Adaptive design of experiments for calibration of complex simulators - An application to uncertainty quantification of a mature oil field

D Busby\textsuperscript{1}, M Feraille\textsuperscript{1}
\textsuperscript{1} IFP (Institut Fran\c cais du P\&eacute;trole, Rueil Malmaison 92500, France
E-mail: daniel.busby@ifp.fr

Abstract. In this work we provide a practical approach to the inverse problem arising when observational data are used to calibrate parameters of an expensive simulation model. Our main application is the history matching problem in oil reservoir forecasting. In such and similar applications, the resulting inverse problem is generally ill-posed, the number of parameters to invert can be very high and the simulation time is very long (up to a few days). In this work a probabilistic approach is adopted to solve the inverse problem; this results in finding a posterior distribution of the simulator uncertain parameters. This posterior distribution is then used to reduce the uncertainty of future forecasts. To reduce the number of simulations, a cheap surrogate of the expensive simulator (an emulator) is build from a limited number of simulations using a Gaussian process regression (kriging) method. A new hierarchical experimental design method is proposed to refine the emulator in the proximity of possible solutions of the inverse problem. These solutions are explored using a Markov Chain Monte Carlo method. At each design iteration, only some accurately chosen points of the obtained posterior sample are simulated. The objective of this design is to iteratively remove the bias of the posterior calculation due to the emulator uncertainty and inaccuracy. This novel methodology is tested on a synthetic water flooded reservoir model that has been used in several previous works: the Imperial College fault model.

1. Introduction
In reservoir engineering initial building of the numerical model for one particular reservoir is based on the "static" data such as well log values, Special Core Analysis (SCAL) results, fluid measured properties (PVT), etc. These "static" data are used to determine or infer many "static" parameters to build the initial reservoir model, such as the geological model, facies variograms, permeability and porosity distributions, etc.. Many of these static parameters are uncertain. This "a priori" uncertainty is generated either by the lack of data in some area of the reservoir either by measurement uncertainty. When the reservoir production is initiated then "dynamic" data (or production history) are available such as measured pressures, oil/water/gas rates at the wells, 4d seismic, etc. By default the initial model simulated results do not cope with these "dynamic" data. Therefore to ensure that the reservoir numerical model reproduces correctly the "dynamic" data we need to calibrate the uncertain "static" parameters. Historically, optimisation techniques were used to obtain one single history matched model that reproduced the production history. It is however well known that this complex, ill-posed inverse problem could lead to several equiprobable matched models.
A probabilistic approach [1] is then adopted. In this approach a posterior distribution of the “static” parameters is computed that takes into account both their “a priori” uncertainty and their matching of the dynamic data.

This “a posteriori” uncertainty leads naturally to the probabilistic computation of the production forecasts (oil, gas and water). This information is then used to take future development decisions.

In oil reservoir applications simulation time can be extremely long (from a few hours to a few days), thus the number of simulations required by the probabilistic approach using direct Monte Carlo methods (usually several thousands) are prohibitive. The idea of using response surface methods and experimental design to solve inverse problems has been proposed in previous works [2, 3, 4] to reduce the number of required simulator runs. However, obtaining an accurate response surface model is usually critical, thus this modeling error has to be taken into account. Iterative methods have been proposed to refine the response surface model in all the input domain [5], or using clustering techniques to select restricted input domains for deterministic optimization [6]. In this method the response surface is obtained using kriging [7] which is a regression method based on Gaussian Processes (GP). GP based regression methods are well known in the analysis and design of computer experiments [8, 10] and are used in many industrial applications. The novelty of this approach is in the adaptive experimental design strategy that reduces the response surface uncertainty on the solutions of the inverse problem generated by Markov Chain Monte Carlo. The final objective of the method is to produce a posterior distribution which is not biased by response surface errors. The method is tested on challenging history matching problem the Imperial College (IC) fault model [11].

2. Probabilistic approach to model parameter calibration

Consider a complex mathematical model implemented in a numerical simulator, which is used to study a large physical system such as an oil reservoir. The model is supposed to produce multiple outputs:

\[ y^i = f^i(x), i = 1, \ldots M \]

where \( x \in \Omega \subset \mathbb{R}^d \) are uncertain parameters with some given prior distributions with density \( P(x) \). Note here the different outputs \( y^i \) with \( i = 1, \ldots M \) can also represent the same output at a different time steps. For example, in reservoir forecasting applications, typical outputs are production rates or pressures, for different wells and at different time steps \( t^1, \ldots t^F \).

Now, suppose to have a set of observations \( d \in \mathbb{R}^m \) of \( m \leq M \) outputs of the physical system. Typically in our applications, the set \( d \) corresponds to observations of outputs at early time steps \( t^1, \ldots t^f \) with \( f < F \). We want to use this information to reduce our uncertainty on \( x \) in order to obtain better predictions on the remaining outputs. Note that if we assume no modeling errors and no observation errors then we could obtain the value of the uncertain parameters by solving the inverse problem \( d^i = f^i(x^*) \) with \( i = 1, \ldots M \). We are usually not able to estimate modeling errors and therefore in this approach we will assume them as negligible respect to observation errors. Assuming the measurements \( d^i \) to follow a normal distribution with zero mean and variances \( \sigma^2_d^i \), the solution of the probabilistic inverse problem [1] is a posterior distribution \( P(x|d) \) given by Bayes’ rule on conditional probabilities:

\[
P(x|d) = \frac{P(d|x)P(x)}{P(d)}
\]

where \( P(d|x) \) is the likelihood function, which is the conditional probability of observing the measurements \( d \), for a given value \( x \) of the parameters. Considering the observation errors as independent and defining the objective function \( O(x) \) as:

\[
O(x) = \sum_{i=1}^{m} \left( \frac{(f^i(x) - d_i)^2}{2\sigma^2_d^i} \right)
\]
the likelihood is then given (up to a normalization constant) by:

\[ P(d|x) \sim \exp(-O(x)). \]  \hspace{1cm} (3)

To obtain a sample of the posterior distribution a very high number of simulations is required. To reduce the number of necessary simulations the objective function \( O(x) \) can be replaced by a proxy model also referred to as a metamodel or an emulator. An emulator is usually composed by a predictor, which is a statistical approximation of the unknown function and also by a measure of uncertainty of the predictor. Gaussian process emulators [2] provides a complete distribution of the function \( f \).

Denoting by \( \hat{O}(x) \) the predictor of a given emulator of \( O(x) \) then the likelihood function \( P(d|x) \) of Eq. (3) can be rewritten as:

\[ P_e(d|x) \sim \exp(-\hat{O}(x)) \]

By taking into account also the emulator uncertainty one obtains the following formula:

\[ P_e(d|x) \sim \int \exp(-\hat{O}(x))P(\hat{O}(x))dP \]  \hspace{1cm} (4)

Note that the more the emulator uncertainty is large and the more Eq. (4) will differ from Eq. (3). It is evident from this equation that in order to have Eq. (4) as close as possible to Eq. (3), the emulator uncertainty has to be reduced as much as possible and more particularly wherever \( O(x) \approx 0 \). Note that although reducing significantly the emulator uncertainty in the entire input domain \( \Omega \) can be computationally expensive, reducing it only where \( O(x) \approx 0 \) will require much less simulations.

3. A Hierarchical Emulator

The construction of the emulator used in this approach is based on modeling the simulator function \( f \) by a stochastic field with a Gaussian process prior distribution. This idea was initially used in geostatistics in the kriging method [7] and then by Sacks et al. [8] for the analysis of computer experiments. A machine learning perspective is given in [9], a Bayesian approach can be found in [2] while for a general review we refer to [10].

Given a deterministic response of our computer model, say, \( f(x) : \mathbb{R}^d \rightarrow \mathbb{R} \), and a set of evaluations \( f_X = \{f(x_1), \ldots, f(x_n)\} \) computed at some design configurations \( X = \{x_1, \ldots, x_n\} \), the following stochastic model for \( f \) is considered:

\[ s(x) = \sum_{j=1}^{k} \beta_j h_j(x) + Z(x). \]  \hspace{1cm} (5)

where \( h_1, \ldots, h_k \) are preselected real-valued functions with coefficients \( \beta = (\beta_1, \ldots, \beta_k)^T \in \mathbb{R}^k \), where \( h_1 \) is a constant. Moreover, \( Z \) is assumed to be a Gaussian random process with mean zero and covariance

\[ \text{cov}[x, y] = \mathbb{E}[Z(x)Z(y)] = \sigma^2 R(x, y) \]  \hspace{1cm} (6)

between \( Z(x) \) and \( Z(y) \), where \( \sigma^2 \) denotes the process variance and \( R(x, y) \) the correlation function. Here \( s(x) \) is a stochastic process which corresponds to our current knowledge about \( f(x) \) (which is a deterministic function). The following correlation function is adopted

\[ R(x, y) = r(x - y) = \exp \left( -\sum_{j=1}^{d} \frac{x_j - y_j}{\theta_j}^p \right) \]  \hspace{1cm} (7)
with \( \theta_j > 0 \), \( 1 \leq j \leq d \), and \( 0 < p \leq 2 \). The parameters \( \theta_j, p, \sigma, \beta \) can be determined by maximum likelihood estimation \([5, 10]\). Then given the trial points \( X \) and the corresponding response values \( f_X \) the mean of the stochastic process for an untried point \( x \) is given by
\[
\hat{s}(x) = h^T(x)\hat{\beta} + r^T(x)R^{-1}(f_X - H\hat{\beta}),
\]
where
\[
h(x) = (h_1, h_2(x), \ldots, h_k(x))^T \in \mathbb{R}^k,
\]
\[
H = (h_j(x_k))_{1 \leq k \leq n; 1 \leq j \leq k} \in \mathbb{R}^{n \times k},
\]
\[
R = (R(x_i, x_j))_{1 \leq i, j \leq n} \in \mathbb{R}^{n \times n},
\]
\[
r(x) = (R(x_1, x), \ldots, R(x_n, x))^T \in \mathbb{R}^n.
\]
and where
\[
\hat{\beta} = (H^TR^{-1}H)^{-1}H^TR^{-1}f_X
\]
Moreover, the variance of the GP process \( s(x) \) is given by
\[
\text{var}[s(x)] = \sigma^2 \left[ 1 - |r(x), h^T(x)| \right]^{-1} \left[ \begin{array}{c} r(x) \\ h(x) \end{array} \right]
\]
Here a sequential design of experiments is adopted where the objective is to reduce the posterior uncertainty by constructing a sequence \( s_1, s_2, \ldots, s_L \) of emulators from samples of \( f \) taken at scattered locations \( X = \{x_1, \ldots, x_n\} \). The construction of the emulators relies on a data hierarchy
\[
X_1 \subset X_2 \subset \cdots \subset X_L \subset X
\]
of nested subsets of \( X \), where \( L \) denotes the number of iterations. This data hierarchy is built according to a sequential design strategy, where the emulator at a level \( l \), is used to determine the next dataset \( X_{l+1} \). The first emulator \( s_1 \) is computed on the basis of an initial design \( X_1 \), which includes only very few data and which is usually a maximin latin hypercube design (LHS) \([10]\). Before discussing how to construct the next iterations we first briefly review the MCMC method used to sample the posterior distribution of Eq. (3).

4. MCMC method to sample the posterior
The MCMC method used in this approach is based on a Metropolis-Hastings algorithm \([12]\). We remind that in MCMC samples of a complex distribution \( P(x) \) are obtained by the stationary distribution of a suited markov chain. After an initial transient the markov chain start reproducing samples of the posterior distribution. The Metropolis algorithm is used to construct the Markov chain and it is based on the following trial and error strategy:

(1) Given a current state \( x^t \), propose a random variation given by a transition function \( T(x, y) \)
(2) Compute the probability ratio \( P(y)/P(x^t) \)
(3) Generate a random number \( r \) from Uniform\([0, 1]\). Let \( x^t = y \) if
\[
r \leq P(y)/P(x^t)
\]
Different choices of transition probabilities \( T(x, y) \) can be adopted depending on the problem, for a review on MCMC methods we refer to \([12]\). Note that a fundamental advantage of the MCMC method is that one does not have to compute the normalization constant of the likelihood density function \( P(d|x) \).
5. Sequential Design for Uncertainty Reduction

5.1. Objective of the sequential design

We remind that our objective is to produce accurate samples of the posterior distribution given by Eq. (3). To this end we build an emulator of the objective function $O(x)$. Given an emulator $s_1$ of $O$ we have to reduce as much as possible its variance particularly in the area of the input space where the most probable solutions of the inverse problem are expected. As a result, we expect our approximated likelihood distribution Eq. (4) to converge toward the real likelihood Eq. (3). Now supposing that the variance measure given by the emulator is correct, one expects that the posterior distribution obtained from Eq. (4) will be much wider than the one given by Eq. (3). Then reducing the emulator variance by an adaptive experimental design will also reduce the posterior uncertainty of the input parameters and consequently the uncertainty of future forecasts.

5.2. Sequential design method

To build an experimental design that follows the objective discussed above, the following method is proposed. Supposing to have an emulator $s_1$ we first obtain a sample of the of the inverse problem solutions by MCMC. Given a reasonable sample of the posterior distribution $X^l_N$, the obtained points are used as candidate points for the experimental design. The new simulation points are selected among the candidate points $X^l_N$ by finding the ones which minimize the average emulator mean squared error on the posterior sample:

$$IMSE = \sum_{i=1}^{N} \text{var}[s_1(x_i)]$$

where $x_i \in X_N$. The emulator variance is indeed locally reduced by adding simulations. Note that a number $N_s$ of points can be added at each step of the design. The number $N_s$ can reflect for instance the number of parallel machines available for the simulations. The algorithm scheme is summarized below:

**Uncertainty Reduction Sequential Design Algorithm**

1. Perform an initial design $X_1$ by maximin LHS and run the simulator at these points.
2. Compute the emulator of the objective function using simulations results of step (1)
3. FOR $\ell = 2, 3, \ldots$ DO
   3a. Compute a sample $X_N$ of size $N$ of the posterior distribution using MCMC. Select the $N_s$ points $X_\ell$ among $X_N$ minimizing the resulting integrated mean squared error $\text{IMSE} = \sum_{i=1}^{N} \text{var}[s_1(x_i)]$ where $x_i \in X_N$.
   3b. Run the simulator at the new points $X_\ell$ and compute a new emulator using all the simulated points.
   3c. Plot quantiles of the posterior distribution of the parameters ($X_N$). Exit when the quantiles of all parameters stabilize

To monitor the effects of the sequential design, the quantiles of the objective function and the quantiles of the input parameters can be plotted as a function of the number of iterations. After an initial number of iterations these quantiles should normally converge toward the real posterior quantiles. This can be considered the stopping criterion of the sequential design after which adding more simulations does not increase our information. An application of the method is discussed in the next section.
6. Application to an oil reservoir

To show the applicability and advantages of the above methodology, we have tested it on a rather difficult test case taken from reservoir forecasting, called the IC Fault Model [11]. The IC Fault Model is a synthetic reservoir model where it has been shown that traditional deterministic history matching approaches can produce models with no predictive value [11]. The methodology described above is used here to calibrate three uncertain reservoir parameters by integrating observed data from the first 3 years of production. We show that the method can be used to effectively reduce uncertainty on future production forecasts using only a few hundreds simulations.

6.1. IC Fault Model description

The IC Fault model is a cross-sectional model of a layered reservoir. The geological model consists of six layers of alternating good and poor quality sands. The width of the model is 1000 feet, with a simple fault at the mid-point, which offsets the layers. There is a water injector well at the left-hand edge, and a producer well on the right-hand edge. Both wells are completed on all layers, and operated at fixed bottom hole pressures. The simulation model is 100x12 grid blocks, with each geological layer divided into two simulation layers with equal thicknesses, each grid block is 10 feet wide. The means for the porosities are: good quality sand, 0.30; poor quality sand, 0.15. The means of the permeabilities are: good quality sand, 158.6 mD; poor quality sand, 2.0 mD. A picture of the reservoir is shown in Fig. (1).

![Figure 1. IC Fault model](image)

6.2. Numerical Results

Three uncertain input parameters are considered: the fault throw $h \in \mathbb{U}[0, 50]$ feet, the good and the poor sand permeability multipliers $k_g \in \mathbb{U}[100, 200]$ and $k_p \in \mathbb{U}[0, 50]$.

In this example, the probabilistic inversion is performed using observations of the oil production rates (OPR) at the time steps $t = 1, 2, \ldots, 11, 12, 24, 36$ months. To compare our method with the example given in [11], the history data is generated using the same simulator and the same values of the best input $x_0$, given by $h_0 = 10.4$, $k_{g0} = 131.6$, and $k_{p0} = 1.3$. Moreover, each observation error $\sigma_d$ is assumed to be equal to 3% of the measurement value. The resulting history data $d$ is obtained by running the simulator at $x_0$. The output we want to predict is the total oil production observed after 10 years (FOPT10). The reservoir simulator is first run on a maximin LHS consisting of $|X_1| = 20$ different configurations for the initial
The uncertainty reduction algorithm is stopped after 285 simulations. In Figure 2 it is shown the resulting posterior input distributions of the three variables $h$, $k_g$, $k_p$. It is important to note that Figure 2 shows only the projection of the posterior distribution on each parameter, thus it is not showing the complete uncertainty reduction obtained by the method. A first remark by analysing each input distribution independently is that the uncertainty on $h$ is only slightly reduced, even though the most probable value is very close to the real value $h_0 = 10.4$ feet. On the other hand the uncertainty on $k_g$ is highly reduced and the real value $k_{g0} = 131.6$, is very close to the distribution peak. Finally for $k_p$, the initial uncertainty has been considerably reduced, however the correct value does not correspond to the center of the distribution.

The results of the uncertainty reduction algorithm on the output FOPT10 is shown in the histograms of Figure 3. Histogram (1) shows the prior distribution while histograms (2) shows the posterior. Note that the posterior distribution of FOPT10 is more peaked and this peak is much closer to the real solution than the prior peak. By considering only the most probable value for decision making, one would conclude that the FOPT10 is equal to 1000000 barrels of oil by looking at the prior distribution whether it would be 890000 barrels by considering the posterior. As a result the prediction error considering the prior distribution is of 180000 barrels, while the prediction error using the posterior is 70000 barrels. The prediction error is then more than halved in this case using this approach.

Our analysis of the IC Fault model also confirms the conclusions presented in [11], that a deterministic history matching approach is usually unreliable, whether a probabilistic approach is more robust and can still provide useful information for decision making. Finally note that the above analysis required 285 simulations, which is usually affordable in such and similar applications and it is by orders of magnitude less than the simulations required by traditional Monte Carlo methods and by other previous methods for the same case study [13]. Nevertheless for less pathological cases we expect this approach to be more efficient.

![Histograms](image)

**Figure 2.** Posterior distributions of $h$ (1), $k_g$ (2) and $k_p$ (3). Note that the values used to generate the history data are $h = 10.4$, $k_g = 131.6$, $k_p = 1.3$.

### 7. Conclusions

We have shown that a calibration approach based on the solution of a probabilistic inverse problem, can be computationally tractable using advanced response surface modeling and adaptive experimental design.
The proposed sequential approach concentrates the simulator runs in regions where possible solutions may be found, while discarding the other regions. As a result, the emulator becomes more accurate at each iteration in the interesting regions of the input space.

To monitor the effectiveness of the design, the posterior distribution of the parameters is computed at each iteration by MCMC. This feature together with the high efficiency of the proposed design strategy represent the main advantages over previous methods, where the emulator uncertainty and inaccuracy are usually not considered.

The numerical application on a low dimension but rather challenging test case indicates that this approach can represent a useful tool for uncertainty reduction and decision-making for a wide class of engineering problems.

References

[1] Tarantola A 2004 *Inverse Problem Theory and methods for model parameter estimation* (Philadelphia: SIAM)
[2] Kennedy M C and O’Hagan A 2000 Bayesian calibration of computer models *J. Roy. Statist. Soc. Ser. B* 63 425–464
[3] Landa J L and Guyaguler B 2003 A Methodology for History Matching and the Assessment of Uncertainties Associated with Flow Prediction *SPE* 84465-MS
[4] Feraille M and Roggero F 2003 Uncertainty quantification for mature field combining the bayesian inversion formalism and experimental design approach, presented at ECMOR IX Conf. Cannes France
[5] Busby D, Farmer C L and Iske A 2007 Hierarchical nonlinear approximation for experimental design and statistical data fitting *SIAM J. Scientific Computing* 29, No.1, 49–69
[6] Li B and Friedman F 2006 Semiautomatic Multiple Resolution Design for History Matching *SPE* 102277
[7] Matheron G Principles of geostatistics 1963 *Economic Geology* 58 1246–1266
[8] Sacks J, Welch W J, Mitchell T J and Wynn H P 1989 Design and analysis of computer experiments *Statist. Sci.* 4 409–435
[9] Rasmussen C E and Williams C K I 2006 *Gaussian Processes for Machine Learning*, (MIT Press)
[10] Santner T J, Williams B J and Notz W I 2003 *The Design and Analysis of Computer Experiments*, (New York: Springer-Verlag)
[11] Tavassoli Z, Carter J N and King P R 2004 Errors in history matching *SPE J.* SPE 86883
[12] Liu J S 2001 *Monte Carlo strategies in Scientific Computing* (New York Berlin Heidelberg: Springer-Verlag)
[13] Erbas D and Christie M 2007 How does sampling strategies Aect Uncertainty Estimations? *Oil & Gas Science and Technology- Rev. IFP* 62 No.2

Figure 3. FOPT10 distributions obtained from the prior (1) and posterior (2) input distribution. Note that the value corresponding to the history data is FOPT10=820000