Large-scale Impact of CO2 Storage Operations: Dealing with Computationally Intensive Simulations for Global Sensitivity Analysis

Jeremy Rohmer, Benoit Issautier, Christophe Chiaberge, Pascal Audigane

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Energy Procedia 37 (2013) 3883 – 3890

Energy Procedia

GHGT-11

Large-scale impact of CO_2 storage operations: dealing with computationally intensive simulations for global sensitivity analysis

Jeremy Rohmer[a]*, Benoit Issautier[a], Christophe Chiaberge[a], Pascal Audigane[a]

[a]*BRGM, 3 av. Claude Guillemin, 45060 Orléans Cédex 2, France*

Abstract

Assessing the potential impacts associated with CO_2 storage operations implies using large-scale models characterized by a very large number of grid cells (>500,000) and high computation time cost (> several hours). Yet, investigating the influence of the input parameters on the model results requires multiple simulations (>1,000), which might become impracticable due to the computation burden. A meta-modelling strategy is then proposed, basically consisting in approximating the long running model by a costless-to-evaluate model, for instance a Gaussian Process, based on a very limited number of simulations (e.g., 50). This strategy is tested to investigate the sensitivity of the overpressure induced by an industrial-scale CO_2 injection into a fluvial heterogeneous reservoir, to the properties of the shale formation using a 3-dimensional long running multiphase flow model (with CPU time > 5 days).

Selection and/or peer-review under responsibility of GHGT

Keywords: Industrial-scale CO_2 injection; Sobol' indices; Monte-Carlo-based technique; Long running simulations; Meta-model;Gaussian process.

1. Introduction

CO_2 capture and geological storage is seen as a promising technology in the portfolio of measures required to mitigate the effects of anthropogenic greenhouse gas emissions (IPCC [1]). Yet, a pre-requisite to its wide scale implementation is demonstrating safety. Assessing the potential risk events and associated impacts resulting from CO_2 storage operations (e.g., CO_2 leakage; brine displacements; far-field pressurization; damaging of the "natural" safety barriers,...) is typically supported by large-scale

* Corresponding author. Tel.: +33-2-3864-3092; fax: +33-2-3864-3689.
E-mail address: j.rohmer@brgm.fr.

1876-6102
Selection and/or peer-review under responsibility of GHGT
doi:10.1016/j.egypro.2013.06.286

numerical simulations i.e. dynamic modeling at the spatial scales of the storage complex, as advocated for instance by the recent European directive on geological storage of carbon dioxide (directive 2009/31/EC). Such numerical tools offer the advantages of capturing complex geological architecture of storage formation (based on the static model), as well as coupling processes (e.g., multiphase flow transport, geochemical, mechanical, etc.) with regards to specific features, along with complex injection scenario (leakage, buoyancy inside dipping aquifers, induced convective dissolution…).

In practice, such numerical models are based on a large variety of input parameters and hypotheses. Each of these input parameters is associated with uncertainty, so that the European directive has advocated the need for measuring the influence of these sources of uncertainty and providing a ranking procedure for an appropriate decision for risk management (directive 2009/31/EC Annex I Step 3.2 Sensitivity characterization), especially to guide future lab or in site characterizations and studies, but also to "simplify" the model by fixing the input parameters with negligible influence. Variance-based global sensitivity analysis (GSA), relying on the Sobol' indices (i.e. sensitivity indices), can provide such valuable information (see Saltelli et al. [2]). This analysis presents the advantages of exploring the sensitivity to input parameters over their whole range of variation (i.e. in a global manner contrary to other local approaches relying on the calculation of partial derivatives), of fully accounting for possible interaction between the input parameters and of being applicable without introducing a priori assumptions on the mathematical formulation of the numerical model (e.g., linearity, monotonicity, etc.).

Conducting GSA is hindered by a major difficulty: the different algorithms available for the estimation of the Sobol' indices require a large number of model evaluations (of the order of thousands, see Saltelli et al. [2]). Yet, numerical models for large-scale impact assessment of CO_2 storage operations are generally characterized by high number of grid cells (> 500,000) and with high CPU time (> several hours) of a single simulation. To overcome this computation challenge, the objective of the present study is to explore the applicability of the combination of an appropriate grid computing architecture and of the meta-modeling technique (Storlie et al. [3]). The latter technique basically consists in replacing the numerical model by a "costless-to-evaluate approximation" (the meta-model also named response surface or surrogate model or reduced model).

The remainder is organized as follows. In a first section, we describe the steps of the strategy combining meta-modelling technique for conducting GSA. In a second section, we apply such a methodology on a simple 1-d analytical model, with low CPU time (Manceau and Rohmer [4]). Using this analytical example enables us to compare the results of the meta-model with the "true" ones. In a third section, the methodology is applied to a long running 3-d multiphase flow model, with CPU time of several days (Issautier et al. [5]), to rank in terms of importance the multiphase properties of the shale rock formation embedding complex porous sandstone bodies (fluvial heterogeneous reservoir).

2. A meta-modelling strategy

2.1. Principles

To overcome the computation challenge related to the estimation of Sobol' indices using a long-running model, we rely on the meta-modelling technique (Storlie et al. [3]). The basic idea of meta-modelling is to replace the long running numerical model f by a mathematical approximation (denoted g) referred to as "meta-model" (also named "response surface", or "surrogate model"). The meta-model corresponds to a "costless-to-evaluate" function aiming at reproducing the behaviour of the "true" model f in the domain of model input parameters $\boldsymbol{x}$ and at predicting the model responses $y=f(\boldsymbol{x})$ with a negligible CPU time (y can be for instance the gaseous saturation value at a given distance from the injection zone). The main steps of the methodology are summarized in Table 1.

Table 1. Description of the meta-modelling strategy for conducting global sensitivity analysis

Step	Description
1	Generate n_0 different values for the input parameters $\boldsymbol{x}$ using a LHS technique and simulate the corresponding model outputs y;
2	Based on this training data, construct a meta-model and assess the approximation and the predictive quality using cross-validation procedure;
3	Using the "costless-to-evaluate" meta-model, compute the Sobol' indices and analyse the importance of each of the input parameters. Compute a confidence interval on each sensitivity index to account for uncertainty in the construction of the approximation.

2.2. Step 1

The first step is to run f for a limited number n_0 of different configurations (named training samples) of m-dimensional vectors of input parameters $\boldsymbol{x}_i=(x_1 ; x_2 ; \ldots, x_m)$ with i=1,2,…n_0. To choose them, a trade-off should be found between maximizing the exploration of the input parameters' domain and minimizing the number of simulations, i.e. a trade-off between the accuracy of the approximation (directly linked with n_0) and the CPU cost. To fulfill such requirements, we propose to randomly select the training samples by means of the Latin Hypercube Sampling LHS method (see e.g., McKay et al. [6]) in combination with the "maxi-min" space filling design criterion.

For each of the randomly selected training sample $\boldsymbol{x}_i$, the corresponding model output y_i is calculated by running the computationally intensive model. The set of n_0 pairs of the form $\{\boldsymbol{x}_i ; y_i\}$, with i=1,2,… n_0, constitute the training data on which the meta-model is constructed in step 2.

2.3. Step 2

Using the training data, f can then be approximated by a meta-model g so that $y=f(\boldsymbol{x})\approx g(\boldsymbol{x})$. Several types of meta-models exist: simple polynomial regression techniques, non-parametric regression techniques, Gaussian process, etc. See Storlie et al. [3] for a recent review. The choice of the meta-model type is guided by the a priori non-linear functional form of f, as well as the number of input parameters. In the following, we will more specifically focus on the meta-model class of Gaussian processes. For full details, the interested reader can refer to Gramacy and Herbert [7] and references therein.

As the methodology involves replacing f by an approximation g, it introduces a new source of uncertainty. Two issues should be addressed: 1. the approximation quality, i.e. to which extent g manages at reproducing the observed y_i, i.e. the ones calculated based on the set of different long running simulations; 2. the predictive quality, i.e. to which extent g manages at predicting y at "yet-unseen" input parameters' configurations.

Regarding the first issue, the differences between the approximated and the true quantity of interest (i.e. the residuals) are usually used. On this basis, the coefficient of determination R^2 can be computed so that if R^2 is close to one, the approximation can be considered of good quality.

Regarding the second quality issue, a first approach would consist in using a test sample of new data. Though the most efficient, this might be often impracticable as additional numerical simulations are costly to collect. An alternative relies on cross-validation procedures (see e.g., Hastie et al. [8]). This technique involves: 1. randomly splitting the initial training data into q equal sub-sets; 2. removing each of these sub-sets in turn from the initial set; fitting a new meta-model using the remaining q-1 sub-sets; 3. the sub-set removed from the initial set constitutes the validation set , which is estimated using the new meta-model. The procedure corresponds to the "leave-one-out" cross validation, if each sub-set is

composed of a single observation. Using the residuals computed at each iteration of the procedure, a coefficient of determination R^2 can be estimated to be used as a metric of predictive quality.

2.4. Step 3

Once validated, the costless-to-evaluate meta-model can be used to estimate y at any "yet-unseen" values of the input parameters and can be used to conduct the GSA using the Sobol' indices. We introduce hereafter the basic concepts of GSA. For a more complete introduction and full derivation of equations, the interested reader can refer to (Saltelli et al. [2] and references therein).

Considering the m-dimensional vector $\boldsymbol{X}$ as a random vector of independent random variable X_i (i=1,2,…,m), then the output $Y=f(\boldsymbol{X})$ is also a random variable (as a function of a random vector). A variance-based sensitivity analysis aims at determining the part of the total unconditional variance Var(Y) of the output Y resulting from each input random variable X_i. This analysis relies on the functional analysis of variance (ANOVA) decomposition of f based on which the Sobol' indices (ranging between 0 and 1) can be defined:

$$S_i = \frac{\mathrm{Var}\left[\mathrm{E}\left(Y|X_i\right)\right]}{\mathrm{Var}\left(Y\right)}, \qquad S_{ij} = \frac{\mathrm{Var}\left[\mathrm{E}\left(Y|X_i, X_j\right)\right]}{\mathrm{Var}\left(Y\right)} - S_i - S_j \qquad (1)$$

The first-order S_i is referred to as "the main effect of X_i" and can be interpreted as the expected amount of Var(Y) (*i.e.* representing the uncertainty in Y) that would be reduced if it was possible to learn the true value of X_i. This index provides a measure of importance useful to rank in terms of importance the different input parameters (Saltelli et al. [2]). The second order term S_{ij} measures the combined effect of both parameters X_i and X_j. Higher order terms can be defined in a similar fashion.

3. Illustration of the meta-modelling technique

3.1. Description of the analytical model

We aim at illustrating step 1 and 2 of the afore-described methodology using the one-dimensional version of the analytical model developed by Manceau and Rohmer [4] with low CPU time (<2 seconds). This model is used to compute the time duration T necessary to trap the whole amount of CO_2 stored (~1.5 Mt) in a 20m thick, 1200m deep aquifer reservoir formation. We consider here the residual trapping associated with the natural groundwater flow acting within the aquifer formation at ~3.5 m/year.

We restrict the analysis to the gaseous and liquid residual saturation (respectively denoted S_{gr} and S_{lr}) of the aquifer formation. Both uncertain input parameters respectively vary between 0.05 and 0.30 and between 0.30 and 0.5. The other input parameters are set at the values described in Manceau and Rohmer [4].

3.2. Application

To assess the efficiency of the meta-modelling strategy, we first calculate the time duration T considering different pairs of $\{S_{gr}\ ;\ S_{lr}\}$ using a grid design of 21x21. In total, 441 simulations using the analytical model were performed. Figure 1 depicts the evolution of T in this 2-dimensional domain (red straight lines). This constitutes the reference, which we aim at approximating.

In this purpose, we generate 10 training data selected through the LHS approach (outlined by a black dot in Figure 1A). On this basis, we construct a meta-model of the form of a Gaussian Process (see e.g., Gramacy and Herbert [7]). The validation phase (through leave-one-out cross validation technique)

provides a satisfactory coefficient of determination of ~95%. The approximation is depicted in Figure 1A by black dashed lines. We can see that in the region where training data where simulated, the reference and the approximation are in good agreement. On the other hand, in regions where poor knowledge has been gathered (in particular in the vicinity of S_{gr}~0.25 and S_{lr}~0.50), we can notice discrepancies. When using more training data (20 samples), we show that the agreement is very good (Figure 1B).

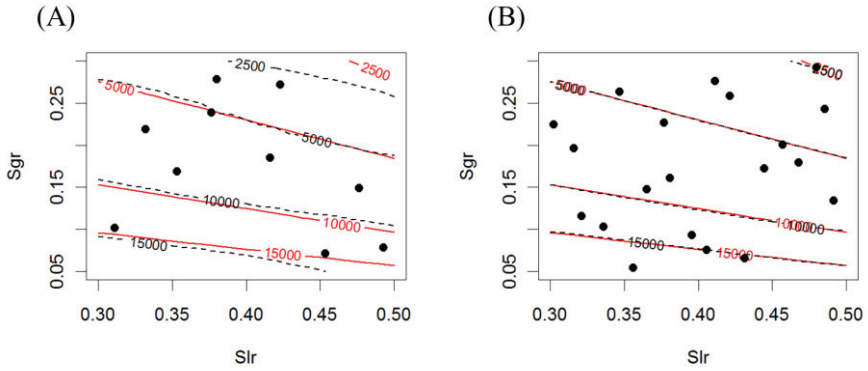

Fig. 1. Residual trapping time duration T (expressed in years) as a function of S_{lr} and S_{gr}. The red straight lines correspond to the "reference" solution computed directly using the analytical model (441 simulations). (A) The black dashed lines correspond to the approximation provided by the Gaussian-Process meta-model (mean predictive estimate) constructed using only 10 samples (black dots). (B) The approximation is constructed using 20 training data. See text for further details.

This simple example highlights the effect of approximation uncertainty when replacing the "true" model by the meta-model to conduct the GSA. In the case of long running simulations, the most often encountered situation is likely to be the one of Figure 1A, i.e. the total computation time cost dictates the affordable number of simulations. Therefore, the uncertainty in the construction of the approximation should be accounted for when presenting the results calculated using the meta-model. Different approaches can be proposed: bootstrap technique, as proposed by Storlie et al. [3] or in the Bayesian framework. In the following, we will concentrate on the second approach, because this formalism is a natural framework for estimating the (hyper-) parameters of Gaussian Processes (Gramacy and Herbert [7]).

4. Long-running application case

4.1. Model set-up and parameters

We aim at applying the methodology using the very long running 3- dimensional multiphase flow model (Issautier et al. [5]) used to simulate an industrial-scale CO_2 injection (i.e. > 1 Mt/y) into a 1200m deep, 60m thick fluvial heterogeneous reservoir seen as a complex layout of highly heterogeneous sandy sedimentary bodies with varying connectivity. Figure 2 provides an overview of the sedimentary bodies embedded in a shaly floodplain. The system corresponding to a 25 km x 25 km x 60 m heterogeneous aquifer formation (open lateral boundaries) is represented by a grid mesh of more than 840,000 cells (with a refined zone at 80m x 80m x 2m in the central part where the injection is conducted). The objective is to investigate the sensitivity of the over- pressure induced by the injection at a distance of 5km from the injector, i.e. at the transition between the sedimentary bodies and the floodplain (along the profile

outlined in Figure 2 by a red straight line). Five input parameters characterizing the shale are considered: the pore compressibility β and the four multiphase flow properties, namely the parameters of the Van-Genuchten's models for capillary pressure and relative permeability $n(VG)$ and P_0, and the gaseous and liquid residual saturation S_{gr} and S_{lr}. Table 2 gives the assumptions for the upper and lower bounds of these parameters.

Table 2. Lower and Upper bound associated to each input parameter

Property	Lower bound	Upper bound
1 Pore compressibility β	9.e-10 Pa^{-1}	4.5e-10 Pa^{-1}
2 Van-Genuchten' s parameter $n(VG)$	0.46	0.60
3 Liquid residual saturation S_{lr}	0.20	0.50
4 Gaseous residual saturation S_{gr}	0.05	0.35
5 Van Genuchten's parameter P_0	5 bar	50 bar

Note that we intentionally choose a problem for which we know before-hand the intuitive solution, namely: the pore compressibility β should have the strongest influence on the variation of the over-pressure. This allows us to lesser extent verifying the results of the sensitivity analysis using the meta-model.

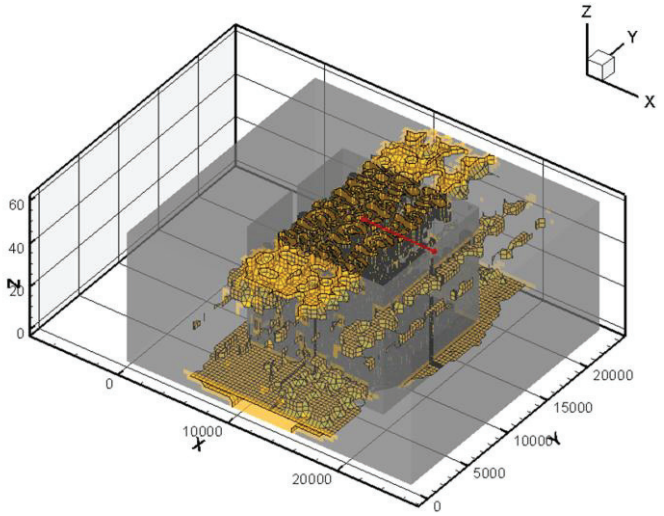

Fig. 2. Overview of the heterogeneous reservoir. The sedimentary bodies are outlined in yellow and the floodplain in grey. The straight red line corresponds to the profile along which the analysis is conducted.

4.2. Application of the meta-modelling strategy

We generate 25 different configurations of input parameters using the LHS technique. Simulations of 10 years of injection (at a constant injection pressure of ~50 % the initial pore pressure) are conducted using the massively parallelized version TOUGH2-MP simulator (Zhang et al. [9]). The CPU time for a single simulation ranges from 5 to 10 days (depending on the values of the input parameters) using 35 CPU running in parallel. The over- pressure at 5 km from the injector ranges from ~3 to ~5 bars.

Using the training data, we approximate the over- pressure at 4 km using a Gaussian process. More precisely, due to the complex relationship between the input parameters and the over- pressure, we use a

Gaussian process of type "treed" as recently introduced by Gramacy and co- authors (see Gramacy and Herbert [7] and references therein). Figure 3A depicts the comparison for the cross-validation procedure between the observed over-pressure values (i.e. the ones simulated with the long running model) and the approximated ones. The closer the dots from the first bisector (outlined in blue), the better the approximation. This validation phase provides a coefficient of determination of around ~87%, which can be considered satisfactory despite the poor estimations of some observations (in particular at low over-pressure, see the leftmost part of Figure 3A).

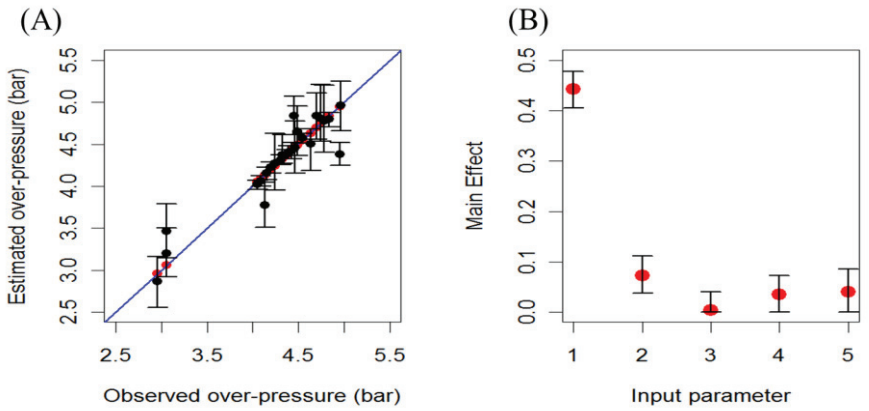

Fig. 3. (A) Comparison between the observed and the approximated (estimated) over- pressures (bar) for the cross-validation procedure. The error- bars correspond to the 95% confidence interval provided by the (treed) Gaussian Process. (B) Main effects estimated for the 5 input parameters using the (treed) Gaussian process. The red dots represent the mean of the main effects and the error bars correspond to the 95% confidence interval representing the uncertainty associated with the construction of the Gaussian Process. See Table 2 for description of the input parameter N°1 to 5.

Finally, using the validated meta-model, we conduct GSA and evaluate the main effects of each of the 5 input parameters on the over- pressure. A Monte-Carlo-based approach is used requiring a total number of 14,000 runs. Obviously, directly using 3-dimensional model would not have been achievable regarding the CPU time of a single simulation (>5 days).

Treating the Gaussian- Process in the Bayesian framework allows us to account for the uncertainty in the construction of the approximation, which is summarized by a 95% confidence interval assigned to each sensitivity measure (Figure 3B). As expected, the pore compressibility (input parameter n°1) has the strongest influence with a main effect of ~0.45, while the others remain below 0.1. Due to the uncertainty introduced by the approximation, the ranking of the input parameters N°2 to 5 is made difficult, because of the overlapping confidence intervals.

Noteworthy, a small value for the main effect does not necessarily mean that the input parameter has a negligible influence on the results. The sum of all the sensitivity measures reaches ~0.60 indicating the high non- linearity of the relationship between the over- pressure and the 5 input parameters. Computing the total effects (see Saltelli et al. [2]) indicates that none of the input parameters can be neglected, because they all reach values above 0.10 (not shown for sake of space).

5. Concluding remarks and further work

In the present study, we investigated the applicability of the meta-modelling technique to conduct global sensitivity analysis of large scale models sued to assess large-scale impacts associated with CO_2

storage operations. Such models are characterized with a very large number of grid cells (>500,000) and high computation time cost (> several hours). The application case is the computationally intensive multiphase flow modeling (with CPU time > 5 days for a single simulation) of an industrial-scale CO_2 injection (i.e. > 1 Mt/y) into a fluvial heterogeneous reservoir. We showed how to estimate the sensitivity measures of each of the input parameters using only a low number of simulations (of the order a few tens). One limitation of the proposed is to introduce a new kind of uncertainty associated with the construction of the approximation. Thus, the need for accounting for the approximation uncertainty in the sensitivity results is highlighted.

The analysis is conducted using a scalar model output (the value of the over- pressure at 5 km from the injector). The next challenge is to deal with spatially-varying or/and time-dependent outputs (e.g., temporal evolution of the over- pressurized area). Further works on this issue can for instance rely on the recent developments of Marrel et al. [10].

Acknowledgements

This study was supported by the BRGM's Research project "S.I.N.G.E." integrated in the Directorate of Research project CSCR03.

References

[1] IPCC (Intergovernmental Panel on Climate Change), 2005. IPCC Special Report on Carbon Dioxide Capture and Storage. Cambridge University Press, New York, USA.

[2] Saltelli A, Ratto M, Andres T, Campolongo F, Cariboni J, Gatelli D, Saisana M, Tarantola S. *Global sensitivity analysis: The Primer*. Chichester (UK): Wiley; 2008.

[3] Storlie CB, Swiler LP, Helton JC, Sallaberry CJ. Implementation and evaluation of nonparametric regression procedures for sensitivity analysis of computationally demanding models. *Reliability Engineering and System Safety* 2009; **94**:1735–1763.

[4] Manceau JC, Rohmer J. Analytical Solution Incorporating History-Dependent Processes for Quick Assessment of Capillary Trapping During CO2 Geological Storage. *Transp Porous Med* 2011; **90**:721-740.

[5] Issautier B, Viseur S, Audigane P. Impacts of fluvial sedimentary heterogeneities on CO2 storage performance AGU Fall Meeting 2011, Session H24B. Heterogeneity and Geologic Storage of CO2 II, San Franciso : USA 2011.

[6] McKay MD, Beckman RJ, Conover WJ. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* 1979; **21**:239–245.

[7] Gramacy RB, Herbert KHL. Adaptive Design and Analysis of Supercomputer Experiments. *Technometrics* 2009; **51**:130-145.

[8] Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer-Verlag; 2009.

[9] Zhang K, Wu YS, Pruess K, User's Guide for TOUGH2-MP. A massively parallel version of the TOUGH2 code, LBNL-315E, Lawrence Berkeley National Laboratory Report 2008.

[10] Marrel A, Iooss B, Jullien M, Laurent B, Volkova E. Global sensitivity analysis for models with spatially dependent outputs. *Environmetrics* 2011; 22:383-397.