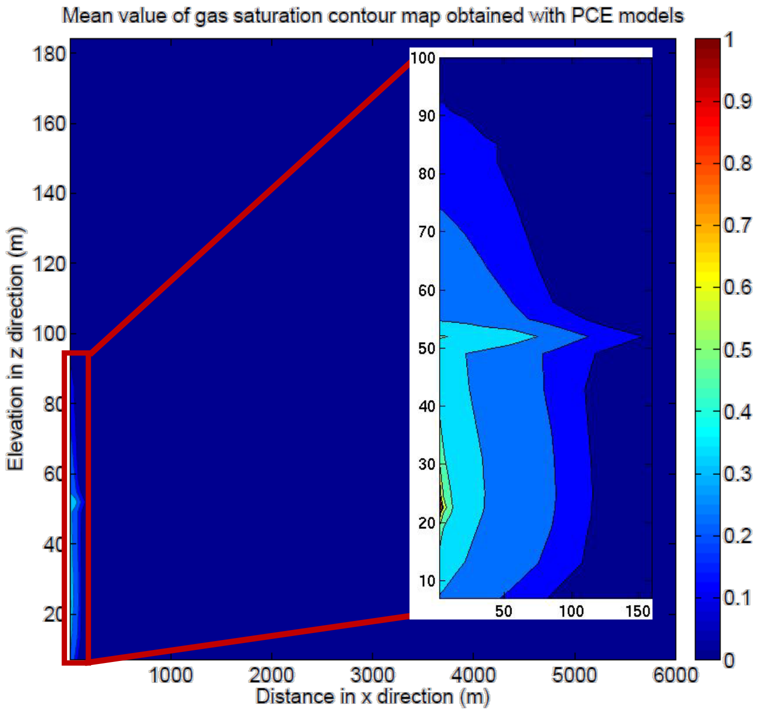


Developing Surrogate Models for CO_2 Sequestration Using Polynomial Chaos Expansion

13 September 2013

Office of Fossil Energy

NRAP-TRS-III-004-2013

Disclaimer

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference therein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed therein do not necessarily state or reflect those of the United States Government or any agency thereof.

This report (NRAP-TRS-III-004-2013) has been reviewed by NETL and approved for public release.

Cover Illustration: Gas saturation contour map obtained with polynomial chaos expansion.

Suggested Citation: Zhang, Y.; Sahinidis, N. V. *Developing Surrogate Models for CO_2 Sequestration Using Polynomial Chaos Expansion*; NRAP-TRS-III-004-2013; NRAP Technical Report Series; U.S. Department of Energy, National Energy Technology Laboratory: Morgantown, WV*,* 2013; p 20.

An electronic version of this report can be found at: www.netl.doe.gov/nrap

Developing Surrogate Models for CO_2 Sequestration Using Polynomial Chaos Expansion

Yan Zhang and Nikolaos V. Sahinidis[1]

[1]Department of Chemical Engineering, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213

NRAP-TRS-III-004-2013

Level III Technical Report Series

13 September 2013

This page intentionally left blank

Table of Contents

List of Figures

Acronyms and Abbreviations

Term	Description
LHS	Latin hypercube sampling
MC	Monte Carlo simulation
MIP	Mixed integer programming
PCE	Polynomial chaos expansion
PDF	Probability density function
SACROC	Scurry Area Canyon Reef Operators Committee

Acknowledgments

This work was completed as part of National Risk Assessment Partnership (NRAP) project. Support for this project came from the Department of Energy's (DOE) Office of Fossil Energy's Crosscutting Research program. The authors wish to acknowledge Robert Romanosky (NETL Strategic Center for Coal) and Regis Conrad (DOE Office of Fossil Energy) for programmatic guidance, direction, and support.

The authors also wish to acknowledge George Guthrie for technical direction guidance, and Grant Bromhal and members of NRAP for comments and suggestions that helped significantly improve the presentation of the material this report.

EXECUTIVE SUMMARY

This report describes a novel technique for deriving surrogate models from simulations. Surrogate models are built iteratively using polynomial chaos expansion (PCE) and detailed numerical simulations of carbon sequestration systems. The technique has been applied to simulations of a 2-D benchmark problem and the Scurry Area Canyon Reef Operators Committee (SACROC) oilfield. In both cases, output variables from a numerical simulator were approximated as polynomial functions of uncertain parameters.

Classical full PCE models are expensive to derive unless the number of terms in the expansion is moderate. To cope with this limitation, 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 number of terms small in the expansion, thus reducing the number of required simulations.

Our numerical experiments show that, in comparison with prior PCE techniques, simpler PCE models are obtained with high accuracy by the proposed approach. The resulting maximum relative error for the benchmark problem is 6%. The derived PCE models can therefore be used effectively in place of the numerical simulator and decrease simulation times by several orders of magnitude.

1. INTRODUCTION

The dynamics of an underground CO_2 plume can be predicted by solving the governing heat and mass balance equations and Darcy's equation. Analytical solutions to these equations can be obtained under simplified assumptions (Nordbotten et al., 2005; LeNeveu, 2008; Juanes and MacMinn, 2009). On the other hand, those governing equations can be solved with numerical simulators such as ECLIPSE or TOUGH2 for much more complicated and heterogeneous conditions. The typical situation encountered in the real world involves incomplete knowledge or limited measurement ability (particularly for porosities and permeabilities) that brings uncertainty into the use of those governing equations. Uncertainties in these input parameters usually have a significant effect on the output of a model, which raises the question of how to reliably quantify the risks of injecting CO_2 underground when such uncertainties are present.

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 using expected frequency histograms/distributions of model input values to obtain frequency histograms/distributions of model outputs. However, numerical models are generally computationally expensive for repeated simulations, especially when a single realization of the simulation requires hours or days of CPU time. As an alternative, we can first approximate the detailed model output of interest using a reduced-order model, such as 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 representing uncertain behavior in a complex system. One category of PCE methods is nonintrusive approaches, which require no manipulation of the equations in the simulator. The unknown coefficients in the expansion are usually evaluated through interpolation or regression. These nonintrusive approaches are easy to implement and generalize to complex systems. For this reason, we focus on nonintrusive PCE methods based on regression.

2. POLYNOMIAL CHAOS EXPANSION

If some model parameters (inputs) are uncertain and can be characterized with probability density functions (PDFs), such as normal or uniform distributions, then model outputs would also behave as random variables and follow distributions as well (see Figure 1).

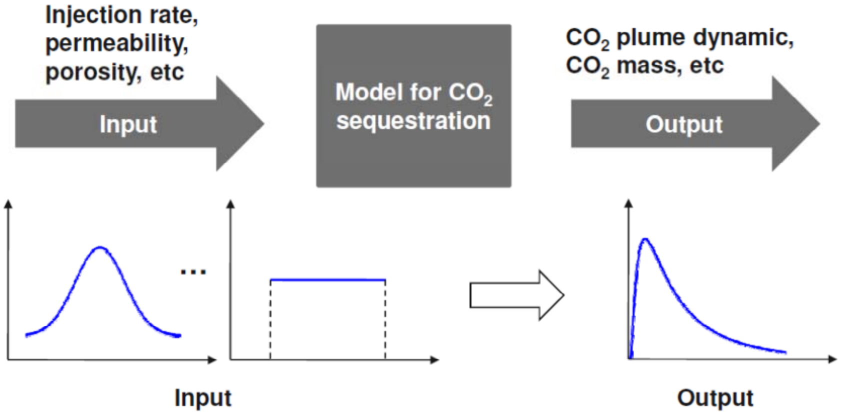

Injection rate, permeability, porosity, etc
Input
Model for CO_2 sequestration
CO_2 plume dynamic, CO_2 mass, etc
Output
...
Input
Output

Figure 1: Uncertainty propagation.

If we assume that a particular model output *y* has finite variance, then *y* can be represented by the following polynomial chaos expansion with respect to uncertain input *x* (Wiener, 1938):

$$y = \sum_{d=0}^{+\infty} B_d \alpha_d$$

where the α's are coefficients, the B's are multivariate polynomial basis functions that are orthogonal with respect to the joint PDF of *x,* and *d* is the degree of basis functions. Details of this definition and a discussion of the generation of multivariate polynomial basis functions can be found in a paper by Zhang and Sahinidis, 2013. In practice, this infinite expansion is truncated at a finite degree *d*. The number of the terms N_t in the expansion can be calculated as $(M+d)!/M!/d!$, where *M* is the number of uncertain inputs, and *d* is the degree of the polynomial basis functions. This number grows rapidly as the number of inputs and degree increase, e.g., for a ten-input expansion truncated at degree six, N_t is 8008.

The coefficients α in the expansion can be estimated through regression, using the following process. First, a few sample points N_p are selected in the domain of uncertain inputs. Then, an $N_p \times N_t$ feature matrix *B* is obtained by evaluating polynomial basis functions under those sample points. Also, a set of N_p values of a model output *y* is obtained by running the detailed simulation with the selected sample points. Then, coefficients are computed by solving the following linear system:

$$\boldsymbol{B\alpha} = \boldsymbol{y}$$

Generally, the number of sample points N_p is greater than N_t so that this linear system is over-determined with a closed form solution for $\boldsymbol{\alpha}$, i.e., $\boldsymbol{\alpha} = (\boldsymbol{B^T B})^{-1}\boldsymbol{B^T y}$. Note that for numerical simulators that discretize space and time in the governing partial differential equations, PCEs are developed for each gridblock. Therefore, the coefficient vector $\boldsymbol{\alpha}$ is temporally and spatially dependent.

The R^2 statistic can be used to measure the goodness of fit for different polynomial basis functions; i.e., the closer R^2 is to one, the better fit it is. Another generally used measure is cross-validation error, e.g., leave-one-out cross validation Q^2. This is often a better estimator to use, as it reduces the chance of overfitting.

However, optimizing the number of sample points brings a potential problem. It means that the numerical simulator needs to be run at least N_t times to estimate the coefficients, which would become computationally prohibitive for multi-dimensional inputs and a high degree of expansion. To cope with this issue, forward and backward stepwise regression techniques can be used, for example as proposed by Blatman and Sudret (2010). However, in this kind of stepwise regression method, the synergistic effect of basis functions is ignored. For example, a 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 previously discarded terms, the resulting truncated polynomial expansion may not be the best subset of the basis set. This issue is addressed with a mixed-integer programming formulation.

STEPWISE REGRESSION WITH MIXED-INTEGER PROGRAMMING FORMULATION

In Zhang and Sahinidis (2013), we proposed a new method of best subset selection based on mixed integer programming (MIP) to build PCE surrogate models. The detailed formulation can be found in the paper by Zhang and Sahinidis (2013). A schematic flowchart of the proposed MIP-based stepwise regression is presented in Figure 2.

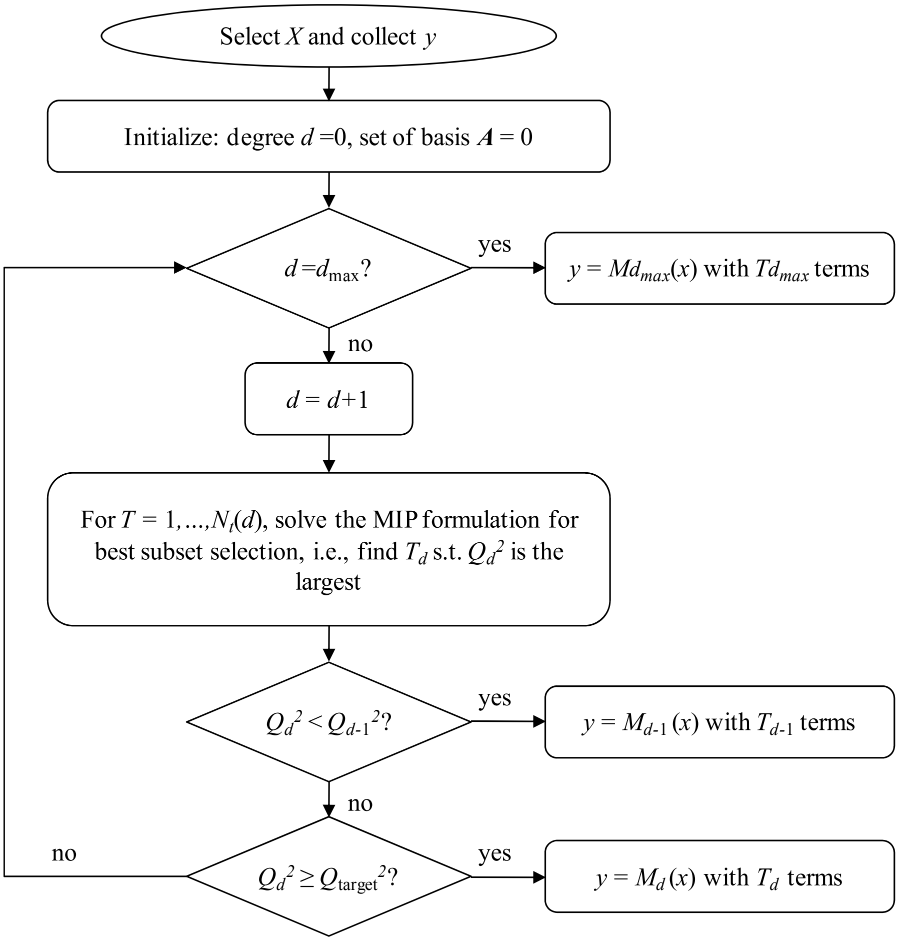

Select X and collect y
Initialize: degree d =0, set of basis $\boldsymbol{A}$ = 0
d = $d_{\max}$?
yes
$y = Md_{max}(x)$ with Td_{max} terms
no
$d = d+1$
For $T = 1,...,N_t(d)$, solve the MIP formulation for best subset selection, i.e., find T_d s.t. Q_d^2 is the largest
$Q_d^2 < Q_{d\text{-}1}^2$?
yes
$y = M_{d\text{-}1}(x)$ with $T_{d\text{-}1}$ terms
no
no
$Q_d^2 \geq Q_{\text{target}}^2$?
yes
$y = M_d(x)$ with T_d terms

Figure 2: Flowchart of MIP-based stepwise regression method.

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 (number of terms in the expansion) 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 of basis functions. This gives us the best subset that considers the synergistic effects of basis functions. Also, the introduction of cross-validation error (which is recalculated at each step in the fitting of the orthogonal polynomial expansion) helps to determine the appropriate degree and number of terms in the expansion and avoid overfitting (see Figure 3).

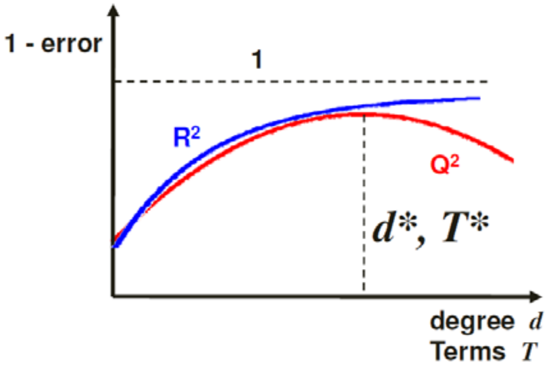


Figure 3: Fitness measure by R^2 and Q^2. T is optimized where Q^2 is a maximum.

In computational results, Zhang and Sahinidis (2013) 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. This proposed adaptive PCE method is applied to a benchmark problem presented in the following sections.

3. MODEL INITIATION

A benchmark of CO_2 injection into a 2-D layered brine formation has been simulated using TOUGH2 (Problem no. 4 in Pruess, 2005). This benchmark problem was based on the first industrial-scale CO_2 disposal project with approximately one million tons of CO_2 per year injecting into a saline aquifer through a horizontal well.

A 2-D vertical half space section was modeled assuming each permeable formation is homogeneous and isotropic (see Figure 4). The domain is discretized into 29×34 (986) gridblocks. The simulation by TOUGH2 provides results such as the pressure and the CO_2 distribution profile (e.g., mass and gas saturation) through all the formation layers.

We are mainly interested in quantifying the impact of the uncertain parameters such porosity and permeability on the model outputs. For illustration purposes, the model outputs will be approximated as polynomial functions of porosity and permeability, allowing us to perform MC simulation with the PCE approximation later. The approach can be easily extended to higher dimensions and include, for instance, injection rate and time as model inputs.

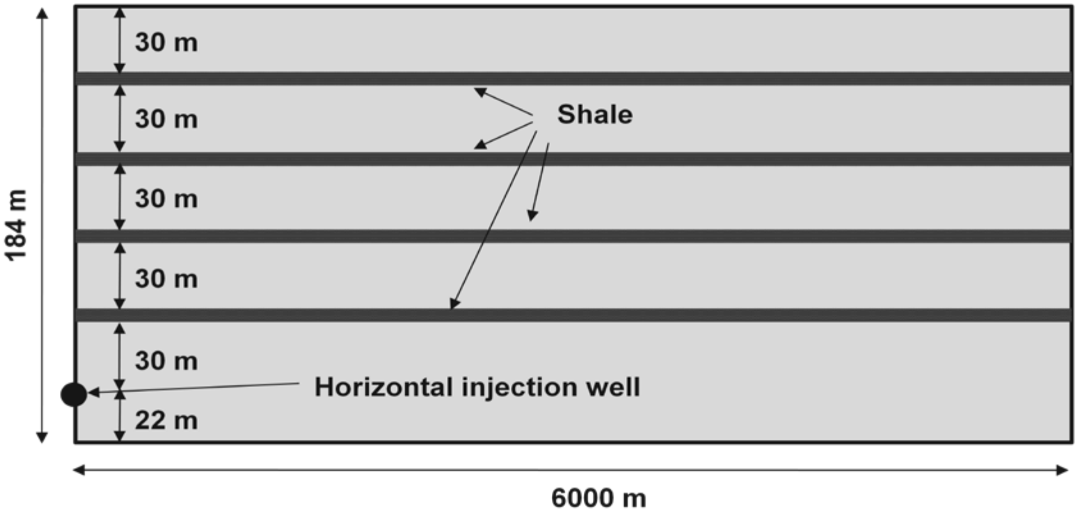
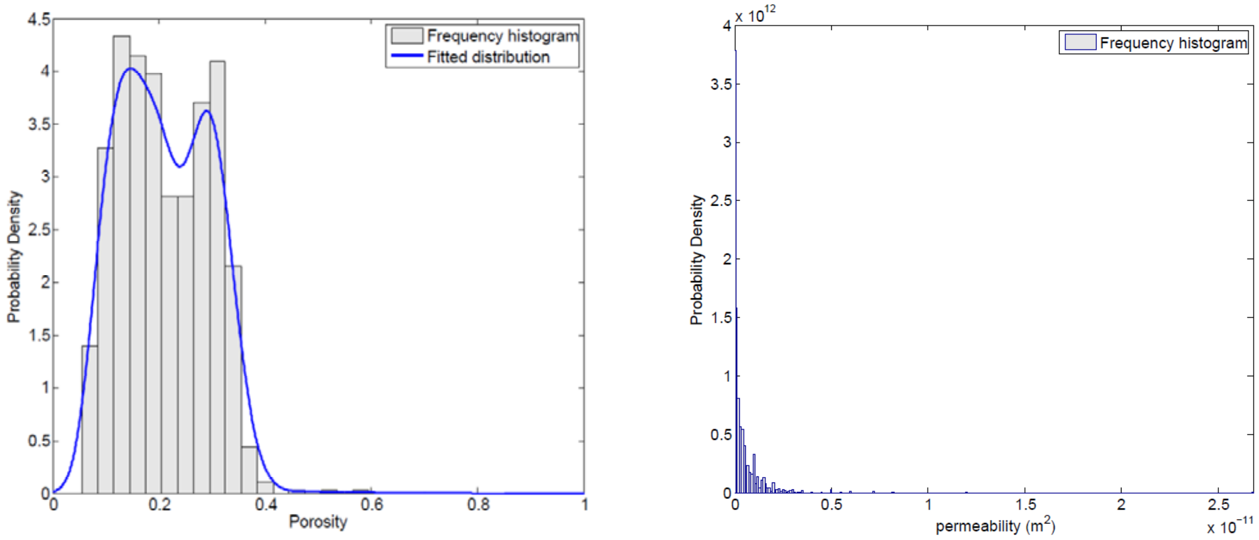


Figure 4: Schematic geometry of CO_2 injection into saline aquifer (half space).

4. PCE SURROGATE MODEL DEVELOPMENT

To determine the polynomial basis functions with respect to uncertain porosity and permeability, the probability distributions of the two parameters need to be specified. As an illustration, the National Petroleum Council database for over one-thousand reservoirs in the United States is used. The marginal distributions of the two parameters are in Figure 4.



(a)

(b)

Figure 5: (a) Probability distribution function of porosity, and (b) Probability distribution function of permeability.

Since neither of the two parameters follows a standard parametric distribution, for example, the fitted distribution of porosity obtained by kernel density estimation is not a standard distribution. Moreover, there exists a strong correlation between the two parameters, a transformation is performed to translate the two correlated parameters into uncorrelated standard normal random variables ξ.

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 random variables, the polynomial basis functions are as follows:

$$B_0(\xi_1, \xi_2) = 1$$

$$B_1(\xi_1, \xi_2) = \xi_1$$

$$B_1(\xi_1, \xi_2) = \xi_2$$

$$B_2(\xi_1, \xi_2) = \xi_1^2 - 1$$

$$B_2(\xi_1, \xi_2) = \xi_1 \xi_2$$

$$B_2(\xi_1, \xi_2) = \xi_2^2 - 1$$

$$\vdots$$

The model output is then a PCE approximation in terms of ξ with the coefficients α's left as unknowns. To estimate unknown coefficients, a fixed design of experiments is chosen, i.e., 100 Latin hypercube sampling (LHS) design for ξ. Then those polynomial basis functions are evaluated for these 100 sample points. 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. Among these, the best subset selection method using the MIP formulation is solved to find a PCE approximation for a model output with relatively large R^2 and Q^2 values. Figure 6, which shows a response surface for mass of gas phase CO_2 in the caprock, is one example of the polynomial surface fitted to 100 random samples with high accuracy, i.e., $R^2 = 0.98$ and $Q^2 = 0.98$. In this case, a fourth-order expansion is used.

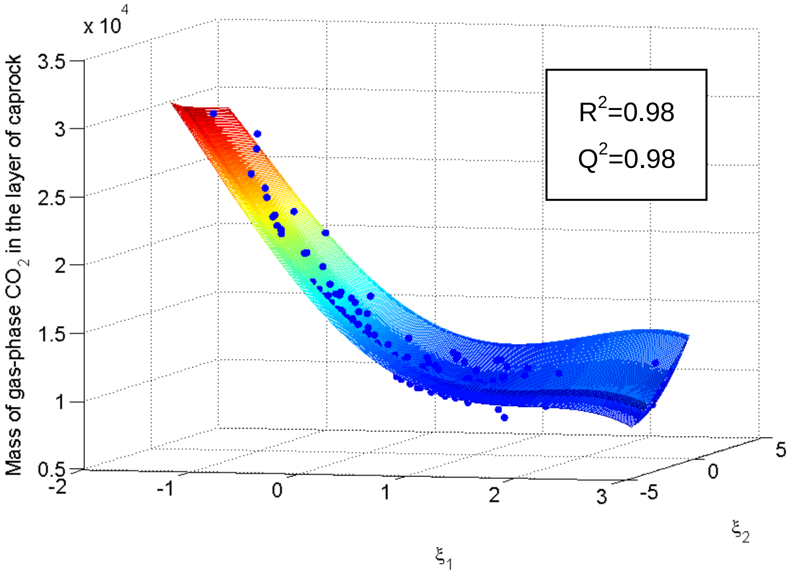
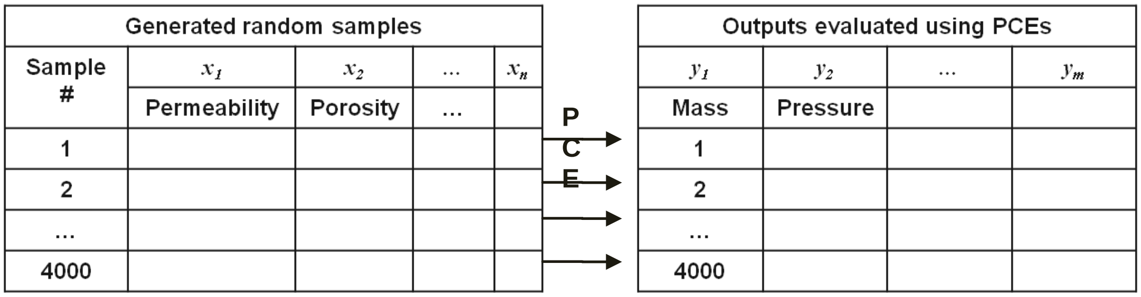


Figure 6: An example of PCE surrogate model, fitting to a fourth-order expansion.

5. UNCERTAINTY ANALYSIS BASED ON PCE SURROGATE MODELS

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 parameters are correlated). An 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 (see Figure 7). Statistical analysis can then be performed for this specific model output *y* for uncertainty analysis.



Generated random samples				
Sample #	x_1	x_2	…	x_n
	Permeability	Porosity	…	
1				
2				
…				
4000				

Outputs evaluated using PCEs			
y_1	y_2	…	y_m
Mass	Pressure		
1			
2			
…			
4000			

Figure 7: Monte Carlo simulation with PCEs.

For the benchmark problem here, it takes about 15 min to perform one TOUGH2 simulation, which means 4,000 MC simulations would take over 1 CPU-month of time. With our developed PCE models, the average time of running simulations is less than 0.1 second. In this case, we performed 1,000 simulations with the TOUGH2 model to obtain distributions of model outputs to use for validating the results of the PCE models. Developing a single PCE model takes a few seconds with MATLAB implementation and running a single PCE simulation takes less than one second. 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 1,000 MC simulations with TOUGH2 directly. Figure 8 shows the cumulative distribution functions for the mass fraction of CO_2 in the layer of saline aquifer after 30 days of injection. The distribution obtained with a second-order polynomial expansion (dashed red) is very close to the distribution (solid blue) obtained by running TOUGH2. Figure 9 shows a comparison between gas saturations predicted by PCE and TOUGH2 models after 30 days of injection. The two models demonstrate strong agreement in this case, which is representative of most of the simulations completed.

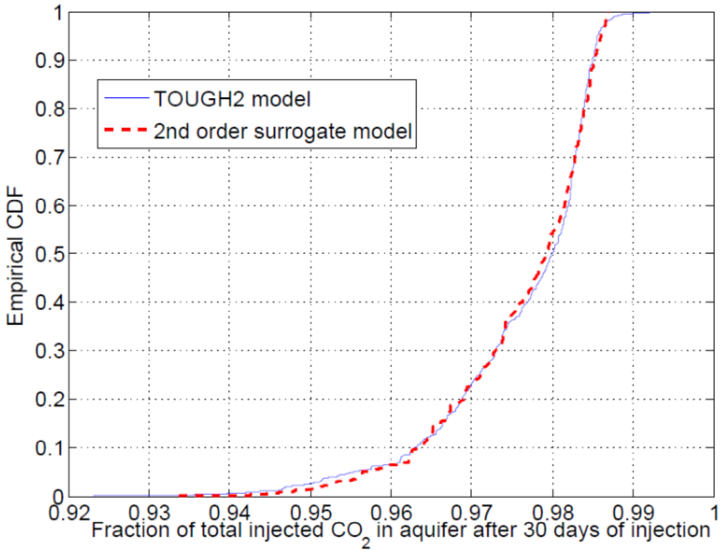


Figure 8: The cumulative distribution functions for the mass fraction of CO_2 in the layer of saline aquifer after 30 days of injection.

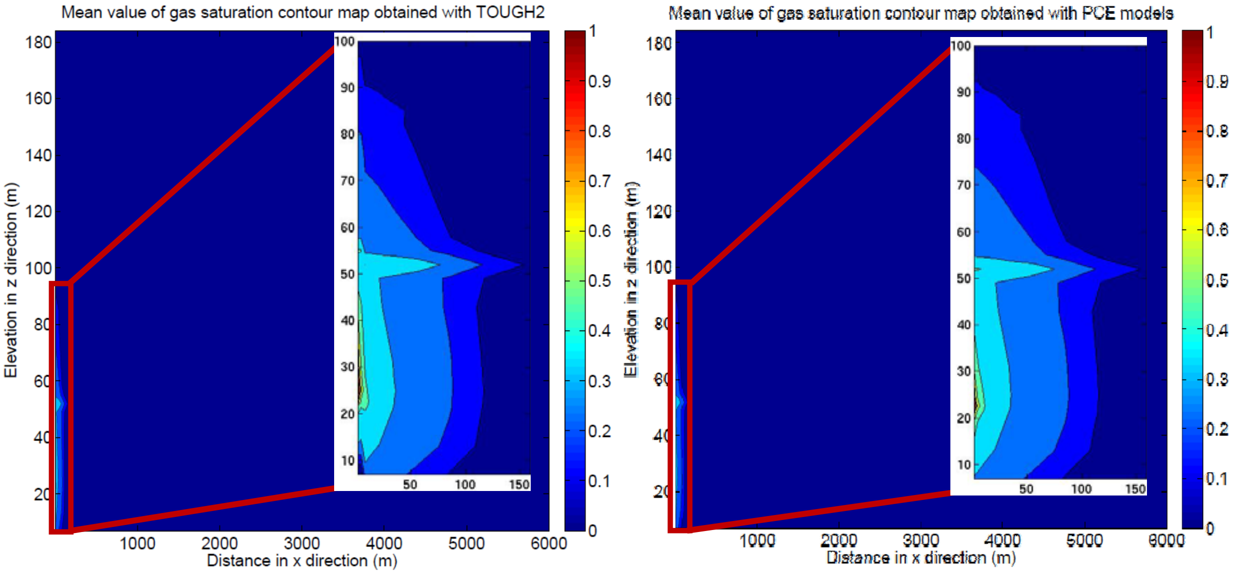


Figure 9: Gas saturation contour map in average obtained with TOUGH2 (left) and with PCEs (right).

6. CONCLUSION

We have presented a MIP-based best subset selection method to iteratively build PCE models for predicting subsurface conditions (e.g., pressure, saturation) during geologic CO_2 storage. This 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 surfaces of model outputs obtained with the PCE surrogate models match well with those obtained with detailed simulations with TOUGH2. We further utilized the PCE models for uncertainty quantification. 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). Using the PCE models reduced the time needed for simulations by orders of magnitude.

7. REFERENCES

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.

Juanes, V.; 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.

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

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 Porous Media* **2005,** *58*, 339−360.

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

Wiener, N. The homogeneous chaos. *Am. J. Math.* **1938,** *60*, 897−936.

Zhang, Y.; Sahinidis, N. V. Uncertainty quantification in CO_2 Sequestration using surrogate models from polynomial chaos expansion. *Ind. Eng. Chem. Res.* **2013,** *52*, 3121–3132.

This page intentionally left blank

NRAP

National Risk Assessment Partnership

NRAP is an initiative within DOE's Office of Fossil Energy and is led by the National Energy Technology Laboratory (NETL). It is a multi-national-lab effort that leverages broad technical capabilities across the DOE complex to develop an integrated science base that can be applied to risk assessment for long-term storage of carbon dioxide (CO_2). NRAP involves five DOE national laboratories: NETL Regional University Alliance (NETL-RUA), Lawrence Berkeley National Laboratory (LBNL), Lawrence Livermore National Laboratory (LLNL), Los Alamos National Laboratory (LANL), and Pacific Northwest National Laboratory (PNNL). The NETL-RUA is an applied research collaboration that combines NETL's energy research expertise in the Office of Research and Development (ORD) with the broad capabilities of five nationally recognized, regional universities—Carnegie Mellon University (CMU), The Pennsylvania State University (PSU), University of Pittsburgh (Pitt), Virginia Polytechnic Institute and State University (VT), and West Virginia University (WVU)—and the engineering and construction expertise of an industry partner (URS Corporation).

Technical Leadership Team

Jens Birkholzer
LBNL Technical Coordinator
Lawrence Berkeley National Laboratory
Berkeley, CA

Grant Bromhal
NETL Technical Coordinator
Lead, Reservoir Performance Working Group
Office of Research and Development
National Energy Technology Laboratory
Morgantown, WV

Chris Brown
PNNL Technical Coordinator
Pacific Northwest National Laboratory
Richmond, WA

Susan Carroll
LLNL Technical Coordinator
Lawrence Livermore National Laboratory
Livermore, CA

Josh White
Lead, Induced Seismicity Working Group
Lawrence Livermore National Laboratory
Livermore, CA

Diana Bacon
Lead, Groundwater Protection Working Group
Pacific Northwest National Laboratory
Richmond, WA

Tom Daley
Lead, Strategic Monitoring Working Group
Lawrence Berkeley National Laboratory
Berkeley, CA

George Guthrie
Technical Director, NRAP
Office of Research and Development
National Energy Technology Laboratory
Pittsburgh, PA

Rajesh Pawar
LANL Technical Coordinator
Lead, Systems/Risk Modeling Working Group
Los Alamos National Laboratory
Los Alamos, NM

Tom Richard
Deputy Technical Director, NRAP
The Pennsylvania State University
NETL-Regional University Alliance
State College, PA

Brian Strazisar
Lead, Migration Pathways Working Group
Office of Research and Development
National Energy Technology Laboratory
Pittsburgh, PA

Sean Plasynski
Deputy Director
Strategic Center for Coal
National Energy Technology Laboratory
U.S. Department of Energy

Jared Ciferno
Director
Office of Coal and Power R&D
National Energy Technology Laboratory
U.S. Department of Energy

Susan Maley
Technology Manager
Crosscutting Research
National Energy Technology Laboratory
U.S. Department of Energy

Regis Conrad
Director
Division of Cross-cutting Research
Office of Fossil Energy
U.S. Department of Energy

NRAP Executive Committee

Cynthia Powell
Director
Office of Research and Development
National Energy Technology Laboratory
U.S. Department of Energy

Alain Bonneville
Laboratory Fellow
Pacific Northwest National Laboratory

Donald DePaolo
Chair, NRAP Executive Committee
Associate Laboratory Director
Energy and Environmental Sciences
Lawrence Berkeley National Laboratory

Melissa Fox
Program Manager
Applied Energy Programs
Los Alamos National Laboratory

Roger Aines
Carbon Fuel Cycle Program Leader
Lawrence Livermore National
Laboratory

George Guthrie
Technical Director, NRAP
Office of Research and Development
National Energy Technology Laboratory

NRAP Technical Report Series