Uncertainty quantification in CO_2 sequestration using surrogate models from polynomial chaos expansion

Yan Zhang[†,‡] and Nikolaos V. Sahinidis[*,†,‡]

National Energy Technology Laboratory, Pittsburgh, PA, US, and Department of Chemical Engineering, Carnegie Mellon University, Pittsburgh, PA, US

E-mail: sahinidis@cmu.edu

Phone: +1 412-268-3338. Fax: +1 412-268-7139

Abstract

In this paper, surrogate models are iteratively built using polynomial chaos expansion (PCE) and detailed numerical simulations of a carbon sequestration system. Output variables from a numerical simulator are approximated as polynomial functions of uncertain parameters. Once generated, PCE representations can be used in place of the numerical simulator and often decrease simulation times by several orders of magnitude. However, PCE models are expensive to derive unless the number of terms in the expansion is moderate, which requires a relatively small number of uncertain variables and a low degree of expansion. To cope with this limitation, instead of using a classical full expansion at each step of an iterative PCE construction method, we introduce a mixed-integer programming (MIP) formulation to identify the best subset of basis terms in the expansion. This approach makes it possible to keep the

[*]To whom correspondence should be addressed

[†]National Energy Technology Laboratory, Pittsburgh, PA, US

[‡]Department of Chemical Engineering, Carnegie Mellon University, Pittsburgh, PA, US

number of terms small in the expansion. Monte Carlo (MC) simulation is then performed by substituting the values of the uncertain parameters into the closed-form polynomial functions. Based on the results of MC simulation, the uncertainties of injecting CO_2 underground are quantified for a saline aquifer. Moreover, based on the PCE model, we formulate an optimization problem to determine the optimal CO_2 injection rate so as to maximize the gas saturation (residual trapping) during injection, and thereby minimize the chance of leakage.

Introduction

Due to increased recent concerns about the effects of CO_2 emissions on global warming, various forms of carbon emission reduction/elimination technologies are under intensive research around the world.[1] Among those, carbon capture and storage (CCS) is considered as one of the promising technologies to reduce CO_2 levels in the atmosphere.[2] CCS usually involves capturing CO_2 from large stationary facilities, such as power plants, and isolating the captured CO_2 from the atmosphere over a long time period. One option for CO_2 storage is to inject CO_2 into deep geological formations, such as saline aquifers, depleted oil/gas reservoirs, or deep unmineable coal seams. It is often assumed that there is a layer of impermeable formation above the storage layer, which seals the sequestration system and prevents CO_2 from escaping.

Extensive research has been done to understand the physical processes in CO_2 sequestration. These processes are modeled by solving governing equations for mass and heat balance, using analytical formulas under simplified assumptions,[3–5] or detailed numerical simulators, such as ECLIPSE[6] or TOUGH2.[7] A benchmark study recently compared several numerical models of CO_2 storage.[8]

Because of incomplete knowledge or limited measurement ability, parameters such as porosities and permeabilities in the equations governing CO_2 plume dynamic often remain uncertain. These uncertainties usually have a substantial effect on the output of the model, which raises the question of what are the risks of injecting CO_2 underground when these models are used for decision making under these uncertainties. Analysis of uncertainties is needed to quantify their

impact on model predictive capabilities as well as to better understand the potential risks of CO_2 storage. One way to quantify these uncertainties is to combine a detailed model, usually a numerical simulation model, with Monte Carlo (MC) simulation that involves repeated simulations to obtain frequency histograms/distributions of model outputs. However, numerical models are generally computationally expensive for repeated simulations, especially when a single realization of a simulation requires hours or days of CPU time. As an alternative, we can first approximate the detailed model output of interest using polynomial chaos expansion (PCE) with respect to the uncertain parameters, and then use the derived PCE approximation to perform MC simulation.

PCE methods can provide efficient and accurate ways of analyzing uncertain behavior in a complex system.[9–16] These methods mainly fall into two categories: intrusive approaches and non-intrusive ones. Intrusive methods involve substituting the PCE approximations into the governing equations and using a Galerkin technique[17] and a discretization scheme to solve for the coefficients in the expansion.[12] Non-intrusive methods evaluate the coefficients in the expansion using a small number of model simulations and require no manipulation of underlying partial differential equations.[9–11,13–16] Therefore, the latter approaches are easier to implement and generalize to complex systems. For this reason, we focus on non-intrusive PCE methods.

In non-intrusive PCE approaches, the coefficients of the polynomial expansion can be computed by projection,[11,14] which calculates an inner product for each coefficient that involves computing multi-dimensional integrals. However, it is a non-trivial task to compute these integrals as it becomes computationally prohibitive for many system inputs (uncertain parameters) and/or high-order expansion cases. Another way to calculate the coefficients is through solving a linear system that is constructed after evaluating the model on a set of samples from the uncertainty space. Depending on the number and type of points being chosen, the resulting model output surface can be either interpolating these points (i.e., choosing collocation points of the polynomial roots and the same number as coefficients in the expansion),[9,13,16] or minimizing a least squares error (i.e., choosing random sampled points with high probabilities and performing regression).[10,14,15] In the first approach, collocation points come from the combination of the roots of next high-order poly-

nomials, which means that these points are different for different degrees of expansion and the corresponding model output needs to be evaluated at all these points. In other words, if a proper PCE model satisfying a pre-set accuracy is going to be built iteratively by increasing its degree one at a time, the total computation is going to increase rapidly as it not only involves the model evaluations under the current set of collocation points, but also the previous evaluations for low-order collocation points. This explosion of CPU time limits applicability of PCE techniques. Therefore, we have adopted approaches that solve for the expansion coefficients based on a fixed set of random samples with high probabilities. By choosing samples with high probabilities, the approximation values of a model output are expected to have small deviations from the exact values in the most probable area of the uncertain parameters.

PCE based on interpolation of collocation points has been applied to the modeling of CO_2 sequestration by Oladyshkin et al..[16,18] Parametric distributions were assumed to fit the observed data of uncertain parameters in order to utilize the corresponding orthogonal polynomial basis.[16] However, using a hypothesis test, Kopp et al.[19] have shown that the raw data used by Oladyshkin et al.[16] fail to follow a normal or a lognormal distribution. In addition, using collocation points[16] to solve for the expansion coefficients limits application to low-dimensional PCEs. In related work,[18] PCE models were developed without parametric assumptions but the correlations of uncertain parameters were not explicitly addressed.

The contributions of the current paper are in two areas:

1. In the context of PCE methodology, we introduce the use of mixed-integer optimization techniques for systematic selection of the coefficients in the polynomial expansion. In contrast to existing forward selection and backward elimination PCE techniques,[15] our approach captures interactions between low- and high-order terms of the polynomial expansion. As a result, simpler and more accurate surrogates can be built with the same amount of input/output information (detailed model simulations).

2. In the context of CO_2 sequestration, we apply the proposed PCE method to obtain PCE models that relax the earlier assumptions on distributions and account for correlations between

uncertain permeabilities and porosities. Our computations demonstrate that PCE surrogates result in highly accurate approximations of a benchmark problem. In addition, we propose a stochastic optimization model that utilizes the PCE surrogates to identify optimal CO_2 injection rates that maximize CO_2 immobilization under parameter uncertainty while ensuring that caprock pressure stays low.

The remainder of the paper is organized as follows. In the following section, we introduce the mathematical concepts of PCE representation. Then, we describe the method of building PCE surrogate models with the MIP-based best subset selection. Using the PCE model, we are able to incorporate operational variables into the proposed framework and formulate an optimal operational problem. A detailed illustration on a benchmark CO_2 storage modeling problem is presented here. Computational results of the uncertainty quantification and optimal design using these PCEs are presented next. Finally, conclusions are provided in the last section.

Polynomial chaos expansion approximation

Let us assume a physical model $\mathscr{Y} = \text{model}(x)$, where $x = \{x_1, ..., x_M\}^T \in \mathbb{R}^M$, $M \geq 1$ is a vector of parameters (model inputs), $\mathscr{Y} = \{y_1, ..., y_N\}^T \in \mathbb{R}^N$, $N \geq 1$ is a vector of model outputs of interest. If the parameter vector x is uncertain and can be characterized with some probability density function (PDF), it is expected that each model output $y_i \in \mathscr{Y}$ is also a random variable. To keep the notation simple, y_i is denoted as y from this point forward. Assuming a particular model output y has finite variance, then y can be represented by the following polynomial chaos expansion:[20]

$$\begin{aligned} y = P(x) = \alpha_0 B_0 + \sum_{j=1}^{M} \alpha_j B_1(x_j) + \sum_{j=1}^{M} \sum_{k=1}^{j} \alpha_{jk} B_2(x_j, x_k) \\ + \sum_{j=1}^{M} \sum_{k=1}^{j} \sum_{h=1}^{k} \alpha_{jkh} B_3(x_j, x_k, x_h) + ... \end{aligned} \tag{1}$$

where the α's are coefficients, and the B's are multivariate polynomial basis functions that are orthogonal with respect to the joint PDF of x.

In practice, this PCE is truncated at a finite number of basis functions. The number of the terms N_t in the expansion is

$$N_t = \frac{(M+d)!}{M!d!} \tag{2}$$

where M is the number of model inputs (length of x) and d is the degree of the expansion, i.e., the highest degree of the polynomial basis functions.

Orthogonal polynomial basis

The orthogonal polynomials for a single random variable x are defined as follows. Assume $f(x)$ is the PDF of x, and $\{\phi_d = x^d + \text{lower degree terms}, d = 0,1,2,...\}$ are polynomial functions of x. Then, the polynomial ϕ_d is defined to be orthogonal to polynomial ϕ_c if

$$\int_\Omega \phi_d \phi_c f(x)\, dx = 0, \quad d \neq c \tag{3}$$

where Ω is the support of the random variable x.

If x can be characterized with one of the most commonly known PDFs, such as a normal or uniform distribution, then the corresponding orthogonal polynomials can be chosen through the Askey scheme[21] or generalized options.[12] For example, a Hermite polynomial sequence corresponds to a standard normally distributed random variable. For x with an arbitrary PDF, the orthogonal polynomials are generated numerically.[22] These orthogonal polynomials are optimal with respect to the PDF of x, in the sense that the error computed as a difference between the exact value of a model output y and the value approximated by Eq. (1) with $\{B_i = \phi_i, i = 0,...,d\}$ converges to zero exponentially as d increases linearly.[12]

For the case of an input vector x with M random variables, Eq. (3) can be applied to define the multidimensional orthogonal polynomials. If the x_i's are independent of each other, the joint PDF of x is then a product of the marginal distributions of each x_i. Thus, the multidimensional

orthogonal polynomial of degree d, i.e., B_d, can be obtained by multiplying the corresponding individual orthogonal polynomial for each x_i:

$$B_d(x) = \prod_{i=1}^{M} \phi_{m_i}(x_i), \quad \text{where } \sum_{i=1}^{M} m_i = d, \quad \phi_0(x_i) = 1,\ i = 1,...,M \tag{4}$$

In the case of correlated x_i's, an appropriate transformation, such as the Nataf transformation, is needed before Eq. (4) can be applied to generate the orthogonal polynomial functions as discussed next.

Transformation for correlated inputs

The Nataf transformation[23] is a nonlinear transformation that translates correlated random variables to uncorrelated standard normal random variables. This transformation is often used for the cases when the marginal PDF of each random variable is known, but not the joint PDF. The transformation can be broken down into the following two steps. First, the correlated x_i's are transformed to correlated standard normal z_i's by setting their cumulative distribution functions (CDFs) equivalent:

$$z_i = \Phi^{-1}(F(x_i)) \tag{5}$$

where $F(x_i)$ is the marginal CDF of the random variable x_i, and $\Phi^{-1}(\cdot)$ is the inverse of the marginal CDF for z_i. These correlated z_i's are then transformed to uncorrelated ξ_i's:

$$z = L\xi \tag{6}$$

where L is obtained by Cholesky decomposition of the correlation matrix of z.

With the Nataf transformation, the correlated input in x-space is transformed to an independent standard normal space. The model output is then approximated as an expansion of the orthogonal polynomial basis of ξ, i.e., a summation of multiplications of Hermite polynomials.

Coefficient estimation

The next step is to estimate the coefficient vector $\alpha \in \mathbb{R}^{N_t}$ in a truncated expansion $M_d(x)$ at degree d with N_t polynomial basis terms. We have used linear regression to solve for the coefficients. First, N_p points are selected for x and denoted by the set $X = \{x^1, ..., x^{N_p}\}$. Then, the coefficients are computed by solving the following linear system:

$$B\alpha = y \tag{7}$$

where the $N_p \times N_t$ matrix B is formed by evaluating the polynomial basis functions at the N_p selected points and are normalized to avoid ill conditioning due to power operations. The vector $y = \{y^1, ..., y^{N_p}\}^T$ contains N_p values of a model output y, which are obtained by running the detailed simulation with the selected points as input values. Generally, N_p is greater than N_t (e.g., $N_p = 2N_t$ is suggested by Hosder et al.[24]) so that Eq. (7) is an over-determined linear system. Then, α is solved for by minimizing the training error $||B\alpha - y||_2^2$, which gives $\alpha = (B^T B)^{-1} B^T y$ provided that $(B^T B)^{-1}$ exists.

The choice of X can be based on random sampling, such as Latin Hypercube Sampling (LHS), from the joint PDF of x. These realizations/points are sorted in the order of descending joint probability densities. Those with higher probability densities are finally selected for the model evaluations to solve for α.

Goodness of fit measures

Now that we have a degree-d PCE approximation $M_{d,X}$ built with a particular choice of X, we can compute errors to see how well $M_{d,X}$ performs. In order to estimate the training error, we employ the most frequently used statistical measure R^2. We also use the leave-one-out cross validation error Q^2 to keep track of overfitting:

$$Q^2 = 1 - \frac{\frac{1}{N_p}\sum_{i=1}^{N_p}(y^i - M_{d,X\setminus\{x^i\}}(x^i))^2}{\sigma^2(y)}$$

In this formula, $M_{d,X\setminus\{x^i\}}$ is a PCE approximation built with N_p-1 points in the set X. One point x_i is left out, and the residual is calculated. Although Q^2 looks similar to R^2, the calculation of Q^2 involves constructing N_p PCE approximations $\{M_{d,X\setminus\{x^1\}},\ldots,M_{d,X\setminus\{x^{N_p}\}}\}$ to compute the residuals. Cross validation techniques are sensitive to overfitting since Q^2 may decrease if excessive higher-degree terms are added to the expansion.

Building PCE approximations

Building an appropriate PCE for a model output is an iterative procedure with the above steps. We start from a degree-d PCE ($d=0$ at iteration 1), compute the coefficients by regression, and calculate R^2 and Q^2. If, at a given iteration, the error (1-Q^2) starts to increase and/or approaches zero (e.g., 0.01), the expansion of degree d is considered as a best fit of the current choice of X. Otherwise, we add one-degree higher polynomial terms to the expansion of the current iteration.

From Eq. (2), we can see that the number of polynomial terms N_t grows very fast when the number of uncertain inputs M and/or the degree of expansion d increase. For example, $N_t=8008$ with $d=6,M=10$. As at least as many as N_t model simulations are needed to get y for regression, a large value of N_t conflicts with our initial goal of reducing computational effort. To cope with this issue, forward and backward stepwise regression techniques[25] can be used, for example as proposed by Blatman and Sudret.[15]

The main idea of the stepwise regression is that, instead of adding all the one-degree higher terms to the expansion of degree d at each iteration, the degree-$(d+1)$ expansion is enriched by adding one polynomial term of degree $d+1$ at each step. If the added term helps to increase model accuracy significantly, it is retained in the degree-$(d+1)$ expansion; otherwise it is discarded. The expansion is updated with or without the addition of this new term. This step is repeated until all $N_t(d+1)-N_t(d)$ degree-$(d+1)$ basis functions are screened. The benefit of this method is that the number of the basis terms is minimized by neglecting those terms with very small α's in the original full PCE expansion; therefore, the number of model simulations is kept as small as

possible. However, one issue of this stepwise approach is that it ignores the synergistic effect of basis functions. For example, basis function B_k discarded at iteration k may become significant in future iterations after the addition of new basis functions. Since the stepwise scheme does not allow the reentry of previous discarded terms, the resulting truncated polynomial expansion may not be the best subset of the basis set. This issue is addressed next.

Best subset selection based on MIP

In our work, we propose a new method of best subset selection based on mixed-integer programming (MIP) to build PCE surrogate models. The formulation is as follows:

$$\begin{aligned} \min \quad & \|B\alpha - y\|_2^2 \\ \text{s.t.} \quad & \sum_{j \in \mathscr{B}} w_j = T \\ & \alpha^L w_j \leq \alpha_j \leq \alpha^U w_j, \ \forall j \in \mathscr{B} \\ & w_j = \{0,1\}, \ \forall j \in \mathscr{B}, \end{aligned}$$

where $\mathscr{B}$ is the set of basis functions up to degree d, α^L and α^U are lower and upper bound of the coefficient α_j, T is the number of terms that is allowable in the polynomial expansion. If the integer variable w_j is one, it indicates that the basis function B_j is allowed to be present in the expansion. If w_j is zero, the basis function B_j is not permitted in the expansion and α_j is forced to zero by the second constraint. By the first constraint, the number of terms in the expansion will be T. For each given T, the above MIP is solved by minimizing the training error $\|B\alpha - y\|_2^2$. The value of the parameter T is chosen by minimizing the cross validation error, i.e., by minimizing $1 - Q^2$.

This MIP formulation preserves the advantages of the stepwise regression, i.e., keeps the number of terms in the expansion small as long as the tuning parameter T is chosen to be small. In fact, if T is set to be N_t, the solution of this MIP also recovers the full classic polynomial chaos

expansion. This flexibility of manipulating the number of terms enables us to construct a general PCE model that is either a full expansion or an expansion of a subset of basis functions. In addition, the optimal set of the basis functions obtained by solving the MIP problem is based on a complete search over the set $\mathscr{B}$. This gives us the best subset that considers the synergistic effects of basis functions. In our computational results, we do observe that the MIP-based method results in smaller subsets of basis functions in comparison to the stepwise method based on forward selection and backward elimination.

The approach of constructing PCE models iteratively is outlined in Figure 1. In comparison to existing PCE methods, the forward and/or backward stepwise regression is replaced with the best subset search using the MIP formulation.

Uncertainty analysis and injection optimization with PCEs

Once we have PCE approximations of the original numerical model as a function of the uncertain parameters, we can then perform uncertainty analysis with the PCE models. The effect of parameter uncertainties can be quantified using MC simulation. Generally, in MC simulation, values of the uncertain parameters are randomly sampled from their respective PDFs, if parameters are independent, or from their joint PDF, if correlated. A LHS technique is used to increase the likelihood that the space of the uncertain parameters is covered sufficiently by these sampled points. By substituting the random values of uncertain parameters into the PCE approximation, the corresponding y values are obtained. Statistical analysis can then be performed for this specific model output y for uncertainty analysis.

In the work by Oladyshkin et al.,[16] design variables are integrated into the PCE approximations, i.e., each design variable or parameter is assigned with a hypothetical distribution and treated as the uncertainty input (parameters). Consequently, the model output becomes some polynomial function of the model input and the design variables. We have extended this integrative idea to formulate an optimal design problem based on the PCEs. In other words, the implicit relationships

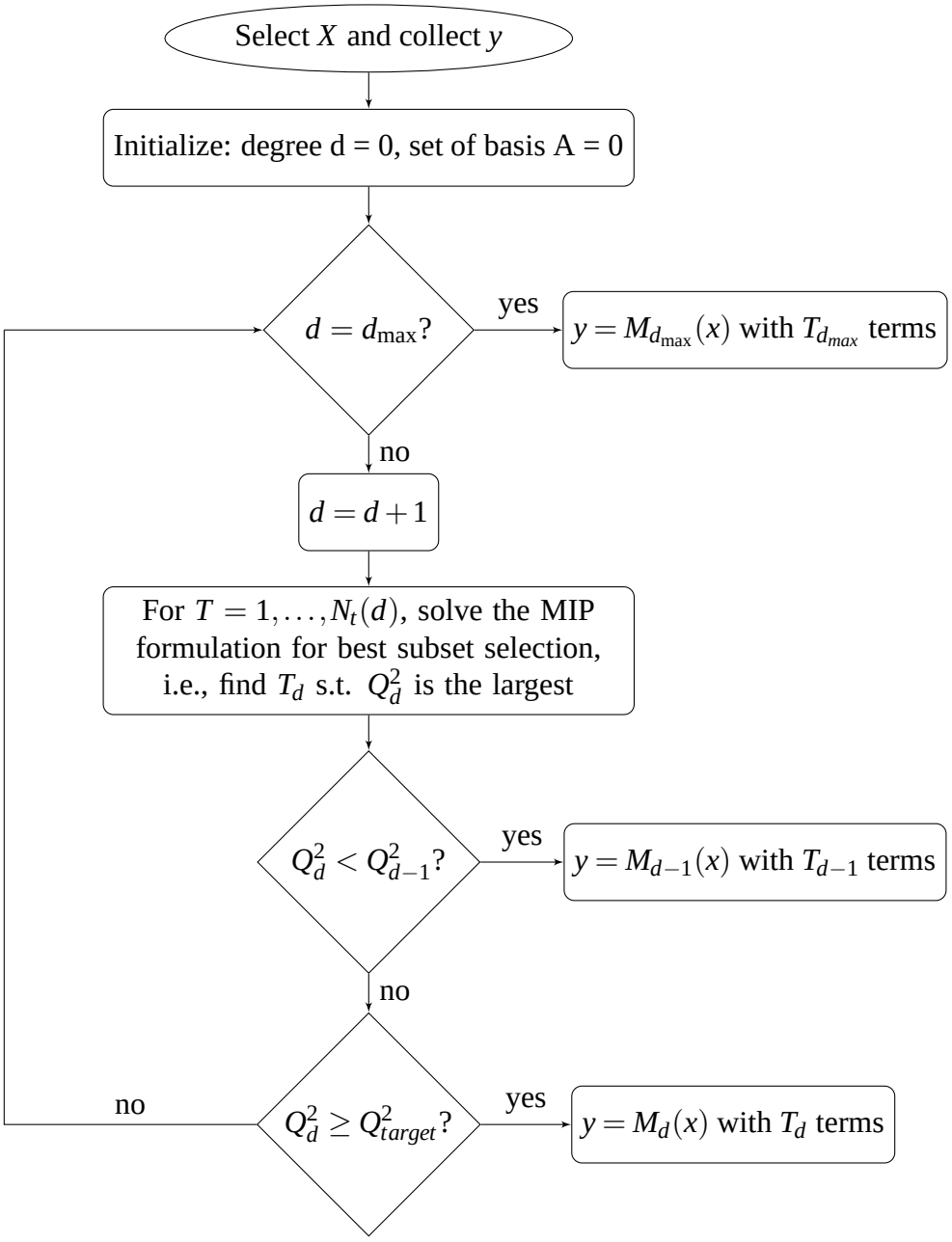

Figure 1: Flowchart of the MIP-based PCE method

between the objective/constraints and the model input and design variables hidden under the black-box numerical simulator are explicitly approximated by the polynomial expansions. The resulting optimization problem under uncertainty takes the following form:

$$\begin{aligned} \min_{z} \quad & f(z,\theta) \\ \text{s.t.} \quad & g(z,\theta) \leq 0 \end{aligned}$$

where z is a vector of design variables, and θ is a vector of uncertain parameters. The objective $f(z,\theta)$ and the constraints $g_i(z,\theta)$ are the polynomial approximations for some model outputs of interest. This optimization problem can be approached with various techniques, including different models of uncertainty that account for worst-case or average-case scenarios.[26]

Case study

We apply the adaptive PCE method to a benchmark CO_2 sequestration problem. As CO_2 is injected into a saline aquifer, it spreads out around the injection well, pushed by the high injection pressure, displacing brine water to its irreducible saturation.[3] The upward movement of CO_2 due to buoyancy stops when CO_2 reaches an overlying impermeable caprock, where CO_2 is physically trapped (*structural* trapping). After injection stops, the CO_2 plume continues to move slowly laterally due to the formation fluid's natural flow, and gradually gets trapped in the pore space, where the plume is eventually entirely immobilized by this *residual* trapping. Solubility trapping and mineral trapping are much slower trapping mechanisms and are not considered as significant in a short simulation period of the nature considered in this case study (two years).

One possible leaking scenario occurs during the lateral movement of the CO_2 plume when CO_2 is not entirely immobilized yet. Leakage happens when the plume encounters faults and/or broken wells. Kumar[27] determined well settings to maximize residual trapping and argued that structural trapping is not desirable before CO_2 is entirely immobilized because any loss of the seal integrity

due to fractures, faults, or broken intercepting wells will cause the mobile CO_2 to leak. Residual trapping is preferred because it restricts the lateral movement of CO_2 and eventually immobilizes the entire CO_2 plume.

Figure 2 shows a simplified sketch of CO_2 injected into a saline aquifer. The modeling of CO_2 injection into a geological formation mainly involves mass balances and two-phase flow dynamics. The following governing equations need to be solved for the dynamics of injected CO_2 plume:

$$\frac{\partial \phi \rho_i S_i}{\partial t} + \nabla \cdot (\rho_i v_i) = Q_i \tag{8}$$

$$v_i = \frac{K k_i}{\mu_i}(\nabla p + \rho_i g \nabla z) \tag{9}$$

$$\sum S_i = 1 \tag{10}$$

where $i = \{CO_2\text{-rich gas phase, brine water phase}\}$; ϕ is the porosity, ρ_i is the density, S_i is the fluid saturation, v_i is the fluid velocity, Q_i is the source (sink) term, K and k_i are the intrinsic and relative permeabilities, p is the fluid pressure, and μ_i is the viscosity.

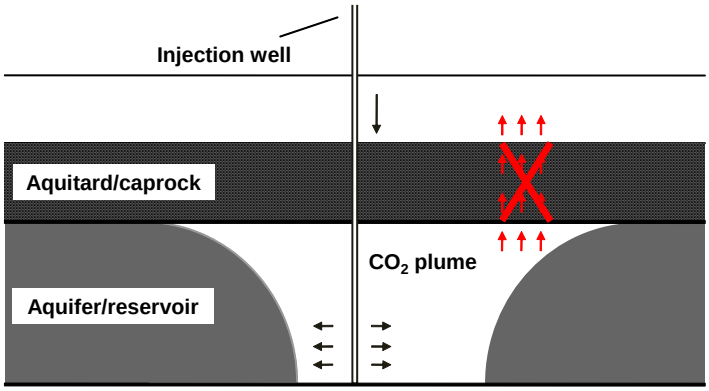

Figure 2: CO_2 injection into a saline aquifer

Problem statement

We consider a CO_2 injection benchmark that has been simulated using TOUGH2 (Problem No. 4).[28] TOUGH2 is a numerical simulator for modeling non-isothermal multiphase flow in fractured

porous media, and has been under development since 1980. In this particular benchmark problem, CO_2 is injected into a saline aquifer through a horizontal well. The injection rate is about 10 kiloton every year. In a 2D vertical section, geological formations are composed of alternating permeable and impermeable layers. The simulation by TOUGH2 provides results such as the pressure of the injection layer and the CO_2 distribution profile (e.g., mass and gas saturation) along all the formation layers.

We are mainly interested in quantifying the impact of the uncertain parameters such as porosity and permeability on the model outputs. Therefore, the model outputs will first be approximated as polynomial functions of porosity and permeability, allowing us to perform MC simulation with the PCE approximation later.

Uncertainty characterization of model input

Instead of using hypothetical parametric distributions such as normals and log-normals, we have used the raw data from field observations. As an illustration, the National Petroleum Council database for over one thousand reservoirs in U.S.[29] is used. The joint distribution as well as the marginal PDFs of the input parameters (porosity and permeability) are plotted in Figure 3. It can be seen from the two marginal histograms that neither of the two distributions follows a standard parametric distribution. This means that the orthogonal polynomials known for standard distributions (e.g., Askey scheme polynomials) cannot be applied as basis functions directly in the expansion. Further, the value of the correlation coefficient ρ ($= 0.8$) between the two parameters implies a strong correlation between the two inputs. Thus, the Nataf transformation is needed before the polynomial basis functions can be generated with Eq. (4). These two correlated parameters are transformed into two uncorrelated standard normals with Eq. (5) and Eq. (6).

Construction of PCEs

Once we have the standard normal random variables, we can utilize the series of Hermite polynomials known as orthogonal polynomials for a single standard normal variable. For the case of two

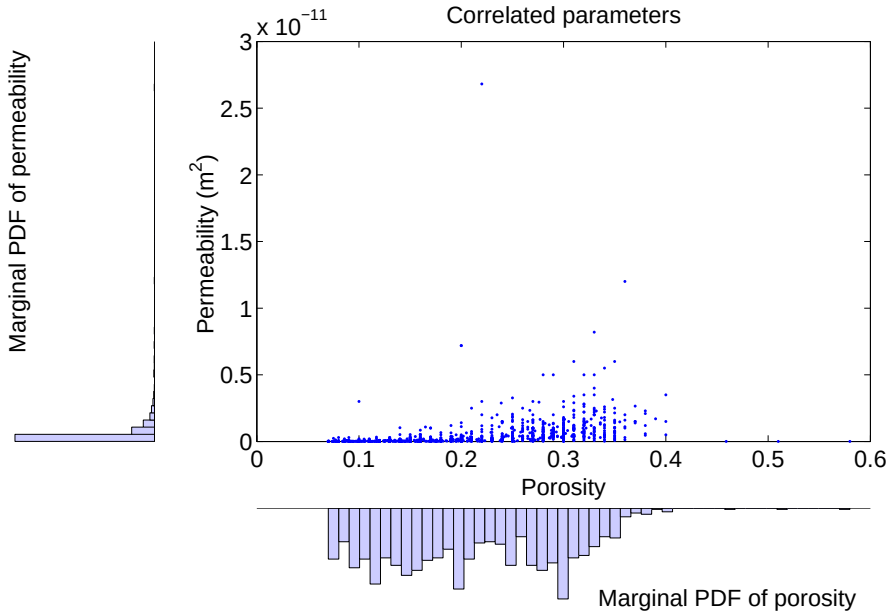

Figure 3: Correlated input parameters

random variables, the polynomial basis functions are as follows:

$$
\begin{aligned}
B_0(\xi) &= \varphi_0(\xi_1)\varphi_0(\xi_2) &&= 1 \\
B_1(\xi) &= \varphi_1(\xi_1)\varphi_0(\xi_2) &&= \xi_1 \\
B_1(\xi) &= \varphi_0(\xi_1)\varphi_1(\xi_2) &&= \xi_2 \\
B_2(\xi) &= \varphi_2(\xi_1)\varphi_0(\xi_2) &&= \xi_1^2 - 1 \\
B_2(\xi) &= \varphi_1(\xi_1)\varphi_1(\xi_2) &&= \xi_1\xi_2 \\
B_2(\xi) &= \varphi_0(\xi_1)\varphi_2(\xi_2) &&= \xi_2^2 - 1 \\
&\vdots
\end{aligned}
$$

The model output is then a PCE approximation in terms of ξ with the coefficients (α's) left as unknowns. To solve for Eq. (7), we need more than N_t samples. Here, a fixed design of experiments, i.e., 100 LHS design for ξ is chosen. B is then evaluated with these random samples. The model evaluation vector y is obtained by first performing the reverse of the Nataf transformation

to translate the random ξ back to the corresponding x values, and running the model with these x samples. With all these, the best subset selection method using the MIP formulation is solved to find a PCE approximation for a model output satisfying relatively large R^2 and Q^2. Figure 4 and Figure 5 are two examples of the polynomial surface fitted to 100 random samples with $R^2 = 0.98$ and $Q^2 = 0.98$. The response surface of the fraction of CO_2 in the caprock fits a third-order expansion, while the response surface of the fraction of gas-phase CO_2 in the caprock fits a fourth-order expansion. The amount of CO_2 in the caprock tends to decrease when the porosity and the permeability of the aquifer become larger (larger ξ_1 and ξ_2). This is because more CO_2 would stay in the aquifer if the porous space in the aquifer is larger (larger porosity) and the flow movement is easier (larger permeability).

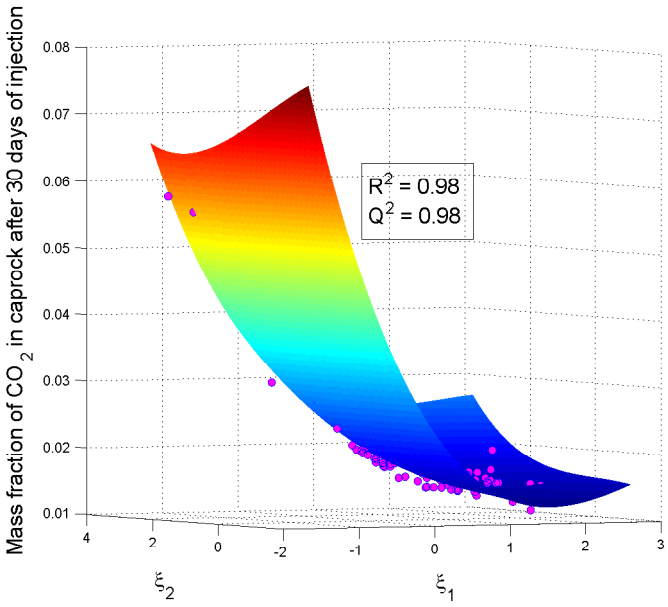

Figure 4: PCE model example 1, fitting a third order expansion

The pressure of the injection layer (i.e., saline aquifer) is going to be disturbed by the injection activity. The pressure buildup due to injection is of particular interest as we want to know whether the buildup would fracture the upper layer seal (i.e., caprock) during injection, under uncertain geological parameters. Another selected model output is gas saturation, which is defined the fraction of the pore space of the geological formation that is occupied by CO_2. A contour map of gas saturation for the 2D domain gives us an idea of how the CO_2 plume is distributed along all formations

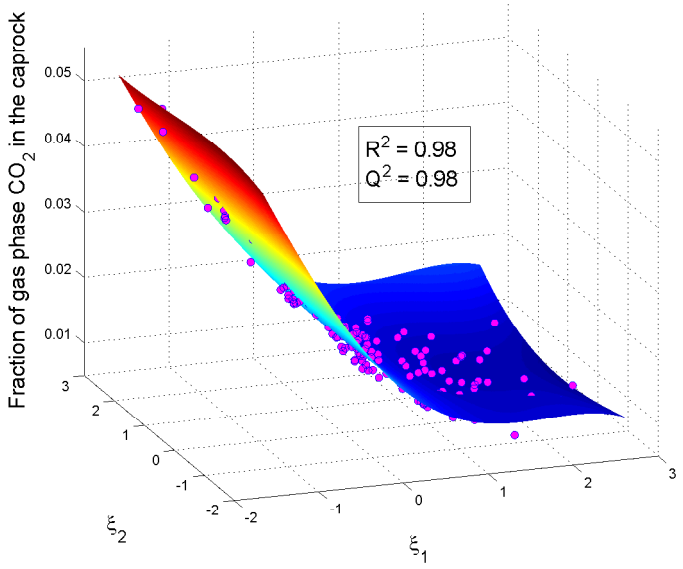

Figure 5: PCE model example 2, fitting a fourth order expansion

after injection for a certain duration. These two model outputs are functions of space and time. The space domain (half space) of the problem is discretized with a mesh of 34×29 (986) grid blocks. The mesh grid is not uniform but refined near the injection well. For each block at a specific time, we construct PCE models for the two output variables, i.e, the coefficients in the PCE model are dependent on space and time. Although this process involves generating a large number of PCE models, the overall computation time is relatively small. In addition, the availability of these PCE models facilitates computation of probabilities for CO_2 leakage as a function of space and time.

Figure 6 and Figure 8 are obtained by running one deterministic TOUGH2 simulation and present the contour map of pressure and gas saturation after 30 days of injection. A pressure buildup is clearly observed around the injection well in Figure 6. In the gas saturation map, the CO_2 plume in the half space is observed. Figure 7 and Figure 9 are obtained by substituting the same model input values used in TOUGH2 simulation into the 986 PCE approximations. Although it takes 100 TOUGH2 simulations to first construct the PCE models, the advantage of the PCE method is in the gain of speedups in conjunction with MC simulation, compared to applying MC simulation with the TOUGH2 simulator directly.

In the two plots for pressure, the overall contours look similar. The main feature of pressure

transition in Figure 6 has been differentiated by the PCE simulation in Figure 7. Notice that there are some wigglings on the boundaries of the contours. This may be due to the oscillating feature of polynomial terms in the expansion. For the gas saturation map, 89 out of 239 constructed gas saturation PCEs show small negative values. These negative values are numerical errors due to the approximations using the polynomial terms. The errors are on the order of 6% of the true values on average. Based on the locations where these small negative values are observed, the negative gas saturation is set to zero since the CO_2 plume will not have arrived at these locations after 30 days of injection. The resulting gas saturation map looks almost identical to the map obtained using TOUGH2.

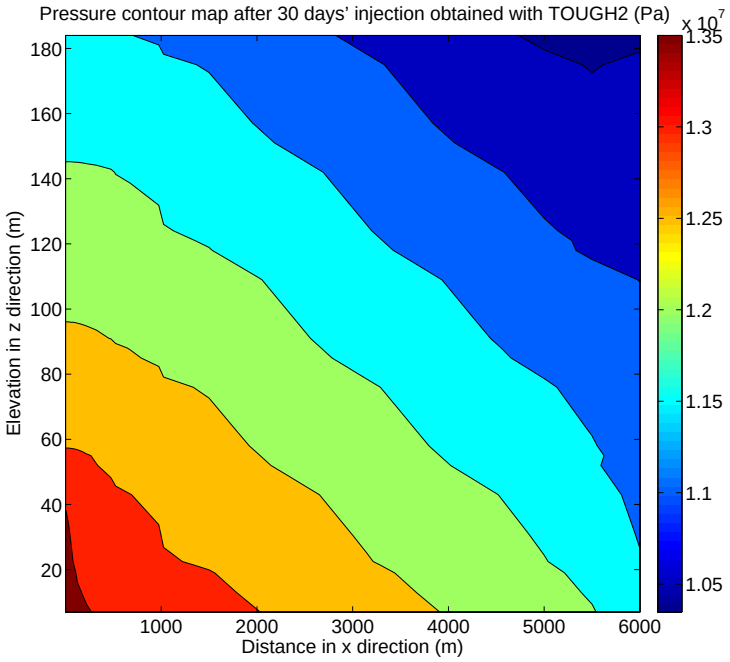

Figure 6: Pressure contour map with TOUGH2 simulation

For comparison, we also implemented the forward and backward stepwise regression method to construct PCE models that achieve a similar Q^2 ($\simeq 0.98$) for the same outputs. As mentioned earlier, the stepwise regression method does not allow the re-entry of previously discarded basis terms, which however might be selected by the MIP-based method. As a result, the subset selected by stepwise regression and even the number of subsets are not necessarily the same as the

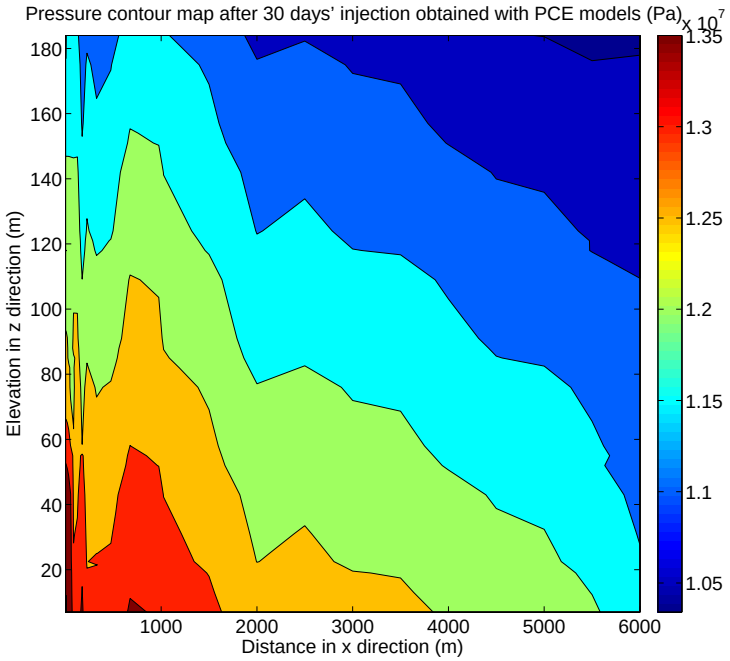

Figure 7: Pressure contour map with PCE approximation

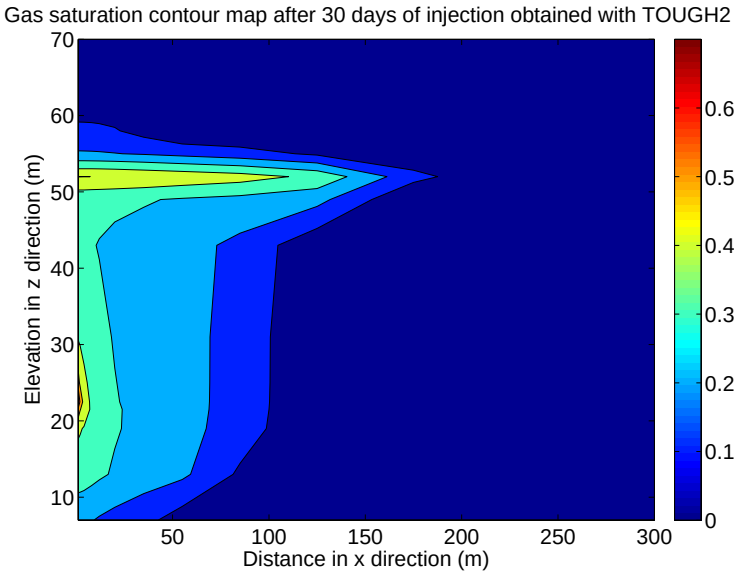

Figure 8: Gas saturation contour map obtained with TOUGH2 simulation (gas saturation is zero beyond the scope of the figure)

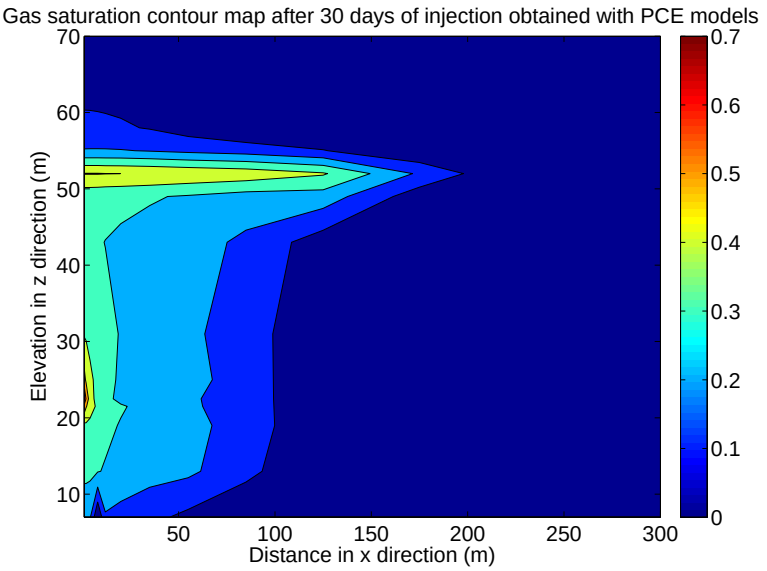

Figure 9: Gas saturation contour map with PCE approximation (gas saturation is zero beyond the scope of the figure)

ones selected by the MIP-based method. From Table 1, we can see that, among the PCE models constructed for the 904 model outputs using both by stepwise regression and MIP-based selection, 81.37% of the models end up with smaller or equivalent degree of expansion when the MIP-based method is used. In addition, all of these lower-degree expansion models require a smaller or equivalent number of terms in the expansions when the MIP-based selection is used. Although 18.63% of the models are in higher-degree expansions if constructed with the MIP-based method, the number of terms in the expansion are still smaller most of the times. We conclude that 93.9% of the time the MIP-based method performs better than forward and backward stepwise regression in the sense that it yields simpler models based on the same number of simulations. Clearly then, this approach would avoid overfitting and require fewer simulations to obtain equally accurate PCE approximations.

Table 1: Performance comparison of MIP-based subset selection and stepwise regression

	$\%\left(\text{Terms}_{\text{MIP}} \leq \text{Terms}_{\text{stepwise}}\right)$	$\%\left(\text{Terms}_{\text{MIP}} > \text{Terms}_{\text{stepwise}}\right)$
$\%\left(\text{Degree}_{\text{MIP}} \leq \text{Degree}_{\text{stepwise}}\right)$	81.37	0
$\%\left(\text{Degree}_{\text{MIP}} > \text{Degree}_{\text{stepwise}}\right)$	12.53	6.10

Correlated sampling for MC simulation

According to the Dvoretzky-Kiefer-Wolfowitz inequality,[32] with probability at least $1-\alpha$,

$$\sup_{t\in\mathbb{R}} |\hat{F}_n(t) - F(t)| \leq \sqrt{\frac{1}{2n}\log\frac{2}{\alpha}},$$

where $\hat{F}_n(\cdot)$ is the empirical CDF based on n samples and $F(\cdot)$ is the true CDF. Therefore, for a 99% confidence band ($\alpha = 0.01$) of $\hat{F}_n$, the maximum error between $\hat{F}_n$ and F is at most 0.05 if the sample size n is larger than 1000. For this reason, we have chosen 1000 samples for Monte Carlo simulation.

The bottom plot in Figure 10 represents the original dataset for porosity and permeability. The top one shows 1000 pairs of correlated random samples used for Monte Carlo simulation. These plots are generated through an inverse of the Nataf transformation, where the empirical distribution functions (i.e., empirical CDFs) are used to be the estimator of the true CDFs for porosity and permeability. From Figure 10, we can see that the simulated samples reproduce the original dataset very well. These 1000 pairs of samples are substituted into the derived PCE approximations for the corresponding y values, and the resulting distributions are shown in the next section.

Results of uncertainty analysis

For the benchmark problem here, it takes about 15 minutes to perform one TOUGH2 simulation, which is still moderate for MC simulation. Therefore, we also perform 1000 simulations with the TOUGH2 model to obtain distributions of model outputs that we can use for validating the results of the PCE models.

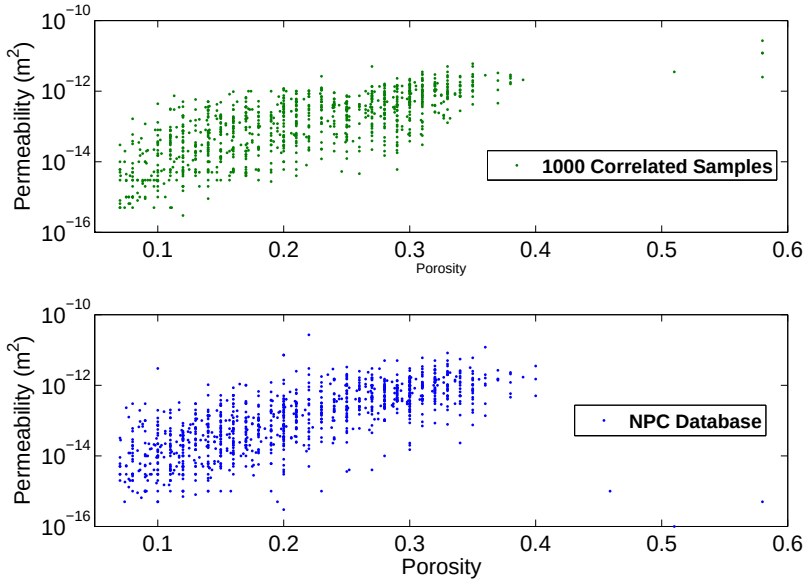

Figure 10: 1000 simulated correlated samples (top) in comparison to NPC database (bottom)

Figure 11 shows the CDFs for the mass fraction of CO_2 in the caprock after 30 days of injection. The distribution obtained with a third-order polynomial expansion (dashed red) is very close to the distribution (solid blue) obtained by running TOUGH2. Similarly, we can see a good match for the mass fraction of gas-phase CO_2 in the caprock in Figure 12. Note that the number of numerical simulations is much smaller using the PCE method. In our study, the choice of random samples is fixed to 100. However, it can be kept as small as six in order to solve the linear system (Eq. (7)). The computation for finding a PCE approximation using the adaptive PCE method takes a few seconds. The computation of MC simulation with PCE models takes about three seconds, which is also negligible. Therefore, the overall computation time of using the adaptive PCE method for MC simulation is mainly due to the 100 numerical simulations at sampled points. This time is about 10% of the time for running 1000 MC simulations with TOUGH2 directly.

Figure 13 is a contour map with the average values of the pressure for the 2D domain of the injection problem. This map is generated by taking an average of 1000 TOUGH2 simulation results. The map shows how the hydrostatic pressure field is disturbed by the injection activity in an average sense. The brighter area near the injection well represents high values of pressure. The pressure differential is primary due to pressure buildup during CO_2 injection which could lead to

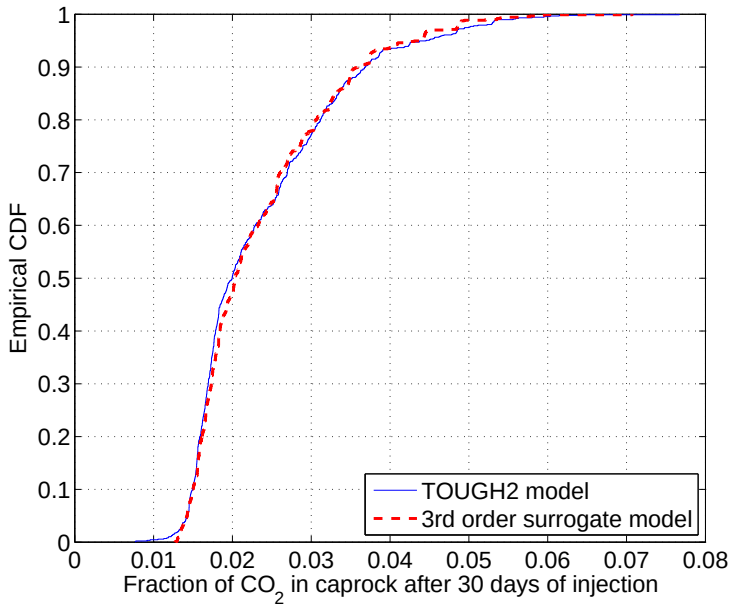

Figure 11: Fraction of CO_2 in the caprock after 30 days of injection

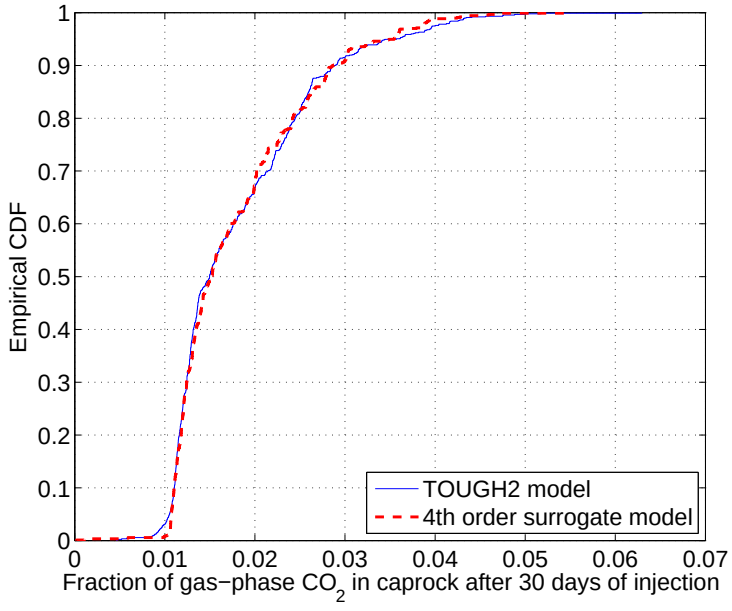

Figure 12: Fraction of gas-phase CO_2 in the caprock after 30 days of injection

higher pressure than the original hydrostatic pressure. We have a very similar map obtained with the polynomial model. Again, we see some differences in the absolute values for some locations.

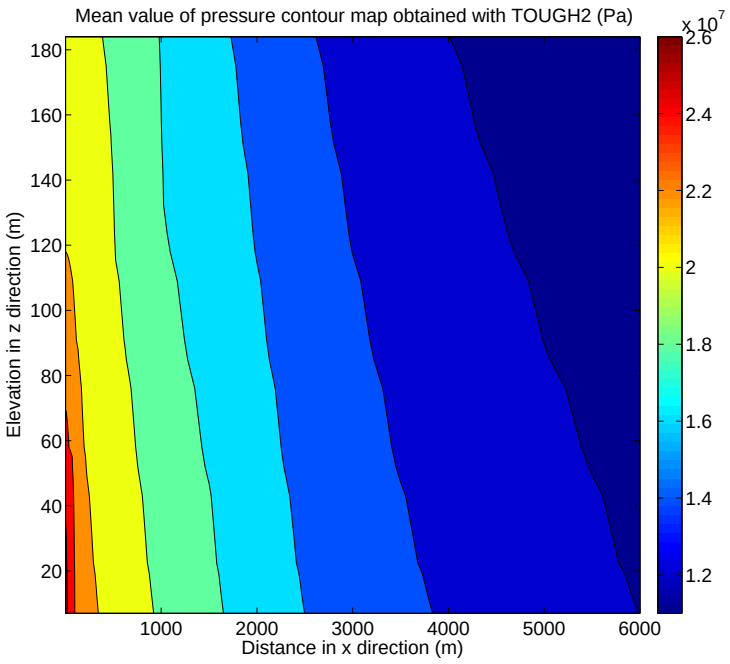

Figure 13: Pressure contour map in average obtained with TOUGH2 simulation

Figure 15 is the contour map of the gas saturation after 30 days of injection obtained with TOUGH2. Almost exactly the same map is obtained with the PCE approximation, see Figure 16.

We are also interested in knowing the pressure in the caprock since we want to know whether the pressure would fracture the caprock during injection. For each simulation, the maximum pressure along the caprock is found and a distribution is then derived for this maximum pressure. If the pressure limit that would break the caprock formation is known, the probability of overpressure can be found in this distribution. For instance, if the fracture pressure is about 500 bar, from Figure 17, the overpressure probability is about 0.001.

Optimal injection rate under uncertainty

To reduce the risk of leakage, we would like to increase the amount of CO_2 trapped by residual trapping as much as possible, while making sure that the buildup pressure during injection does

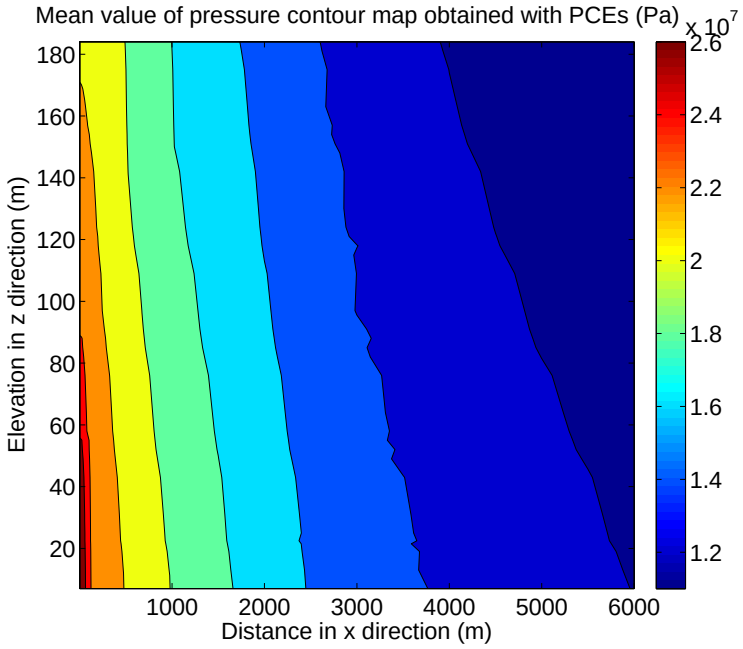

Figure 14: Pressure contour map in average with PCE approximation

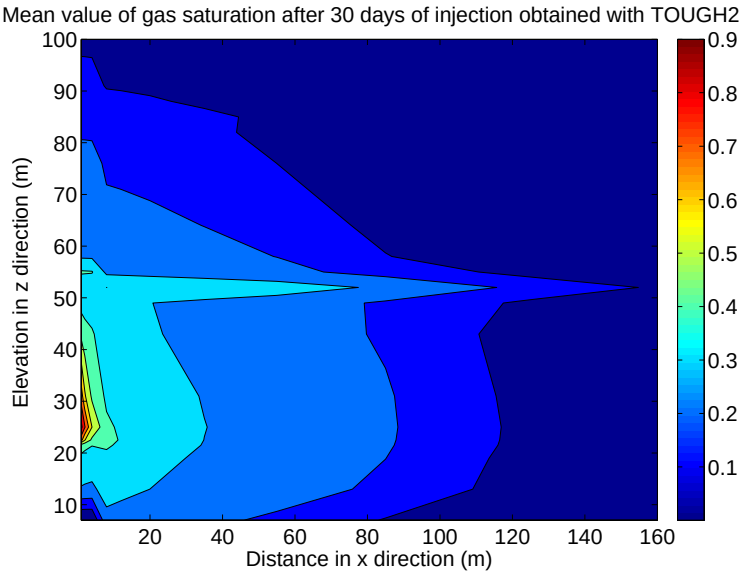

Figure 15: Gas saturation contour map in average obtained with TOUGH2 simulation (gas saturation is zero beyond the scope of the figure)

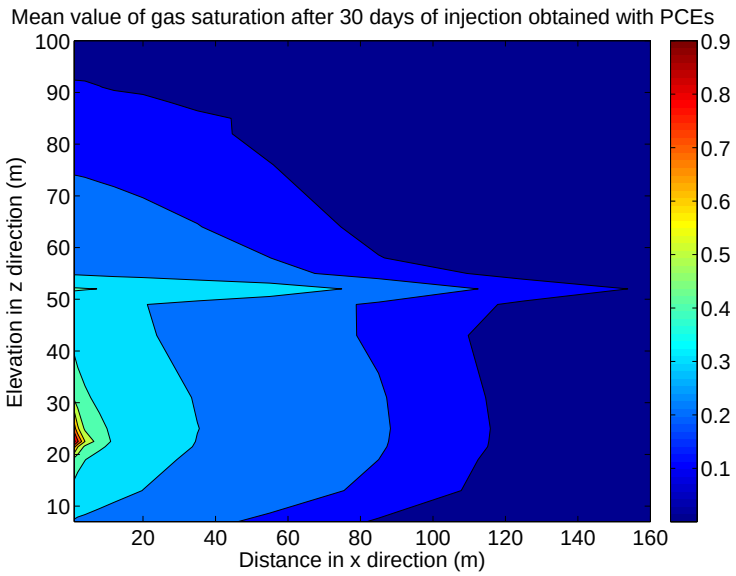

Figure 16: Gas saturation contour map in average obtained with PCE approximation (gas saturation is zero beyond the scope of the figure)

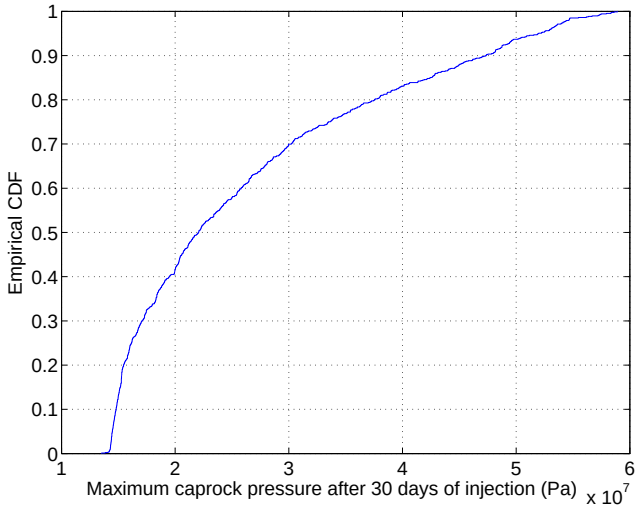

Figure 17: Maximum pressure distribution

not fracture the overlying seal. An optimization problem is formulated for this purpose based on the PCE approximations including the operational variable (the injection rate). In our study, a uniform random variable is used to represent the injection rate during the PCE development step. The reason of choosing a uniform distribution for the design variable is that we assume no preference and place equal weights for values in the design interval. The mean of the uniform distribution equals the value of the injection rate in the original 2D simulation, and the random variable varies between half of the mean and 1.5 times the mean. The polynomial basis for the uniform random variable as well as the two transformed normal variables are then used to get the PCE approximations of model outputs. The optimization problem takes following form:

$$\begin{aligned} \max_{z} \quad & \mathbb{E}_{\xi_1,\xi_2\in\Omega}\left(\sum_{k=1}^{986} S_{g_k}(z,\xi_1,\xi_2)\right) \\ \text{s.t.} \quad & p_k(z,\xi_1,\xi_2) \le p_{\text{limit}}, \qquad k=1,\ldots,986 \end{aligned} \tag{P1}$$

where ξ_1 and ξ_2 are the transformed random parameters for porosity and permeability, z is the operational variable, the injection rate, $\mathbb{E}(\cdot)$ is the expectation function with respect to uncertain ξ_1 and ξ_2, Ω is the support of the uncertain parameters, $S_{g_k}(z,\xi_1,\xi_2)$ and $p_k(z,\xi_1,\xi_2)$ are the polynomial functions for gas saturation and caprock pressure at each grid block k, and p_{limit} is the upper bound for the maximum caprock pressure, which is set to 4×10^7 pa when solving this optimization problem.

The objective is to maximize the expectation of the residual trapping which immobilizes the CO_2 plume, and therefore the proportion of the mobile plume is minimized so that the probability of the mobile plume encountering faults and/or broken wells is reduced. Gas saturation is selected as the measure of residual trapping. Because of pressure buildup during injection, we are concerned about whether the increase of the pressure may break the upper formation. So the maximum caprock pressure is constrained by an upper bound. An optimal injection rate z is determined to maximize the gas saturation (minimizes the chance of leakage) under uncertain parameters ξ_1 and ξ_2, constrained by the pressure limit. This one-stage stochastic problem is solved using a scenario-

based approach, i.e., under 1000 realizations of uncertain parameters. Problem (P1) now becomes:

$$\begin{aligned} \max_{z} \quad & \frac{1}{1000}\sum_{j=1}^{1000}\left(\sum_{k=1}^{986} S_{g_k}(z,\xi_1^j,\xi_2^j)\right) \\ \text{s.t.} \quad & p_k(z,\xi_1^j,\xi_2^j) \le p_{\text{limit}}, \qquad k=1,\ldots,986; \quad j=1,\ldots,1000 \end{aligned} \tag{P2}$$

In this formulation, the expectation $\mathbb{E}(\cdot)$ has been replaced by the average of the summation of 1000 realizations. As the uncertain parameter is revealed (realized), we also solved the following problem: $\forall$ realizations $j \in \{1,\ldots,1000\}$,

$$\begin{aligned} \max_{z} \quad & \sum_{k=1}^{986} S_{g_k}(z,\xi_1^j,\xi_2^j) \\ \text{s.t.} \quad & p_k(z,\xi_1^j,\xi_2^j) \le p_{\text{limit}}, \qquad k=1,\ldots,986 \end{aligned} \tag{P3}$$

Because of the polynomial features (high nonlinearity thus nonconvex) in both the objective function and constraints, we have used the deterministic global optimization solver BARON to solve the problem to global optimality.[30,31]

Figure 18 shows the optimal injection rates from solving (P3). The optimal injection rate that maximizes the expectation of residual in (P2) is shown at the valley of the top left plot, i.e., the lower bound. From Figure 18, we find that the optimal injection rate reaches the maximum possible value that it can take (1.5 times the mean) in most scenarios except when the permeability and porosity are relatively small. This is expected because, in small porosity and permeability areas, the ability of CO_2 to move is small. As a result, CO_2 injected into such areas tends to result in large pressure buildup. Thus, the pressure constraint is more likely to be active and restrict the optimal injection rate for cases of small porosity and permeability. In comparison to the results of the previous simulation, we see from Figure 17 that, if the injection rate is fixed, the probability of the maximum caprock pressure to exceed 4×10^7 pa is 0.17. This is a quite large failure probability from the perspective of safety. The optimal injection rate shown at the bottom part of the plot is a robust operation that is feasible over the entire uncertain domain. Thus, we conclude that the

probability of overpressure under optimal injection rate is asymptotically zero.

The optimal injection rate under uncertainty was obtained from the formulation (P2) using PCE models constructed after 30 days of injection. It is worthwhile to investigate how this optimal injection rate changes over longer time periods, such as one year and two years. The numerical studies for one and two years of injection show that the optimal injection rate remains unchanged, i.e., at its lower bound. The optimal injection rate under uncertainty hits its lower bound at 30 days of injection, which means the pressure constraint is active and restricts the injection rate. Clearly, then, for longer time periods of injection, the pressure constraint will remain active as the pressure buildup during injection becomes larger as times goes by. Therefore, the optimal injection rate that is feasible under the domain of uncertain parameters remains the same under the longer time horizons.

Conclusions

In this work, we have proposed a MIP-based best subset selection method to iteratively build polynomial chaos expansion models for the numerical simulation of CO_2 geological sequestration. The particular PCE method is able to capture synergistic effects between low- and high-order polynomial terms, thus providing high accuracy and computational efficiency. In our study, correlated uncertain parameters are considered without assumptions of parametric distributions, thereby reducing the error introduced by subjectively fitting raw data to parametric distributions. The response surface of model outputs obtained with the PCE surrogate models match well those obtained with detailed simulations with TOUGH2.

We further utilized the PCE models for uncertainty quantification and optimal operation. In uncertainty analysis, the probability distributions from Monte Carlo simulation with PCE approximations are very close to the true distribution functions (those obtained with Monte Carlo simulation using TOUGH2). We gain orders of magnitude speed-up by using the PCE models. In the optimal operational problem, we have provided a rigorous way for determining the injection rate

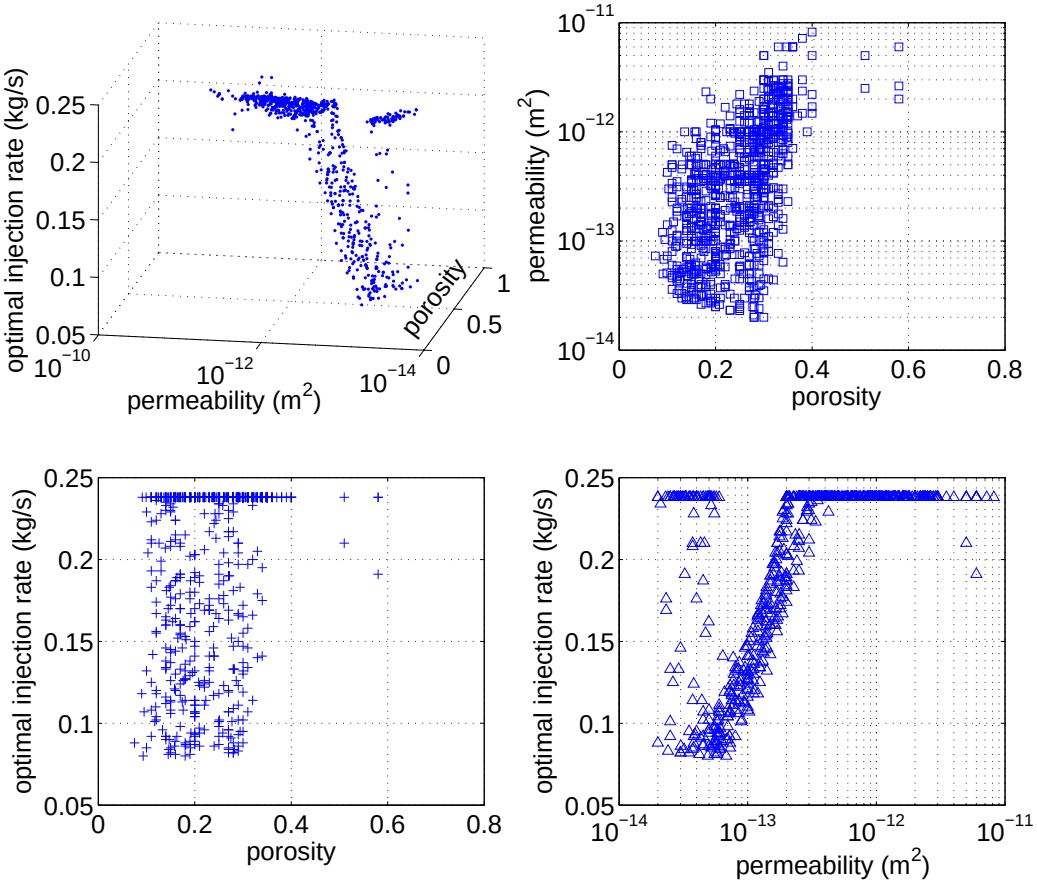

Figure 18: Optimal injection rates with respect to uncertain porosity and permeability: 3D plots of optimal injection rates under 1000 realizations of uncertain parameters (top left); Scatter plot of 1000 realizations of uncertain parameters (top right); Optimal injection rates vs. porosity (bottom left); Optimal injection rates vs. permeability (bottom right)

under uncertainty with minimum leakage risks.

Acknowledgement

As part of the National Energy Technology Laboratory's Regional University Alliance (NETL-RUA), a collaborative initiative of the NETL, this technical effort was performed under the RES contract DE-FE0004000.

Disclaimer

References

1. Pacala, S.; Socolow, R. Stabilization wedges: Solving the climate problem for the next 50 years with current technologies. *Science* **2004**, *305*, 968–972.

2. Metz, B., Davidson, O., de Coninck, H., Loos, M., Meyer, L., Eds. *IPCC Special Report on Carbon Dioxide Capture and Storage*; Cambridge University Press: New York, 2005.

3. Juanes, R.; MacMinn, C. W. A mathematical model of the footprint of the CO_2 plume during and after injection in deep saline aquifer systems. *Energy Procedia* **2009**, *1*, 3429–3436.

4. LeNeveu, D. M. CQUESTRA—A risk and performance assessment code for geological sequestration of carbon dioxide. *Energy Conversion and Management* **2008**, *49*, 32–46.

5. Nordbotten, J. M.; Celia, M. A.; Bachu, S. Injection and storage of CO_2 in deep saline aquifers: analytical solution for CO_2 plume evolution during injection. *Transport in Porous Media* **2005**, *58*, 339–360.

6. Schlumberger, `http://www.slb.com/services/software/reseng/compositional.aspx`.

7. Pruess, K. The TOUGH codes—A family of simulation tools for multiphase flow and transport processes in permeable media. *Vadose Zone Journal* **2004**, *3*, 738–746.

8. Class, H. et al. A benchmark study on problems related to CO_2 storage in geologic formations. *Computational Geosciences* **2009**, *13*, 409–434.

9. Webster, M.; Tatang, M. A.; McRae, *Application of the probabilistic collocation method for an uncertainty analysis of a simple ocean model*; 1996; MIT Joint Program on the Science and Policy of Global Change.

10. Isukapalli, S. S.; Roy, A.; Georgopoulos, P. G. Stochastic response surface methods (SRSMs) for uncertainty propagation: Application to environmental and biological systems. *Risk Analysis* **1998**, *18*, 351–363.

11. Ghiocel, D. M.; Ghanem, R. G. Stochastic finite element analysis of seismic soil structure interaction. *Journal of Engineering Mechanics* **2002**, *128*, 66–77.

12. Xiu, D.; Karniadakis, G. E. Modeling uncertainty in flow simulations via generalized polynomial chaos. *Journal of Computational Physics* **2003**, *187*, 137–167.

13. Li, H.; Zhang, D. Probabilistic collocation method for flow in porous media: Comparisons with other stochastic methods. *Water Resources Research* **2007**, *43*, 44–56.

14. Eldred, E. S.; Burkardt, J., Comparison of non-intrusive polynomial chaos and stochastic collocation methods for uncertainty quantification. *47th AIAA Aerospace Sciences Meeting including the New Horizons Forum and Aerospace Exposition, Orlando, FL, AIAA paper 2009-0976* **2009**

15. Blatman, G.; Sudret, B. An adaptive algorithm to build up sparse polynomial chaos expansions for stochastic finite element analysis. *Probabilistic Engineering Mechanics* **2010**, *25*, 183–197.

16. Oladyshkin, S.; Class, H.; Helmig, R.; Nowak, W. An integrative approach to robust design and probabilistic risk assessment for CO_2 storage in geological formations. *Computational Geosciences* **2011**, *15*, 565–577.

17. Villadsen, J.; Michelsen, M. L. *Solution of differential equation models by polynomial approximation*; Prentice-Hall: New Jersey, 1978.

18. Oladyshkin, S.; Class, H.; Helmig, R.; Nowak, W. A concept for data-driven uncertainty quantification and its application to carbon dioxide storage in geological formations. *Advances in Water Resources* **2011**, *34*, 1508–1518.

19. Kopp, A.; Binning, P.; Johannsen, K.; Helmig, R.; Class, H. A contribution to risk analysis for leakage through abandoned wells in geological CO_2 storage. *Advances in Water Resources* **2010**, *33*, 867–879.

20. Wiener, N. The homogeneous chaos. *American Journal of Mathematics* **1938**, *60*, 897–936.

21. Askey, R.; Wilson, J. *Some basic hypergeometric orthogonal polynomials that generalize Jacobi polynomials*; American Mathematical Society: Providence, 1985.

22. Gautschi, W. Algorithm 726: ORTHPOL—A package of routines for generating orthogonal polynomials and Gauss-type quadrature rules. *ACM Transactions on Mathematical Software* **1994**, *20*, 21–62.

23. Kiureghian, A. D.; Liu, P. Structural reliability under incomplete probability information. *Journal of Engineering Mechanics* **1986**, *112*, 85–114.

24. Hosder, S.; Walters, R. W.; Balch, M., Efficient sampling for non-intrusive polynomial chaos applications with multiple uncertain input variables. *9th AIAA Non-Deterministic Approaches Conference* **2007**

25. Weisberg, S. *Applied linear regression*, 3rd ed.; John Wiley & Sons: New Jersey, 2005.

26. Sahinidis, N. V. Optimization under uncertainty: state-of-the-art and opportunities. *Computers and Chemical Engineering* **2004**, *28*, 971–983.

27. Kumar, D. Optimization of well settings to maximize residually trapped CO_2 in geologic carbon sequestration. M.Sc. thesis, Stanford University, Stanford, MA, 2007.

28. Pruess, K. ECO2N: A TOUGH2 fluid property module for mixtures of water, NaCl, and CO_2. Lawrence Berkeley National Laboratory: Berkeley, 2005.

29. National Petroleum Council Public Database, `http://www.netl.doe.gov/technologies/oil-gas/software/database.html`.

30. Sahinidis, N. V. BARON: A general purpose global optimization software package. *Journal of Global Optimization* **1996**, *8*, 201–205.

31. Tawarmalani, M.; Sahinidis, N. V. A polyhedral branch-and-cut approach to global optimization. *Mathematical Programming* **2005**, *103*, 225–249.

32. Dvoretzky, A.; Kiefer, J.; Wolfowitz, J. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Annals of Mathematical Statistics* **1958**, *27*, 642–669.