria, A. Semenikhin, Application of mathematical optimization techniques for well pattern selection, in: SPE Russian Oil and Gas Exploration & Production Technical Conference and Exhibition, Moscow, Russia, 2014.

[15] Na Zhang, Yating Wang, Qian Sun, Yuhe Wang, Multiscale mass transfer coupling of triple-continuum and discrete fractures for flow simulation in fractured vuggy porous media, Int. J. Heat Mass Transf. (ISSN 0017-9310) 116 (2018) 484–495, https://doi.org/10.1016/j.ijheatmasstransfer.2017.09.046, http://www.sciencedirect.com/science/article/pii/S0017931017316964.

[16] Qiuqi Li, Yuhe Wang, Maria Vasilyeva, Multiscale model reduction for fluid infiltration simulation through dual-continuum porous media with localized uncertainties, J. Comput. Appl. Math. (ISSN 0377-0427) 336 (2018) 127–146, https://doi.org/10.1016/j.cam.2017.12.040, http://www.sciencedirect.com/science/article/pii/S0377042718300086.

[17] I. Yucel Akkutlu, Yalchin Efendiev, Maria Vasilyeva, Yuhe Wang, Multiscale model reduction for shale gas transport in a coupled discrete fracture and dual-continuum porous media, in: Multiscale and Multiphysics Techniques and their Applications in Unconventional Gas Reservoirs, J. Nat. Gas Sci. Eng. (ISSN 1875-5100) 48 (2017) 65–76, https://doi.org/10.1016/j.jngse.2017.02.040, http://www.sciencedirect.com/science/article/pii/S1875510017300938.

[18] J.E. Killough, M. Rame, A new approach to flow simulation in highly heterogeneous porous media, SPE Form. Eval. (ISSN 0885-923X) 7 (1992), https://doi.org/10.2118/21247-PA.

[19] Pascal Audigane, Martin J. Blunt, Dual mesh method for upscaling in waterflood simulation, Transp. Porous Media 55 (01) (2004) 71–89, https://doi.org/10.1023/B:TIPM.0000007309.48913.d2.

[20] D. Khoozan, B. Firoozabadi, D. Rashtchian, M.A. Ashjari, Analytical dual mesh method for two-phase flow through highly heterogeneous porous media, J. Hydrol. (ISSN 0022-1694) 400 (1) (2011) 195–205, https://doi.org/10.1016/j.jhydrol.2011.01.042, http://www.sciencedirect.com/science/article/pii/S0022169411000679.

[21] Maria Vasilyeva, Eric T. Chung, Siu Wun Cheung, Yating Wang, Georgy Prokopev, Nonlocal multicontinua upscaling for multicontinua flow problems in fractured porous media, J. Comput. Appl. Math. (ISSN 0377-0427) (2019), https://doi.org/10.1016/j.cam.2019.01.024, http://www.sciencedirect.com/science/article/pii/S0377042719300408.

[22] Michael Christie, Upscaling for reservoir simulation, J. Pet. Technol. 48 (11) (1996) 1004–1010, https://doi.org/10.2118/37324-MS.

[23] Dasheng Qi, Tim Hesketh, An analysis of upscaling techniques for reservoir simulation, Pet. Sci. Technol. 23 (7–8) (2005) 827–842, https://doi.org/10.1081/LFT-200033132.

[24] Shing Chan, Ahmed H. Elsheikh, A machine learning approach for efficient uncertainty quantification using multiscale methods, J. Comput. Phys. (ISSN 0021-9991) 354 (2018) 493–511, https://doi.org/10.1016/j.jcp.2017.10.034.

[25] Yalchin Efendiev, Juan Galvis, Eduardo Gildin, Local–global multiscale model reduction for flows in high-contrast heterogeneous media, J. Comput. Phys. (ISSN 0021-9991) 231 (24) (2012) 8100–8113, https://doi.org/10.1016/j.jcp.2012.07.032, http://www.sciencedirect.com/science/article/pii/S0021999112004160.

[26] Kevin Carlberg, Youngsoo Choi, Syuzanna Sargsyan, Conservative model reduction for finite-volume models, J. Comput. Phys. (ISSN 0021-9991) 371 (2018) 280–314, https://doi.org/10.1016/j.jcp.2018.05.019, http://www.sciencedirect.com/science/article/pii/S002199911830319X.

[27] Moritz Gosses, Wolfgang Nowak, Thomas Wöhling, Explicit treatment for Dirichlet, Neumann and Cauchy boundary conditions in POD-based reduction of groundwater models, Adv. Water Resour. (ISSN 0309-1708) 115 (2018) 160–171, https://doi.org/10.1016/j.advwatres.2018.03.011, http://www.sciencedirect.com/science/article/pii/S0309170817307467.

[28] J. Nagoor Kani, Ahmed H. Elsheikh, Reduced-order modeling of subsurface multi-phase flow models using deep residual recurrent neural networks, Transp. Porous Media (ISSN 0169-3913) (2018) 1–29, https://doi.org/10.1007/s11242-018-1170-7.

[29] Shaoxing Mo, Yinhao Zhu, Nicholas Zabaras, Xiaoqing Shi, Jichun Wu, Deep convolutional encoder-decoder networks for uncertainty quantification of dynamic multiphase flow in heterogeneous media, Water Resour. Res. 55 (1) (2019) 703–728, https://doi.org/10.1029/2018WR023528, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018WR023528.

[30] Simeon Agada, Sebastian Geiger, Ahmed Elsheikh, Sergey Oladyshkin, Data-driven surrogates for rapid simulation and optimization of WAG injection in fractured carbonate reservoirs, Pet. Geosci. (ISSN 1354-0793) (2016), https://doi.org/10.1144/petgeo2016-068, https://pg.lyellcollection.org/content/early/2016/12/10/petgeo2016-068.

[31] Laureline Josset, Vasily Demyanov, Ahmed Elsheikh, Ivan Lunati, Accelerating Monte Carlo Markov chains with proxy and error models, in: Statistical Learning in Geoscience Modelling: Novel Algorithms and Challenging Case Studies, Comput. Geosci. (ISSN 0098-3004) 85 (2015) 38–48, https://doi.org/10.1016/j.cageo.2015.07.003, http://www.sciencedirect.com/science/article/pii/S009830041530011X.

[32] Roland Schöbi, Bruno Sudret, Global sensitivity analysis in the context of imprecise probabilities (p-boxes) using sparse polynomial chaos expansions, Reliab. Eng. Syst. Saf. (ISSN 0951-8320) (2018), https://doi.org/10.1016/j.ress.2018.11.021, http://www.sciencedirect.com/science/article/pii/S0951832017306099.

[33] Alejandra Camacho, Alvaro Talavera, Alexandre A. Emerick, Marco A.C. Pacheco, João Zanni, Uncertainty quantification in reservoir simulation models with polynomial chaos expansions: Smolyak quadrature and regression method approach, J. Pet. Sci. Eng. (ISSN 0920-4105) 153 (2017) 203–211, https://doi.org/10.1016/j.petrol.2017.03.046, http://www.sciencedirect.com/science/article/pii/S0920410517303960.

[34] Xiaojing Wu, Weiwei Zhang, Shufang Song, Zhengyin Ye, Sparse grid-based polynomial chaos expansion for aerodynamics of an airfoil with uncertainties, Chin. J. Aeronaut. (ISSN 1000-9361) 31 (5) (2018) 997–1011, https://doi.org/10.1016/j.cja.2018.03.011, http://www.sciencedirect.com/science/article/pii/S1000936118301031.

[35] Jun Xu, Fan Kong, A cubature collocation based sparse polynomial chaos expansion for efficient structural reliability analysis, Struct. Saf. (ISSN 0167-4730) 74 (2018) 24–31, https://doi.org/10.1016/j.strusafe.2018.04.001, http://www.sciencedirect.com/science/article/pii/S0167473017303922.

[36] Pramudita Satria Palar, Lavi Rizki Zuhal, Koji Shimoyama, Takeshi Tsuchiya, Global sensitivity analysis via multi-fidelity polynomial chaos expansion, Reliab. Eng. Syst. Saf. (ISSN 0951-8320) 170 (2018) 175–190, https://doi.org/10.1016/j.ress.2017.10.013, http://www.sciencedirect.com/science/article/pii/S0951832016304872.

[37] S. Abraham, M. Raisee, G. Ghorbaniasl, F. Contino, C. Lacor, A robust and efficient stepwise regression method for building sparse polynomial chaos expansions, J. Comput. Phys. (ISSN 0021-9991) 332 (2017) 461–474, https://doi.org/10.1016/j.jcp.2016.12.015, http://www.sciencedirect.com/science/article/pii/S0021999116306684.

[38] Negin Alemazkoor, Hadi Meidani, A preconditioning approach for improved estimation of sparse polynomial chaos expansions, Comput. Methods Appl. Mech. Eng. (ISSN 0045-7825) 342 (2018) 474–489, https://doi.org/10.1016/j.cma.2018.08.005, http://www.sciencedirect.com/science/article/pii/S0045782518303918.

[39] Vahid Abolghasemi, Saideh Ferdowsi, Saeid Sanei, A gradient-based alternating minimization approach for optimization of the measurement matrix in compressive sensing, Signal Process. (ISSN 0165-1684) 92 (4) (2012) 999–1009, https://doi.org/10.1016/j.sigpro.2011.10.012, http://www.sciencedirect.com/science/article/pii/S0165168411003665.

[40] G. Li, Z. Zhu, D. Yang, L. Chang, H. Bai, On projection matrix optimization for compressive sensing systems, IEEE Trans. Signal Process. (ISSN 1053-587X) 61 (11) (June 2013) 2887–2898, https://doi.org/10.1109/TSP.2013.2253776.

[41] Jerrad Hampton, Alireza Doostan, Coherence motivated sampling and convergence analysis of least squares polynomial chaos regression, Comput. Methods Appl. Mech. Eng. (ISSN 0045-7825) 290 (2015) 73–97, https://doi.org/10.1016/j.cma.2015.02.006, http://www.sciencedirect.com/science/article/pii/S004578251500047X.

[42] Serhat Hosder, Robert W. Walters, Michael Balch, Point-collocation nonintrusive polynomial chaos method for stochastic computational fluid dynamics, AIAA J. 48 (12) (dec 2010) 2721–2730, https://doi.org/10.2514/1.39389.

[43] Dongbin Xiu, Liang Yan, Ling Guo, Stochastic collocation algorithms using L1 minimization, Int. J. Uncertain. Quantificat. (ISSN 2152-5080) 2 (3) (2012) 279–293.

[44] Leo Wai-Tsun Ng, Michael Eldred, Multifidelity uncertainty quantification using non-intrusive polynomial chaos and stochastic collocation, in: 53rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference, 23 April 2012 – 26 April 2012, Honolulu, Hawaii, ISBN 978-1-60086-937-2, 2012.

[45] Ali Alkhatib Masoud Babaei, Indranil Pan, Robust optimization of well location to enhance hysteretical trapping of CO2: assessment of various uncertainty quantification methods and utilization of mixed response surface surrogates, Water Resour. Res. 51 (12) (2015) 9402–9424, https://doi.org/10.1002/2015WR017418, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015WR017418.

[46] Indranil Pan, Ali Alkhatib Masoud Babaei, Robust optimization of subsurface flow using polynomial chaos and response surface surrogates, Comput. Geosci. (ISSN 1573-1499) 19 (5) (Oct 2015) 979–998, https://doi.org/10.1007/s10596-015-9516-5.

[47] Kurt R. Petvipusit, Ahmed H. Elsheikh, Tara C. Laforce, Peter R. King, Martin J. Blunt, Robust optimisation of CO2 sequestration strategies under geological uncertainty using adaptive sparse grid surrogates, Comput. Geosci. (ISSN 1573-1499) 18 (5) (Oct 2014) 763–778, https://doi.org/10.1007/s10596-014-9425-z.

[48] Ahmed H. Elsheikh, Ibrahim Hoteit, Mary F. Wheeler, Efficient Bayesian inference of subsurface flow models using nested sampling and sparse polynomial chaos surrogates, Comput. Methods Appl. Mech. Eng. (ISSN 0045-7825) 269 (2014) 515–537, https://doi.org/10.1016/j.cma.2013.11.001, http://www.sciencedirect.com/science/article/pii/S004578251300296X.

[49] Fotios Petropoulos, Nikolaos Kourentzes, Konstantinos Nikolopoulos, Enno Siemsen, Judgmental selection of forecasting models, J. Oper. Manag. (ISSN 0272-6963) 60 (2018) 34–46, https://doi.org/10.1016/j.jom.2018.05.005, http://www.sciencedirect.com/science/article/pii/S0272696318300251.

[50] Jerrad Hampton, Alireza Doostan, Basis adaptive sample efficient polynomial chaos (BASE-PC), J. Comput. Phys. (ISSN 0021-9991) 371 (2018) 20–49, https://doi.org/10.1016/j.jcp.2018.03.035, http://www.sciencedirect.com/science/article/pii/S0021999118301955.

[51] Hamid Bazargan, Mike Christie, Ahmed H. Elsheikh, Mohammad Ahmadi, Surrogate accelerated sampling of reservoir models with complex structures using sparse polynomial chaos expansion, in: Data Assimilation for Improved Predictions of Integrated Terrestrial Systems, Adv. Water Resour. (ISSN 0309-1708) 86 (2015) 385–399, https://doi.org/10.1016/j.advwatres.2015.09.009, http://www.sciencedirect.com/science/article/pii/S030917081500216X.

[52] Katerina Konakli, Bruno Sudret, Polynomial meta-models with canonical low-rank approximations: numerical insights and comparison to sparse polynomial chaos expansions, J. Comput. Phys. (ISSN 0021-9991) 321 (2016) 1144–1169, https://doi.org/10.1016/j.jcp.2016.06.005, http://www.sciencedirect.com/science/article/pii/S0021999116302303.

[53] Jin Meng, Heng Li, An efficient stochastic approach for flow in porous media via sparse polynomial chaos expansion constructed by feature selection, Adv. Water Resour. (ISSN 0309-1708) 105 (2017) 13–28, https://doi.org/10.1016/j.advwatres.2017.04.019, http://www.sciencedirect.com/science/article/pii/S030917081630625X.

[54] Ling Guo, Akil Narayan, Tao Zhou, A gradient enhanced ℓ1-minimization for sparse approximation of polynomial chaos expansions, J. Comput. Phys. (ISSN 0021-9991) 367 (2018) 49–64, https://doi.org/10.1016/j.jcp.2018.04.026, http://www.sciencedirect.com/science/article/pii/S0021999118302420.

[55] Mishal Thapa, Sameer B. Mulani, Robert W. Walters, A new non-intrusive polynomial chaos using higher order sensitivities, Comput. Methods Appl. Mech. Eng. (ISSN 0045-7825) 328 (2018) 594–611, https://doi.org/10.1016/j.cma.2017.09.024, http://www.sciencedirect.com/science/article/pii/S0045782517306539.

[56] Kai Cheng, Zhenzhou Lu, Adaptive sparse polynomial chaos expansions for global sensitivity analysis based on support vector regression, Comput. Struct. (ISSN 0045-7949) 194 (2018) 86–96, https://doi.org/10.1016/j.compstruc.2017.09.002, http://www.sciencedirect.com/science/article/pii/S0045794917305047.

[57] Srikara Pranesh, Debraj Ghosh, Cost reduction of stochastic Galerkin method by adaptive identification of significant polynomial chaos bases for elliptic equations, Comput. Methods Appl. Mech. Eng. (ISSN 0045-7825) 340 (2018) 54–69, https://doi.org/10.1016/j.cma.2018.04.043, http://www.sciencedirect.com/science/article/pii/S0045782518302287.

[58] Xiangfeng Guo, Daniel Dias, Claudio Carvajal, Laurent Peyras, Pierre Breul, Reliability analysis of embankment dam sliding stability using the sparse polynomial chaos expansion, Eng. Struct. (ISSN 0141-0296) 174 (2018) 295–307, https://doi.org/10.1016/j.engstruct.2018.07.053, http://www.sciencedirect.com/science/article/pii/S014102961830511X.

[59] Géraud Blatman, Bruno Sudret, Adaptive sparse polynomial chaos expansion based on least angle regression, J. Comput. Phys. (ISSN 0021-9991) 230 (6) (2011) 2345–2367, https://doi.org/10.1016/j.jcp.2010.12.021, http://www.sciencedirect.com/science/article/pii/S0021999110006856.

[60] Xiu Yang, Huan Lei, Nathan A. Baker, Guang Lin, Enhancing sparsity of Hermite polynomial expansions by iterative rotations, J. Comput. Phys. (ISSN 0021-9991) 307 (2016) 94–109, https://doi.org/10.1016/j.jcp.2015.11.038, http://www.sciencedirect.com/science/article/pii/S0021999115007780.

[61] Qiujing Pan, Daniel Dias, Sliced inverse regression-based sparse polynomial chaos expansions for reliability analysis in high dimensions, in: Special Section: Applications of Probabilistic Graphical Models in Dependability, Diagnosis and Prognosis, Reliab. Eng. Syst. Saf. (ISSN 0951-8320) 167 (2017) 484–493, https://doi.org/10.1016/j.ress.2017.06.026, http://www.sciencedirect.com/science/article/pii/S095183201630864X.

[62] Trevor Hastie, Rob Tibshirani, Regularization paths for generalized linear models via coordinate descent, in: Special Section: Applications of Probabilistic Graphical Models in Dependability, Diagnosis and Prognosis, J. Stat. Softw. (ISSN 1548-7660) 33 (2010) 1–22, https://www.ncbi.nlm.nih.gov/pubmed/20808728.

[63] Ling Guo, Yongle Liu, Tao Zhou, Data-driven polynomial chaos expansions: a weighted least-square approximation, J. Comput. Phys. (ISSN 0021-9991) 381 (2019) 129–145, https://doi.org/10.1016/j.jcp.2018.12.020, http://www.sciencedirect.com/science/article/pii/S0021999119300014.

[64] Arun Kaintura, Tom Dhaene, Domenico Spina, Review of polynomial chaos-based methods for uncertainty quantification in modern integrated circuits, Electronics (ISSN 2079-9292) 7 (3) (2018) 21, https://doi.org/10.3390/electronics7030030.

[65] Milton Abramowitz, Irene A. Stegun, Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, Dover, New York, 1964, ninth Dover printing, tenth GPO printing edition.

[66] J.-C. Cortés, J.-V. Romero, M.-D. Roselló, R.-J. Villanueva, Improving adaptive generalized polynomial chaos method to solve nonlinear random differential equations by the random variable transformation technique, Commun. Nonlinear Sci. Numer. Simul. (ISSN 1007-5704) 50 (2017) 1–15, https://doi.org/10.1016/j.cnsns.2017.02.011, http://www.sciencedirect.com/science/article/pii/S1007570417300588.

[67] Albert N. Shiryaev, Probability, Springer-Verlag, New York, ISBN 978-1-4757-2539-1, 1996.

[68] Christine De Mol, Ernesto De Vito, Lorenzo Rosasco, Elastic-net regularization in learning theory, J. Complex. (ISSN 0885-064X) 25 (2) (2009) 201–230, https://doi.org/10.1016/j.jco.2009.01.002, http://www.sciencedirect.com/science/article/pii/S0885064X0900003X.

[69] Dhammika Amaratunga, Javier Cabrera, Analysis of data from viral DNA microchips, J. Am. Stat. Assoc. 96 (456) (2001) 1161–1170, https://doi.org/10.1198/016214501753381814.

[70] Murray Rosenblatt, Remarks on a multivariate transformation, Ann. Math. Stat. 23 (3) (1952) 470–472, https://doi.org/10.1214/aoms/1177729394.

[71] E. Torre, S. Marelli, P. Embrechts, B. Sudret, Data-driven polynomial chaos expansion for machine learning regression, preprint, arXiv:1808.03216, 2018, https://arxiv.org/abs/1808.03216.

[72] J. Manceau, J. Rohmer, Post-injection trapping of mobile CO2 in deep aquifers: assessing the importance of model and parameter uncertainties, Comput. Geosci. (ISSN 1573-1499) 20 (2016) 1251–1267, https://doi.org/10.1007/s10596-016-9588-x.

[73] Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, Least angle regression, Ann. Stat. 32 (2) (2004) 407–499, https://doi.org/10.1214/009053604000000067.

[74] Ron Rubinstein, Michael Zibulevsky, Michael Elad, Efficient implementation of the K-SVD algorithm using Batch orthogonal matching pursuit, CS Technion 40 (2008).

[75] David Ackley, A Connectionist Machine for Genetic Hillclimbing, Springer US, ISBN 978-1-4612-9192-3, 1987.

[76] Xiang Ma, Nicholas Zabaras, Kernel principal component analysis for stochastic input model generation, J. Comput. Phys. (ISSN 0021-9991) 230 (19) (2011) 7311–7331, https://doi.org/10.1016/j.jcp.2011.05.037, http://www.sciencedirect.com/science/article/pii/S0021999111003494.

[77] Ahmed Tarek, Principles of waterflooding, Chapter 14, in: Ahmed Tarek (Ed.), Reservoir Engineering Handbook, fifth edition, Gulf Professional Publishing, ISBN 978-0-12-813649-2, 2019, pp. 901–1107, http://www.sciencedirect.com/science/article/pii/B9780128136492000141.

[78] Nélio Henderson, Luciana Pena, Simulating effects of the permeability anisotropy on the formation of viscous fingers during waterflood operations, J. Pet. Sci. Eng. (ISSN 0920-4105) 153 (2017) 178–186, https://doi.org/10.1016/j.petrol.2017.03.047, http://www.sciencedirect.com/science/article/pii/S0920410517304102.

[79] R.H. Brooks, A.T. Corey, Hydraulic properties of porous media, Hydrol. Pap. 3 (1964).