Abstract

Reservoir simulation models incorporate physical laws and reservoir characteristics. They represent our understanding of sub-surface structures based on the available information. Emulators are statistical representations of simulation models, offering fast evaluations of a sufficiently large number of reservoir scenarios, to enable a full uncertainty analysis. Bayesian History Matching (BHM) aims to find the range of reservoir scenarios that are consistent with the historical data, in order to provide comprehensive evaluation of reservoir performance and consistent, unbiased predictions incorporating realistic levels of uncertainty, required for full asset management. We describe a systematic approach for uncertainty quantification that combines reservoir simulation and emulation techniques within a coherent Bayesian framework for uncertainty quantification.

Our systematic procedure is an alternative and more rigorous tool for reservoir studies dealing with probabilistic uncertainty reduction. It comprises the design of sets of simulation scenarios to facilitate the construction of emulators, capable of accurately mimicking the simulator with known levels of uncertainty. Emulators can be used to accelerate the steps requiring large numbers of evaluations of the input space in order to be valid from a statistical perspective. Via implausibility measures, we compare emulated outputs with historical data incorporating major process uncertainties. Then, we iteratively identify regions of input parameter space unlikely to provide acceptable matches, performing more runs and reconstructing more accurate emulators at each wave, an approach that benefits from several efficiency improvements. We provide a workflow covering each stage of this procedure.

The procedure was applied to reduce uncertainty in a complex reservoir case study with 25 injection and production wells. The case study contains 26 uncertain attributes representing petrophysical, rock-fluid and fluid properties. We selected phases of evaluation considering specific events during the reservoir management, improving the efficiency of simulation resources use. We identified and addressed data patterns untracked in previous studies: simulator targets, e.g. liquid production, and water breakthrough lead to discontinuities in relationships between outputs and inputs. With 15 waves and 115 valid emulators, we ruled out regions of the searching space identified as implausible, and what remained was only a small
proportion of the initial space judged as non-implausible ($\sim 10^{-11\%}$). The systematic procedure showed that uncertainty reduction using iterative Bayesian History Matching has the potential to be used in a large class of reservoir studies with a high number of uncertain parameters.

We advance the applicability of Bayesian History Matching for reservoir studies with four deliveries: (a) a general workflow for systematic BHM, (b) the use of phases to progressively evaluate the historical data; and (c) the integration of two-class emulators in the BHM formulation. Finally, we demonstrate the internal discrepancy as a source of error in the reservoir model.

**Keywords:** Uncertainty Reduction, Bayesian History Matching, Emulation, Simulation Targets, Systematic Procedure

**Introduction**

One of the biggest challenges for the energy industry is how to deal with many sources of uncertainty. Reservoir simulation models describe the understanding and interpretation of sub-surface structures, incorporating available data and technology. Reservoir model calibration is an inverse problem based on historical reservoir data: a high dimensional, ill-posed, non-linear problem.

The ultimate goal of a calibration process is to provide background for well informed and efficient decisions. Reservoir calibration can reduce, but not eliminate, the uncertainty in calibrated models. Finding the whole class of scenarios capable of representing the reservoir historical behaviour is essential as it adds value to an asset by giving a realistic evaluation of reservoir performance and consistent predictions with corresponding uncertainties. Calibrated models drive, for example, recovery strategies optimisation and risk quantification.

Multiple possible solutions are inherent in inverse problems for imperfect models possessing a large number of uncertain attributes (inputs) and outputs (production data observed with uncertainty). Advanced calibration techniques have been developed and applied in the energy industry referred to as History Matching (Oliver and Chen 2011, Oliver et al. 2008), Data Assimilation (Evensen 2009; Carrassi et al. 2018) and Uncertainty Quantification and Reduction (Smith 2014).

Challenges in the calibration of reservoir models that we address with this study are:

- **Preservation of variability** in calibrated models/scenarios while considering several sources of uncertainty involved in the calibration process;
- **High dimensional input space** implying that a large number of simulator evaluations would be required to provide a consistent representation of the uncertainties and for the maintenance of geological realism;
- **Time and resources** to evaluate a large number of simulations required in a calibration process (i.e. ideally an exhaustive assessment of all possible scenarios would be made);
- **High dimensional output space** requiring advanced analytic techniques to suitably capture patterns, trends, and associations of data with diverse characteristics.

Craig et al. (1995) made significant progress in this area by formulating an alternative calibration technique, combining simulation models and emulators under a Bayesian framework, referred to as Bayesian History Matching (BHM). The originality of this approach is seen in the identification of the whole range of solutions which are concurrently compatible with the historical reservoir performance, given major sources of uncertainty such as observation error and model discrepancy.

This class of technique is qualified as Bayesian mainly for (a) the uncertainty quantification, which is based on prior knowledge about the model and process characteristics, and (b) the uncertain nature of the reservoir parameters, attributed to our lack of knowledge about the subsurface (in contrast to a random
process, as in a frequentist approach). The traditional nomenclature for inverse problems in the petroleum industry is reformulated, building the term *Bayesian History Matching*.

A central element of the BHM approach is the combination of simulation models and emulators. We present their main features in Table 1 and follow with additional definitions.

<table>
<thead>
<tr>
<th><strong>Table 1</strong>—Main characteristics of simulation models and emulators, central elements of the Bayesian History Matching approach.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Simulation models</strong></td>
</tr>
<tr>
<td>Definition</td>
</tr>
<tr>
<td>State</td>
</tr>
<tr>
<td>Cost</td>
</tr>
</tbody>
</table>

- **Data set** consists of several possible reservoir model scenarios (inputs) and the corresponding quantities of interest (outputs); the scenarios are sampled in order to cover the input parameter space;
- **Training data set** is a data set used to construct emulators (*e.g.* select active variables and fit the statistical model); in this study, training data are plotted in reddish;
- **Testing data set** is an independent data set applied to validate and select concurrent emulators constructed with the training set for a given quantity of interest; plotted in bluish.

**Bayesian History Matching**

A fundamental characteristic distinguishes Bayesian History Matching (BHM) approaches from other calibration techniques. In order to find the whole range of solutions for the calibration problem, the focus under a BHM is on the following question:

*Which parts of the input space are unlikely to lead to acceptable fits between the model outputs and historical data?*

From a practical perspective, the rationale for solving the inverse problem is reformulated. We look for what is not the solution to the problem in order to rule this part out of the input space and identify appropriate solutions. This process is far more effective than directly looking for good solutions as often large regions of the input space can be ruled out by considering only small numbers of outputs.

The core strategy of this approach relies on three elements (Figure 1). Firstly, a restricted number of scenarios is carefully sampled from the search space of interest (the input parameter space formed from the uncertain attributes). In Figure 1.a, a sample of scenarios from a search space of two dimensions (uncertain attributes named $\phi$ and $k_r$) is illustrated via a pair plot.

Using the reservoir simulator, we evaluate the sample set to acquire outcomes corresponding to measurable physical quantities (*e.g.* pressure and production rates). These outcomes are usually called quantities of interest or reservoir outcomes. We construct emulators (Figure 1.b) based on the data set from the simulated samples by applying appropriate statistical techniques. Because emulators are fast, a large number of scenarios can then be evaluated, allowing exhaustive exploration of the search space.

The implausibility measure (Figure 1.c) examines the difference between the historical value and the expected outcome from the emulator, compared to all the major uncertainties that arise: from using the emulator as a representation of the simulator, from using the measurement data to represent the real data and from using the reservoir model to represent the real physical reservoir. The implausibility measure is important because it identifies which parts of the input space are unlikely to lead to acceptable fits (*i.e.* the implausible regions) and answers the question stated above. We rule out regions of the search space identified as implausible, and what remains is only the proportion of the original space currently judged as non-implausible. Because a large number of scenarios are evaluated via the emulator, we can draw detailed images of the implausible regions projected into two-dimensional subspaces, referred to as implausibility pairs plots (red, Figure 1.c).
An iterative process is implemented in order to narrow the searching space sequentially. A new wave refocuses the search space on the current non-implausible regions (green, Figure 1.c), from where a limited number of carefully designed scenarios is sampled. This data set is used to construct additional and more accurate emulators leading to further space reduction. The iterative process through waves sequentially discard regions of the input space, continually refocusing our search on the regions judge as non-implausible (Vernon et al. 2018; Williamson et al. 2017). Jointly, the three elements of Figure 1 are able to mimic complex data structures with simple emulators in a reliable way.

The motivation for exploring this approach in the context of petroleum reservoirs includes: (a) efficient use of simulation and associated computational cost; (b) extensive exploration of reservoir scenarios to secure the representativeness of the whole class of models capable of describing the reservoir historical data; (c) integration of diverse sources of uncertainty coming from observed data and an imperfect simulation model; and (d) qualitative and quantitative insights about the reservoir description and performance.

BHM has been successfully employed across a diverse set of scientific domains, e.g. galaxy formation (Vernon et al. 2010), petroleum engineering (Moreno et al. 2018; Ferreira et al. 2014), climate modelling (Williamson et al. 2017) and system biology (Vernon et al. 2018). The main challenges for the application of BHM for reservoir models remain in (a) high dimensionality of inputs (e.g., spatial uncertain attributes linked to porosity, permeability and facies maps) and outputs (several observed measures to be used in the calibration process); (b) identification of structures, dependency and interdependency in data sets; (c) modelling structures of discrepancy between reservoir models and the real reservoir; and (d) time to evaluate scenarios through simulation.

Objective
We aim to advance the applicability of Bayesian History Matching (BHM) for reservoir studies by offering:

1. **A systematic workflow** to structure BHM techniques. The workflow is designed to (a) scale up to high dimensional input and output data, (b) secure flexibility on combining diverse emulation techniques, and (c) perform stages automatically, centring users’ focus on analysis and synthesis;

2. **Emulators for two-class quantities of interest** applied to mimic outputs from the reservoir suitably labelled as binary (e.g. outputs related to simulation targets such as liquid production rate);
3. **Phases of evaluation** which split historical data into physically meaningful periods to gradually introduce data into the analysis to take advantage of information from early time;

4. **Model discrepancy** demonstration, showing that noisy data defined as simulator target during the historical period leads to a discrepancy between the reservoir model and the real reservoir.

This work focuses on the development of methodology.

We apply our procedures in a case study to discuss analytical aspects and to demonstrate its applicability for the analysis of petroleum reservoirs. This case study considers (a) 26 uncertain reservoir attributes, (b) 25 active wells and (c) 4018 days of historical data. The hypothetical reality is one of the combinations of the uncertain attributes considered. The aim is to test the potential of the procedure for a complex reservoir model under a controlled situation while illustrating the main steps of the methodologies developed.

**Statistical Methodology**

We now introduce the standard form of Bayesian History Matching (BHM), as described by Vernon et al. (2018). This framework is then extended by the incorporation of two-class emulators to address specific structures in reservoir data sets, and subsequently by the use of phases of evaluation. We provide the background information for the applied statistical models and emulator validation.

**Formulation of Bayesian History Matching**

Measurable quantities from the real reservoir are denoted by the vector $z$ (e.g. well pressure in a given time). They result from the sum between the corresponding $y$ quantities from the real physical reservoir and observational errors $e$. Sources of observational errors depend on field's well surveillance programme and the measurement process in place, including (a) equipment calibration (random and systematic types); (b) chemical analysis for gas-oil-ratio; (c) apportionment of field production to well production and production testing, and (d) data manipulation.

$$ z = y + e $$

(1)

The reservoir simulation model computes a corresponding vector of quantities via a function $f$ of $x$, where $x$ is a vector of input parameter values representing a reservoir scenario (i.e. a combination of the uncertain attributes). We represent the difference between the real reservoir and the reservoir model as the model discrepancy term $\epsilon$. Equation 2 indicates that even evaluating the most appropriate scenario $x^*$ through $f(x^*)$, there is a difference $\epsilon$ between the reservoir model and the real reservoir.

$$ y = f(x^*) + \epsilon $$

(2)

Model discrepancy arises, for example, from simplifications in physical laws modelling the phenomena in place (e.g. multi-phase flow in porous media), reservoir conditions and characteristics.

As an example, the case study described on page 14 has $x^*$ as the hypothetical reality, which is a vector of known uncertain attributes. In real applications, $x^*$ is undefined: it is very unlikely that a simulation model perfectly represents the real field behaviour. Each element $i=1,...,q$ of the vector $f(x)$ corresponds to measurable quantities of interest of the reservoir and is defined as $f_i(x)$.

Emulators $f^*(x)$ are statistical approximations of simulation models.

They offer fast evaluations of sufficiently large numbers of scenarios from the input space and enable a full uncertainty analysis (Vernon et al. 2010; Craig et al. 1997; Craig et al. 1995). For any input scenario, an emulator provides a mean and distribution describing how close it is likely to be to the simulator output. The expected outcome is a plausible interpolation (or extrapolation) of the training data. The distribution around the mean is a reasonable expression of emulator uncertainty (O'Hagan 2004). Distinct from simulators, emulators do not directly incorporate reservoir conditions, characteristics or physical laws. Table 1 summarised complementary characteristics between simulators and emulators.
We employ an implausibility measure to describe which parts of the input space are unlikely to lead to acceptable fits between the model output and observed data. Equation 3 standardises the difference between the emulator expectation \( E(f_i^*(x)) \) for the \(-i\)th considered output and corresponding historical data \( z_i \) by all uncertainties identified in the process. They are expressed in terms of the variance of the emulator \( \text{Var}(f_i^*(x)) \), the variance of model discrepancy \( \text{Var}(\epsilon_i) \) and the variance of the observation error \( \text{Var}(e_i) \).

\[
I_i^2(x) = \frac{[E(f_i^*(x)) - z_i]^2}{\text{Var}(f_i^*(x)) + \text{Var}(\epsilon_i) + \text{Var}(e_i)}
\]  

For each scenario \( x \), the implausibility measures \( I_i(x) \) calculated from each output emulated, \( i \neq [1,q] \), can be combined in various ways. Among the diverse possibilities, the first, second or third maximal implausibility can be selected, depending on the number of emulators and the understanding about the uncertainties in the calibration process. In this application, we use the maximal implausibility as in Equation 4. Alternative options are discussed by Vernon et al. (2010).

\[
I_M(x) = \max_{i \in Q} I_i(x)
\]  

In parallel, the cut-off \( \omega \) addresses an assessment about the appropriateness of the assumptions made to compute implausibility and the associated mean distribution of \( I_M \). Pukelsheim (1994) states that for all continuous unimodal distributions (e.g. normal, double exponential, chi-squared, t, lognormal), it holds that 95% coverage is obtained within 2.98 standard deviations from the mean. For example, we defined for our application an implausibility cut-off \( \omega = 3 \) considering that the assumptions of continuous and unimodality are reasonable for the distribution of \( I_M \). The cut-off \( \omega \) defines a boundary at each iteration or wave, to label regions of the search space as either:

- **Implausible scenarios** when \( I_M(x) > \omega \), specifying that the combination of inputs \( x \) is unlike to lead to acceptable fits, i.e. the scenario response is sufficiently far from the historical data to be ruled out from the search space;
- **Non-implausible scenarios** when \( I_M(x) \leq \omega \), which result from a good fit to historical data (the numerator in Equation 3), or high variance of \( f_i^*(x) \), \( \epsilon_i \) or/and \( e_i \) (the denominator of Equation 3). A high emulator variance can be resolved in later waves, where more accurate emulators are constructed.

Continuous quantities of interest are traditionally modelled using an emulator given in Equation 5. Active variables \( x_{A_i} \) are a subset of inputs selected as the most influential for a given quantity of interest \( i \). They are used to construct the corresponding emulator:

\[
f_i^*(x) = \sum_j \beta_{ij} g_{ij}(x_{A_i}) + u_i(x_{A_i}) + w_i(x)
\]  

The first term is a regression term, where \( g_{ij} \) are known deterministic functions of the active variables \( x_{A_i} \) and \( \beta_{ij} \) are unknown scalar regression coefficients. A common choice is low order polynomials, e.g. a regression with first and second order, and interactive terms (Vernon et al. 2018). The other terms, \( u_i(x_{A_i}) \) is a Gaussian process over \( x_{A_i} \) and its associated nugget \( w_i(x) \), which is related to the fact that only a subset of uncertain attributes are included in the emulator as active variables.

Choosing appropriate statistical models to construct emulators is a strategic step while performing Bayesian History Matching (BHM). The choice of emulator models can influence computational effort.
required, number of waves and, ultimately, the quality of the resulting calibrated model. Their performance to properly mimic simulators also depends on the size, dimensionality, quality and nature of the data used.

In our analysis, we identified two categories of quantities of interest: continuous and binary, the latter having not been previously employed in a BHM setting. Concurrently, we choose algorithms capable of modelling them: regression and two-class classification models. We recall some features of these models in the next sections. James et al. (2013) offer a didactic understating of these models with useful application in R code.

Indicators for diagnostics support the analysis and validation of emulators. Table 2 summarises the main features of selected indicators. Combined, they assess emulators for the continuous and binary quantities of interest. We expand the description for information index, credible interval diagnostics, positive and negative predictive value in the next sections and demonstrate their application on page 23. For further information about adjusted-$R^2$ and Normalised Root Mean Square Error (RMSE$_n$) for validation of emulators, we recommend Moreno et al. (2018). Other indicators can be used when simulating a test set is not affordable (e.g. leave-one-out diagnostics). For more details, see Bastos and O‘Hagan (2009).

### Table 2—Summary of selected indicators - their combination enables to evaluate and select emulators for continuous and two-class quantities of interest in a comprehensive way.

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Set</th>
<th>Quantity of Interest</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information index ($D_{info}$)</td>
<td>Training &amp; test</td>
<td>Continuous &amp; binary</td>
<td>Proportion of scenarios expected to be implausible</td>
</tr>
<tr>
<td>Credible Interval Diagnostics ($D_{cI}$)</td>
<td>Training &amp; test</td>
<td>Continuous</td>
<td>Proportion of scenarios for which simulation outcome is covered by the emulator credible interval</td>
</tr>
<tr>
<td>Positive Predictive Value (PPV)</td>
<td>Training &amp; test</td>
<td>Continuous</td>
<td>Proportion of implausible scenarios correctly classified</td>
</tr>
<tr>
<td>Negative Predictive Value (NPV)</td>
<td>Training &amp; test</td>
<td>Binary</td>
<td>Proportion of non-implausible scenarios correctly classified</td>
</tr>
<tr>
<td>Adjusted-$R^2$</td>
<td>Training</td>
<td>Continuous</td>
<td>Proportion of the variance of the output explained by the regression</td>
</tr>
<tr>
<td>RMSE$_n$</td>
<td>Test</td>
<td>Continuous</td>
<td>Normalised Root Mean square Error</td>
</tr>
</tbody>
</table>

### Information index
Information index $D_{info}$ (Equation 6) is an estimation of the proportion of the remaining input space that can be ruled out (as implausible) in the process. It is computed through the implausibility measure and implausibility cut-off $\omega$, where $\mathbf{1}(\cdot)$ is an indicator function. The higher the information index, the better: it means that we are able to identify a large proportion of implausible scenarios.

$$D_{info} = \frac{1}{n_{sce}} \sum_{k=1}^{n_{sce}} \mathbf{1}(I(x^k) > \omega)$$  \hspace{1cm} \text{(6)}$$

The set of size $k=1, \ldots, n_{sce}$ input locations can be a training set, test set, or a collection of new scenarios (for implausibility analysis, for example). Therefore, we define the information index (a) $D_{info,i}$ for an emulator of the output $i$, based on the implausibility $I_i(x^k)$, and (b) $D_{info,M}$ for a set of emulators, based on the implausibility $I_M(x^k)$. In Equation 6, both $I_i(x^k)$ or $I_M(x^k)$ are based on the emulator expected outcome and emulator variance through Equations 3 and 4.

Alternatively, we can estimate an adapted information index $\widetilde{D}_{info}$ based on outcomes from simulated scenarios. We define this indicator by making two changes to the original measure:

1. The emulator expectation $E(f^*_i(x))$ is replaced by $f_i(x)$, the simulation outcome;
2. The emulator variance $Var(f^*_i(x))=0$.

In this way, we have:
\[ I_i^2(x) = \frac{[f_i(x) - z_i]^2}{\text{Var}(\epsilon_i) + \text{Var}(\epsilon_i)} \]  

(7)

The corresponding adapted information index \( \bar{D}_{\text{info},i} = 1/n_{sce} \sum_{i=1}^{n_{sce}} 1(\bar{I}(x^k) > \omega) \). We provide examples of the application of \( \bar{D}_{\text{info},i} \) in the Results and Discussions (to evaluate the results of the application of the uncertainty reduction procedure) and in Appendix B (for the selection of outputs to emulate).

**Continuous quantities of interest**

The statistical model that we apply to construct emulators for continuous quantities of interest (e.g. oil production, pressure) is a simple version of Equation 5: low order polynomial regression. Multiple indicators and qualitative analysis of diagnostics plots aid the validation of emulators. Besides the information index, we assess the Credible Interval Diagnostics (Bastos 2010) and monitor adjusted-R^2 and RMSEn.


**Credible Interval Diagnostics.** Based on Bastos and O’Hagan (2009), we denote by \( CI_i(\alpha) \) a 100\( \alpha \)% credible interval for the simulator output \( f_i(x^k) \) at \( k=1, \ldots, n_{sce} \) input locations denoted \( x^k_{A_i} \), with \( 0 \leq \alpha \leq 1 \).

The Credible Interval Diagnostics is defined in Equation 8, where \( I(\cdot) \) is an indicator function. It computes the proportion of scenarios for which the simulation outcome is covered by the credible interval provided by the emulator.

\[ D_{CI}(f_i(x)) = \frac{1}{n_{sce}} \sum_{k=1}^{n_{sce}} 1 \left(f_i(x^k_{A_i}) \in CI^k(\alpha)\right) \]  

(8)

Credible Interval Diagnostics is an appropriate indicator to determine whether the uncertainty estimation for an emulator corresponds to its actual uncertainty. We would expect the observed value for \( D_{CI}(f_i(x)) \) to be close to \( \alpha \) when the emulator uncertainty is appropriately estimated. In fact, an appropriate emulator is neither (a) under-confident, with too large uncertainty on expected outcomes, nor (b) overconfident, with too small uncertainty on expected outcomes.

We are cautious about parts of the input space which are wrongly excluded based on the emulator’s expectation. Therefore, the evaluation of an emulator is critical to guard against the possibility of being overconfident: it could lead to wrong conclusions about the implausible space. A cut-off value describes the lower \( D_{CI} \) limit to validate an emulator (e.g. \( D_{CI} > 0.85 \)) and we also select concurrent emulators based on this indicator.

**Binary quantities of interest**

Binary quantities of interest have two distinct classes of outcomes and present discontinuous behaviour across two discrete regions separated by a boundary. In reservoir studies, binary outcomes result from a classification of continuous outcomes from simulations into binary data. We identified a two-class pattern in well data such as cumulative liquid and water production for a given time and average bottom hole pressure within a time window. Therefore, we introduce classification models into our statistical formulation. Two-class emulators recognise to which of these regions a new scenario belongs to, before simulating the new scenario.

The first stage classifies the simulator outcome in two classes, obtaining a binary outcome. In Figure 2.a, a pairs plot for three illustrative uncertain attributes (\( \phi, k_x \) and \( k_z \)) displays two distinct regions: the green region is classified as 0; the red region is classified as 1.
To obtain these two regions, a transformation function maps \( f_i \), the simulator output, onto a binary outcome \( B_i \). Equation 9 defines \( B_i \) by considering the difference between \( f_i(x) \) and observed data \( z_i \) with respect to a tolerance \( tol_i \). An alternative formulation is proposed with an indicator function \( 1(\cdot) \). The magnitude of the tolerance denotes the uncertainty due to the observational error \( e_i \) and the model discrepancy \( \epsilon_i \). Note that our convention labels as 0 the non-implausible and as 1 the implausible regions.

\[
B_i(x) = \begin{cases} 
0, & \text{if } |z_i - f_i(x)| \leq tol_i \\
1, & \text{otherwise} 
\end{cases}
\]

We focus on removing one of the discrete regions, the non-implausible one. Nevertheless, we emphasise that a similar approach is adaptable for other data patterns; for example, a data structure that requires the identification of both discrete regions because it is convenient to emulate them separately.

In the second stage, Figure 2.b, we apply classification models (also called classifiers) to construct emulators. Classifiers provide the probability of a given new scenario \( x \) to be labelled as 1. Therefore, before simulating new scenarios, we are able to predict this probability. Logistic regression and linear discriminant analysis are classical two-class classifiers. They are implemented in R with lda and glm.

The logistic function (Equation 10) in the logistic regression keeps the predicted probabilities in the interval between 0 and 1. This statistical model is fitted based on maximum likelihood (James et al. 2013).

\[
P(B_i = 1|x) = \frac{e^{\sum_j \beta_{ij}g_{ij}(x_{A_i})}}{1 + e^{\sum_j \beta_{ij}g_{ij}(x_{A_i})}}
\]

We derive Equation 11, having in the left-hand side the log-odds or logit and on the right-hand side a formulation equivalent to linear regression models. Therefore, we model the conditional distribution of the response \( B_n \), given the predictors \( x_A \) as:

\[
\log \left( \frac{P(B_i = 1|x)}{1 - P(B_i = 1|x)} \right) = \sum_j \beta_{ij}g_{ij}(x_{A_i})
\]

With linear discriminant analysis, the distribution of the predictors \( x_A \) is modeled separately in each of the response classes (i.e. given \( B_n \), the possible number for classes \( k \) is two: 0 or 1). Bayes’ theorem is used to flip these around into estimates for \( P(B_i = k|X = x_{A_i}) \). When these distributions are assumed to be normal,
it turns out that the model is very similar in form to logistic regression. From James et al. (2013), reasons
to consider a linear discriminant analysis as an alternative to logistic regression include:

- When the classes are well-separated, the parameter estimates for the logistic regression model are
  surprisingly unstable. Linear discriminant analysis does not suffer from this problem.
- If the size of the training set is small and the distribution of the predictors $x_A$ is approximately
  normal in each of the classes, the linear discriminant model is again more stable than the logistic
  regression model.

The flexibility to integrate diverse statistical models is a strength of our BHM systematic procedure,
presented on page 12. With quantitative and qualitative diagnostics, we select the best among concurrent
emulators for the same quantity of interest. The cross-plot of Figure 2.b compares binary data from simulated
scenarios with the corresponding probabilities predicted by the emulator. A threshold probability called the
decision boundary establishes a connection between Figures 2.b and 2.c. The labels $B_i*$ for all the scenarios
to the left of the decision boundary are 0; for all the scenarios to the right of the decision boundary are 1.

In Equation 12, this boundary links the probability estimated by the emulator $P(B_i=1|x)$ with the labels
$B_i*$. We define $\omega*$ larger than the implausibility cut-off $\omega$, and this operation rescales the outcome from the
emulator defining a new implausibility measure for the two class emulators.

$$I_i(x) = B_i^*(x) = \begin{cases} 1 \cdot \omega*, & \text{if } P(B_i = 1|x) \geq \text{decision boundary} \\ 0 \cdot \omega*, & \text{if } P(B_i = 1|x) < \text{decision boundary} \end{cases} \quad (12)$$

In the third stage, we can classify training and test scenarios with the emulator. Figure 2.c presents
the classification of the training set by the emulator, highlighting only one scenario wrongly excluded.
The discussion about predictive values in the next section emphasises that among the scenarios ‘wrongly
excluded’, ‘wrongly kept’, ‘correctly kept’ and ‘correctly excluded’, we are mainly concerned with
scenarios wrongly excluded.

With this formulation for binary quantities of interest, we integrate uncertainties of the calibration process
with tol, (incorporating observational error and model discrepancy) and the decision boundary (emulator uncertainty).

**Positive and Negative Predictive Value**

Predictive values estimate the probability of the emulator labels being correct. For a data set, the proportion
of scenarios correctly classified as (a) implausible - label 1 - is the Positive Predictive Value (PPV); (b)
non-implausible – label 0 - is the Negative Predictive Value (NPV) (Altman and Bland 1994). Equations
13 and 14 define:

$$PPV = \frac{\text{True 1}}{\text{True 1} + \text{False 1}} \quad (13)$$
$$NPV = \frac{\text{True 0}}{\text{True 0} + \text{False 0}} \quad (14)$$

The elements of these indicators are possible combinations between emulator and simulator labels. In
Figure 3.a, we have: True 0 are the scenarios correctly kept by the emulator; True 1 are the ones correctly
excluded; False 0 are the scenarios wrongly kept, and False 1 are the ones wrongly excluded. While PPV
judges the proportion of scenarios correctly excluded by the emulator, the NPV controls the emulator
efficiency to keep suitable scenarios. Importantly, the iterative nature of BHM enables us to rule out in later
waves scenarios that are wrongly kept. Nevertheless, regions of search space wrongly excluded are critical.
Therefore, we are extra cautious about PPV.
Our objective is to choose an optimal decision boundary that maximises the PPV while keeping NPV as high as possible. The graph in Figure 3.b facilitates this choice. It plots PPV and NPV versus each possible decision boundary. For this example, an optimal decision boundary named cutpoint is approximately 0.85, where PPV is the maximal.

We set a decision boundary linked to the training data set. Our aversion to wrongly exclude scenarios is implemented via two strategies: (a) increment the training set cutpoint to accommodate small dissimilarities between training and test sets characteristics; (b) choose a fixed unbalanced decision boundary setting that a very high emulator probability (e.g. 0.95) is required to exclude a scenario.

**Description and Application of the Systematic Procedure**

The proposed systematic workflow consists of a sequence of 20 steps collected in six groups where simulation and emulation techniques are combined to reduce uncertainty in a petroleum reservoir. The main features of the procedure are: (a) **repeatability**, due to the sequential nature of the steps, logically associated; (b) **flexibility**, since the steps of the high-level structure are adaptable to project requirements; (c) **scalability** to higher dimensions, as specific steps are planned to accommodate techniques for dimensionality reduction. Several steps are automatable and enable the team to focus on the analysis and synthesis of the project. Eleven steps (1 to 4, 7, 8, 12, 15, 17, 18 and 20) concentrate such activities.

We overview these six groups before detailing the description of the systematic workflow (Figure 4):

- **Definition of case study**: we set the case study by combining our knowledge, information and data about the petroleum reservoir, models, uncertainties and observational error in historical data;
- **Definition of strategy for Bayesian History Matching (BHM) and Uncertainty Reduction (UR)**: we plan the data analysis in order to conduct the calibration process efficiently. Two objectives are explored: (a) we identify outputs to emulate considering data structure, and (b) we select phases of evaluation considering specific events during the reservoir management (e.g. wells and field behaviour);
- **Data preparation**: we build independent training and test sets of size $m$ and $n$ scenarios respectively. The test set is used for the selection and validation of emulators; nevertheless, the workflow is adaptable to studies which an independent test set is not affordable (e.g. simulation cost). The estimation of model discrepancy incorporates the uncertainty due to the model being an imperfect representation of the real reservoir. Although a full description of model discrepancy is not the main focus of this paper, we demonstrate the impact of a particular form of model inaccuracy related to simulation targets;
• **Construct and validate emulators**: We select outputs to emulate targeting the most informative ones. Complex or simple emulators are constructed as needed. Implausibility analysis only considers valid emulators;

• **Evaluation of scenarios and uncertainty reduction**: All valid emulators are applied simultaneously in order to evaluate scenarios through the implausibility measures;

• **Decision for phase and wave** (definitions below): We check the need for a new wave, a new phase or both. We anticipate our comments on the criteria that we applied. In *STEP 15*, a minimum number of scenarios in the training set $\lambda.m$ is required to construct new emulators (*e.g.* $0.5m$ scenarios); In *STEP 17* ‘Criteria to change phase met?’, we applied the minimum proportion of remaining space ruled out with the emulators constructed until a given wave (*e.g.* $D_{\text{info,M}} \geq 70\%$). In *STEP 18* – ‘Criteria to end calibration met’, we defined the criteria that *phase* is the last phase of evaluation. It indicates that our uncertainty reduction ends when the last phase is evaluated and when $D_{\text{info,M}} \leq 70\%$.

---

**Figure 4**—Systematic workflow for uncertainty reduction applying the Bayesian History Matching approach; six groups represent the 20 steps; steps 1 to 4, 7, 8, 12, 15, 17, 18 and 20 concentrate activities for analysis and synthesis; the test set is independent and optional; each step of this high-level structure can be planned to answer requirements specific for a study; waves imply on new simulations of scenarios; phases increment the amount of historical information considered in the process.
We highlight two core definitions in this systematic procedure:

- **Wave**: an iterative portion of the procedure whereby we simulate a limited number of scenarios. The corresponding quantities of interest are computed, emulated and used to reduce input space. To iterate through waves allows one to sequentially discard regions of the input parameter space, refocussing our search on the remaining non-implausible inputs (Vernon et al. 2018). Waves greatly improve the efficiency of the uncertainty reduction process.

- **Phase of evaluation** (or phase): a distinct period in the historical period which is chosen considering reservoir behaviour and operational conditions. Calibrating models through phases allows one to gradually introduce historical data in the analysis in order to (a) explore distinct relationships between uncertain attributes and reservoir outputs, (b) reduce the computation effort spent in simulations and (c) gradually incorporate the information available.

To introduce the reader to the conceptual mechanism of waves and phases, we propose an illustrative example in Figure 5 for the training set. In Wave 1, we have 100 scenarios simulated until the end of the Phase of evaluation 1. Four valid emulators are constructed in this wave. The implausibility analysis of the set result in 18 non-implausible scenarios (smaller than $\lambda.m=50$). Since $D_{\text{info,M}}$ of Wave 1 is larger than 70%, the criteria in **STEP 17** is not met yet, we keep evaluating the same Phase 1 and move onto Wave 2.

![Figure 5—Conceptual mechanism of waves and phases in the systematic procedure](image)

Wave 2 supplements the final training set of Wave 2 with 100 new non-implausible scenarios, resulting in an initial training set size of 118 scenarios. They enable the construction of two valid emulators, classifying as implausible 103 out of the 118 initial scenarios in Wave 2. Note that the row ‘number of emulators’ accounts only for emulators constructed in a given wave. The emulators are evaluated cumulatively, meaning that in Wave 2, the total number of emulators evaluated is $4+2=6$.

The same logic of the transition between Wave 1 and 2 applies in the transition between Wave 2 and 3, which has an initial training set of 115 scenarios. The two emulators constructed in Wave 3 do not rule out a large proportion of the search space (i.e. are not informative enough). Then, **STEPS 17 and 18** call a new phase of evaluation, bringing in further (later in time) outputs for consideration. Wave 4 is the first wave that uses simulations that run until the end of Phase 2. Therefore, only 100 scenarios are available for this phase, but 168 is the size of the initial training set for Phase 1. The following waves keep a similar logic.

**Definition of the case study**

In **STEP 1**, the asset team defines a reservoir simulation model representing available information (e.g. seismic and well log data) and relevant uncertainties in the reservoir. Ranges and distributions define uncertain attributes which we expect to be informative about the reservoir behaviour. The uncertainty on these attributes arises, among other things, from our lack of knowledge of the sub-surface.

Uncertain attributes are classified in several ways: (a) uncertain physical properties themselves (e.g. rock compressibility, fluid viscosity) or pseudo-properties, which are not the actual properties but used to describe
relationships between real-world quantities (e.g. effective permeability for phases, spatial properties of grid blocks); (b) as independent (e.g. fluid properties and rock compressibility) or dependent (e.g. porosity and permeability); (c) continuous and discrete (e.g. facies, PVT tables, fault existence).

Additionally, the asset team can set up the numerical simulator with tuning parameters and numerical approximations, among other things, which can have a significant impact in simulation time (Avansi et al. 2019), and give rise to additional uncertainties.

For the application of the proposed procedure, we established a synthetic black-oil reservoir model with a known hypothetical reality called HR-82 and based on the benchmarking case (Avansi and Schiozer, 2015). We describe the relevant features of HR-82 for this study in the main text. The full description is available in Appendix A, where we detail how we defined the uncertain attributes.

An important feature is a sealing barrier separating the reservoir in two compartments (east and west blocks). The production and injection wells are represented by red and blue rectangles in Figure 6. The reservoir is spatially divided into 13 regions. Given the focus of this study, to describe and demonstrate methodological development, we considered uncertain the parameters for regions 6, 8 and 10, leading to a total of 26 parameters, a sufficient number for our purposes. We highlight some steps that are planned to enable scalability on the number of dimensions (see STEPS 3 and 10).

We defined all parameters as continuous and uniformly distributed, and we divide the 26 uncertain parameters of Table 3 into four classes:

- **Global parameters** (first two lines in blue, two uncertainties) influence the whole reservoir model;
- **Regional parameters** (four lines in green, twelve uncertainties) modify petrophysical properties of each the uncertain regions (regions 6, 8 and 10 in this study) with four parameters;
- **Sector parameters** (three lines in orange, three uncertainties) model attributes of the east block;
- **Local parameters** (last line in yellow, nine uncertainties) are independent multipliers of the Well Index (WI) for the nine wells located in the regions 6, 8 and 10.
Table 3—Uncertain parameters considered in the case study applied to demonstrate the systematic BHM procedure – parameters in blue have global influence; in green, regional influence (four attributes for each of the three regions); in orange, sector parameters influence only the east block; in yellow, local influence in nine wells. Appendix A provides the complete description.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Ranges</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c_p$</td>
<td>Rock compressibility</td>
<td>[10, 96] E-6</td>
</tr>
<tr>
<td>$c_{cv}$</td>
<td>Relative permeability</td>
<td>[0.86, 1.28]</td>
</tr>
<tr>
<td>$m_{phi}$</td>
<td>Porosity multiplier</td>
<td>[0.75, 1.25]</td>
</tr>
<tr>
<td>$A$</td>
<td>$k_x$ angular coefficient</td>
<td>[0.135, 0.175]</td>
</tr>
<tr>
<td>$B$</td>
<td>$k_x$ linear coefficient</td>
<td>[-0.4, 1.1]</td>
</tr>
<tr>
<td>$m_{kz}$</td>
<td>$k_z$ multiplier</td>
<td>[0.1, 0.5]</td>
</tr>
<tr>
<td>$PVT_{co,EB}$</td>
<td>Oil compressibility</td>
<td>[1.40, 1.62] E-3</td>
</tr>
<tr>
<td>$PVT_{o,EB}$</td>
<td>Oil viscosity related</td>
<td>[2.5, 50.0] E-4</td>
</tr>
<tr>
<td>$WOC_{eq}$</td>
<td>Water-Oil-Contact</td>
<td>[3169, 3179]</td>
</tr>
<tr>
<td>$W_{in}$</td>
<td>Well index factor (9)</td>
<td>[0.6, 1.4]</td>
</tr>
</tbody>
</table>

In STEP 2, we select and estimate the error of the historical data. A critical analysis of the available data is performed in order to assess the levels of uncertainty in the information (see examples of sources of error in the measurement process on page 5). In our application, the historical data of HR-82 is based on a hypothetical reality obtained from one of the simulated scenarios. The output from this scenario is noised up by adding both random and systematic errors. Table 4 presents the corresponding intervals of uncertainty in terms of ±3σ, where σ stands for the relevant standard deviations. Note that both contributions to the error are multiplicative in relation to the hypothetical measured data. We compute the corresponding variances of observational error $Var(e_i)$ to integrate into the implausibility measures.

Table 4—Error in the historical data. In our case study, it corresponds to the noise added to the production data of the hypothetical reality; random and systematic errors are defined in terms of three standard deviations and are multiplicative.

<table>
<thead>
<tr>
<th>Observed data</th>
<th>Random error ($±3σ_{e_i}^{ran}$)</th>
<th>Systematic error ($±3σ_{e_i}^{sys}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liquid production rate ($q_l$)</td>
<td>±0.06</td>
<td>±0.03</td>
</tr>
<tr>
<td>Water injection rate ($i_w$)</td>
<td>±0.06</td>
<td>±0.00</td>
</tr>
<tr>
<td>Water production rate ($q_w$)</td>
<td>±0.05</td>
<td>±0.05</td>
</tr>
<tr>
<td>Bottom hole pressure of production and injection wells ($P_{Din}, P_{bbh}$)</td>
<td>±0.02</td>
<td>±0.02</td>
</tr>
</tbody>
</table>

Define the strategy for the BHM and UR process

We dedicate STEPS 3 and 4 to the definition of the strategy for uncertainty reduction. These activities are critical (a) to remain efficient in the use of computational resources, (b) to enable the dimensional scalability of our proposed workflow, and ultimately (c) to succeed in the quantification of uncertainty.

In STEP 3, we identify inputs and outputs to emulate and we assess their relationship.

Firstly, our understanding of reservoir characteristics is used to perform informed judgements about physical relationships between outputs and uncertain attributes. This assessment avoids the consideration of spurious uncertain attributes as possible active variables for constructing emulators for some quantities of interest. For high dimensional case studies, we can also take advantage of (a) transformations in data, if coherent with physical laws and phenomena in place, background and examples available in Hoaglin et al. (1983), (b) supervised and unsupervised statistical learning techniques for dimensionality reduction.

Secondly, the selection of a sub-group of outputs to be considered during the iterative process makes sense for three reasons: (a) a high dimensional output space can be available (e.g. well data and seismic surveys) and trying to emulate all of them can be inefficient and unnecessary; (b) data quality can be diverse within the set available; (c) the assumption that all the quantities are independent is certainly not valid; therefore, the data structure can be considered viable for an effective dimensional reduction. For examples of dependency, see Fricker (2010).
We perform an initial evaluation of the appropriate outcomes to emulate. They can be of diverse natures, for example, single values (e.g. cumulative or rate at a given time); time series (e.g. coefficients of production curves); average values (e.g. average pressure on a given period); indicators of misfit between measured and simulated data (e.g. normalized quadratic deviation with sign, Almeida et al. (2014)).

For our case study, we identified inputs considering the sealing fault as the main reservoir feature, defining two non-communicating compartments. Uncertain attributes from the east block only influence outputs from wells in this compartment, similarly for west block inputs and outputs. Additionally, we assign the uncertain attributes with local impact to outputs from their respective wells. Therefore, emulators for outputs from the west block have a maximum of 11 active variables (two global, eight regional and one local attributes); in the east block, this maximum is 10 active variables.

In STEP 3, as quantities of interest, we have identified average pressure (of production $\bar{p}_{pbh}$ and injection well $\bar{p}_{ibh}$, water breakthrough time and cumulative quantities (liquid production $L_p$, water production $W_p$ and oil production $N_p$, and water injection $W_i$) at the end of phases of evaluation. We approach the concept of phases in STEP 4. These quantities capture the reservoir behaviour related to material balance, communication between wells and the start of production of a new fluid phase in the wells (water). The average pressure is computed in the period between two consecutive phases of evaluation while the well is active (e.g. for Phase 2, the window is between 518 and 1,461 days). In parallel, we derive observational error from Table 4 corresponds appropriately to these quantities.

In Table 5, we summarise the pre-selected quantities of interest used to construct emulators. We explain why and how we emulate some quantities as binary outputs in the sections named ‘Pattern 1: Simulator targets’ and ‘Pattern 2: Breakthrough Time’.

Table 5—Summary of quantities of interest identified in STEP 3, corresponding emulation techniques and models.

<table>
<thead>
<tr>
<th>Quantities of interest emulated</th>
<th>Emulation technique</th>
<th>Emulation model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cumulative quantities ($L_p, W_i$)</td>
<td>Two-class classification</td>
<td>Logistic regression and Linear discriminant analysis</td>
</tr>
<tr>
<td>Breakthrough Time (BT)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average pressure ($\bar{p}<em>{pbh}, \bar{p}</em>{ibh}$)</td>
<td>Regression</td>
<td>Linear and quadratic regression</td>
</tr>
<tr>
<td>Cumulative quantities ($W_p, N_p$)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In STEP 4, we define the phases of evaluation. Phases of evaluation are windows in the historical period. They enable exploration of relationships between inputs and outputs while considering the evolution of time-based physics governing flow in porous media. Phases of evaluation formalise our understanding about temporal changes in the reservoir, for example, number of phases flowing, drainage area, recovery mechanism from early to late production stages.

For the case study, the definition of phases of evaluation considers schedule of wells, water breakthrough, maintenance stops and field behaviour. Table 6 presents characteristics of the phases and illustrates the corresponding time window in a plot for field oil production versus time.
Table 6—Characteristics of the five phases of evaluation defined for the case study; the window for each phase is from time zero to the one given in days; the plot of field oil production rate $q_o$ versus time illustrates the time window.

<table>
<thead>
<tr>
<th>Phase</th>
<th>Time</th>
<th>Characteristics</th>
<th>Field $q_o$ versus time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>518 days</td>
<td>Four active production wells in the WB, early stage, one well with water BT, before 1st maintenance stop</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1461 days</td>
<td>Last time with only 4 active production wells, two wells with water breakthrough (BT)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2710 days</td>
<td>All 25 wells active, 21 wells opened for less than 700 days, 9 wells with BT, end of oil production plateau</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>3256 days</td>
<td>Oil production plateau ends, 13 wells with water BT, before 2nd maintenance stop</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>4018 days</td>
<td>Last time with historical data available, 14 wells with water breakthrough</td>
<td></td>
</tr>
</tbody>
</table>

Additionally, part of the strategy for BHM defines (a) how to combine the implausibility measure of the valid emulators; (b) the implausibility cut-off; and (c) all other design choices and decision statements in the workflow. The first two points depend on the stage of the reservoir modelling and calibration processes, the characteristics of the uncertain attributes (how much we know we do not know?) among others. For (a) and (b), we applied the choices discussed in the Statistical Formulation, on page 5. We discuss the remaining choices within the corresponding steps.

**Pattern 1: Simulator targets.** We usually specify observed production and injection rates as simulation targets for wells while simulating the historical period. Additionally, minimum and maximum pressures are set as boundary conditions for producers and injectors. The main objectives are to avoid unphysical fluid behaviour and PVT table extrapolation. Usually, pressure limits are much wider than the operational window, and not expected to be observed in the real field. That is, the lower pressure limit is much lower than the lower pressure required to produce fluid from the reservoir to the surface; the maximal pressure limit is much higher than rock fracture pressure. Figure 7 illustrates these settings in plots for liquid production rate $q_l$ and bottom hole pressure $p_{phb}$ versus time.

![Figure 7—Settings for reservoir simulation in historical period – (a) the observed liquid production rate of the well is the simulation target; (b) limit pressures are much wider than the operational window.](image)

The simulator calculates the bottom hole pressure required for producing or injecting at the given simulation target (*i.e.* the target is the primary control). The production rate of other flowing phases is dependent on the reservoir conditions surrounding the well. We usually perform the calibration process by a joint analysis of the ratios of fluid phases and bottom hole pressure. Nevertheless, when the simulation reaches a pressure limit in an attempt to deliver the targets, the simulation control of the well changes from target to the attained pressure limit.

This mechanism impacts the simulation outcomes in a particular manner described by two situations:
1. While a production well attains the lower pressure limit during the simulation, the production rate of the simulator is smaller than the specified target;
2. While an injection well attains the upper pressure limit during the simulation, the injection rate is smaller than the specified target.

We recognised that when one of these situations occurs, it induces a pattern in the simulation outcomes. The two distinct regions of behaviour characterise different relationships between inputs and outputs. Figure 8 illustrates these regions and relationships with a univariate example, i.e. only a porosity attribute is considered uncertain for the simulation runs. Figure 8.a shows whether the simulator reaches the simulation target during the whole time window (region 1) or not (region 2). This evaluation is made by the cumulative liquid production at the end of the time window. This binary behaviour is distinctly driven by the condition of the bottom hole pressure (Figure 8.b): in region 2, the bottom hole pressure has stagnated at the lower pressure limit, which again, is much lower than the lower operational condition. We will capture this binary behaviour using two-class classification emulation techniques described above.

Figure 8—Conceptual description of the two distinct regions in the simulation outcome; one uncertain attribute related to porosity is plotted against (a) cumulative liquid production in a time window, and (b) well bottom hole pressure. Region 2 presents an abnormal behaviour because the pressure is stagnated on the lower pressure limit. Note that the lower pressure limit is much lower than the lower operational condition.

The mechanism of simulation control change is also relevant while considering multiple uncertain attributes. Before exploring an example, we introduce a type of graphic frequently used in our analysis.

- **Cross-plots** provide qualitative diagnostics of multi-variate emulators. Simulation outputs and emulation outputs are respectively plotted in the x and y-axis. The axes usually have the same scale. For continuous quantities of interest, a 45-degree line shows where simulation output is equal to emulation output. Cross-plots usually contain a scatter plot of one or two data sets (in the last case, training and test sets can be compared, for example). With these elements, we can evaluate patterns in the residuals between emulator and simulator output. Alternatively, vertical bars can demonstrate additional information related to the emulator, for example, emulator uncertainty given by the standard deviation of the emulator. In this case, we indicate in the caption of the figure what the bars represent.

Figure 9 illustrates cross-plots of emulators for two continuous quantities of interest. These emulators are constructed with a regression model of first order terms only and the training set of our first wave. We recognise a straightforward pattern in Figure 9.a: the cumulative liquid production of simulations is limited to the simulation target, but some scenarios produce much less than it. In Figure 9.b, the lower limit pressure is reached by these same scenarios with abnormal behaviour related to liquid production. Two regions are distinguishable in each of the figures and comparable with the regions described in Figure 8.

We recall Equation 9 which defines binary quantities from continuous outputs of the reservoir. When sufficiently large pressure limits are set in the simulator, we apply $tol_i=0$ or a very small value (for example,
considering possible effects in rounding or significant figures in the simulator output). When calibrating reservoir models for real fields, upscaling in the vicinity of the well may have a relevant role for the discrepancy in the injectivity and productivity of the well. This local discrepancy can justify the pressure reaching the lower pressure limit for a short period when opening the well. In this case, we suggest to estimate \( tol \), in order to encompass this short period that targets are not met.

![Figure 9](image)

Figure 9—Cross-plot of emulators constructed for two continuous quantities of interest based on a training set of Wave 1, Phase 1. This a multi-variate example where: (a) the cumulative liquid production reaches the simulation target for the scenarios vertically aligned but fails for the others, which is deemed abnormal behaviour; (b) the bottom hole pressure stagnates at the lower pressure limit for the scenarios with abnormal behaviour.

Figure 10 illustrates the classification of simulated continuous quantities into binary quantities for \( tol_i=0 \). On the one hand, on the top of Figure 10, all the scenarios classified as non-implausible scenarios produce the simulation target during the whole history, and \( B_i(x)=0 \). For these scenarios, the corresponding pressure is not adjusted (top of Figure 10.b), but at least kept higher than the lower pressure limit at all times. On the other hand, all the scenarios classified as implausible do not produce the simulation target during the whole history period because they reached the lower pressure limit (bottom of Figure 10.b), \( B_i(x)=1 \). In fact, because the lower pressure limit is much lower than the lower operation pressure limit, we also want to rule out this part of the input space which gives rise to scenarios with abnormal behaviour.

![Figure 10](image)

Figure 10—Simulated continuous quantities of interest classified in binary quantities; scenarios reaching the simulation target are currently classified as non-implausible \( B_i(x)=0 \), note that the variability in the pressures is high; emulators for continuous quantities can later evaluate the pressure in the region with smooth behaviour. Scenarios which do not reach the simulation target are classified as implausible \( B_i(x)=1 \) (red), it is due to pressures at the level of the lower pressure limit.
**Pattern 2: Breakthrough Time.** Water breakthrough is the first time when water reaches the production well. In field management, this measured time and subsequent water-oil ratio trends are usually key performance indicators. They can also be indicative of channelling and bypassing problems in the field (Baker 1998).

Besides being a critical aspect for reservoir management, Breakthrough Time (BT) leads to a discontinuity in simulation outputs which are related to water production. In a three-phase system (oil, water and gas), it defines two distinct regions:

- Region 1: BT did not occur; the cumulative oil production is equal to cumulative liquid production;
- Region 2: BT occurred; cumulative oil and water production add up to cumulative liquid.

Figure 11.a is a conceptual description of two distinct regions in cumulative water production given a unique uncertain attribute (related to porosity). To create one statistical model to represent each one of the trends would be more efficient than to construct a single model representative of both regions. When we have the historical data indicating whether the BT occurred in the real well, we can define one of the regions to consider. Figure 11.b provides an example of a cross-plot for the well PROD024 highlighting the physical boundary \( W_p = 0 \) for scenarios in which no BT occured.

![Figure 11—a Conceptual description of the two distinct regions in the simulation outcome; one uncertain attribute related to porosity is plotted against cumulative water production in a time window; the question raised is answered by the historical data of the well; (b) Cross-plot simulator output versus emulator output for cumulative water production of PROD024A before the application of a two-class emulator; a physical boundary draws a pattern for scenarios with no breakthrough.](image)

Our objective is to classify the two distinct regions highlighted in Figure 11.a. Therefore, instead of selecting the BT itself (i.e. a date) as a quantity of interest, we identify scenarios with similar behaviour as the historical data (did the breakthrough occur or not in historical data?). The proposed approach is suitable because of the iterative nature of BHM. Firstly, we identify Regions 1 and 2 ("No BT" or "BT occurred"). Secondly, we construct emulators able to identify these distinct regions and rule out the one which is dissimilar from the historical data. Finally, we remain with a region having smooth behaviour related to cumulative water production, which a far simpler emulator can model (when compared to the emulator required to represent both Regions 1 and 2).

We present the conceptual description of the labelling process in Figure 12 and illustrate it with examples from Phase 3 of our application. The labelling process is coherent with Equation 9 presented in the Statistical Methodology section. The legend of Figure 12 shows the elements considered in the classification process: timeline indicates a time scale; observed BT is the historical data \( z_i \) in the Statistical Formulation; BT tolerance considers our evaluation about the observational error \( e_i \) and model discrepancy \( \epsilon_i \), describing our
uncertainty about the observed BT; \( tol_i \) is the tolerance applied in Equation 9, derived from BT tolerance; \( \text{simulated time} \) is the last time that scenarios were evaluated with the simulator.

The BT of the simulated scenarios - \( f_i(x) \) in the Statistical Formulation - is represented in four distinct groups, a combination between the labels: green for non-implausible, red for implausible scenario; circle for scenarios which BT is identified during the simulated period; question mark for scenarios which BT is later than the simulated period (i.e. we do not know the exact BT of these scenarios, we only know that it is later than the simulated time).

Within this framework, we highlight three distinct cases to systematically classify simulation scenarios as implausible or non-implausible (see Figure 12):

a. **Scenarios with late BT**: we classify as implausible the scenarios for which the water breakthrough did not occur during the simulated period, while the observed breakthrough occurred before the simulated time minus the breakthrough tolerance. These implausible scenarios are in Region 1 "No BT" in Figure 11. They lead to a discontinuity in the evaluation of cumulative water production. In this case, the value of the tolerance \( tol_i \) in Equation 9 is between zero and the simulated time and can be written as:

\[
    tol_i(x) = \begin{cases} 
    z_i, & \text{if } z_i - f_i(x) \geq 0 \\
    \text{simulated time} - z_i, & \text{otherwise}
    \end{cases}
\]  

b. **Uncertain situation**: we classify as non-implausible all the simulated scenarios when the observed water breakthrough of a well occurred within the tolerance around simulated time (purple arrows in Figure 12), that is, it is uncertain which of the two regions in Figure 11 we should care. In this case, the value of the tolerance \( tol_i \) in Equation 7 can take any non-negative value, and \( tol_i \rightarrow \infty \);

c. **Scenarios with early BT**: we classify as implausible the scenarios for which the water breakthrough occurred during the simulated time, while the observed breakthrough occurred later than the simulated time plus the breakthrough tolerance. These implausible scenarios are in the region labelled as "BT occurred" in Figure 11. They lead to a spurious evaluation of cumulative water production since no water has been produced until the simulated time. In this case, the value of the tolerance \( tol_i \) in Equation 9 is greater than the simulated time and can be written as:

\[
    tol_i = z_i - \text{simulated time}
\]  

Figure 12—Conceptual description and examples for the classification of water Breakthrough Time (BT) into a binary quantity of interest. We aim to define scenarios in the two distinct regions 'NO BT' or 'BT occurred' - (a) Scenarios with late BT are represented by the red question marks. For these scenarios, we do not know when BT occurs, we only know that is sufficiently later than the observed BT; therefore they are labelled as implausible; (b) When the observed BT is near the latest simulated time, we have an uncertain situation about the region of interest, and all scenarios are non-implausible; (c) Implausible are the scenarios with BT occurring earlier than the simulated time, when the latter insufficiently earlier than the observed BT.
Once we set our tolerance $tol_i$ for each case above, we define conditional tolerances. The definition of the water BT for observed data and simulated scenarios is discussed by Formentin et al. (2019), Almeida et al. (2018) and Lawal et al. (2007).

**Data preparation**

In *STEP 5*, the objective is to sample a set of scenarios which is representative of the search space. The size of the training set ($m$) is a compromise between the affordability of simulations (*e.g.* a sample size as small as possible) and the accuracy of emulators (*e.g.* sufficiently large to enable the construction of informative emulators of sufficient accuracy in the current wave). In *STEP 5.a*, we design scenarios for the test set (sample size $n$). The test set is only used for the selection and validation of emulators. The workflow is adaptable to a situation where an independent test set is not affordable. Bastos (2010) describes several design possibilities, pointing differences between sampling strategies for training and test sets.

The specification of the design decision $m$ depends on the number of uncertain attributes considered, the expected number of active inputs $n_{Ai}$, the complexity of emulated outcomes, and other practical aspects. To balance this decision, we can consider $(n_{Ai} + 2)(n_{Ai} + 1)/2$, which is the minimum number of points necessary to construct a quadratic response surface (Busby 2007).

For our demonstration, we apply the space filling Latin Hypercube sampling technique for generating samples from a multidimensional distribution. We defined $m=n=100$ scenarios because we can run 100 simulations in parallel in the cluster available. It is a reasonable size: the largest number of expected active variables in our application is 11, referred to in ‘*STEP 3-Identify inputs and outputs to emulate*’.

*STEP 6* and *6.a* include the preparation of simulation files, simulation of scenarios and extraction of readable files in table format from the simulator.

In *STEP 7*, model discrepancy estimates the uncertainty about the simulation model in representing the real field (Equation 3). Goldstein et al. (2013) distinguish two types of model discrepancy:

i. Internal discrepancy: This relates to any aspect of structural discrepancy whose magnitude we may assess by experiments on the computer simulator. Internal discrepancy analysis gives a lower bound on the structural uncertainty that we must introduce into our model analyses;

ii. External discrepancy: This relates to inherent limitations of the modelling process embodied in the simulator. There are no experiments on the simulator which may reveal this magnitude. It is determined by a combination of expert judgements and statistical estimation.

In reservoir engineering, the model discrepancy arises, for example, from lack of sufficient data or techniques to explore it, inaccuracy in reservoir size dimensions, in net-to-gross calculations, in reservoir architecture, spatial properties, upscaling models, fluid properties. After Ringrose and Bentley (2015), the aim in defining model uncertainties is to place our models within a framework that can overcome data limitations and personal bias and give us a useful way of quantifying forecast uncertainty. The need for considering model discrepancy and model error while performing history matching is studied in recent works related to petroleum reservoirs (Evensen 2018a; Evensen 2018b; Evensen and Eikrem 2018).

In our case study, we introduce a recurrent component of the model discrepancy due to the inherent errors in the liquid and injection rates used as simulation targets. These uncertainties arise in real case studies but are often neglected. Here we noised up the hypothetical reality outputs to mimic the real world situation. Figure 13 represents this effect: *error-free* (brown line) is the direct outcome from the simulation of the hypothetical reality; *reference* (black dots) is the error-free outcome with independent random and systematic noise added (with characteristics described in Table 4); *target=ref* (blue line) is the outcome of the simulation having the reference data as target.
In Figure 13.a for liquid production rate versus time, the reference and target=ref. are coincident, as the earlier is the simulation target. In Figure 13.b – bottom hole pressure versus time, we verify differences between these three simulations. Target data is reached in Figure 13.a in detriment of an additional error on pressure; this additional error needs to be included in the calibration process.

Figure 13.b is a demonstration of internal discrepancy caused by the systematic error in the liquid production rate of HR-82. In our application, we considered the discrepancy in a straightforward way, but we emphasize the need to account for the model discrepancy in detail in reservoir history matching. Further discussions are available in Goldstein et al. (2013) and we leave a more advanced treatment to future work.

**Construct and validate emulators**

Selection of outputs to emulate (STEP 8) is critical for the effective use of information from a high dimensional output space and, ultimately, to the efficiency of the Bayesian History Matching (BHM) process. We make a deliberate decision considering (a) the selection of all available quantities of interest, which would incorporate all the information available in the process; (b) the total time invested for emulation (construct, validate and evaluate new scenarios with the valid emulators), the most expensive decision; and (c) dependence between quantities of interest. By selecting a limited number of outputs to emulate, we aim to identify the largest implausible region to rule out of our analysis while keeping the number of emulators low.

Because it involves some further subtleties, our workflow for STEP 8 is fully described in Appendix B. Firstly, we select the specific types of outputs to emulate. Secondly, we use the adapted implausibility measures $\hat{I}_i$ and $\hat{D}_{info}$ (Equation 13) to estimate which combination of outputs has the highest potential to become the most informative. The strategy of selection enables the emulation of outputs from earlier phases of evaluation, as exemplified in Figure 5. Each output selected in STEP 8 is expected to respond more strongly to a sub-set of uncertain attributes – the active variables, and ideally, in a smooth way.

STEP 9 is a checkpoint whether outputs are selected or not. It is mainly relevant when, in STEP 8, we try to select quantities of a phase $\phi$ which is earlier than the last phase simulated (phase), see Appendix B for further details of $\phi$ and phase.

For STEP 10, classical model fitting criteria such as AIC and BIC (using the "step" function in R) can be applied to select active variables for each selected output (Vernon et al. 2018, Vernon et al. 2010). The use of linear regression comes from our expectation that relevant relationships between attributes and outputs can be identified by linear operators as, if an input is at all active for a particular output, it will most likely induce some linear effect on that output, even if its actual functional dependence is far more complex and non-linear (Vernon et al., 2010)

With these active variables, STEP 11 constructs several competitive emulators for each selected quantity. This step provides flexibility to our procedure. The time invested in constructing each competitive emulator
can be estimated beforehand, generally depending on the statistical model and the number of scenarios in the training set.

Figure 14 details our workflow where: **STEP 11.1** defines the possible statistical models to be considered depending on the type of output (see Table 5). We can integrate multiple and diverse emulation techniques based on the requirements of the study. In our application, for two-class models \( n = 3 \) encompassing (1) linear discriminant analysis, (2) logistic regression with linear terms and (3) with linear and quadratic terms. We iterate in **STEPS 11.2 to 11.5** until all the competitive emulators are constructed.

In **STEP 12**, after constructing competitive emulators, we select the best emulator and validate them based on positive and negative predictive values (for two-class quantities) and on credible interval diagnostics and information index (for continuous quantities). These indicators are calculated for training and test sets independently: while indicators for test sets are effectively used to select and validate emulators, we also monitor training sets.

We now detail the choices made for binary quantities of interest in our application. Logistic regression models have a fixed boundary decision on 0.95; for linear discriminant models, we define the decision boundary from the training set (e.g. a larger value than the cutpoint defined in Figure 3).

As validation criteria, binary quantities of interest need a minimum Positive Predictive Value of 0.90, e.g. at least 90% of the emulator’s implausible scenarios correspond to the simulator’s implausible scenarios. This unbalanced PPV threshold \( PPV_{\text{thres}} \gg 0.50 \) represent our aversion to rule out wrongly regions of the search space. Simultaneously, we consider a smaller threshold for the Negative Predictive Value \( NPV_{\text{thres}} = 0.50 \), since bad scenarios kept in the analysis can be ruled out at later waves. When two concurrent emulators meet the validation criteria, the one with higher NPV is selected.

We show an example of selection of concurrent emulators for a binary quantity. Emulator 1 model is a logistic regression (Figures 15.a and b); Emulator 2, a linear discriminant analysis (Figures.c and d). Training and test sets are in reddish and blueish, respectively. Figures 15.a and c are traditional diagnostics plots. They compare binary simulation outputs with the emulator probability \( P(B_i=1|x) \). The grey vertical lines are the boundary decision applied. Figures 15.b and d are alternative plots, comparing the continuous output for the simulator (i.e. cumulative water production in 2,710 days, Phase 3) with the probability estimated by the emulator. The vertical line places the historical data. It indicates that we should rule out the region ‘No BT’ and care about the region ‘BT occurred’. Scenarios in the grey regions of these plots are ruled out by the emulator; these scenarios are coloured in orange and light blue to enable the distinction.
Both scenarios have PPV equal to 1, indicating that neither of them wrongly exclude scenarios based on the emulator evaluation. The drawback of these emulators is that they keep several scenarios in the ‘No BT’ region (e.g. NPV < 1). In this situation, we choose logistic regression, the more informative emulator, i.e. the one with higher NPV.

For continuous quantities of interest, we define a threshold for the information index $D_{info,thres} = 0.05$, e.g. at least 5% of remaining search space needs to be ruled out by a valid emulator. An emulator is valid when $D_{info,i} \geq 0.05$ and the Credible Interval Diagnosis with $\alpha = 95\%$ is larger than 0.85, i.e. the 95% credible interval defined by the emulator covers at least 85% of the simulation outcomes. These criteria combined validate an emulator as informative and with accurate uncertainty estimation. If more than one concurrent emulators are valid, the emulator with larger $D_{CI}$ is selected, justified by our caution about wrongly ruling out non-implausible scenarios.

We present two examples of cross-plot diagnostics of concurrent emulators for continuous quantities of interest (Figure 16 and 17). The plots are for the training (a and c) and test (b and d) sets. The concurrent emulators are regression models with (a) terms of first order only and (b) terms of first and second order. Note that we omit labels for y-axis of the cross plots for the test sets: they are the same as for the training sets.

The first example (Figure 16) highlights the potential of the information index as a selection indicator. The vertical bars account for $\pm \omega [Var(f^*_i(x)) + Var(e_i) + Var(e_i)]^{1/2}$, which is the implausibility cut-off $\omega$ multiplied by the denominator of the implausibility measure. The cross-plots of the first emulator (Figures 16.a and.b) have emulator expectation with a curved trend, which indicates, for example, that it missed some quadratic relationship. Both emulators have $D_{CI}$ of training and test sets higher than 90%, indicating that the level of uncertainty is appropriately estimated. Nevertheless, the information index $D_{info}$ of training and test sets are lower than 5% expressing the large uncertainty related to this emulator.
The cross-plots of the second emulator (Figures 16.c and d) show emulator expected outputs very close to the 45-degree line, indicating that the second order terms more accurately captured relationships between active variables and the average pressure of the well RJS19 with a coherent uncertainty estimation. The information index is above 30%, therefore we select the regression model containing terms of first and second order.

With Figure 17, we compare (a) an emulator with appropriate estimation of uncertainty with (b) an over-confident emulator. For each plot, the vertical bars represent the 95% credible interval estimated around the expected outcome \(E(f^*_i)\). Both emulators of Figure 17 are highly informative (\(D_{\text{info},1}=0.46\) and \(D_{\text{info},2}=0.67\)). Nevertheless, a \(D_{\text{CI}}\) of the test set close to our threshold 0.85 indicates that the second emulator tends to estimate a small uncertainty, in this case on the limit of over-confidence. The \(D_{\text{CI}}\) of the training set is very high (99%), and this difference with the \(D_{\text{CI}}\) of the test set also reveals this over-confidence characteristic. Considering the principles of BHM, we select the emulator with terms of first order only, which has higher \(D_{\text{CI}}\) and is sufficiently informative.

The last step to construct and validate emulators, \textit{STEP 13}, ensures that only valid emulators are applied for the implausibility analysis.
Evaluation of scenarios and input space reduction
In *STEP 14*, all valid emulators are used to evaluate the implausibility for the training and test sets, checking how many non-implausible scenarios are available for each of them.

*STEP 15* uses the number of non-implausible scenarios in the training set to identify the need for a new wave, which adds new non-implausible scenarios to the analysis. The value $\lambda_m$ – a minimum number of scenarios required to construct emulators - is a design option (see discussion related to *STEP 6*). For our case study, at least 50 scenarios in the training set are required to construct new emulators. If we have sufficient non-implausible scenarios in the training set (*STEP 15*), we have the possibility to construct new emulators.

In *STEP 16*, we follow with an implausibility analysis to evaluate the proportion of remaining space classified as implausible for all the emulators constructed up to this wave. As emulators are fast to evaluate, we can afford to evaluate a large number of new scenarios (#$10^3$).

We described *STEPS 17* and *18* with Figure 4. The criteria to change phase (*STEP 17*) is set as $D_{info,M} < 70\%$ and the criteria to end calibration (*STEP 18*) is the last phase to be evaluated (jointly with the previous condition $D_{info,M} < 70\%$).

*STEP 19* (design of new non-implausible scenarios) is performed in the situation where the workflow identifies the need for a new wave. These new scenarios are non-implausible for all the emulators constructed until the current stage of the analysis.

Apply non-implausible scenarios
When *STEP 18* indicates that we completed the uncertainty reduction, we proceed with the application of non-implausible scenarios. One possible application is for reservoir forecasting, which involves several additional concepts as described in Busby et al. (2007), Craig et al. (2011), Goldstein and Rougier (2006) and Craig et al. (1996).

Results and Discussions
Here we discuss important features resulting from the application of our systematic procedure to the case study HR-82. There are four figures of the process that we would like to highlight:

- we used 15 waves for 5 phases of evaluation;
- we evaluated 3,000 scenarios with the simulator with different stopping times; the simulation time would be equivalent to 1,637 full simulations without phasing;
- we obtained valid emulators for 115 quantities of interest out of a total of 198 selected quantities;
- we reduced the original space to 3.58e-11\% (non-implausible space) by the end of the calibration process.

This indicates that this procedure may be more efficient and far less expensive than traditional methods. We note that traditional methods often do not attempt to identify all input points consistent with historical data, which are essential for an accurate, unbiased prediction. We now look at a breakdown for each phase of evaluation in Table 7 and the performance of this process for different perspectives. It is worth noting that the cost-benefit between simulation time and space reduction in earlier phases is quite low.
Phases of evaluation enable the efficient use of resources for reservoir simulation. We consider that the median of the time to simulate one scenario is representative of the process. The time to simulate scenarios in each phase is plotted in boxplots (Figure 18.a). From the total number of scenarios simulated, 20% were until the end of Phase 1 where the simulation time of each scenario corresponds to 12% of the median simulation time for scenarios until the end of Phase 5. The short simulations of Phase 1 allowed emulators to be constructed that ruled out 97.86% of the search space as implausible, i.e. very unlikely to match with the historical data.

The time spent to evaluate a new scenario with emulation is a fraction of the time spent to evaluate a new scenario through simulation. Figure 18.b presents the number of emulators constructed in each phase, highlighting that each phase considers all the emulators constructed since the beginning of the process (it is cumulative). Figure 18.c presents the time spent to evaluate scenarios with emulators: for Phase 5, in less than 25 seconds, we evaluate 100,000 scenarios with emulators, while it took 565 seconds to evaluate only one scenario through simulation, giving a speed increase of six orders of magnitude (note: we can run 100 scenarios in parallel in the cluster).

The potential of combining simulation and emulation techniques is evidenced: using simulation we capture physical phenomena in place in the reservoir, with emulators, we speed up the parts of the history matching process requiring a sufficiently large number of evaluations to be consistent from a statistical perspective.

For each wave, Figure 19 presents (a) information index $D_{info}$ computed from a large number of scenarios during the implausibility analysis (STEP 16) and represents the proportion of remaining space judged as implausible, and (b) the proportion of initial space considered non-implausible, remaining in the analysis.

Table 7—Breakdown of figures from our application based on phases of evaluation. For each phase of evaluation, we present the period simulated (a window between zero and the time given in seconds), the total number of scenarios simulated, the median time of the time invested in simulation, the number of emulators constructed in the phase, the mean time to evaluate 100,000 scenarios with emulators and the remaining proportion of the original space given as non-implausible.

<table>
<thead>
<tr>
<th>Phase</th>
<th>Period simulated</th>
<th>Scenarios simulated</th>
<th>Median time simulation</th>
<th>Number of emulators</th>
<th>Aver. time emulators (100,000 scenarios)</th>
<th>Non-implausible space</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>518 s (3 waves)</td>
<td>600</td>
<td>66 s (12% of the max)</td>
<td>8 (7%)</td>
<td>1.41 s</td>
<td>2.14%</td>
</tr>
<tr>
<td>2</td>
<td>1,461 s</td>
<td>200 (1 wave)</td>
<td>237 s (42% of the max)</td>
<td>3 (3%)</td>
<td>1.71 s</td>
<td>1.48%</td>
</tr>
<tr>
<td>3</td>
<td>2,710 s</td>
<td>1,600 (3 waves)</td>
<td>339 s (60% of the max)</td>
<td>73 (63%)</td>
<td>18.26 s</td>
<td>3.15e-9</td>
</tr>
<tr>
<td>4</td>
<td>3,256 s</td>
<td>400 (2 waves)</td>
<td>456 s (81% of the max)</td>
<td>21 (18%)</td>
<td>21.83 s</td>
<td>8.68e-11</td>
</tr>
<tr>
<td>5</td>
<td>4,018 s</td>
<td>200 (1 wave)</td>
<td>565 s (max)</td>
<td>10 (9%)</td>
<td>24.81 s</td>
<td>3.58e-11</td>
</tr>
</tbody>
</table>
(note the log scale). In Figure 19.a, we highlight the cut-off of 70% defined as criteria to change phase in the *STEP 17*: each wave that was unable to achieve this threshold led to an increment in the phase of evaluation for the following wave. This happened in our application in Waves 3, 4, 12 and 14. Wave 15 evaluated Phase 5 and failed to reach 70%, leading to the end of the procedure by this wave.

Wave 4, evaluating Phase 2, had the lowest proportion of space cut-out among all the waves. Indeed, the same four wells of Phase 1 are active in Phase 2, leading this particular phase to a modest role in the uncertainty reduction. Phase 3 kept being highly informative for seven consecutive Waves (5 to 11). That level of uncertainty reduction is due to the 21 additional active wells in both east and west block of the reservoir. By the end of Wave 12, the proportion of the original space considered non-implausible is $3.15 \times 10^{-9}$.

Phase 4 was highly informative only for Wave 13, which is coherent with the characteristics of this phase (i.e. the water breakthrough occurs for four additional wells). The fifth and last phase of the process did not offer new events to the reservoir historical data, leading to an unremarkable performance in terms of uncertainty reduction.

By the end of the calibration process, the hypothetical reality is also judged as non-implausible: a useful consistency check. The final proportion of input space remaining as non-implausible was $3.58 \times 10^{-11}$. We see that this substantial uncertainty reduction is a direct result of the interplay between the simulator behaviour and the specification of all the required uncertainties, although we note that in high dimensional space such large reductions are expected. Note also that our approach should not suffer the problems of "ensemble collapse", as can be seen by the widely dispersed input points by the end of Wave 12, Phase 3, in Figure 20 below.

**Uncertainty Reduction in search space and quantities of interest**

The iterative process implemented narrowed the search space sequentially. Each new wave re-established the search space on non-implausible regions, from where a limited number of scenarios is sampled and used to construct additional emulators. Figure 20 shows pairs plots for 18 uncertain attributes by the end of Wave 3 (left plots) and Wave 12 (right plots). The description of each attribute is presented in Table 3, but we highlight that they are more relevant to regions 6, 10 and 8 respectively.
Figure 20—Non-implausible space for 18 uncertain attributes mostly relevant for regions 6 (top), 10 (middle) and 8 (bottom). The left side plots are by the end of Wave 3 (before start of Phase 2); the right side plots by the end of Wave 12 (before start of Phase 4); diagonals with the marginal density, upper panels the minimised implausibility and lower panels show the intensity of non-implausible scenarios. All plots indicate the hypothetical reality which correctly remains non-implausible.

For each pairs plot of Figure 20, the diagonal provides the marginal density of each attribute in the non-implausible space, providing a sense about the dispersion of the uncertain attributes compared to the initial distribution (uniform along the limits plotted). We highlight with dotted vertical lines the hypothetical
reality. For some attributes, the hypothetical reality value is on (or very close to) the boundary of the initial range and superposed by the box of the diagonal plots (e.g. mKz(R6) in the plot for region 6).

The upper panel (above diagonal) plots the minimised implausibility using an appropriate scale and show non-implausible regions in green (light and dark green represent implausibility smaller or equal to 3). In Equation 10, we defined a high factor $\omega^*$, allowing the identification of the regions in purple that were ruled out by two-class emulators constructed in each phase. In reddish are the regions judge as implausible by emulators for continuous outputs. Grey (in the legend named NA) indicates regions ruled out up to Wave 3. Therefore, these regions do not make part of the search space of later phases. For both upper and lower panels, we plot the hypothetical reality as a black triangle.

The lower panel plots (below diagonal) give the intensity (or optical depth) of non-implausible points for each pair of attributes (Vernon et al. 2010). Dark blue indicates that a high concentration of non-implausible scenarios is evaluated in the region. This plot provides a sense of the depth of the non-implausible space.

Firstly, we analyse the plots from the end of Wave 3, which is the last wave evaluated until Phase 1. We highlight that in this phase (up to 528 days), only four vertical wells are active. We observe that most reduction of search space is on the attributes related to the porosity of Region 6 (mpOR, A and B, left upper panel) and on global attributes (mainly $c_{krw}$ related to water relative permeability, but also slightly rock compressibility, left middle panel). Intensity plots (lower panel) also reveal that the emulators captured the influence of other uncertain attributes (well index WI of RJS19, mpor and B of region 10).

Until Phase 1, two-class emulators are only applied for cumulative liquid production. The purple colour highlights that the regions ruled out by two-class emulators are far from the hypothetical reality (which is coherent with the arguments provided in the section ‘Pattern 1: Simulator targets’). The uncertain parameters of the east block (left lower panel) do not present any indication of uncertainty reduction, which is expected since no well of this region is active until 528 days.

Secondly, we focus on the plots for the end of Wave 12 (right), which is the last wave only evaluating until Phase 3 (2,710 days). Figure 19 showed that most of the uncertainty reduction occurs in the evaluation of Phase 3. Figure 20 highlights the regions of the search space remaining in the analysis. The diagonal reveals that the ranges of several attributes concentrate around the hypothetical reality.

Nevertheless, the concentration of points is not always symmetric around the hypothetical reality, nor should we expect them to be, especially when it falls in the corners of the non-implausible regions (e.g. A and B of region 6, in the upper plots) or in the limit of the ranges defined in STEP 1 of the systematic procedure. The noise added to the target data, and its corresponding model discrepancy will also contribute to shifting the cloud of non-implausible points around the true hypothetical reality.

The upper panels highlight several regions of the search space ruled out by the two-class emulators, especially noticeable for mpor(R10). Finally, the lower panels emphasise that the remaining non-implausible space is a fraction of the original space.

**Successive waves and phases of evaluation**

We dedicate this section to discuss the iterative principle of BHM applied in our systematic procedure. To iterate in waves allows the discarding of regions of the input parameter space sequentially and the application of simple emulators to mimic a complex simulator and its input space in a reliable way.

Figure 21 compares two emulators constructed for cumulative water production of the well RJS19 in Phase 3 (2,710 days). On the one hand, Figure 21.a plots a design of scenarios representing the original search space with five uncertain attributes in light blue. Note that this is only an illustrative example, supposing the case that we did not use phases of evaluation; this training set is not used in the calibration process (i.e. in Wave 1 we do not evaluate any scenario until Phase 3). This design of the original search space is used to construct an emulator, the cross-plot of which is presented in Figure 21.a in light blue. We observe a non-linear pattern, i.e. the simulated and emulated outcomes do not follow a straight line.
Moreover, the discontinuous behaviour due to a late water breakthrough time concentrates several points following the zero water production from the simulator output. Finally, the emulator uncertainty plotted in error bars $\pm 3\sigma(f_i(x))$ is very large. These diagnostics suggest that a more complex statistical model is required to accurately represent this level of complexity in the relationships between input and outputs existent in Figures 21.a and .b which would require many more simulator runs to train.

On the other hand, Figure 21.d plots a design of scenarios representing the search space in Wave 3 with five uncertain attributes in purple. Some regions of the space were already ruled out in previous waves, which is evident for mpor, B and $c_{krw}$. This design is used to construct an emulator, with the cross-plot presented in Figure 21.c in dark purple.

We compare this emulator with the first one through a zooming box relating Figures 21.b and c. We observe a linear pattern, i.e. the simulated and emulated outcomes follow a straight line, with a higher $R^2_{adj}$ which by definition indicates the proportion of the variance of the output explained by the regression (Table 2). The absence of simulated scenarios with no water breakthrough time ($W_p=0$) is coherent with a limited range of the uncertain attribute related to water relative permeability $c_{krw}$. Finally, the emulator uncertainty plotted in bars are far narrower.

These diagnostics indicates that this simple statistical model is able to properly represent the relationships between input and outputs over the current non-implausible space, demonstrating the strength of the iterative approach: smaller volumes of input space in later waves are substantially easier to emulate as we have removed the erratically behaved and physically irrelevant regions from the search space.

Reducing the search space through the waves corresponds to reduce the uncertainty about the behaviour of the reservoir. In Figure 22, we present indicators for six emulators constructed for the average pressure of PROD024A ($\bar{p}_{pwh}$) in Phase 3 over Waves 6, and 8 to 12. Figure 22.a highlights the sequential reduction of the standard deviation (sd) of the quantity of interest, reduction of emulator uncertainty (residual standard deviation) and information index.
The corresponding pressure in time (Figure 22.b) corroborates the idea that we are narrowing down to the regions of interest in our input space and calibrating the model. Wave 6 is the first wave that this quantity of interest was selected as output to emulate (STEP 8). The high standard deviation of the average pressure (40 kgf/cm²) is aligned with the spread of the corresponding pressure curves, in a range between the lower pressure limit (definition in Figure 7) and around 350 kgf/cm². The emulator uncertainty is sufficient to allow a reduction of 81% of the remaining search space (e.g. a very informative emulator).

In Wave 7, this quantity of interest is not chosen as output to emulate in STEP 8 (Figure 4). The emulator of Wave 8 is constructed over a narrower search space, and the corresponding standard deviation of the average pressure is smaller, highlighting that no scenarios reached the lower limit boundary. The emulator uncertainty is smaller and the information index keeps on a high level (62%).

The following emulators keep the trend. Comparing the figures for the emulator of Waves 11 and 12, we observe that the standard deviations of the average pressure are very similar (10 and 9), but the fact the emulator uncertainty is 6 times smaller in the emulator of Wave 12 leads to a higher information index for it (20% of the remaining space is ruled out).

**Analysis of uncertainty reduction for production and injection wells**

To show the potential of the systematic procedure proposed, we present the adapted implausibility measure $\tilde{I}_t$ (Equation 7) in boxplots. This measure compares the simulation output against the historical data, a simplification of the traditional implausibility measure.

Figure 23 shows these results for production wells in Phase 5 (i.e. 4018 days). Plots from the original search space (Figure 23.a) are presented with their corresponding boxplots from the end of the calibration process (Figure 23.b). Note: we simulated scenarios from the original searching space until Phase 5 for comparison purposes only, i.e. this scenarios were not applied in the analysis. The grey regions highlight the interval which scenarios are considered non-implausible, but also provide a sense of scale since the y-axis from the right plots are much smaller than the ones from the right side (i.e. the outputs are much closer than the historical data after the calibration process). The x-axis presents the names of the wells, which are the same for all production wells (labels provided on the last plot for production wells). We highlight specific wells in the box plots and present the corresponding well data.
The first pair of plots in Figure 23 are for cumulative oil production ($N_p$). Initially, these quantities of interest are distant from the historical data (e.g. several wells have the median above the threshold of three). Several outliers (mainly for wells NA1A and RJS19) indicates scenarios poorly representing the reservoir behaviour. We present the well data oil production rate of NA1A. After the calibration process, all quantities of interest are within the grey region, have a very low median and the interquartile range describes much more calibrated scenarios.
The second pair of plots in Figure 23 are for cumulative water production ($W_p$). By the end of the calibration process, the medians are closer to the threshold of three. Nevertheless, the fact the median for PROD024A is above the grey region indicates that this well could be subject of further evaluations, possibly with new waves. For that, the criteria established in STEP 17 and/or 18 could be replaced by the adapted implausibility measure of all quantities of interest evaluated, for example. We plot the well data water production rate of PROD024A. In our case study, higher medians $W_p$ than $N_p$ are related to the fact that both components of observational error are proportional to the production rate (Table 4). In this way, we have a smaller variance of the observational error for water production, which is inversely related to the adapted implausibility measure.

The third pair of plots is for cumulative liquid production. Five wells (NA1A, PROD012, PROD014, PROD024A and RJS19) initially do not meet the liquid target informed to the simulator. By the end of the procedure, all wells meet the target informed to the simulator (e.g. no scenario reaches the lower limit pressure in the regions of the wells). We plot the liquid production rate of the well RJS19.

Average pressure of production wells for the window between Phases 4 and 5 shows a definite improvement in the representation of the behaviour of the wells by the end of the historical period. We present the bottom hole pressure for NA1A.

For the injection wells (Figure 24), initially six wells (INJ003, INJ005, INJ010, INJ15, INJ22 and INJ23) could not inject as much as the simulation target. This lower volume injected is associated with scenarios reaching the upper pressure limit, demonstrated in Figure 7. We illustrate liquid production rate and bottom hole pressure of the well INJ015 as examples. All the scenarios evaluated by the end of the calibration process meet the injection rate target. The average pressure of injection wells for the window between Phases 4 and 5 indicates that all scenarios have adapted implausibility measure below the threshold of 3.

Finally, we present data for the well PROD021. Figure 25.a shows that liquid production rate (Figure 25.a) is the same as the simulator target during all the production time since the start of the calibration.
process. We observe a bias in the bottom hole pressure (Figure 25.b) of calibrated scenarios: all of the green lines are below the reference data. We emphasise that this specific well presented a clear internal discrepancy due to the noisy simulator target, as shown in Figure 25.c.

Conclusions

The ultimate goal of a calibration process is to provide a background for well informed and efficient decisions. Finding the whole class of scenarios capable of representing the reservoir historical behaviour is essential in order to give a realistic evaluation of reservoir performance and consistent, unbiased predictions incorporating realistic levels of uncertainty, required for full asset management.

We presented a procedure for systematic uncertainty reduction for petroleum reservoirs combining reservoir simulation and Bayesian emulation techniques. We explored Bayesian History Matching techniques in order to provide an alternative and more rigorous tool for reservoir studies dealing with probabilistic uncertainty reduction. Challenges addressed by our systematic procedure are related to the consideration of several sources of uncertainty involved in a calibration process, efficient use of simulation time and high dimensional input and output spaces.

The procedure for systematic uncertainty reduction was applied to calibrate a case study with 26 uncertain attributes and 14 production and 11 injection wells. We defined the case study and took one of the scenarios to be the hypothetical reality (i.e. the known answer). The aim was to test the potential of the procedure for a complex reservoir model under a controlled situation while illustrating the main steps of the procedure.

In total, we evaluated 4,018 days of historical date with five phases of evaluation; ran 15 waves sequentially; simulated 3,000 scenarios for training and test sets (the simulation time would be equivalent to 1,637 full simulations without phasing); and constructed valid emulators for 115 quantities of interest out of a total of 198 selected quantities. By the end of the calibration process and after removing the implausible regions, a small fraction in the order of $10^{-11}\%$ of the input space remained as non-implausible, demonstrating a substantial uncertainty reduction. The main novel contributions of this work to the Bayesian History Matching framework for reservoir analysis are as follows.

A systematic workflow to structure BHM techniques is presented. We deliver scalability to higher dimensions in input and outputs spaces by (a) using informed judgements of the asset team to identify inputs and outputs to emulate (STEP 3); (b) proposing a procedure for output selection which considers the combination of quantities of interest considered as the most informative (STEP 8); and (c) planning the selection of active variables (STEP 10), where the sufficient technique currently applied (stepwise selection) can be substituted by techniques more appropriate for higher dimensions, if required.
We secure flexibility for the integration of diverse emulation techniques. The workflow for the construction of concurrent emulators (STEP 11) for a given quantity of interest is capable of integrating diverse emulation techniques. In our application, we combined techniques for continuous and two-class quantities of interest.

We enable repeatability by developing a sequence of logically associated steps. The main workflow describes a high-level structure; for which individual steps can be designed in coherence with project requirements. It also allows the performing of several stages automatically, centring the users’ focus on activities related to data analysis and synthesis.

**Phases of evaluation** split the historical data and progressively add information to the process. Phases allow us to explore more straightforward relationships from the early period of well production, and to exploit these relationships using emulation to rule out large regions of input space, based only on early time information. This promotes an efficient use of resources for reservoir simulation. In our application, for example, Phase 1 (until 518 days) running time is estimated at 12% of the full historical period (4,018 days) running time. Three waves until Phase 1 rule out 97.86% of the input space (i.e. implausible regions), leaving only 2.14% of the space to evaluate in later phases, making subsequent emulation in later phases easier and more efficient.

**Two-class emulators** accurately model discontinuous behaviour identified for some quantities of interest typically found in reservoir simulations. We formulated a version of the implausibility measure for two-class emulators, integrating binary outcomes. In our application, we recognised data patterns from simulator targets and water breakthrough that required two-class emulators to construct appropriate statistical representations. This class of emulators allows one to rule out parts of the input space that lead to discontinuities, leaving regions with smoother relationships between inputs and outputs that can be subsequently analysed using standard emulators for continuous smooth quantities.

The **internal discrepancy** of the simulation model is demonstrated and related to the noisy targets added to simulate the historical period in our synthetic case study. The example provided suggests that the community should explore this important aspect of uncertainty in calibration studies, independently of the calibration technique applied. We aim in the future to report further, more detailed analysis on how to formally represent and evaluate model discrepancy, including this major source of internal discrepancy for petroleum reservoirs. We emphasise that the systematic workflow for BHM is designed to account for the model discrepancy in its statistical formulation.

The combination of each of these features in the systematic procedure is a novel approach to Bayesian History Matching. The application in the case study is a demonstration of the potential of the iterative nature of Bayesian History Matching combined with phases of evaluation and two-class emulators in addressing challenging problems of reservoir calibration, including the identification of the set of all inputs consistent with historical data, required for realistic predictions and asset management. The uncertainty reduction procedure was demonstrated as a powerful technique to search high dimensional space using substantially less computational time than using the complex simulation model alone.

**Acknowledgement**

This research was carried out in association with the ongoing project registered as ANP 19708-7, "Fomento à Formação de Recursos Humanos em Gestão de Incertezas e Tomada de Decisão: Um Programa BG Fellowship" (UNICAMP/Shell Brazil/ANP), co-funded by CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico – Brazil) through Science Without Borders Programme and Shell Brazil under the ANP R&D levy as "Compromisso de Investimentos com Pesquisa e Desenvolvimento".

The authors thank ENERGI SIMULATION, UNISIM, DE-FEM-UNICAMP, CEPETRO and Department of Mathematical Sciences (Durham University) for supporting this work and CMG, Emerson and Schlumberger for software licenses. The authors also thank our colleagues who contributed to this
paper: Alessandra Davolio and Forlan la Rosa Almeida (discussions about simulation targets), Carla Ferreira (discussions about visualization and noisy in historical data), Daniel Carvalho (IT support), Gilson Moura (definition of well trajectory applied in the case study), Jonathan Owen (discussions about implausibility), Ricardo Moreno (discussions about emulation).

**Nomenclature**

\[ B = \text{indicator function of a transformation of simulation output} \]
\[ B^* = \text{indicator function obtained from the emulator output} \]

BHM = Bayesian History Matching

BT = Breakthrough Time

CI = Credible Interval

\[ D_{\text{info}} = \text{information index} \]
\[ \bar{D}_{\text{info}} = \text{adapted information index} \]

\[ D_{CI} = \text{Credible Interval Diagnostics} \]

\[ \epsilon = \text{model discrepancy, the difference between the real reservoir and the reservoir model} \]

\[ E = \text{expectation operator} \]

\[ e = \text{vector of observational errors} \]

\[ f = \text{function of reservoir simulation model that computes a vector of quantities of interest} \]

\[ f^* = \text{emulator function} \]

\[ g = \text{known deterministic function} \]

\[ I = \text{implausibility measure} \]

\[ I_{M} = \text{maximal implausibility measure} \]

\[ \bar{I}_{\text{im}} = \text{adapted implausibility measure} \]

\[ i_{w} = \text{water injection rate} \]

\[ L_{p} = \text{cumulative liquid production} \]

\[ m = \text{number of reservoir scenarios simulated to increment the training set in each new wave} \]

\[ n = \text{number of reservoir scenarios simulated to increment the test set in each new wave} \]

\[ n_{A} = \text{number of active variables} \]

\[ n_{sce} = \text{total number of scenarios} \]

\[ N_{p} = \text{cumulative oil production} \]

NPV = Negative Predictive Value

phase = maximum phase of evaluation simulated in a given training set

\[ p_{bh} = \text{bottom hole pressure of injection wells} \]

\[ \bar{p}_{bh} = \text{average bottom hole pressure of injection wells} \]

\[ p_{pbh} = \text{bottom hole pressure of production wells} \]

\[ \bar{p}_{pbh} = \text{average bottom hole pressure of production wells} \]

PPV = Positive Predictive Value

\[ q_{l} = \text{liquid production rate} \]

\[ q_{o} = \text{oil production rate} \]

\[ q_{w} = \text{water production rate} \]

RMSE = Normalised Root Mean Square Error

tol = tolerance applied to compute \( B \)

\[ u = \text{Gaussian process} \]

\[ Var = \text{variance} \]

\[ W_{i} = \text{cumulative water injection} \]

\[ W_{p} = \text{cumulative water production} \]
\( w \) = nugget process
\( \omega \) = implausibility cut-off
\( x \) = vector of input parameter values representing a reservoir scenario
\( x^* \) = most appropriate vector of uncertain attributes
\( y \) = vector of quantities from the real physical reservoir
\( z \) = vector of measurable quantities from the real reservoir
\( \alpha \) = proportion covered by the credible interval
\( \beta \) = unknown scalar regression coefficients
\( \phi \) = a phase of evaluation to select outputs to emulate \( \phi[#1, \text{phase}] \)
\( \lambda \) = proportion of \( m \) providing a training set sufficiently large to construct emulators
\( \omega^* \) = factor to rescale the indicator function of two-class emulators

**Subscripts**
- \( A \) = active variables
- \( i \) = a measurable quantity of interest of the reservoir, \( i[#1, q] \)
- \( j \) = index corresponding to a regression term in the emulator equation

**Superscripts**
- \( k \) = scenario of the reservoir model, \( k[#1, n_{\text{sce}}] \)

**References**


Appendix A

Description of HR-82

The study case HR-82 derives from one of the geological realisations of the UNISIM-I-H model (Avansi and Schiozer 2015). A review of the inputs and outputs of the case study HR-82 differs it from the original benchmarking case. We firstly highlight the main characteristics of inputs and outputs from the HR-82 and then present additional details.

1. 82 attributes model all uncertain parameters as continuous variables:
   a. Geological uncertainty: continuous variables represent geological uncertainties related to porosity and permeability (instead of geological maps in UNISIM-I-H);
   b. Tables of fluid properties and relative permeability: coefficients of equations model the relationships of the original tables in continuous attributes;
   c. Uncertain attributes added near the well: well index multipliers;
   d. All the considered uncertain parameters of the HR-82 have uniform distribution through an interval of values.

2. The historical data is based on a hypothetical reality defined as a known combination of uncertain parameters:
   a. The original simulation outputs of the hypothetical reality (without noise) are added of random and systematic errors;
   b. The model discrepancy was assessed to incorporate errors in simulator target values during the historical period.

Table A-1 summarises all the uncertain attributes of HR-82. All of them are uniformly distributed and have a value in their ranges which is considered the hypothetical reality.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
<th>Ranges</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c_p$</td>
<td>Rock compressibility</td>
<td>[10, 96] E-6</td>
<td>1</td>
</tr>
<tr>
<td>$c_{krw}$</td>
<td>Relative permeability</td>
<td>[0.86, 1.28]</td>
<td>1</td>
</tr>
<tr>
<td>mphi</td>
<td>Porosity multiplier</td>
<td>[0.75, 1.25]</td>
<td>13</td>
</tr>
<tr>
<td>A</td>
<td>k\textsubscript{x}, angular coefficient</td>
<td>[0.135, 0.175]</td>
<td>13</td>
</tr>
<tr>
<td>B</td>
<td>k\textsubscript{x}, linear coefficient</td>
<td>[-0.4, 1.1]</td>
<td>13</td>
</tr>
<tr>
<td>mkz</td>
<td>kz multiplier</td>
<td>[0.1, 0.5]</td>
<td>13</td>
</tr>
<tr>
<td>PVT\textsubscript{oil},E3</td>
<td>Oil compressibility</td>
<td>[1.40, 1.62] E-3</td>
<td>1</td>
</tr>
<tr>
<td>PVT\textsubscript{oil},E4</td>
<td>Oil viscosity related</td>
<td>[2.5, 50.0] E-4</td>
<td>1</td>
</tr>
<tr>
<td>WOC\textsubscript{EO}</td>
<td>Water-Oil-Contact</td>
<td>[3169, 3179]</td>
<td>1</td>
</tr>
<tr>
<td>Wir</td>
<td>Well index factor (9)</td>
<td>[0.6, 1.4]</td>
<td>25</td>
</tr>
</tbody>
</table>

The two first attributes of Table A-1 have an impact in all the reservoir model. The parameter $c_{krw}$ describes uncertainty related to relative permeability using regression for three tables of the UNISIM-I-H. Figure A-1 shows that water relative permeability grows as an exponential function of water saturation multiplied by $c_{krw}$ varying on an interval.
Figure A-1—Relative permeability as a function of water saturation. Three tables from the benchmarking case UNISIM-I-H are plotted. A simple equation models the relationship with one uncertain attribute, the $c_{krw}$.

Table A-1 follows with four lines (in green) describing additional 52 ($4 \times 13$ regions) uncertain parameters modelled. Geological properties of porosity and permeability are based on a map of porosity $\phi_{map}$. The reservoir model is divided on the 13 regions – Figure A-2 - as in Maschio and Schiozer (2016), each of them having 4 uncertain parameters associated:

- a porosity multiplier $mphi_{region \ i}$;
- two coefficients ($A_{region \ i}$ and $B_{region \ i}$) to model the horizontal permeability;
- one multiplier for vertical permeability $mkz_{region \ i}$.

Figure A-2—Two-dimensional aerial view of the reservoir model highlighting (a) the 13 regions composing the reservoir, (b) a sealing fault diving the reservoir into two compartments (east block with regions 8, 11 and 12, west block with the other regions), (c) the position of production and injection wells, in red and blue respectively.

The Equations 16 to 18 represent the porosity of a region as a function of its multiplier, the horizontal permeability as a function of the porosity, and the vertical permeability as a function of horizontal permeability.

\[
\phi_{region \ i} = mphi_{region \ i} \ast \phi_{map} \tag{16}
\]

\[
kx_{region \ i} = 10^{100 \ast A_{region \ i} \ast \phi_{region \ i} - B_{region \ i}} \tag{17}
\]

\[
kz_{region \ i} = mkz_{region \ i} \ast kx_{region \ i} \tag{18}
\]

Properties of three PVT tables were analysed and modelled as a function of two additional parameters. PVT$_{co,EB}$ directly represent oil compressibility; PVT$_{ai,EB}$ models oil viscosity through an equation with exponential decay in the interval of interest (Figure A-3).
Figure A-3—Oil viscosity dependence on pressure; the three PVT tables from the UNISIM-I-H benchmarking case are plotted. The PVT table is an uncertainty identifies in the east block only. When we consider the range of pressure under interest, we can model the oil viscosity by an equation with exponential decay and with one uncertain attribute \( a_i \).

Remaining properties defined in the PVT tables (\( F_i \)) were estimated based on a regression over the interval of pressure, with no uncertainty associated with them (Figure A-4).

Finally, each of the 25 production and injection wells has a multiplier of well index associated to represent local effects impacting pressure exclusively around the wells.

Other relevant aspects include the use of well trajectories to define the wells; pressure limits of 36-450 kgf/cm\(^2\) during history period; operational conditions for forecasting are the same as the ones applied in UNISIM-I-H.
Appendix B

Step 8 – Selection of outputs to emulate

In STEP 8, we aim to select the number of outputs to emulate in order to balance the computational effort for constructing and evaluating new scenarios through emulators. In this way, we try to construct informative emulators to be considered in the analysis progressively. We propose a procedure for output selection which considers the combination of quantities of interest in order to estimate which one is the most informative. Before STEP 8, STEP 15 in the high-level workflow (Figure 4) guarantees that we have sufficient scenarios in the training set to construct emulators.

The workflow presented in Figure B-1 is applied in our case study to select outputs to emulate. With Figure 5, we illustrated the mechanism of the systematic procedure, showing that we can construct emulators for quantities of interest from all phases already evaluated. The starting point of the selection process enables this flexibility: STEP 8.1 is a conditional statement to decide whether we should evaluate quantities of interest of Phase 1 or more advanced phases already simulated. Note that in the diagram of Figure B-1 phase is the latest phase evaluated through simulation (coherent with the main workflow in Figure 4), and \( \phi \) is the phase from which quantities of interest can be selected. If a new wave was simulated, we set \( \phi = 1 \) with STEP 8.2, else we choose a more advanced wave in STEP 8.3 (i.e. an successive increment in \( \phi \) until the latest phase evaluated with simulators).

![Figure B-1—Workflow to select outputs to emulate in STEP 8. In the first three steps, the phase of evaluation \( \phi \) is defined. STEPS 8.4 to 8.6 identify the class of outputs to be considered in the selection. STEPS 8.8 to 8.12 select a combination of outputs with a potential to have the higher information index.](image)

In STEPS 8.4. to 8.6, we check specific classes of outputs to emulate. Two-class outputs allow identifying scenarios reaching simulation targets or physical boundaries (water breakthrough). We recommend firstly to construct emulators for these binary quantities (STEPS 8.4 and 8.5). They can rule out regions of the input space that lead to discontinuous behaviour, which is beneficial to emulate continuous outputs (STEP 8.6). For continuous outputs, we use the formulation of adapted implausibility measure \( I_i \) and \( \bar{D}_{info} \). Equation
13. This choice is a consequence that the *STEP 8* occurs previously than the construction of emulators in the main workflow. We highlight that this formulation implies an imprecision in the output selection, which consists in all outputs being considered equally challenging to emulate (which is not expected to be true).

When any quantity of interest offers (a) enough variability and (b) potential to rule out implausible regions of the space (i.e. $\hat{D}_{info} \geq D_{info,thres}$ in our application $D_{info,thres}=5\%$), no quantity of interest is selected in $\varphi$. In *STEP 8.7*, we consider the case where $\varphi=\text{phase}$, a condition that would terminate our selection process with no output to emulate based on the training set available (*STEP 8.13*).

In *STEP 8.8*, we select relevant outputs following *STEPS 8.4* to 8.6. In *STEP 8.9*, the adapted information index for all possible combinations of outputs is calculated, which we plot in Figure B-2. We highlight the combinations with five outputs (Figure B-2.a) to clarify that each circle is a possible combination of five outputs. These plots were constructed to select outputs from Phase 3 in the Waves 6 and 15. They show a trend to increase the adapted information index as the number of output in the combinations increase. In Figure B-2.a, the maximal possible $\hat{D}_{info}$ is reached by all combinations of 15 outputs, for example.

![Figure B-2—Adapted information index for combinations of continuous outputs of Phase 3 (STEP 8.9) – (a) in Wave 6, 19 outputs are available, the combinations of five outputs are highlighted; with 15 outputs the adapted information index reaches the limit of 100%; (b) in Wave 15, five outputs are available; the maximal adapted information index is approximately 85%. The comparison between the two waves emphasises the uncertainty reduction along the waves.](image)

In *STEP 8.10*, we select a number of outputs (e.g. 15). A design option stands for the number of outputs to be emulated. One should account for the possibility to miss informative outputs in the stage of the process, or for increasing the computation cost demanded to construct and evaluate additional outputs. Then, in *STEP 8.11*, we identify the combinations with the highest information index $\hat{D}_{info}$. In Figure B-2.a, all combinations of 15 outputs provide the maximal possible $\hat{D}_{info}$. A coherent way to select one of these combinations and the corresponding outputs to select (*STEP 8.12*) is to choose the combination that offers the highest maximal implausibility $\hat{I}_M$.

We add Figure B-2.b in our analysis to emphasise that, in later waves, both (a) the maximal possible $\hat{D}_{info}$ (approximately 85% in Wave 15 versus 100% in Wave 6), and (b) the number of outputs possible to emulate tends to decrease.