



This is a repository copy of *Gaussian process emulators for quantifying uncertainty in CO2 spreading predictions in heterogeneous media*.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/115569/>

Version: Accepted Version

Article:

Tian, L., Wilkinson, R.D. orcid.org/0000-0001-7729-7023, Yang, Z. et al. (3 more authors) (2017) Gaussian process emulators for quantifying uncertainty in CO2 spreading predictions in heterogeneous media. *Computers and Geosciences*, 105. pp. 113-119. ISSN 0098-3004

<https://doi.org/10.1016/j.cageo.2017.04.006>

Article available under the terms of the CC-BY-NC-ND licence (<https://creativecommons.org/licenses/by-nc-nd/4.0/>)

Reuse

This article is distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) licence. This licence only allows you to download this work and share it with others as long as you credit the authors, but you can't change the article in any way or use it commercially. More information and the full terms of the licence here: <https://creativecommons.org/licenses/>

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



eprints@whiterose.ac.uk
<https://eprints.whiterose.ac.uk/>

1 Gaussian Process Emulators for Quantifying
2 Uncertainty in CO₂ Spreading Predictions in
3 Heterogeneous Media

4 Liang Tian^a, Richard Wilkinson^b, Zhibing Yang^a, Henry Power^c, Fritjof
5 Fagerlund^a, Auli Niemi^a

6 ^a*Air, Water and Landscape Sciences, Department of Earth Sciences, Villavägen 16, SE-752*
7 *36 Uppsala University, Sweden*

8 ^b*School of Mathematics and Statistics, University of Sheffield, UK*

9 ^c*Faculty of Engineering, University of Nottingham, University Park, Nottingham NG7*
10 *2RD, UK*

11 **Abstract**

12 We explore the use of Gaussian process emulators (GPE) in the numerical sim-
13 ulation of CO₂ injection into a deep heterogeneous aquifer. The model domain
14 is a two-dimensional, log-normally distributed stochastic permeability field. We
15 first estimate the cumulative distribution functions (CDFs) of the CO₂ break-
16 through time and the total CO₂ mass using a computationally expensive Monte
17 Carlo (MC) simulation. We then show that we can accurately reproduce these
18 CDF estimates with a GPE, using only a small fraction of the computational cost
19 required by traditional MC simulation. In order to build a GPE that can pre-
20 dict the simulator output from a permeability field consisting of 1000s of values,
21 we use a truncated Karhunen-Loève (K-L) expansion of the permeability field,
22 which enables the application of the Bayesian functional regression approach.
23 We perform a cross-validation exercise to give an insight of the optimization
24 of the experiment design for selected scenario: we find that it is sufficient to
25 use 100s values for the size of the training set and that it is adequate to use as
26 few as 15 K-L components. Our work demonstrates that GPE with truncated
27 K-L expansion can be effectively applied to uncertainty analysis associated with
28 modeling of multiphase flow and transport processes in heterogeneous media.

29 *Keywords:* Permeability heterogeneity, Karhunen-Loève expansion, Monte
30 Carlo, Uncertainty analysis, Gaussian process emulator

1 **1. Introduction**

2 Planning and operation of a carbon dioxide capture and storage (CCS)
3 project requires reliable model predictions concerning the fate of the stored
4 CO₂. Carefully conducted numerical simulations are critical for the under-
5 standing of the associated coupled physical and chemical processes (Pruess and
6 García, 2002; Juanes et al., 2006; Doughty, 2007; Dai et al., 2016; Bacon et al.,
7 2016; Xiao et al., 2016). An important additional complication arises from the
8 geological heterogeneity of the target formation, such as stratigraphic archi-
9 tecture and facies distribution, which is difficult to estimate from the limited
10 number of observations available (i.e., from the sparse networks of primarily
11 vertical investigation wells) in a deterministic manner (Ambrose et al., 2007;
12 Tsang et al., 2008; Gershenson et al., 2015; Ritzi et al., 2016; Tian et al., 2016b;
13 Ampomah et al., 2016). Therefore, robust and computationally effective meth-
14 ods for dealing with the uncertainty arising from the geological heterogeneity
15 are in great need. In general, two components contribute to the modelling
16 uncertainty for CO₂ geological storage: (1) input uncertainty, including the
17 aforementioned parameter uncertainties (unknown geology), and (2) model un-
18 certainty, or “structural uncertainty” according to the conventional hydrological
19 modelling terminology (Renard et al., 2010), as modelling approaches are de-
20 veloped under different conceptual and methodological frameworks, involving
21 various approximations and simplifications. An example on the latter is the
22 work reported by Nordbotten et al. (2012), where a benchmark simulation case
23 was run with various numerical codes and effort was made to evaluate the signif-
24 icance of deviated solutions from various modelling strategies and assumptions.
25 In the present work, we focus on the input uncertainty.

26 Standard geostatistical techniques are used to resolve the input uncertainty
27 when evaluating reservoir CO₂ storage performance. For example, the Um-
28 brella Point power plant model (based on the Frio formation) was created using
29 TProGs program by Doughty and Pruess (2004) where multiple two-dimensional

1 stochastic representations of fluvial depositional settings were picked deliber-
2 ately to reproduce realistic three-dimensional geologic structures. A sequential
3 indicator simulation approach was used by Flett et al. (2007) to create realistic
4 shale facies distribution for 3-D notional marine sand system models with vary-
5 ing net-sand-to-gross-shale ratios. A sequential Bayesian simulation technology
6 was used by Claprood et al. (2014) in constructing a porosity distribution for a
7 3-D model of Beauharnois Formation to understand its CO₂ storage potential.
8 In terms of the characterization of the spatial permeability distribution, Han
9 et al. (2010) created multiple two-dimensional permeability fields with inclusion
10 of low permeability lenses using a sequential Gaussian simulation approach. Dis-
11 cussions on effects of the permeability heterogeneity include the contributions
12 from Jahangiri and Zhang (2011) with a focus on the plume distribution, and
13 from Lengler et al. (2010) with a focus on small-scale heterogeneity ($< 100m$).
14 Using a macroscopic invasion percolation model, Yang et al. (2013) performed a
15 detailed parametric sensitivity study on upscaled capillary pressure-saturation-
16 relative permeability relationships for CO₂ migration in multimodal heteroge-
17 neous media. A more recent sensitivity study was reported by Tian et al. (2016a)
18 where the parameters controlling the spatial correlation structures of the per-
19 meability fields were systematically analysed so as to understand their effects
20 on CO₂ storage performance.

21 A Monte Carlo simulation method is normally used when a deterministic
22 description of the model input cannot be used (James, 1980). In this approach,
23 multiple, mutually different but equiprobable realizations of the parameter field
24 are generated, the model problem simulated for all of them, and the output
25 analysed in terms of the statistics of the outputs. The method has been proved
26 viable for the simulation of geological storage of CO₂ (Jahangiri and Zhang,
27 2011; Deng et al., 2012; Dai et al., 2014; Tian et al., 2016a). However, an obvi-
28 ous limitation for the method is the high computational cost, which limits the
29 number of possible runs for large-scale, long-term simulations of CO₂ migration
30 in 3-D heterogeneous medium. This in turn violates the underlying criteria of
31 the Monte Carlo approach, which require the model to be run at many input

1 configurations in order to accurately infer the uncertainty in the model pre-
2 dictions. Therefore, new reduced-order models that can capture the essential
3 behaviour of the fully physically based models, yet avoiding the prohibitive com-
4 putational cost of them are of great interest. A general overview on surrogate
5 modelling in water resources was given by Razavi et al. (2012). More recently,
6 Liu et al. (2013) developed geostatistical reduced order models (GROMs) in
7 the parameter domain to solve under-determined inverse problems addressing
8 subsurface multiphase transport.

9 In this paper, we propose a Bayesian approach for uncertainty analysis (UA),
10 that is, the forward propagation of uncertainty through a model. We focus
11 on simulators such as TOUGH2 / ECO2N (Pruess et al., 1999; Pruess and
12 Spycher, 2007), which are used for the numerical simulation of CO₂ injection
13 into deep heterogeneous aquifers. These numerical models (called the *simulator*)
14 are deterministic, meaning they will always produce the same output if the input
15 is known exactly, and thus can be regarded as mathematical functions $f(\cdot)$. As
16 we are uncertain about the input Z (i.e., the true permeability is unknown),
17 this uncertainty is transferred to $f(Z)$, so that we are uncertain about the best
18 prediction. The objective of uncertainty analysis is therefore to estimate the
19 distribution of $f(Z)$, given a distribution for inputs Z .

20 **2. Methodology**

21 We present the modeling problem and describe the quantities of interest in
22 Section 2.1. In Section 2.2, we present the method to simulate the random
23 permeability field. In Section 2.3, we describe the Gaussian process emulation
24 (GPE) methodology and its application to our problem. A complete procedure
25 to our implementation of GPE is given in Section 2.4. In Section 2.5 we describe
26 the use of GPE for uncertainty analysis.

27 *2.1. Modelling of CO₂ migration in a heterogeneous aquifer*

28 We consider supercritical CO₂ injection from a vertical borehole, and we
29 simulate CO₂ migration until the CO₂ plume front reaches the monitoring well

1 at the far end of the domain (Fig. 1). The simulations are performed using the
 2 TOUGH2/ECO2N code (Pruess et al., 1999; Pruess and Spycher, 2007). The
 3 quantities of interest are the breakthrough time (BT) and the total mass (TM)
 4 of the injected CO₂. For the numerical experiments where we want to address
 5 the uncertainty caused by heterogeneity, we vary the correlation length of the
 6 randomly generated permeability fields, but use a fixed standard deviation (see
 7 2.2). A more detailed description is given in the *Supplementary Information*(SI).

8 In this work, we use the notation Z to denote the permeability spatial field
 9 and want to find the distribution of $f(Z)$ given the distribution of Z , where $f(\cdot)$
 10 represents the simulator output (e.g., either the total mass or the breakthrough
 11 time of the CO₂). In other words, our objective is to estimate the cumulative
 12 distribution functions (CDFs)

$$F(y) = \mathbb{P}(f(Z) \leq y). \quad (1)$$

13 The CDFs can be estimated using a Monte Carlo (MC) approach if sufficient
 14 computer power is available. If Z_1, \dots, Z_n is a large sample from log-Gaussian
 15 random field (log-GRF) we are using to model the heterogeneous permeability
 16 field, then the empirical CDF (ECDF),

$$\hat{F}(y) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{f(Z_i) \leq y}, \quad (2)$$

17 is an unbiased estimator of the CDF. Here, \mathbb{I}_A is an indicator function taking
 18 value 1 if event A occurred and 0 otherwise.

19 2.2. Modelling the heterogeneous permeability field

20 We consider a representation of Z on a two-dimensional mesh grid with a
 21 finite resolution 100×20 . The x in the notation $Z(x)$ is the location coordinate
 22 vector, emphasizing that Z is location dependent. Our prior model for Z is

$$\log Z \sim N(\mu, \Sigma), \quad (3)$$

23 where we specify Σ through a covariance function that describes the permeabil-
 24 ity covariance between any two locations in the domain, i.e., $\Sigma_{ij} = c(x_i, x_j)$



Figure 1: Conceptual model of the simulation domain (Tian et al., 2016a)

1 for some covariance function c , and spatial locations x_i and x_j . Several tech-
 2 niques exist to simulate realisations from this distribution, including circulant
 3 embeddings, Karhunen-Loève expansions and stochastic collocation (Graham
 4 et al., 2011). The method of Karhunen-Loève (K-L) decomposition is used in
 5 our work. The Karhunen-Loève theorem says that $Z(x)$ admits a representation
 6 of the form

$$Z(x) = \sum_{i=1}^{\infty} \xi_i \lambda_i \phi_i(x) \quad (4)$$

7 where the λ_i and $\phi_i(x)$ are the ordered eigenvalues and eigenfunctions of the
 8 covariance function respectively, and the ξ_i are independent $N(0, 1)$ random
 9 variables. Note that if interest lies solely in the value of Z on a finite grid of
 10 n values (as in our case), then this reduces to a finite sum of n terms, and the
 11 K-L decomposition provides an exact decomposition of the correlation function
 12 on the discrete grid (Crevillén-García et al., 2017). To reconstruct $Z(x)$, only
 13 the $\{\xi_i\}_{i=1}^n$ need to be saved, since λ_i and ϕ_i are determined by the covariance
 14 function and thus remain the same throughout the uncertainty analysis. The
 15 simulator is then considered as a function of $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)^\top$ instead of Z , i.e.,
 16 $f(Z) \equiv f(\boldsymbol{\xi})$.

17 In order to calculate the CDFs of the target quantities and evaluate the
 18 performance of the GP emulator, two datasets are generated for each of three
 19 selected scenarios where we vary the correlation-length of the unknown per-
 20 meability fields (Table 1, first three rows). The first dataset consists of 10^4
 21 input-output pairs and is used to produce a MC estimate of the CDF; the sec-

1 ond dataset consists of a smaller number of numerical simulations and is used
 2 for training the emulator. The overall procedure is illustrated in Fig. 2 and is
 3 further explained in the following section.

4 2.3. Gaussian process emulation

5 An emulator (Kennedy and O’Hagan, 2000) is a statistical model that closely
 6 mirrors a simulator. It is built using an ensemble of input-output pairs $\{X_i, y_i\}_{i=1}^N$
 7 and can be used to predict the simulator output for any new input. The most
 8 popular approach to building emulators is to use a Gaussian process (GP) (Ras-
 9 mussen and Williams, 2006), which are equivalent to the *kriging* models used
 10 in geostatistics (Stein, 1999). Gaussian processes describe an infinite collec-
 11 tion of random variables, and can be thought of as distributions over functions
 12 (Rasmussen and Williams, 2006; Crevillén-García et al., 2017). A GP is fully
 13 specified by its mean and covariance functions (Rasmussen and Williams, 2006).

14 In our case, direct application of GP would be computationally costly for that
 15 a 2,000 dimensional input space would require thousands of training samples (as
 16 the *hyperparameters* associated with each input component are estimated from
 17 the simulator data by solving an optimisation problem, e.g., Crevillén-García
 18 et al., 2017). Instead, we can construct a GP emulator by exploiting the spatial
 19 structure in Z provided by the exact decomposition of Z on a discrete grid. If
 20 we order the eigenvalues in Eq.(4) so that $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$, then we can
 21 achieve a form of data compression by truncating the expansion to the first d
 22 terms

$$\tilde{Z}(x) = \sum_{i=1}^d \xi_i \lambda_i \phi_i(x), \quad (5)$$

23 and thus representing the permeability in a lower dimensional space. This trun-
 24 cation explains the most variance and achieves the minimum mean square error
 25 amongst all such approximations. We exploit this truncation in order to build
 26 a reduced order emulator from \tilde{Z} rather than Z , which is equivalent to building
 27 an emulator with input $\boldsymbol{\xi} = (\xi_1, \dots, \xi_d)^\top$.

28 The emulator requires the simulator to be run a small number of times
 29 (n_{train}) at carefully selected inputs (design points) to create a set of *training*

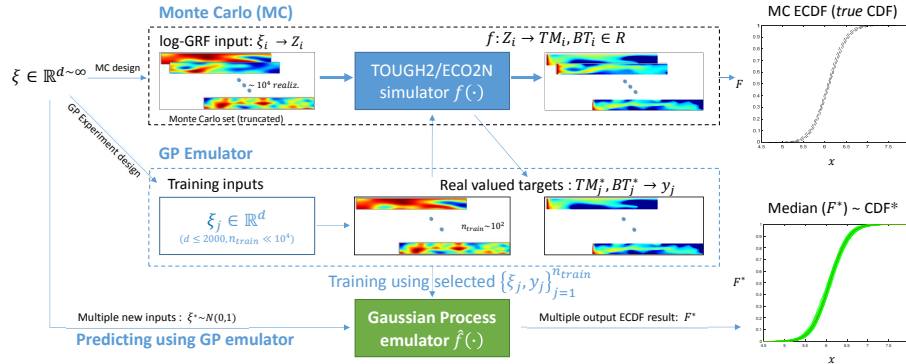


Figure 2: Comparing procedures for estimating CDFs using Monte Carlo simulation (TOUGH2/ECO2N) and Gaussian process emulation. The thickness of the arrow illustrates the relative computational cost.

1 *inputs* (See Fig.2). Because the simulation of Z is based on a truncated K-
2 L expansion, the training ensemble is a set $\{\xi_i, y_i\}_{i=1}^{n_{train}}$ where each $\xi_i \in \mathbb{R}^d$.
3 Space-filling designs (McKay et al., 1979; Morris and Mitchell, 1995) are recom-
4 mended for GP models, as GP predictions essentially interpolate based on the
5 distance to a few of the nearest training points. We use the maximin Latin hy-
6 percube designs which maximise the minimum distance between any two points
7 in the training set. We will examine the optimal value of d and n_{train} using
8 predictive performance measures in Section 4.

9 The implementation of GPs require that we specify prior mean and co-
10 variance functions. We use a constant mean function and choose between the
11 squared exponential and Matérn covariance functions. The *hyperparameters* in-
12 volved in these two terms are estimated through *training* using type II maximum
13 likelihood (Rasmussen and Williams, 2006). We use the GPstuff implementa-
14 tion of Gaussian processes (Vanhatalo et al., 2012), which are a set of MATLAB
15 codes integrating Gaussian process models for Bayesian analysis. Notice that
16 the GP covariance function (also called the kernel) should be distinguished from
17 the one mentioned earlier in describing the spatial correlation of the permeabil-
18 ity field.

1 *2.4. GP emulation with K-L truncation*

2 We summarize the procedure as follows:

- 3 1. Choose design $\xi_{i=1}^n$ using a maximin Latin hypercube design where $\xi \in R^N$
 4 2. Run simulator to obtain training set $\{\xi_i, y_i\}_{i=1}^n$. We then truncate each ξ
 5 to the first d elements. The value of d will be optimized in Step 6.
 3. Pick a prior mean function $m(\xi) = \mathbb{E}[\hat{f}(\xi)]$ and covariance function $k(\xi, \xi') = \text{Cov}(\hat{f}(\xi), \hat{f}(\xi'))$ where $\hat{f}(\cdot)$ is the emulator. For example, the *square exponential* (SE) covariance function is

$$k(\xi, \xi') = \sigma^2 \exp\left(-\frac{1}{2} \frac{|\xi - \xi'|^2}{\lambda}\right)$$

where λ is a length scale hyper parameter, and σ^2 a variance parameter.

We denote the GP prior by:

$$\hat{f}(\xi) \sim \mathcal{GP}(m(\xi), k(\xi, \xi')).$$

4. Update the GP to find the posterior mean (m^*) and covariance functions (k^*) using equations:

$$m^*(\xi) = m(\xi) + t(\xi)^\top K^{-1}(\mathbf{y} - \mathbf{m}),$$

$$k^*(\xi^*, \xi^*) = k(\xi, \xi) - t(\xi)^\top K^{-1}t(\xi)$$

6 where $K_{ij} = k(\xi_i, \xi_j)$ is the Gram matrix, $t(\xi)^\top = (k(\xi_1, \xi), \dots, k(\xi_n, \xi))$,
 7 and \mathbf{m} and \mathbf{y} are the vectors of simulator responses and their prior mean
 8 for the emulator. Note that the posterior is a GP conditioned on the
 9 training set.

- 10 5. Optimize the hyperparameters, such as λ, σ^2 in SE, by maximising the
 11 type II maximum likelihood (see Rasmussen and Williams, 2006).
 12 6. Optimize the choice of d , the covariance function, etc, using cross-validation
 13 to estimate a measure of the predictive performance.

14 *2.5. Using GP for UA*

15 Once we have a GP emulator of the simulator, we can use it to predict the
 16 simulator CDF and to quantify the uncertainty in our estimate. To estimate

1 the CDFs, we use the procedure suggested in Oakley and O’Hagan (2002). This
 2 involves drawing sample functions $\{f_j\}_{j=1}^L$ from the GP that are consistent with
 3 the training data by adding in new design points $\{\xi_i^*\}_{i=1}^{1000}$, and simulating a value
 4 for the response from the GP emulator. We then update the emulator to take
 5 into account the fake simulated data. The placement and number of additional
 6 design points is chosen so as to make the uncertainty in the simulated functions
 7 f_j essentially zero. We then estimate the CDF for each simulated function
 8 using Monte Carlo in the usual manner, giving us L realizations F_1^*, \dots, F_L^* .
 9 From this we use the median of the CDFs as a point estimate, and can calculate
 10 uncertainty about our estimates using the ensemble of CDFs.

11 3. Results

12 3.1. Estimating the CDF

13 Each quantity of interest (total mass (TM) or breakthrough-time (BT)) from
 14 each of the three cases (three different models for the unknown permeability
 15 field) is considered as a standalone problem. As the training set is based on a
 16 Latin hypercube design, we use a fixed number of training points (Table 1) to
 17 construct each of the three GP structures. For each emulated ECDF curve, 1,000
 18 random sample points are first generated using a pseudorandom number (vector)
 19 generator in Matlab assuming a dimension corresponding to $d_{train} = 30$ (Case
 20 1) or $d_{train} = 20$ (Case 2 and 3). Then, this set of random inputs, together with
 21 the corresponding training pairs, were used to feed the designated GP structure
 22 in order to produce / draw one sample from the posterior distribution. For each
 23 quantity of interest, 100 posterior samples ($L = 100$) were used to calculate the
 24 median ECDF. Note that this is computationally cheap as it does not involve
 25 running the TOUGH2/ECO2N simulator

26 Fig. 3 shows the breakthrough time for *Case 1*. The GP curve is the median
 27 CDF calculated from the 100 posterior samples. The confidence intervals of
 28 the MC CDF are omitted for visual clarity. The dashed lines (posterior credible
 29 intervals) indicate that the MC CDF is enveloped within the emulator confidence

Table 1: Case specifications and results for model selection

Case No.		1	2	3
Correlation length		0.075	0.15	0.30
size of MC set	N_{MC}	10,000	10,000	10,000
size of training set	n_{train}	800	400	400
dimension of the training set	d_{train}	30	20	20
$CRPS_{BT,Matérn}$		0.00640	0.00193	0.00153
$CRPS_{BT,SE} (d_{train} = 20)$		0.00108	0.00187	0.00135
$CRPS_{TM,Matérn}$		0.00490	0.00766	0.00975
$CRPS_{TM,SE} (d_{train} = 20)$		0.02489	0.02508	0.02534

1 intervals. Excellent matches are observed: for all cases examined, the median
 2 GP curves replicate the MC ones almost exactly. The mean CRPS (Continuous
 3 Rank Probability Score, see the SI) for the three correlation length cases are
 4 0.00640, 0.00193 and 0.00153, respectively. A similar procedure was used for
 5 the total CO₂ mass (TM) at the breakthrough time. The TM ECDF curves
 6 from the MC are also well predicted by the median GP results. The TM result
 7 exhibits a slightly less good match in comparison to the observation from the
 8 BT, especially for the lower and upper tail of the ECDF. However, the 5th to
 9 the 95th percentiles of the GP prediction agree closely with the MC results. The
 10 CRPSs for three tested cases are, respectively, 0.00490, 0.00766 and 0.00975.

11 Note that for TM smaller CRPSs are observed for Case 1 in compari-
 12 son to the other cases (Table 1) due to a larger number of training points
 13 ($n_{train,case1} = 800$) and the higher dimension of the training inputs ($d_{case1} = 30$
 14 KL components). Note also that the CRPSs for BT are noticeably smaller in
 15 comparison to the TM ones (one order of magnitude). Excellent agreement is
 16 observed for BT results (Fig. 3). For Case 2 and Case 3, the results are visually
 17 similar to Case 1 and are therefore not included for space considerations.

18 3.2. Cross validation

19 At the initial stage of the experimental design, two key factors are very dif-
 20 ficult to determine beforehand, namely the size of the training set (n_{train}) and

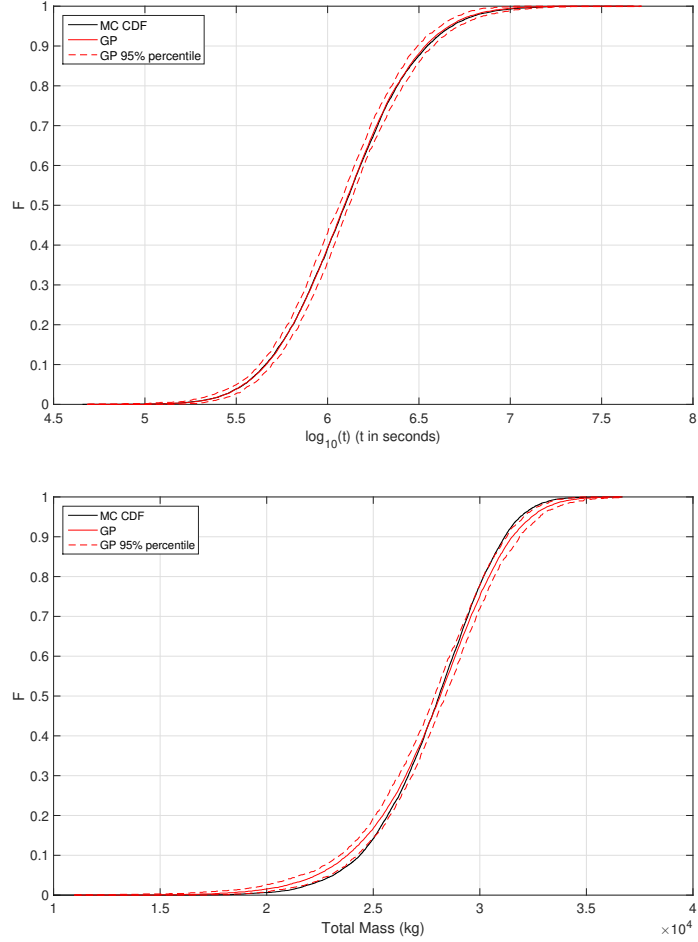


Figure 3: Comparison of GP emulation vs. Monte Carlo simulations. Top: breakthrough time (BT) recorded in seconds; bottom: the total mass of CO₂ (TM).

1 its dimension (d_{train} , the number of K-L components retained for the predic-
 2 tion). Using leave-one-out cross validation (LOO-CV, see also SI) can guide us
 3 in tackling these issues. For each GP, LOO-CV has been performed to estimate
 4 the predictive accuracy of the emulator in two steps: Step 1, a training set with
 5 fixed size is selected and the predictive performance measured using the Dawid
 6 score (DS), which can be thought of as being similar to the log-likelihood (see

1 Wilkinson et al., 2011, and the SI). This score is then plotted as a function of
2 the number of K-L components; Step 2, the number of K-L components is now
3 fixed and the predictive performance is plotted as a function of the size of the
4 training set.

5 The DS estimated using LOO-CV are plotted as a function of the number of
6 K-L components in Fig. 4. It is found that by using a fixed size of the training
7 set for all cases, the DS score becomes stabilized when using more than 15 K-L
8 components ($d_{train} \geq 15$). When using exactly 15 K-L components for each
9 case to fit the GPs, the DS score appears to become stabilized when using a
10 training set with more than 100 design points ($n_{train} \geq 100$, see Fig. 5).

11 4. Discussion

12 The investigated two dimensional model domain has 2,000 elements rep-
13 resenting a spatially correlated heterogeneous permeability field. Uncertainty
14 analysis using the classical MC method requires that the already computational
15 demanding simulator to be run for as many as 10^4 times. For the GP emulator
16 approach to UA, the main part of computational cost comes from the simulator
17 runs needed for the training inputs. GP posterior sampling has in comparison
18 virtually no computational cost. In this section we discuss the design and the
19 construction of the GP emulator.

20 4.1. Model configuration

21 One very important aspect of using GP emulation is the choice of the covari-
22 ance function that defines the nearness or similarity in the input space (Ras-
23 mussen and Williams, 2006). In other words, how similar $f(\mathbf{x})$ is likely to be to
24 $f(\mathbf{x}')$ when \mathbf{x} is close to \mathbf{x}' . The covariance function can be any positive definite
25 function, so that it generates a valid covariance matrix for any set of inputs.
26 Some of the commonly used functions are the squared exponential covariance
27 function (SE) and the Matérn class of covariance functions. The SE covariance
28 function generates samples that are infinitely differentiable, whereas the Matérn

1 covariance function (with $\nu = \frac{3}{2}$ degrees of freedom) generates samples that are
 2 only once differentiable. It can be hard to judge in advance what the more
 3 appropriate model might be, but we can use CV scores to guide the choice. We
 4 constructed alternative GPs using both for each of the cases examined in Section
 5 3 (see Table 1). The ECDFs calculated using the Matérn covariance function
 6 ($\nu = \frac{3}{2}$) exhibit smaller CRPS values in comparison to the ones calculated using
 7 SE. For the emulation of BT, there is no noticeable difference between using the
 8 SE or Matérn covariance functions. However, for TM the Matérn exhibits much
 9 better predictive performance. Notice that the choice of d_{train} (the dimension
 10 of training points, in our case equivalent to the number of K-L components) will
 11 affect the performance of the GP emulator, depending on the number of train-
 12 ing points (n_{train}). We note that the choice of covariance function can affect
 13 the performance of the GPE, and that more complex covariance functions can
 14 be obtained by combining covariance functions (see Rasmussen and Williams,
 15 2006, for example). A detailed discussion is beyond the scope of the current
 16 work, but can be found in Crevillén-García (2016).

17 4.2. Cross-validation and optimization

18 We would like to use the smallest number of the training inputs possible to
 19 create an emulator that meets our accuracy requirements. To investigate this,
 20 we use the method of cross-validation (CV). The idea is to split the training set
 21 into two disjoint sets, one of which is used for the training and the other is used
 22 for the validation of the emulator. Notice that such splits can be done repeatedly
 23 in multiple ways (k -fold CV), one extreme case is when $k = n$, also known as
 24 leave-one-out cross-validation (LOO-CV). We can use CV scores to choose the
 25 optimum input dimensionality (the number of K-L coefficients, d_{train}) and the
 26 number of design training points (n_{train}) to be used in the GP. The evaluation
 27 is done by looking at the variance of the predicted value in LOO-CV as well as
 28 the Dawid score for the overall prediction error.

29 In our calculations, the size of the training ensemble is 800 for Case 1 but
 30 400 for Case 2 and Case 3. The reason for using more training sets in Case

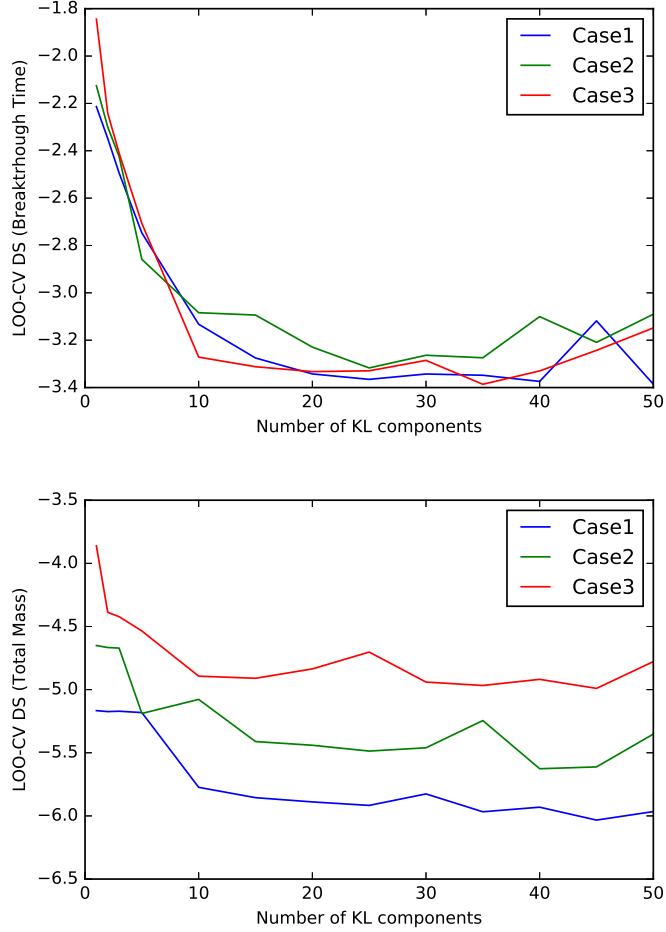


Figure 4: Dawid scores indicating prediction accuracy (estimated using LOO-CV) vs. number of K-L components retained (d_{train}).

1 1 is that the correlation length for the log-Gaussian permeability field model
 2 is smaller in Case 1 than in Cases 2 and 3. Thus, the permeability varies
 3 over shorted distances, and so we need more K-L components to describe the
 4 variation well, and consequently we need a larger training ensemble to build
 5 an adequate emulator. For predicting the BT ECDF (Fig. 4), using 15 K-
 6 L components provides good results, whereas for predicting the TM ECDF,

1 around 20 K-L components is preferred. The indication is that the calculations
 2 of breakthrough time and total mass for the injection simulation of CO₂ are two
 3 very different processes.

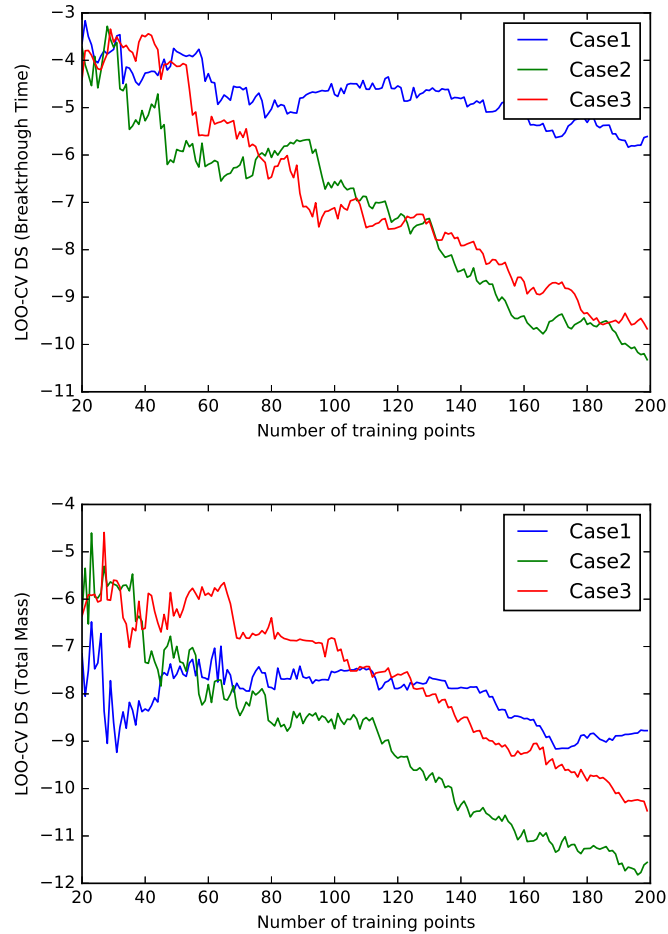


Figure 5: *LOO-CV Scores vs. the size of the training set (n_{train}).*

4 A priori, it is difficult to provide a precise value for an adequate or appropri-
 5 ate number of training points required for a GP, as, to the best of our knowledge,
 6 a priori estimation of the error is not possible for GPs. Optimization of the de-
 7 sign would mean changing the space filling design, which would mean drawing

1 new samples ξ_i from $\mathbb{R}^{d=2000}$. To understand whether this design improved the
 2 GP performance, the simulator (TOUGH2/ECO2N) would need to be rerun
 3 so as to generate the corresponding new training ensemble. In other words,
 4 one would need to build new GPs based on additional simulator runs in order
 5 to understand the potential gain from optimization. This would be extremely
 6 computationally costly, and so a different approach has been used here.

7 Considering Case 1, for example, where we have generated 800 training pairs
 8 ($n_{train} = 800$), we start by building an emulator, $GP_{0,j=20}$, using a random
 9 draw (whilst trying to retain some of the space filling properties of the design)
 10 of $j = 20$ training points from initial set of 800. A first DS score can then
 11 be calculated for $GP_{0,j=20}$ using LOO-CV. By randomly adding one training
 12 point at a time from the remaining training pairs, we can iteratively create
 13 new emulators, $GP_{i,j=20+i}$. The resulting Dawid scores then reflect how the
 14 predictive performance improves as the sample size increases. It should be noted
 15 that Latin-hypercube sampling has been used to create the initial 800 points.
 16 The re-sample of the existing Latin-hypercube set should be path-independent.
 17 Fig. 5 shows the decreasing trend of DS score reflecting that more information
 18 is provided by the training set as the sample size increases. It can be seen that
 19 100 training pairs would be needed for Case 1 when building a GP for BT ECDF
 20 using only 15 K-L components. Note that the pattern of TM LOO-CV result
 21 for Case 1 (Fig. 5, lower panel) is different from the other cases. We further
 22 extended the LOO-CV test for Case 1 and the decreasing trend in the DS score
 23 was confirmed (Fig. 6). This indicates that for heterogeneous domain with a
 24 smaller correlation length, a larger training set may be needed for constructing
 25 the GP so as to achieve a similar predictive performance.

26 *4.3. Using GP for uncertainty analysis*

27 The output from each GP constructed in Section 3 is a collection of ran-
 28 dom variables indexed by ξ . An assumption has been made that the spatial
 29 distribution of the heterogeneous field can be adequately described by ξ . In a
 30 geostatistics perspective, the conventional perception of correlation length (λ),

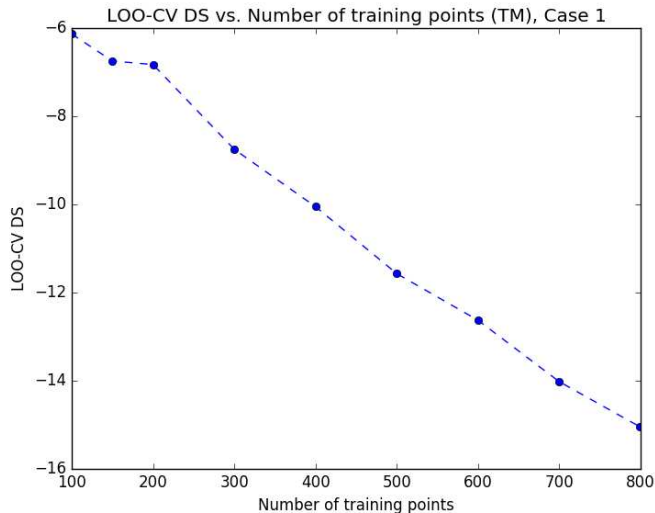


Figure 6: *LOO-CV Scores vs. the size of the training set (n_{train}), Case 1*

1 standard deviation (σ) and the descriptive covariance function (see SI) of the
 2 permeability field can all be interpreted as possible projections of ξ .

3 We use standalone GPs in predicting the ECDF for each uncertain output
 4 of interest. It is worth noting that the two outputs, the breakthrough time
 5 and the total mass, are fundamentally different processes. Fig.3 shows that
 6 the breakthrough time is log-normally distributed, while the total mass follows
 7 a normal distribution. The GP emulator prediction is noticeably better for
 8 $\log_{10}(BT)$ than for TM. This difference in reproducing the MC results may
 9 indicate that the dependence of TM on the underlying permeability field is more
 10 complex than that of BT. Additional metrics apart from the K-L expansion
 11 parameter (or alternative methods) describing the permeability fields may be
 12 needed to improve the uncertainty analysis of the total CO₂ mass.

13 We have shown that the use of GP for UA, in our case exploring the ECDFs
 14 of BT and TM, results in considerably lower computational cost compared to
 15 classical MC analyses. By improving the experimental design, it is possible to
 16 further improve the model performance.

1 5. Concluding remarks

2 We have carried out uncertainty analysis of the simulations of CO₂ injection
3 and migration into a deep heterogeneous saline aquifer using both MC simula-
4 tion and GP emulation. We have shown how GPEs can successfully be used
5 to predict ECDFs of the breakthrough time and total CO₂ mass, replicating
6 the ECDF estimates obtained using Monte Carlo simulation, at only a small
7 fraction of the computational cost. The GPs automatically provide confidence
8 intervals for the estimates of the CDF, which compare well to those calculated
9 from classical MC. Our work demonstrates that GP emulators with truncated
10 Karhunen-Loève expansion can be effectively applied to uncertainty analysis
11 associated with modeling of multiphase flow and transport processes in hetero-
12 geneous media.

13 We have also examined the issues surrounding experimental design, including
14 the possibilities to further optimize the GP. An optimum design may need to
15 re-sample the input space, and therefore need additional simulator runs. To
16 address this, an alternative approach has been taken by down-sampling the
17 training set. The results from the cross-validation exercise indicate significant
18 performance gain from potential optimization. This information provides a good
19 starting point for further applications.

20 We have treated the two outputs, namely the CO₂ breakthrough time and
21 the total CO₂ mass as two independent processes, and built standalone GPs for
22 each one. It is possible to construct a single GP with multiple outputs (Alvarez
23 et al., 2011), and this may provide one future perspective for exploring the
24 internal physical mechanism for a complex system. Another future aspect would
25 be to use simulations of varying fidelity and then to use multilevel emulation to
26 further increase the accuracy of the GPE (cf. multi-level Monte Carlo in Giles
27 et al. (2015)).

28 We have also explored the indication from modelling of heterogeneous media
29 and identified that the conventional perception on correlation length is, from a
30 geostatistic perspective, a matter of parameter bounds and dimensions. Finally,

1 we note that future work is needed to address the limitation associated with the
2 use of truncated Karhunen-Loève expansion, which is a smooth representation
3 of the random field, for application to real reservoirs which often exhibit multi-
4 scale permeability heterogeneity.

5 **Acknowledgement**

6 The authors would like to acknowledge Prof. Andrew Cliffe at the School
7 of Mathematical Sciences, University of Nottingham for his invaluable contri-
8 butions to this work. Unfortunately Prof. Cliffe passed away during the prepa-
9 ration of this manuscript. He will be remembered for his great qualities as a
10 researcher and a friend.

11 The work has been supported by the European Community’s Seventh Frame-
12 work Programme (FP7) MUSTANG (No. 227286), PANACEA (No. 282900)
13 and TRUST (No. 309067) projects. The computations were performed on re-
14 sources at Uppsala Multidisciplinary Center for Advanced Computational Sci-
15 ence (UPPMAX) and at Chalmers Centre for Computational Science and En-
16 gineering (C3SE), both provided by the Swedish National Infrastructure for
17 Computing (SNIC).

18 **References**

- 19 Alvarez, M. A., Rosasco, L., Lawrence, N. D., 2011. Kernels for Vector-Valued
20 Functions: a Review. arXiv preprint arXiv:1106.6251 4 (3), 1–37.
- 21 Ambrose, W. A., Lakshminarasimhan, S., Holtz, M. H., Núñez-López, V., Hov-
22 orka, S. D., Duncan, I., 2007. Geologic factors controlling CO₂ storage capac-
23 ity and permanence: case studies based on experience with heterogeneity in
24 oil and gas reservoirs applied to CO₂ storage. *Environmental Geology* 54 (8),
25 1619–1633.
- 26 Ampomah, W., Balch, R., Cather, M., Rose-Coss, D., Dai, Z., Heath, J., Dew-
27 ers, T., Mozley, P., 2016. Evaluation of co₂ storage mechanisms in co₂ en-

- 1 hanced oil recovery sites: Application to morrow sandstone reservoir. *Energy*
2 & *Fuels* 30 (10), 8545–8555.
- 3 Bacon, D. H., Qafoku, N. P., Dai, Z., Keating, E. H., Brown, C. F., 2016.
4 Modeling the impact of carbon dioxide leakage into an unconfined, oxidizing
5 carbonate aquifer. *International Journal of Greenhouse Gas Control* 44, 290
6 – 299.
- 7 Claprood, M., Gloaguen, E., Sauvageau, M., Giroux, B., Malo, M., 2014.
8 Adapted sequential Gaussian simulations with Bayesian approach to evaluate
9 the CO₂ storage potential in low porosity environment. *Greenhouse Gases:*
10 *Science and Technology* 4 (6), 761–776.
- 11 Crevillén-García, D., 2016. Uncertainty quantification for flow and transport in
12 porous media. Ph.D. thesis, University of Nottingham.
- 13 Crevillén-García, D., Wilkinson, R., Shah, A., Power, H., 2017. Gaussian process
14 modelling for uncertainty quantification in convectively-enhanced dissolution
15 processes in porous media. *Advances in Water Resources* 99, 1 – 14.
- 16 Dai, Z., Stauffer, P. H., Carey, J. W., Middleton, R. S., Lu, Z., Jacobs, J. F.,
17 Hnottavange-Telleen, K., Spangler, L. H., 2014. Pre-site Characterization
18 Risk Analysis for Commercial-Scale Carbon Sequestration. *Environmental*
19 *Science & Technology* 48 (7), 3908–3915.
- 20 Dai, Z., Viswanathan, H., Middleton, R., Pan, F., Ampomah, W., Yang, C.,
21 Jia, W., Xiao, T., Lee, S.-Y., McPherson, B., Balch, R., Grigg, R., White,
22 M., 2016. CO₂ Accounting and Risk Analysis for CO₂ Sequestration at En-
23 hanced Oil Recovery Sites. *Environmental Science & Technology* 50 (14),
24 acs.est.6b01744.
- 25 Deng, H., Stauffer, P. H., Dai, Z., Jiao, Z., Surdam, R. C., 2012. Simulation
26 of industrial-scale CO₂ storage: Multi-scale heterogeneity and its impacts on
27 storage capacity, injectivity and leakage. *International Journal of Greenhouse*
28 *Gas Control* 10 (0), 397–418.

- 1 Doughty, C., 2007. Modeling geologic storage of carbon dioxide: Comparison
2 of non-hysteretic and hysteretic characteristic curves. *Energy Conversion and*
3 *Management* 48 (6), 1768–1781.
- 4 Doughty, C., Pruess, K., 2004. Modeling Supercritical Carbon Dioxide Injection
5 in Heterogeneous Porous Media. *Vadose Zone Journal* 3 (3), 837–847.
- 6 Flett, M., Gurton, R., Weir, G., 2007. Heterogeneous saline formations for car-
7 bon dioxide disposal: Impact of varying heterogeneity on containment and
8 trapping. *Journal of Petroleum Science and Engineering* 57 (12), 106–118.
- 9 Gershenzon, N. I., Ritzi, R. W., Dominic, D. F., Soltanian, M., Mehnert, E.,
10 Okwen, R. T., 2015. Influence of small-scale fluvial architecture on CO₂ trap-
11 ping processes in deep brine reservoirs. *Water Resources Research* 51 (10),
12 8240–8256.
- 13 Giles, M. B., Nagapetyan, T., Ritter, K., 2015. Multilevel monte carlo approx-
14 imation of distribution functions and densities. *SIAM J. Uncertainty Quan-*
15 *tification* 3, 267–295.
- 16 Graham, I., Kuo, F., Nuyens, D., Scheichl, R., Sloan, I., 2011. Quasi-monte carlo
17 methods for elliptic pdes with random coefficients and applications. *Journal*
18 *of Computational Physics* 230 (10), 3668 – 3694.
- 19 Han, W. S., Lee, S.-Y., Lu, C., McPherson, B. J., 2010. Effects of permeability
20 on CO₂ trapping mechanisms and buoyancy-driven CO₂ migration in saline
21 formations. *Water Resour. Res.* 46 (7), W07510.
- 22 Jahangiri, H. R., Zhang, D., 2011. Effect of spatial heterogeneity on plume
23 distribution and dilution during CO₂ sequestration. *International Journal of*
24 *Greenhouse Gas Control* 5 (2), 281–293.
- 25 James, F., 1980. Monte-Carlo Theory and Practice. *Reports on Progress in*
26 *Physics* 43 (9), 1145–1189.

- 1 Juanes, R., Spiteri, E. J., Orr, F. M., Blunt, M. J., 2006. Impact of relative
2 permeability hysteresis on geological CO₂ storage. *Water Resources Research*
3 42 (12), 1–13.
- 4 Kennedy, M. C., O’Hagan, A., 2000. Predicting the output from a complex
5 computer code when fast approximations are available. *Biometrika* 87 (1),
6 1–13.
- 7 Lengler, U., De Lucia, M., Kühn, M., 2010. The impact of heterogeneity on
8 the distribution of CO₂: Numerical simulation of CO₂ storage at Ketzin.
9 *International Journal of Greenhouse Gas Control* 4 (6), 1016–1025.
- 10 Liu, X., Zhou, Q., Birkholzer, J., Illman, W. A., 2013. Geostatistical reduced-
11 order models in underdetermined inverse problems. *Water Resources Research*
12 49 (10), 6587–6600.
- 13 McKay, M. D., Beckman, R. J., Conover, W. J., 1979. A Comparison of Three
14 Methods for Selecting Values of Input Variables in the Analysis of Output
15 from a Computer Code. *Technometrics* 21 (2), 239–245.
- 16 Morris, M. D., Mitchell, T. J., 1995. Exploratory designs for computational
17 experiments. *Journal of Statistical Planning and Inference* 43 (3), 381–402.
- 18 Nordbotten, J. M., Flemisch, B., Gasda, S. E., Nilsen, H. M., Fan, Y., Pickup,
19 G. E., Wiese, B., Celia, M. A., Dahle, H. K., Eigestad, G. T., Pruess, K., 2012.
20 Uncertainties in practical simulation of CO₂ storage. *International Journal of*
21 *Greenhouse Gas Control* 9 (0), 234–242.
- 22 Oakley, J., O’Hagan, A., 2002. Bayesian inference for the uncertainty distribu-
23 tion of computer model outputs. *Biometrika* 89 (4), 769–784.
- 24 Pruess, K., García, J., 2002. Multiphase flow dynamics during CO₂ disposal
25 into saline aquifers. *Environmental Geology* 42 (2-3), 282–295.
- 26 Pruess, K., Oldenburg, C., Moridis, G., 1999. TOUGH2 User’s Guide, Version
27 2.0, Report LBNL-43134, Lawrence Berkeley National Laboratory, Berkeley,
28 California.

- 1 Pruess, K., Spycher, N., 2007. ECO2N - A fluid property module for the
2 TOUGH2 code for studies of CO2 storage in saline aquifers. *Energy Con-*
3 *version and Management* 48 (6), 1761–1767.
- 4 Rasmussen, C. E., Williams, C. K. I., 2006. *Gaussian Processes for Machine*
5 *Learning*. University Press Group Limited.
- 6 Razavi, S., Tolson, B. A., Burn, D. H., 2012. Review of surrogate modeling in
7 water resources. *Water Resources Research* 48 (7), W07401.
- 8 Renard, B., Kavetski, D., Kuczera, G., Thyer, M., Franks, S. W., 2010. Un-
9 derstanding predictive uncertainty in hydrologic modeling: The challenge of
10 identifying input and structural errors. *Water Resources Research* 46 (5),
11 W05521.
- 12 Ritzi, R. W., Freiburg, J. T., Webb, N. D., 2016. Understanding the (co)variance
13 in petrophysical properties of co2 reservoirs comprising sedimentary architec-
14 ture. *International Journal of Greenhouse Gas Control* 51, 423 – 434.
- 15 Stein, M. L., 1999. *Interpolation of Spatial Data*. Springer Series in Statistics.
16 Springer New York, New York, NY.
- 17 Tian, L., Yang, Z., Fagerlund, F., Niemi, A., 2016a. Effects of permeability
18 heterogeneity on CO2 injectivity and storage efficiency coefficient. *Greenhouse*
19 *Gases: Science and Technology* 6 (1), 112–124.
- 20 Tian, L., Yang, Z., Jung, B., Joodaki, S., Erlström, M., Zhou, Q., Niemi, A.,
21 2016b. Integrated simulations of CO2 spreading and pressure response in the
22 multilayer saline aquifer of South Scania Site, Sweden. *Greenhouse Gases:*
23 *Science and Technology* 6 (4), 531–545.
- 24 Tsang, C.-F., Birkholzer, J., Rutqvist, J., 2008. A Comparative Review of Hy-
25 drologic Issues Involved in Geologic Storage of CO2 and Injection Disposal
26 of Liquid Waste. *Journal Name: Environmental Geology; Journal Volume:*
27 *54; Journal Issue: 8; Related Information: Journal Publication Date: 2008*
28 *54 (8), 1723–1737.*

- 1 Vanhatalo, J., Riihimäki, J., Hartikainen, J., Jylänki, P., Tolvanen, V., Vehtari,
2 A., 2012. Bayesian Modeling with Gaussian Processes using the GPstuff Tool-
3 box. *ArXiv e-prints* 14 (1), 48.
- 4 Wilkinson, R. D., Vrettas, M., Cornford, D., Oakley, J. E., 2011. Quantifying
5 Simulator Discrepancy in Discrete-Time Dynamical Simulators. *Journal of*
6 *Agricultural, Biological, and Environmental Statistics* 16 (4), 554–570.
- 7 Xiao, T., McPherson, B., Pan, F., Esser, R., Jia, W., Bordelon, A., Bacon, D.,
8 2016. Potential chemical impacts of CO₂ leakage on underground source of
9 drinking water assessed by quantitative risk analysis. *International Journal of*
10 *Greenhouse Gas Control* 50, 305–316.
- 11 Yang, Z., Tian, L., Niemi, A., Fagerlund, F., 2013. Upscaling of the constitu-
12 tive relationships for CO₂ migration in multimodal heterogeneous formations.
13 *International Journal of Greenhouse Gas Control* 19 (0), 743–755.