Nested sampling algorithm for subsurface flow model selection, uncertainty quantification, and nonlinear calibration

A. H. Elsheikh,1,2 M. F. Wheeler,1 and I. Hoteit3

Received 13 December 2012; revised 30 September 2013; accepted 31 October 2013.

Calibration of subsurface flow models is an essential step for managing ground water aquifers, designing of contaminant remediation plans, and maximizing recovery from hydrocarbon reservoirs. We investigate an efficient sampling algorithm known as nested sampling (NS), which can simultaneously sample the posterior distribution for uncertainty quantification, and estimate the Bayesian evidence for model selection. Model selection statistics, such as the Bayesian evidence, are needed to choose or assign different weights to different models of different levels of complexities. In this work, we report the first successful application of nested sampling for calibration of several nonlinear subsurface flow problems. The estimated Bayesian evidence by the NS algorithm is used to weight different parameterizations of the subsurface flow models (prior model selection). The results of the numerical evaluation implicitly enforced Occam’s razor where simpler models with fewer number of parameters are favored over complex models. The proper level of model complexity was automatically determined based on the information content of the calibration data and the data mismatch of the calibrated model.


1. Introduction

Calibrating subsurface models and accounting for parameter uncertainty have received increasing attention in recent studies [Carrera and Neuman, 1986a, 1986b; Kitanidis, 1997; Efendiev et al., 2005; Dostert et al., 2006; Vrugt et al., 2005; Jiang and Woodbury, 2006; Romayne et al., 2008; Fu and Gomez-Hernandez, 2009; Schöninger et al., 2012]. These techniques can be classified into Bayesian methods based on Markov chain Monte Carlo (MCMC) [Oliver et al., 1997; Efendiev et al., 2005; Elsheikh et al., 2012], gradient-based optimization methods [McLaughlin and Townley, 1996; Carrera et al., 2005; Altaf et al., 2013], stochastic search algorithms [Li and Reynolds, 2011; Elsheikh et al., 2013a, 2013b, 2013c, 2013d], and ensemble Kalman filter methods [Moradkhani et al., 2005; Navdal et al., 2005; Oliver et al., 2008; Luo et al., 2012; Elsheikh et al., 2013e]. It is evident from these studies (among others) that Bayesian statistics provide a general framework for estimating the probability distribution functions of the unknown parameters (i.e., inverse uncertainty quantification). However, a fully Bayesian approach requires a large number of simulation runs and might be impractical for large models [Liu et al., 2009]. In the current study, we are interested in efficient model calibration in a broad sense which includes parameter estimation (i.e., best fit calibration), uncertainty quantification (i.e., posterior distributions of model parameters), and estimation of model selection statistics (i.e., relative weights of different models).

In a Bayesian framework, model selection is based on the marginal likelihood (i.e., Bayesian evidence). However, the Bayesian evidence is a high-dimensional integral over the parameter space and thus is challenging to estimate. Approximate methods were developed for model comparison based on the Akaika information criterion [Akaika, 1974], Bayesian information criterion (BIC) [Schwarz, 1978], and the deviance information criterion [Spiegelhalter et al., 2002]. These approximate information criteria rely on asymptotic approximations to the Bayesian evidence while sharing a common idea of penalizing model complexity. These different information criteria were extensively utilized for Bayesian model averaging (BMA) [Duan et al., 2007; Vrugt and Robinson, 2007; Rojas et al., 2008; Tsai and Li, 2008; Li and Tsai, 2009; Parrish et al., 2012; Neuman et al., 2012], where the model uncertainties are incorporated by averaging predictions over many competing models. The averaging process in BMA utilizes the marginal likelihood to assign the weights to the different models. For example, Li and Tsai [2009] approximated the marginal likelihood for BMA using BIC. Alternatively, the Kashyap model selection criterion [Kashyap, 1982] was utilized by Neuman et al. [2012] within a BMA framework.
Another class of methods for Bayesian model selection is based on MCMC [Metropolis et al., 1953; Hastings, 1970]. Generally, two different techniques can be utilized to sample between different models: reversible jump (RJ) methods [Green, 1995] and product space methods [Carlin and Chib, 1995]. Green [1995] introduced the reversible jump Markov chain Monte Carlo method (RJ-MCMC, also known as trans-dimensional MCMC) as a general framework for sampling parameters from different models. In the RJ-MCMC framework, the parameter space dimension can vary along the Markov chain. Constrains are defined to ensure the reversibility of each jump between different models possibly with different dimensions. This requires defining a reversible projection operator between these different spaces. However, RJ-MCMC methods might be difficult to tune [Han and Carlin, 2001], and for many problems there is no obvious way to construct a reversible jump proposal (interpolation operator) between different models. Product space methods introduced by Carlin and Chib [1995] use a composite search space that encompasses all model spaces jointly. This eliminates the need to define reversible projection operator between models as in the RJ-MCMC framework, but the sampler needs to simultaneously track the parameters for all models.

In the current study, we utilize the nested sampling (NS) algorithm introduced by Skilling [2004, 2006] for simultaneously solving the parameter estimation, uncertainty quantification, and model selection. Nested sampling is an efficient sampling algorithm that is capable of sampling the posterior distribution and providing an estimate of the marginal likelihood (Bayesian evidence). The algorithm can be considered as a global optimization algorithm as it utilizes an active set of samples that covers the prior distribution. The active set of samples is incrementally evolved to climb the likelihood function to the higher likelihood regions without selecting the best sample and thus avoids being trapped in a local minima. The climbing process is done iteratively by replacing the samples with the lowest data fit, one at a time. After several iterations, the active set will contain a set of samples that can be used to obtain a maximum a posteriori estimate of the posterior distribution by picking the sample with the lowest data mismatch. Also, the posterior distribution can be estimated from the set of replaced samples and the final active set after terminating the nested sampling iteration.

The nested sampling algorithm has been successfully applied to several cosmological problems [Mukherjee et al., 2006; Shaw et al., 2007; Vegetti and Koopmans, 2009; Bridges et al., 2009]. Nested sampling is also increasingly applied for statistical inference and Bayesian model selection in different scientific fields [Granqvist et al., 2011; Mthembu et al., 2011; Schuet et al., 2011; Burkoff et al., 2012] and more recently for nonlinear data assimilation [Elsheikh et al., 2013]. In this paper, we report the first application of the nested sampling algorithm for calibration, uncertainty quantification, and prior model selection of subsurface flow models.

The remainder of this paper is organized as follows: section 2 provides a general introduction to Bayesian inference. Section 3 presents in detail the nested sampling algorithm. Section 4 provides an abstract formulation of the multiphase flow problem and a description of Gaussian process regression (GPR) [Rasmussen and Williams, 2006], as it is used to integrate prior knowledge about the unknown fields, followed by a brief discussion of a search space dimension reduction method based on Karhunen-Loève (KL) expansion [Kac and Siegert, 1947; Loève, 1948; Karhunen, 1947]. In section 5, the nested sampling algorithm is applied for calibration and prior model selection of several subsurface flow models. Finally, the conclusions of our work are drawn in section 6.

2. Bayesian Inference

In this section we present a brief introduction to Bayesian inference. Bayesian inference is the process of building a probabilistic model $M$ that characterizes the generative process producing observable data $D$ [Gull, 1988; Hoeting et al., 1999; Sivia and Skilling, 2006]. This process is based on describing the dependencies between the input and output in terms of probabilities. We define a set of models $M^k, k=1,2,\ldots,K$, each depending on a set of model input parameters $\phi^k$, that are assumed to be a set of random variables. We note that the vector of unknown parameters $\phi^k$ may be of different sizes for each model. The major elements of Bayesian inference are the prior, posterior, and the likelihood. The prior distribution $p(\phi^k|M^k)$ describes the probability distribution that expresses beliefs and uncertainties about the true values of the model parameters before the inference step. The inference step then tries to relate the data $D$ to the model by some sort of inversion, resulting in the posterior distribution of the model parameters $p(\phi^k|D,M^k)$. Once a certain set of parameters is chosen, the likelihood of these parameters given the model $M^k$ is defined as $L(\phi^k) = p(D|\phi^k,M^k)$. The posterior is related to the prior via Bayes’ rule

$$p(\phi^k|D,M^k) = \frac{p(D|\phi^k,M^k)p(\phi^k|M^k)}{p(D|M^k)} = \text{Likelihood} \times \text{prior} \times \text{Normalizing constant},$$

which directly follows from the definition of conditional probability. The normalizing constant appearing in the denominator is called the Bayesian evidence or the marginal likelihood and is defined as

$$Z^k = p(D|M^k) = \int p(D|\phi^k,M^k)p(\phi^k|M^k)d\phi^k.$$

The marginal likelihood is used for model selection and comparison [Gull, 1988; Sivia and Skilling, 2006]. In the context of Bayesian model averaging, given a prior probability $p(M^k)$ that $M^k$ is the true model, the posterior probability for a model $M^k$ is defined as

$$p(M^k|D) = \frac{p(D|M^k)p(M^k)}{\sum_{k=1}^{K} p(D|M^k)p(M^k)}.$$

If one is to choose the model that fits the data best, this will lead to selecting an overparameterized model because it can provide better data fit. However,
overparameterized models can overfit the data and will generally exhibit poor predictive performance, as it can exaggerate minor fluctuations in the data. Occam’s razor provides a different view, where simpler models are preferred over unnecessarily complex models. The marginal likelihood automatically and quantitatively embodies Occam’s razor [Gull, 1988] where complex models are automatically penalized under Bayes’ rule as explained by MacKay [1992].

[11] For a practical application of Bayesian inference methods, the modeler needs to specify both the prior and the likelihood distribution. The likelihood is simple to specify by an error metric in a certain norm. However, the choice of the prior is more controversial. In principle, the prior should be set by considering all available information about the parameters $\phi^k$ before collecting the data $D$. This is a nontrivial task, and the prior is usually chosen from a parametric family of distributions or mixtures of those [Sivia and Skilling, 2006]. Typically, a Markov chain Monte Carlo is used to explore the posterior via the Metropolis-Hastings (MH) algorithm [Metropolis et al., 1953; Hastings, 1970]. However, evaluating the evidence $p(D|\mathcal{M}^k)$ of model $k$ requires evaluating a high-dimensional integral given in equation (2). For Bayesian model selection, the Bayesian evidence is evaluated for a set of different models. If all models have the same prior model probability, the model with the largest Bayesian evidence is assumed to better represent the data. In the context of Bayesian model averaging, different models can be weighted by their corresponding Bayesian evidence multiplied by the prior model probability [Burnham and Anderson, 2002].

3. Nested Sampling

[12] Nested sampling is a Bayesian sampling algorithm that can provide an estimate of the Bayesian evidence as well as samples from the posterior distribution. The main idea of nested sampling is to convert the high-dimensional integral in equation (2) into a 1-D integral that is easy to evaluate numerically [Skilling, 2004, 2006]. In doing so, the posterior distribution is built incrementally by having a set of samples that evolves to high likelihood regions of the search space. In this section the model index $k$ will be dropped as the nested sampling algorithm is independently executed on each model.

[13] In nested sampling, the constrained prior mass $X$ is defined as

$$X(\lambda) = \int_{\lambda(\phi) > \lambda(\hat{\phi})} p(\phi|\mathcal{M}) d\phi,$$

where $\hat{\phi}$ can take any value from the prior resulting in $\lambda(\phi) \in [0, 1]$. Using this definition, the integration of the Bayesian evidence can be transformed into

$$Z = p(D|\mathcal{M}) = \int p(D|\phi, \mathcal{M}) p(\phi|\mathcal{M}) d\phi = \int \lambda(\hat{\phi}) dX(\phi),$$

where $dX(\phi) = p(\phi|\mathcal{M}) d\phi$ represents a fraction of the total prior mass. Defining $\lambda = \lambda(\phi)$, the prior mass $X$ can be parameterized as $X(\lambda)$ given by

$$X(\lambda) = \int_{\lambda(\phi) > \lambda(\hat{\phi})} p(\phi|\mathcal{M}) d\phi, \quad (6)$$

with $\lambda \in [0, 1]$. For the case of normalized prior probability, the prior mass $X(\lambda)$ will have a value 1 if $\lambda = 0$ and will decrease to 0 as $\lambda$ approaches 1. The Bayesian evidence integral can then be expressed as

$$Z = \int_0^1 \lambda dX(\lambda). \quad (7)$$

and the problem of calculating the Bayesian evidence becomes a 1-D integral. Nested sampling provides an iterative algorithm to evaluate the fractions of the prior mass $dX(\lambda)$ corresponding to a set of samples $\phi_t$ with likelihood values $\lambda_t = \lambda(\phi_t) = \lambda_j$. Once the fractions of the prior mass are estimated, the Bayesian evidence can be approximated using standard quadrature methods. The mapping of variables in the nested sampling is closely related to the standard cumulative distribution function mapping [Murray, 2007].

[14] If a set of samples $\phi_1, \ldots, \phi_m$ are generated such that the likelihood of these samples is an ordered sequence of increasing values $\lambda_1 < \lambda_2 < \cdots < \lambda_{m-1} < \lambda_m$, then the corresponding prior masses $X_1, \ldots, X_m$ will be a sequence of decreasing values,

$$0 < X_m < \cdots < X_2 < X_1 < 1. \quad (8)$$


$$Z = \sum_{j=1}^{m} Z_j, \quad Z_j = \frac{\lambda_j}{2} (X_{j-1} - X_j). \quad (9)$$

[16] This equation can be easily evaluated if the difference of the prior masses $(X_{j-1} - X_j)$ can be estimated. To obtain $X_j$, a set of $N$ samples are drawn from the prior such that all the samples satisfy the constraint $\lambda(\phi_i) > \lambda_{j-1}$ for all $i \in \{1, 2, \ldots, N\}$. If the sample $\phi_{\text{worst}}$ with the lowest likelihood is selected and $\lambda_j$ is set equal to $\lambda(\phi_{\text{worst}})$, then the prior mass shrinkage factor $t_j = X_j / X_{j-1}$ can be estimated statistically from the distribution $P(t_j) = N(t_j-1)^{N-1}$, where $t_j \in (0, 1)$. This is the probability distribution for the largest of $N$ samples drawn uniformly from the interval $[0, 1]$ (largest in terms of $X$ value and smallest in terms of the likelihood value). This fact is based on order statistics [Balakrishnan and Cohen, 1991] where the $N$th ordered point drawn from a distribution has its cumulative quantity distributed according to the distribution Beta $(N, 1)$ [Skilling, 2006; Chopin and Robert, 2010]. The distribution for $t \sim$ Beta $(N, 1)$ has the following properties: $E(\log t) = -1/N$ and $\sigma(\log t) = 1/N$ for the mean and standard deviation, respectively. Since each $t_j$ is independent, after $j$ iterations the prior volume will shrink down such that $\log X_j \approx -(j-1)/(N)$ As an approximation, $X_j$ can be taken as a deterministic value of $\exp(-j/N)$ [Skilling, 2006].

[17] An iterative algorithm starting from the iteration index $j = 1$ and $X_0 = 1$ and $Z_0 = 0$ can be formulated by sampling the prior space using $N$ samples and evaluating the likelihood of each of these samples. These samples are
called the active set. Following that, the samples in the active set are evolved to the high likelihood region by iteratively replacing the sample \( \phi_{\text{worst}} \) with the lowest likelihood \( \mathcal{L}_{\phi_{\text{worst}}} \) from the active set by a new sample \( \phi_{\text{new}} \) from the prior with a higher likelihood (satisfying the hard constraint \( \mathcal{L}(\phi_{\text{new}}) > \mathcal{L}(\phi_{\text{worst}}) \)). Then the prior mass at the \( j \)th iteration can be estimated statistically or set as \( X_j = \exp (-j/N) \) deterministically. This enables estimating an increment of the evidence \( Z_j = \mathcal{L}(X_j - 1 - X_j)/2 \). After \( M \) iterations the total Bayesian evidence can be estimated by summing the contributions from each iteration.

[18] In the previous steps, we did not draw \( N \) new samples at each iteration. Instead, the samples are reused between the iterations by replacing the sample with lowest likelihood (highest \( X \)) by a new sample while keeping the other \( N - 1 \) samples in the active set unchanged. Effectively, the nested sampling algorithm climbs up the likelihood function, by shrinking the parameter space by a factor, every time a sample is replaced. Ultimately, it samples only the part of the likelihood function close to the maximum likelihood (ML). If \( N \) is large enough, the probability of missing a mode of the likelihood function is reduced. Figure 1 shows the action of the nested sampling algorithm, where the samples discarded during the nested sampling iterations are shown in the physical space in Figure 1a and in the 1-D likelihood space in Figure 1b.

[19] So far, our description of the nested sampling algorithm was focused on evaluating the Bayesian evidence \( \mathcal{Z} \). However, the set of samples discarded during the nested sampling iteration as well as the samples included in the active set after terminating the algorithm can be used for estimating the posterior distribution. These samples contributed to the 1-D curve \( \mathcal{L}(X) \) as shown in Figure 1b. The area under this 1-D curve is already decomposed into steps defined by equation (9). In other words, the existing sequence of discarded samples \( \phi_1, \phi_2, \ldots \) already provides a discrete representation of the posterior probability density function with an importance weight given by \( \mathcal{L}(X_j - 1 - X_j)/2 \). To obtain equally weighted posterior samples, we need to resample (with replacement) a subset of the discarded samples with normalized weights obtained from the area under the 1-D integral [Skilling, 2006].

3.1. Algorithmic Details

[20] The algorithmic description of the nested sampling is listed in Algorithm 1. The active set is stored in the set \( S_{\text{active}} \), where the samples are iteratively replaced to climb the likelihood function. The discarded sample at each nested sampling iteration is stored in the set \( S_{\text{posterior}} \) along with its associated weight. After the last nested sampling iteration, all the samples in the active set are augmented to \( S_{\text{posterior}} \) to account for the remaining part of the 1-D integral after the last discarded sample (the boundary term). Similarly, a boundary term is added to the Bayesian evidence integral as defined by line 14 in Algorithm 1. The Bayesian evidence update value at each iteration can be used to show the convergence of the total Bayesian evidence (integral term) and can be used as a termination criterion. If the current update is small compared to the current evidence value, one can deduce that the subsequent iterations are not likely to significantly contribute to the accumulated value of \( \mathcal{Z} \). Also, one can use an upper bound based on defining a maximum likelihood value \( \mathcal{L}_{\text{max}} \) to terminate the algorithm.

Algorithm 1: Nested Sampling Algorithm

**Input**: Size of the active set \( N \), number of iterations \( M \), model \( \mathcal{M} \), and prior distribution \( p(\phi | \mathcal{M}) \)

1. \( S_0 = 0, X_0 = 1, S_{\text{active}} = \{ \}, S_{\text{posterior}} = \{ \} \).
2. Sample \( \phi_1 \sim p(\phi | \mathcal{M}) \) \( \forall i = 1, \ldots , N \) (generate active set).
3. Set \( S_{\text{active}} = \{ \phi_1, \ldots , \phi_N \} \).
4. for \( j = 1, \ldots , M \) do
   5. Set \( \phi_{\text{worst}} = \text{argmin}_{\phi \in S_{\text{active}}} p(D | \phi, \mathcal{M}) \).
   6. Set \( \mathcal{L} = p(D | \phi_{\text{worst}}, \mathcal{M}) \).
   7. Set \( X_j = \exp (-j/N) \).
   8. Set \( w_j = (X_{j-1} - X_j)/2 \).
   9. Set \( Z_j = Z_{j-1} + w_j \mathcal{L} \).
   10. Set \( S_{\text{posterior}} \leftarrow S_{\text{posterior}} \cup \{ (\phi_{\text{worst}}, w_j \mathcal{L}) \} \).
   11. Sample \( \phi_{\text{new}} \sim p(\phi | \mathcal{M}) \) such that \( p(D | \phi_{\text{new}}, \mathcal{M}) > \mathcal{L} \) (constrained sampling using Algorithm 2).
   12. Set \( S_{\text{active}} \leftarrow (S_{\text{active}} \setminus \{ \phi_{\text{worst}} \}) \cup \{ \phi_{\text{new}} \} \), (update the active set)
13. end
14. Set \( Z = Z_M + (X_M/N)(\mathcal{L}(\phi_1) + \cdots + \mathcal{L}(\phi_N)) \).
15. Set \( S_{\text{posterior}} \leftarrow S_{\text{posterior}} \cup \{ (\phi_j, (X_M/N)\mathcal{L}(\phi_j)) \} \) \( \forall \phi_j \in S_{\text{active}} \).

[21] The main difficulty of the nested sampling algorithm is in the sample replacement step which requires proposing a new sample \( \phi_{\text{new}} \) that conforms to the constraint \( \mathcal{L}(\phi_{\text{new}}) > \mathcal{L}_j \). The problem becomes harder after a few nested sampling iterations as \( \mathcal{L}_j \) reaches a high value and the parameter space satisfying the lower bound constraint shrinks to a smaller region. Skilling [2006] proposed using a random walk Markov chain algorithm to tackle this problem. Following Skilling [2006], a random sample is picked from the active set, excluding the one with the lowest likelihood to start a short Markov chain. This starting sample will satisfy the desired condition of the prior and the likelihood constraint, and the objective is to move away from that sample without violating the likelihood constraint but far enough to lose memory of the starting sample.
Assuming that the MCMC chain starts from the state $\phi_i$, where $\phi_i \in S_{\text{active}}, \phi_i \neq \phi_{\text{worst}}$, we utilize a simple random walk proposal distribution $q(\xi|\phi_i)$ for the MCMC chain of the form

$$\xi = \phi_i + \delta w, \quad w \sim N(0, I),$$  \hspace{1cm} (10)

where $\delta$ is a scaling factor controlling the update step size, $I$ is the identity matrix, and $N$ denotes the normal distribution. The details of the constrained local MCMC sampling are presented in Algorithm 2.

Algorithm 2: Constrained Local Sampling via Metropolis–Hastings Algorithm

Input: Likelihood hard constraint $L(\phi_{\text{worst}})$, starting sample of the MCMC chain $\phi_i$, number of local MCMC iterations $L$, and scaling factor $\delta$

Output: $\phi_{\text{new}}$

1. Set $\theta^0 = \phi_i$
2. for $r = 1, \ldots, L$ do
3. $\theta^r \leftarrow \theta^{r-1}$
4. Propose $\xi \sim q(\xi|\theta^r)$ using $\xi = \theta^r + \delta w, \quad w \sim N(0, I)$
5. Evaluate $L(\xi)$
6. if $L(\xi) > L(\phi_{\text{worst}})$ then
7. Compute $z = \min \left\{ 1, \frac{L(\xi)}{L(\phi_{\text{worst}})} \right\}$
8. Generate $u \sim U(0, 1)$ (i.e., from a uniform distribution)
9. if $u \leq z$ then
10. $\theta^r \leftarrow \xi$
11. end
12. end
13. end
14. Set $\phi_{\text{new}} \leftarrow \theta^L$.

4. Problem Formulation and Parameterization

[22] A two-phase immiscible flow in a heterogeneous porous subsurface region is considered. For clarity of exposition, gravity and capillary effects are neglected. However, the proposed model calibration algorithm is independent of the selected physical mechanisms. The two phases will be referred to as water with the subscript $w$ for the aqueous phase and oil with the subscript $o$ for the nonaqueous phase. This subsurface flow problem is formulated by two equations: the pressure equation and the saturation equation. The pressure equation is

$$-\nabla \cdot (K(S_w) \nabla p) = q,$$ \hspace{1cm} (11)

where $p = p_w = p_o$ is the pressure, $S_w$ is the water saturation, $K$ is the absolute permeability tensor, $q = Q_o/\rho_o + Q_w/\rho_w$ is the normalized source or sink term, $\lambda_o(S_w) = \lambda_w(S_w) + \lambda_o(S_w)$ is the total mobility, and $\rho_w, \rho_o$ are the water and oil...
fluid densities, respectively. The sum of the fluid saturations should add up to one (i.e., \(S_w + S_o = 1\)) as the pore space is filled with fluids. Thus, only the water saturation equation is solved

\[
\frac{\partial S_w}{\partial t} + \nabla \cdot (\mathbf{v}(S_o) \mathbf{v}) = \frac{Q_w}{\phi_w},
\]  

(12)

where \(\Phi\) is the porosity, \(f(S_o)\) is the fractional flow function, and \(v = [\kappa \partial \phi(S_o)] \nabla p\) is the total Darcy’s velocity. The relative mobilities are modeled using polynomial equations of the form

\[
\lambda_w(S_o) = \frac{(S_o)^2}{\mu_w}, \quad \lambda_o(S_o) = \frac{(1-S_o)^2}{\mu_o}, \quad S_{nw} = \frac{S_n - S_{nc}}{1 - S_o - S_{nc}},
\]

(13)

where \(S_{nc}\) is the connate or irreducible water saturation, \(S_{nr}\) is the irreducible oil saturation, and \(\mu_w, \mu_o\) are the water and oil fluid viscosities, respectively. The pressure equation (11) is discretized using a standard finite-volume scheme with the two-point flux approximation [Chen, 2007], and the saturation equation (12) is discretized using a finite-volume scheme in space and backward Euler method in time and solved by a standard Newton iteration [Chen, 2007].

[23] The objective of the subsurface flow model calibration is to estimate the above model parameters. For simplicity, the subsurface permeability field is considered to be the only unknown parameter. The permeability field is assumed to be a lognormal random variable as it is usually heterogeneous and shows a high range of variability. In the numerical test cases, the permeability field is calibrated in two stages, static data integration and dynamic data integration. In the static data integration step, the permeability values at the wells are interpolated stochastically using GPR assuming the Matérn covariance function. The mean field and covariance matrix are then obtained by the GPR and are used to define a reduced search space using the KL expansion. Finally, the dynamic data integration step is performed using the nested sampling algorithm on the reduced search space. In the following subsections, we present a brief review of GPR as it is used for static data integration. Following that, the details of search space dimension reduction using the Karhunen-Loève expansion are presented. The featured application of the GPR and the KL expansion are choices done in this specific study but are in no way required components of the nested sampling technique.

4.1. Gaussian Process Regression

[24] We utilize GPR to perform stochastic interpolation of static data collected at the wells. Gaussian process regression is also known as kriging in the geostatistical literature or as multivariate-Gauss model for the log conductivity, with model-based covariance function and ML covariance parameters [Matérn, 1986; Handcock and Stein, 1993; Diggle and Ribeiro, 2007; Marchant and Lark, 2007; Nowak et al., 2010]. A detailed description of GPR can be found in Rasmussen and Williams [2006] and Elsheikh et al. [2012, 2013e]. Here we highlight some of the important details specific to our numerical test cases.

[25] In the current study, the Matérn covariance function is used [Matérn, 1986; Rasmussen and Williams, 2006]

\[
c_{\text{Matérn}}(r) = \frac{2^{1-v}}{(v \Gamma(v))} \left( \sqrt{2v \frac{r}{\ell}} \right)^v J_v \left( \sqrt{2v \frac{r}{\ell}} \right),
\]

(14)

where \(v\) denotes the gamma function, \(J_v\) is the modified Bessel function of the second kind of order \(v > 0, r = ||x - x'||\) is the distance between two points, and \(l\) is the correlation length. The set of parameters \(v, \ell\) and the global noise variance \(\sigma_n^2\) (see Elsheikh et al. [2013d] for more details) are known as the hyperparameters of the Gaussian process \(\mathcal{H}_P = (v, \ell, \sigma_n^2)\). Finding the optimal set of hyperparameters is a model selection problem. However, for the case of GPR with Gaussian measurement noise the integrals over the parameter space are analytically tractable. Thus, the optimal hyperparameters can be evaluated efficiently by maximizing the logarithm of the marginal likelihood [see Rasmussen and Williams, 2006, algorithm 5.1]. The estimated optimal set of hyperparameters is called the maximum likelihood type II estimate (ML-II). These parameters are found by solving a nonconvex optimization problem using standard methods for unconstrained optimization (e.g., conjugate gradient with random restart) [MacKay, 1999].

4.2. Parameterization Using Karhunen-Loève Expansion

[26] The results of GPR are a real-valued random field \(K\) with mean \(\mu(x)\) and a covariance function \(C(x_1, x_2)\). The KL expansion [Kac and Siegert, 1947; Loève, 1948; Karhunen, 1947], is a classical method for quantization of Gaussian random functions. The logarithm of the permeability field can be parameterized using a finite number of eigenvectors as following

\[
K(x, \xi) = \mu(x) + \sum_{k=1}^{n} \sqrt{\lambda_k} \xi_k \psi_k(x),
\]

(15)

where \(\xi_k\) is a set of random variables, \(\lambda_k\) is a set of nonnegative real numbers, and \(\psi_k(x)\) is an orthonormal set of deterministic functions. The spectral decomposition of the covariance function \(C\) results

\[
C(x_1, x_2) = \sum_{k=1}^{\infty} \lambda_k \psi_k(x_1) \psi_k(x_2),
\]

(16)

where \(\lambda_k > 0\) is the eigenvalue and \(\psi_k\) is the corresponding eigenvector. Preserving the dominant \(n\) eigenvectors \(\psi_k, k = 1, 2, \ldots, n\), and removing all the noisy components corresponding to small eigenvalues \(\lambda_k, k > n\), result in an optimal dimension reduction technique. Different log-permeability realizations \(K\) of the random field \(K\) can be generated for different values of \(\xi_k\). In model calibration, the values of \(\xi_k, k = 1, 2, \ldots, n\), are estimated such that the measured data match the simulation results [Efendiev et al., 2005; Dostert et al., 2006; Elsheikh et al., 2012, 2013e].

5. Numerical Evaluation

[27] The nested sampling algorithm is evaluated on three test cases modeling a water flooding process under different
injection-production patterns. For the parameter estimation problem, the water-cut curves at the production wells are used to define the misfit function and guide the inverse problem solution. Each water-cut curve is sampled at 50 points, and these samples are used for calculating the likelihood. The observation data (water-cut values) are perturbed with zero mean uncorrelated white noise with a variance of 0.01. The unknown log-permeability field is parameterized using the KL expansion, and different parameterization models can be formulated based on the number of retained KL terms.

5.1. Water Flooding Test Case 1

In this test case, the discretized model uses a 2-D regular grid of $41 \times 41$ blocks in the $x$ and $y$ directions, respectively. The size of each grid block is 10 m in each direction and a unit thickness in the $z$ direction. The porosity is assumed to be constant in all grid blocks and equals 0.2. The water viscosity $\mu_w$ is set to 0.3 cp, and the oil viscosity $\mu_o$ is set to 0.3 cp. The irreducible water saturation and irreducible oil saturation are set as $S_{or} = S_{swc} = 0.2$. In the figures, injection wells are shown as black dots and the production wells as white dots.

Figure 2a shows the logarithm of the reference permeability field (permeability units is Darcy $= 9.869233 \times 10^{-13}$ m2), and Figure 2b shows the injection/production pattern for this problem. The reference field was sampled at nine points (shown in Figure 3 as blue dots), and the sampled values are used as an input to the GPR for static data integration. GPR utilizes a Matérn covariance function with an order parameter $v = 3$, and the measurement noise of the static permeability values at the wells is assumed zero. The correlation lengths are optimized using the ML-II estimator, and Figure 3 shows the mean log-permeability field obtained from the GPR along with the $\pm 2$ standard deviations bounds where the reference field is plotted in red, and the interpolated mean value is shown in blue color. Using the covariance matrix obtained by GPR, dimension reduction using the KL expansion is applied, and different models with different number of unknown parameters are obtained by changing the number of retained KL terms.

Figure 4 shows the prior and posterior water-cut curves obtained by the nested sampling algorithm for a model with five KL basis functions as it corresponds to the maximum Bayesian evidence as we will show later. The results are shown in terms of dimensionless time defined by the pore volume injected (PVI). The initial active set contained 25 samples. The posterior production curves, shown in Figure 4b, are obtained after 200 nested sampling iterations, and different colors are assigned to each production well. At each iteration, a local MCMC chain is run for 10 steps with a random perturbation following a normal distribution with a scaling parameter of 0.1, and the measurement error variance used in defining the likelihood is set to 0.01. Figure 5a shows the mean posterior log-permeability fields (in Darcy) obtained by 50 samples from the posterior distribution, and the corresponding posterior standard deviations are shown in Figure 5b where the high uncertainty regions can be easily identified.

Next, we perform a limited comparison of the nested sampling algorithm to a standard random walk MH algorithm for the model with five retained KL terms.
details of the MH algorithm closely follow Algorithm 2 after eliminating step 6 (the constraint condition). Figure 6 shows the trace plot of an adaptive random walk MCMC chain with an acceptance ratio of 0.23. The distribution of the obtained KL weights (after removing 5000 samples as burn-in samples) is shown in Figure 7. Burn-in samples are samples from the initial portion of the Markov chain and are commonly discarded to minimize the effect of chain initial values on the posterior distribution. The distribution of the KL weights obtained by the NS algorithm with an active set of 25 samples and 200 nested sampling iterations is also shown in Figure 7. These distributions are close enough, and the differences can be attributed to sampling errors. The nested sampling algorithm utilized 2025 forward runs and sampling errors are expected in any Bayesian scheme with limited number of forward runs. The computational efficiency of the NS algorithms is evident as it does not require a burn-in phase.

The nested sampling algorithm is run on a set of different models with different number of retained KL terms ranging from 1 KL term up to 20 KL terms. The minimum posterior water-cut error for the different models and the logarithm of the Bayesian evidence are shown in Figure 8. The results are shown for NS runs with different active set sizes and for different number of NS sampling iterations. For each set of parameters the results are averaged over

---

**Figure 4.** Water-cut curves at the production wells for test case 1 and for the model with five KL terms utilizing an active set of 25 samples and using 200 nested sampling iterations. (a) Prior water-cut curves and (b) posterior water-cut curves.

**Figure 5.** Inferred permeability fields at the end of the nested sampling algorithm for test case 1 and the model of five KL terms utilizing an active set of 25 samples and using 200 nested sampling iterations. (a) Mean posterior log-permeability field (in Darcy) and (b) standard deviations of the posterior log-permeability field (in Darcy).
eight different runs where the individual runs are plotted as dotted lines and the mean value is plotted in solid thick line. The results show that the models with one and two KL terms were not capable of matching the data and have a significantly smaller Bayesian evidence compared to other models with more KL terms. As the model complexity increases, the model with five KL terms shows the highest Bayesian evidence, and a slight gradual reduction of the estimated Bayesian evidence is observed with the increase of the number of retained KL terms over five. This result is consistent with Occam’s razor as the models with the fewest parameters should be selected if a good match to the data can be attained and the models with too many parameters should be avoided as they might overfit the data. This interplay between the model complexity and the goodness of fitness is effectively managed by the nested sampling algorithm. Comparing the results obtained by different NS runs with different active set sizes and different nested sampling iterations shows insignificant differences in the estimated logarithm of the Bayesian evidence values. However, the NS run with an active set of 50 samples terminated after 200 nested sampling iterations resulted in a slight increase in the minimum misfit error for models with seven KL terms or more as shown in Figure 8a. However, this early termination (utilizing 200 instead of 400 nested sampling iterations) did not affect the estimated logarithm of the Bayesian evidence value (as shown in Figure 8b). This result demonstrates the robustness of the NS algorithm to the two parameters: active set size and number of nested samples, for this simple problem.

Figure 9 shows the error reduction versus the number of forward runs for different models with different number of retained KL terms. We observe the convergence of all models after 200 nested sampling iterations (each iteration uses a local MCMC with 10 steps for sample replacement). In Figure 9a the minimum water-cut error in the $\ell_2$ for all samples in the active set is shown. This value does not decrease at each iteration as the NS algorithm iteratively replaces the sample with the highest mismatch error in the active set. However, the maximum $\ell_2$ error within the active set decreases at each iteration as shown in Figure 9b. It is also clear that models with three KL terms and more could fit the data to almost the same level; however, the model with five KL terms shows the highest Bayesian evidence as shown in Figure 8b.

5.2. Water Flooding Test Case 2

This problem uses layer 10 of the tenth SPE (Society of Petroleum Engineers) comparative test case [Christie and Blunt, 2001] as the reference permeability field. The rectangular domain is partitioned into a regular grid of 220
60 cells of size $6.096 \times 3.048$ with a thickness of 0.6096 m. The porosity is set to 0.2 over all grid blocks, and the parameters $\mu_w=0.3$ cp and $\mu_o=3$ cp are used. The irreducible water and oil saturations are set as $S_{wc}=S_{or}=0.2$. Figure 10a shows the logarithm of the reference permeability field (in Darcy). The mean log-permeability field obtained from the GPR is shown in Figure 10b. This field was sampled at the 10 well locations shown in Figure 10b as dots (black and white). For this model, GPR utilizes a Matérn covariance function with an order parameter $\nu = 1$, and the measurement noise of the static permeability values at the wells is assumed to be zero. The mean values from the GPR and the resulting covariance matrix were used as an input for KL model reduction to obtain the significant eigenmodes of

![Figure 8](image_url)

**Figure 8.** Model comparison results for test case 1 using different NS runs. (a) Minimum water-cut error in $\ell_2$ norm for samples in the active set at the end of the NS iteration versus the number of KL terms and (b) estimated logarithm of the Bayesian evidence versus the number of KL terms.

![Figure 9](image_url)

**Figure 9.** Error reduction for different prior models versus number of nested sampling iterations for test case 1. (a) Minimum water-cut error for all samples in the active set measured in the $\ell_2$ norm and (b) maximum water-cut error for all samples in the active set measured in the $\ell_2$ norm.
the interpolated log-permeability field. Figure 10b also marks the five injection wells as block dots and the five production wells as white dots.

Figure 11 shows the prior and posterior water-cut curves at the production wells for the model utilizing 30 KL terms as this model shows the highest Bayesian evidence among the competing models. The nested sampling algorithm utilized an active set of 25 samples, and the algorithm was terminated after 600 nested sampling iterations. At each iteration, a local MCMC chain is run for 10 steps with a random perturbation following a normal distribution with a scaling parameter of 0.05, and the measurement error variance used in defining the likelihood is set to 0.01. The dynamic data collected at the five different production wells is successfully matched as shown in Figure 11b. The corresponding mean posterior log-permeability fields (in Darcy) and standard deviation are shown in Figures 12a and 12b, respectively. We observe a large level of

Figure 10. Injection and production patterns for test cases 2 and 3 (black dots for injection wells and white dots for production wells). (a) Reference log-permeability field for test cases 2 and 3 (in Darcy), (b) mean log-permeability field obtained by Gaussian process regression for test case 2 (in Darcy), and (c) mean log-permeability field obtained by Gaussian process regression for test case 3 (in Darcy).

Figure 11. Water-cut curves at the production wells for test case 2 and for the model with 30 KL terms utilizing an active set of 25 samples and using 600 nested sampling iterations. (a) Prior water-cut curves and (b) posterior water-cut curves.
uncertainty that indicates that the information content of the data used to calibrate the model is low.

Figure 13 shows the results of a prior model comparison study utilizing different NS runs with different active set sizes and different number of NS iterations. Figure 13a shows the minimum posterior water-cut error versus the number of retained KL terms, and Figure 13b shows the estimated Bayesian evidence for different models using different number of KL terms to parameterize the log-permeability field. The results for eight different NS runs are presented as follows:

- **Figure 12.** Inferred fields at the end of the nested sampling algorithm for test case 2 with the model of 30 KL terms utilizing an active set of 25 samples and using 600 nested sampling iterations. (a) Mean posterior log-permeability field (in Darcy) and (b) standard deviations of the posterior log-permeability field (in Darcy).

- **Figure 13.** Model comparison results for test case 2 using different NS runs. (a) Minimum water-cut error in $\ell_2$ norm for samples in the active set at the end of the NS iteration versus the number of KL terms and (b) estimated logarithm of the Bayesian evidence versus the number of KL terms.
runs along with mean value are shown in each subfigure. For the minimum posterior mismatch error, it is observed that the model with the highest evidence (30 KL terms) does not correspond to the lowest minimum posterior error. This interplay between model complexity and quality of fitness is effectively managed by the nested sampling algorithm. We want to highlight that all the different models with different number of KL terms were terminated after the same number of nested sampling iterations (either 3000 or 600) in order to make the model comparison neutral to the computational cost. In terms of the estimated logarithm of the Bayesian evidence, Figure 13b shows that NS with

![Figure 14](image1.png)

**Figure 14.** Water-cut curves at the production wells for test case 3 and for the model with 60 KL terms utilizing an active set of 25 samples and using 600 nested sampling iterations. (a) Prior water-cut curves and (b) posterior water-cut curves.

![Figure 15](image2.png)

**Figure 15.** Inferred log-permeability fields at the end of the nested sampling algorithm for test case 3 with the model of 60 KL terms utilizing an active set of 25 samples and using 600 nested sampling iterations. (a) Mean posterior log-permeability field (in Darcy) and (b) standard deviations of the posterior log-permeability field (in Darcy).
different active set sizes and different number of NS iterations produces consistent relative values. Thus, any run will lead to picking the model with 30 KL terms as the best model. The comparison of the results obtained for an active set size of 50 with 300 NS iteration versus 600 iterations shows that the value of estimated evidence might be slightly underestimated for the case of early termination of the NS algorithm.

5.3. Water Flooding Test Case 3

[37] The last test problem utilizes the same reference permeability used in test case 2. However, a different injection/production pattern is used. The data from a total of 15 wells is used to infer the permeability field. The mean log-permeability field obtained from the GPR is shown in Figure 10c. The reference field was sampled at the 15 wells shown in Figure 10c, and the data was fed to the GPR utilizing a Matérn covariance function with an order parameter $\nu = 1$. The mean interpolated values and the corresponding covariance matrix were used as an input for KL dimension reduction to obtain the significant eigenmodes of the interpolated log-permeability field. Different models are built with different number of retained KL terms ranging from 10 KL terms up to 120 KL terms.

[38] Figure 14 shows the prior and posterior water-cut curves at the production wells for the model with 60 KL terms. The results are shown in terms of dimensionless time defined by the PVI. For this problem, the match to the reference data is not perfect, but the major features of the solution are captured by the calibration process. Figure 15 shows the mean posterior log-permeability fields (in Darcy) and the associated standard deviation obtained by the nested sampling algorithm. The estimated levels of uncertainty after the dynamic data integration step are much less

![Figure 16. Samples from the posterior distribution of the log-permeability field (in Darcy) for test case 3 with the model of 60 KL terms (active set size = 25 and number of nested sampling iterations = 600).](image)

![Figure 17. Model comparison results for test case 3 using different NS runs. (a) Minimum water-cut error in $\ell_2$ norm for samples in the active set at the end of the NS iteration versus the number of KL terms and (b) estimated logarithm of the Bayesian evidence versus the number of KL terms.](image)
than those obtained in the previous test case as more data are used for the calibration process. Figure 16 shows four samples from the posterior distribution of the log-permeability fields (in Darcy), and small localized differences are observed that reflect the reduced level of uncertainty after the dynamic data integration step.

Figure 17a shows the minimum posterior water-cut error within the active set after terminating the NS algorithm versus the number of KL terms. In Figure 17b the estimated logarithm of the Bayesian evidence for different models with different number of retained KL terms is shown. The results are shown for different NS runs with different active set sizes and different number of nested sampling iterations. We observe that early termination of the NS iteration affects the quality of the estimated Bayesian evidence as this problem is more challenging than the previous two test cases. We also observe for the runs with larger active set size (50 samples), a slight increase in the minimum posterior water-cut error for the model with 120 KL terms as the models with more KL terms are harder to fit. We note that each curve uses the same number of nested sampling iterations for the different models with different number of KL terms. As for model selection statistics shown in Figure 17b, it might be confusing to decide which is the best model among those retaining 40–80 KL terms. A two-level strategy can be performed to tackle this problem, where an initial inexpensive run with a small active set can be utilized to narrow down the range of models under consideration. Following that, a more comprehensive run for the models with 40, 60, and 80 KL terms could be performed with enough nested sampling iterations to ensure the convergence of the algorithm.

Comparing the results from test cases 2 and 3 we observe that the amount of data used to constrain the inverse problem determines the optimal level of model complexity. For test case 2, data from five production wells were used to constrain the model. In contrast, the information content from 10 production wells were used to calibrate the third test problem, and a more complex model (with 60 KL terms in contrast to 30 for test case 2) produced the highest marginal likelihood (Bayesian evidence).

5.4. Discussion

In this subsection, we study the convergence properties of the NS algorithm on test case 2 and test case 3. Figure 18 shows the error decay (maximum active set value in $\ell_2$ norm) versus the number of forward runs for the NS algorithm with different active set sizes. For the two test cases, NS runs with smaller active set outperformed the runs with the larger active set size in terms of error reduction for the same number of forward runs. This is attributed to the tendency of the NS to perform as an optimization-like algorithm when a small active set size is utilized. In contrast, the runs with larger active set sizes produce good search space exploration with more computational cost. This is easily explained in the limit for a run with an active set of one sample. NS is then a pure randomized optimization algorithm, where at each iteration, the current sample is replaced with another sample uphill the likelihood surface.

6. Conclusions

Nonlinear model calibration and uncertainty quantification are challenging problems. During the formulation of the calibration problem, certain modeling assumptions are made. The effect of these assumptions on the model calibration is often neglected. In the current study, we presented the nested sampling algorithm that tackles model calibration, uncertainty quantification, and model comparison simultaneously. Nested sampling has been successfully applied to the inference of subsurface properties as well as...
selecting the most plausible model from a set of models based on the number of unknown parameters and the information content of the calibration data. However, the nested sampling algorithm can be applied for calibration and model selection of various computer models (i.e., single-phase flow, multicomponent transport, and geomechanical problems).

[41] We obtained various numerical results showing the interplay between the estimated Bayesian evidence (used for model selection) and the minimum posterior error values in the active set after terminating the nested sampling algorithm. It was observed that models with more parameters might need more iterations to obtain a good fit; however, we did choose to terminate all the models after the same number of nested sampling iterations to perform a model comparison study that is neutral to the computation cost. Having said that, the presented results conform to the Occam’s razor as simpler models with fewer number of parameters were selected only if they can match the data. The power of this technique is evident in comparing the results from different examples. For those test problems utilizing calibration data with more information content, models with larger number of unknowns produced higher Bayesian evidence. This dependence of the optimal number of parameters in the model as related to the information content of the data utilized to calibrate the model raises concerns about many of the parameter estimation techniques that define the number of unknowns as the number of grid blocks.

[44] The nested sampling algorithm can be considered as a global optimization algorithm as it utilizes an active set of samples that covers the prior distribution. This active set incrementally climbs the likelihood function to the higher likelihood regions. Climbing the likelihood surface requires sampling from the prior while satisfying a lower bound constraint on the likelihood value. For this step, we utilized a random sampling using a short Markov Chain starting from a randomly selected sample from the active set. The results obtained show the success of this strategy. We are currently investigating alternative efficient methods to perform this constrained sampling step.

[45] Acknowledgments. The authors would like to thank the Associate Editor and the anonymous reviewers for their insightful and constructive comments. We are particularly indebted to Wolfgang Nowak for the extensive and insightful comments which resulted in considerable improvements to this manuscript. Further thanks are due to Albert C. Reynolds for commenting on an earlier version of this manuscript.

References


