Rapid evaluation and optimization of carbon dioxide-enhanced oil recovery using reduced-physics proxy models

Watheq J. Al-Mudhafar1 | Dandina N. Rao2 | Sanjay Srinivasan3 | Hung Vo Thanh4 | Erfan M. Al Lawe1,5

1Department of Reservoirs and Fields Development, Basrah Oil Company, Basrah, Iraq
2Department of Petroleum Engineering, Louisiana State University, Baton Rouge, Louisiana, USA
3Department of Energy and Mineral Engineering, Penn State University, University Park, Pennsylvania, USA
4School of Earth and Environmental Sciences, Seoul National University, Gwanak-gu, Korea
5Subsurface Department, BP, Basrah, Iraq

Abstract

The carbon dioxide (CO2) injection in oil reservoirs is a promising enhanced oil recovery method to reduce carbon emissions into the atmosphere. The CO2 injection simulation and optimization require a large computation time, especially in real large-scale oil reservoirs. This paper integrates experimental design and machine learning to construct a reduced-physics surrogate model alternative to the complex reservoir simulator. This surrogate model was used to evaluate and optimize the CO2-assisted gravity drainage (CO2-GAGD) process, applied to a clastic reservoir in the Rumaila oil field. In the GAGD process optimization, five operational decision parameters controlling the production and injection activities were manipulated to attain optimal future oil production. Hundreds of simulation runs were created by Latin Hypercube Sampling to build a proxy-based optimization. The optimal scenario increased the cumulative oil production by 416 MMSTB at the end of 10-year prediction period. Finally, four machine learning (ML) algorithms were built as proxy surrogates alternative to the complex compositional reservoir simulations: Quadratic Equation (QM), FUzzy logic-GEnetic algorithm (FUzzy-GEnetic), Multivariate Additive Regression Splines (MARS), and Generalized Boosted Modeling (GBM). To show how robust using the ML approaches for the fast CO2-GAGD process optimization, the random subsampling cross-validation was adopted to conclude the optimum proxy metamodel that provides the lowest mismatch between the proxy- and simulator-based cumulative oil production. The GBM algorithm achieved the highest adjusted-\(R^2\) (0.9973) and lowest root mean square prediction error \(3.704 \times 10^6\) to produce the most accurate metamodel. The resulting accurate GBM proxy metamodel can be used for fast optimization of the CO2-GAGD process. Specifically, GBM-metamodel should lead to achieving higher optimal reservoir flow responses, especially with running a large number of simulation runs. Consequently, this proposed trained GBM-ML proxy metamodel could be used for gas injection...
optimization and uncertainty assessment with far-less computational effort than with the conventional approaches, which use a complex flow simulator.

KEYWORDS
CCUS, clastic oil reservoirs, design of experiments, enhanced oil recovery, machine learning, proxy metamodeling

1 | INTRODUCTION

Energy consumption worldwide is vastly increasing annually due to the technology revolution and the population’s significant incrimination. Discoveries of new oil fields have become rare in the last decade. Therefore, boosting oil production by utilizing all available technological methods including enhanced oil recovery (EOR) methods is inevitable. Specifically, enhanced oil recovery is becoming more important than ever, as finding new sources is becoming more costly and time-consuming, especially when the oil price is facing a huge reduction. Gas flooding is considered one of the most promising technologies used in EOR projects. Gas flooding is being implemented through the continuous gas injection (CGI) or water alternating gas (WAG) methods. However, the gas-assisted gravity drainage (GAGD) technique was recently implemented to advance the volume of the recovered oil in both secondary and tertiary stages for immiscible/miscible gas injection methods. Consequently, it was proposed to boost the volume of the recovered oil in the main pay of Zubair formation in the Rumaila oil field, which is located in southern Iraq. The field was selected because it is a mature oil field that has been producing for more than 60 years and waterflooding is no longer effective to increase the recovery of oil.

Elevating oil recovery by implementing gas flooding is crucial in field development through EOR projects, especially when conducted in real oil fields. Many limitations impact reservoir production through the CO₂-EOR processes. These parameters are operational or controllable by field operators, such as well production and injection constraints. Some other factors are uncontrollable, such as the geological properties. These operational factors need to be optimized to identify the optimal solution for the reservoir flow response. It is indispensable to define the ideal limits of the operational parameters that govern the effectiveness of the EOR method. These factors are primarily comprised of operational constraints with the injection and production wells in a certain reservoir. More specifically, how the production constraints are defined to regulate the volume of produced and the injected volumes in the reservoir; consequently, it provides a major influence on the reservoir flow response. Therefore, the enhancement of these parameters can achieve ideal reservoir performance with the time relative to reservoir cumulative oil production as well as net present value (NPV).

Proxy modeling and the design of experiments (DoE), also noun as designed experiments, are statistical techniques, which are integrated to build a response surface methodology (RSM) as an uncomplicated substitute or metamodel for the convoluted models to assess the several constructed experiments in the enhancement approach instead of evaluating the same simulator. The proxy modeling optimization has not been found in the literature when it comes to the CO₂-GAGD technique. However, the proxy model has been implemented in different reservoir studies and EOR modeling, such as oil production optimization, waterflooding processes, gas flooding processes, steam injection, chemical flooding, and history matching. The DoE concepts generate various computer trials (realizations) for the subject by connecting the levels for every parameter. These experiments are evaluated to compute the response factor. The created experiments and response factors are taken into account in the statistical modeling to generate a relationship, which characterizes the proxy or surrogate model. Many DoE approaches have been used in various reservoir simulation studies to build proxy models. The most common DoE approaches are fractional factorial design, Central composite (CC) designs, D-optimality also noun as D optimal design, and Latin Hypercube Design. There are many successful examples of using the proxy models in the literature of reservoir studies, for instance, the second-degree polynomial equation, kriging algorithms, and artificial neural networks algorithm.

Constructing a large and complex compositional reservoir flow simulation for the evaluation and optimization of the CO₂-EOR, especially in a real oil field, is an expensive and time-consuming step, especially in the need to run hundreds of simulation scenarios through the optimization process. Therefore, it is essential to construct a simplified model (metamodel or surrogate)
alternative to the complex flow simulation models. This simplified metamodel could provide very fast and high-quality analysis of the complex reservoir engineering problems. Specifically, metamodeling or surrogate modeling serves as an alternative modeling technique to fast solve complex problems by reducing the computation time along with preserving the accuracy of detailed complex solutions. The possible disadvantages of these metamodels are related to the accuracy of the algorithms used to build the proxy metamodels. However, this point has been overcome by using advanced machine learning approaches that provide very accurate modeling and prediction in comparison to the outcome of complex models. Consequently, this paper combines the DoE and proxy metamodeling for the fast and accurate evaluation and optimization of the CO2-assisted gravity drainage (CO2-GAGD) process in heterogeneous clastic reservoirs of the south Rumaila oil field.

To the best of our knowledge, this study's new contributions in comparison to the previous studies include various points: (i) the fuzzy logic-genetic algorithm and generalized boosted modeling approaches are the first time used to construct the proxy metamodels for fast optimization of the gas injection processes based on large reservoir flow simulation of a real oil field. (ii) the power of integrating experimental design and fuzzy logic-genetic algorithm along with generalized boosted modeling has never been adopted in the optimization of the CO2-GAGD process. (iii) the developed proxy metamodels can adapt for fast evaluation and optimization of other CO2-EOR projects that have similar reservoir characterization. (iv) the resulting accurate proxy metamodels can provide full optimization of CO2 EOR in real oil fields by running thousands or millions of designed simulations within a few seconds. (v) these accurate surrogate models can efficiently replace the computationally expensive compositional simulation models of other real oil fields in the case of a lack of data, which are required to build the simulation models.

The entire evaluation and optimization workflow can be summarized by the following steps, as depicted in Figure 1:

(i) Build a full detailed compositional reservoir flow simulation to assess the reservoir production by the CO2-GAGD flooding during 10 years of future reservoir production.

(ii) Then, the proxy models were reconstructed by manipulating the operational decision parameters that impact CO2 flooding through the DoE-latin hypercube sampling (DoE-LHS) approach.

(iii) More specifically, the DoE-LHS approach was used to create multiple simulation runs by generating multiple levels of each of the controlling gas injection parameters.

(iv) The levels of the controlling parameters are manipulated in a way that ensures the minimum number of descriptive simulation runs that lead to

![Reservoir Simulation & History Matching Model](image1)

![CO2-GAGD Process Evaluation](image2)

![Design of Experiments](image3)

![Run the designed simulations by compositional simulator](image4)

![Generation the database for machine learning modeling](image5)

![Build the Proxy Metamodel](image6)

![Determine the Optimal Proxy Model for Fast Optimization](image7)

**FIGURE 1** General sketch of rapid GAGD process evaluation and optimization problems. GAGD, gas-assisted gravity drainage.
achieving the optimum conditions of the CO\textsubscript{2} injection implementation.

(v) Later and after running hundreds of simulation cases, the DoE-LHS approach was incorporated to generate a simplified surrogate approach (metamodel) alternative to the compositional reservoir model.

(vi) To construct the surrogate metamodels, four machine learning (ML) and artificial intelligence (AI) algorithms were adopted: polynomial proxy model, fuzzy logic-genetic algorithm, multivariate additive regression splines, and generalized boosted modeling.

(vii) The random sub-sampling cross-validation with variance calculations was then used to validate the accuracy of the four proxy models.

2 | GAGD PROCESS

Natural separation of reservoir fluids is imperative to boost the recovery of bypassed oil by the CO\textsubscript{2}-GAGD technique by supplying gravity steady oil sweep. The GAGD technique was patented to improve oil recovery in different production stages including secondary and tertiary for both immiscible and miscible gas injection practices.\textsuperscript{1} The GAGD process is achieved by placing horizontal production wells at the lower part of the target reservoir. Then, the immiscible or miscible gas injection process in a gravity-stable state by the vertical injectors at the top of the reservoir is initiated.\textsuperscript{24} Because of the gravity segregation occurring from the various fluid densities at reservoir conditions, the accumulation of the injected gas will be at the top of the reservoir to generate a gas cap, as illustrated in Figure 2.\textsuperscript{25} This process will supply gravity stable oil sweep that moves the hydrocarbon to the bottom of the reservoir towards horizontal production wells, which helps to achieve improved sweep performance and optimum oil recovery.

Mostly, the CO\textsubscript{2} gas is favored for the GAGD process as it provides optimum volumetric sweep performance with ideal microscopic displacement efficiency, particularly when it is used in miscible injection processes. Furthermore, the ideal volumetric sweep performance enhances retarding CO\textsubscript{2} arrival in production wells.\textsuperscript{1} Slowing down or diminishing the arrival of the injected gas minimizes concurrent gas-liquid flow, and later enhances gas injection to sustain the reservoir pressure.

3 | FIELD DESCRIPTION

The subject oil field was first discovered at the end of 1953 in the south of Iraq. The field is approximately 50 km to the west of Basrah city and located 30 km to the west of the Zubair field.\textsuperscript{26} The field length is 100 km with 12–14 km width and is located 3 km below sea level. Reservoir flanks are dipping with angles that do not exceed 3° while the crest only dips with 1°. South Rumaila oil field consists of many oil-bearing reservoir intervals.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{GAGD_process.png}
\caption{Schematic of the GAGD process. GAGD, gas-assisted gravity drainage.}
\end{figure}
The main prolific oil reservoir is Zubair characterized by the Late Berriasian Albian cycle while its sediments go back to the lower cretaceous age. The thickness of the Zubair reservoir ranges from 280 to 400 m and the sand to shale ratio depicts that the Zubair formation involves five members that are rich in organic contents: upper shale, upper sandstone (main producing), middle shale, lower sand, and lower shale. According to Al-Ansari, Zubair formation is a conventional reservoir that does not contain any complicated geological structures and figures such as faults or fractures. Four key sectors were defined in the South Rumaila field: Janubia, Rumaila, Shamiya, and Qurainat. Rumaila sector and only minor regions of the Shamiya and Janubia sectors will be studied in this study. The selection of these parts was decided to depend on the gathered reservoir data and the capability to characterize the major parts of the reservoir, where the wells are producing and water injection activities are performed, shown in Figure 3A that represents the structural contour map.

It was found that the Zubair formation has two types of boundary conditions that include a no-flow boundary and an aquifer. The no-flow boundary was proposed to cover the reservoir’s northern and southern areas. This assumption is acceptable as it mimics the reality since the reservoir adopted balance production and injection rates. Moreover, the streamlines in the north and south of the reservoir were crossed by the isobaric lines, which are perpendicular to the reservoir boundaries. Therefore, the flow direction is parallel to these boundaries, as shown in Figure 3B. Whereas the east and the west flanks were characterized by flow boundaries that symbolize the natural water drive to mimic the effects of the existing infinite aquifer.

The South Rumaila field was first developed for primary production in 1954 while water injection was commenced in 1980s to maintain the reservoir pressure and to sustain the west flank aquifer support, which is about 20 times stronger than the east flank aquifer. Throughout field development, 40 production wells were drilled in the regions that are under investigation in this study. The minimum well spacing between each of the producers and injectors is 400 meter. The development of the South Rumaila field included shutting of some of the reservoir layers, as they were the reason for high water cut values that reached about 98%. Until 2004, the total volume of the injected water reached approximately 1.1 billion barrels with various rates of injection. The maximum injection rate value was 426,000 BPD for 2 months in 1988. Since the water cut hit 80% in some wells, artificial lift (ESPs) has been used to sustain production. The main pay in South Rumaila has original oil in place (OOIP) of 19.5 billion barrels and the estimated recovery factor is around 55%. The studied sector in this study has an OOIP of 6.13 billion barrels.

4 | COMPOSITIONAL SIMULATION OF GAGD DESIGN

A comprehensive compositional reservoir simulation was constructed to optimize the recovery of oil by the GAGD process in the main pay reservoir through five main steps.
4.1 | The geological modeling

The geological description of the main pay reservoir in South Rumaila depicts that it has three rock types: sand, shale, and shaly sand that are distributed through the reservoir with different permeability ranges. To model lithofacies and petrophysical properties, a full detailed geostatistical model with 1,908,900 grids has been employed using the Multiple Point Geostatistics and Sequential Gaussian Simulation. More specifically, the multiple-point geostatistics has been first used for 3D lithofacies modeling. Then, the sequential Gaussian simulation has been adopted for the 3D horizontal permeability and porosity distribution given each lithotype, as depicted in Figures 4 and 5, respectively.

FIGURE 4  Three-dimensional distribution of the reservoir permeability for the South Rumaila main sector

FIGURE 5  Three-dimensional distribution of the reservoir porosity for the South Rumaila main sector
4.2 | Relative permeability

Since the oil relative permeability is a key factor in the gravity-drainage process, three different relative permeability and capillary curves were conducted into the reservoir flow simulation model as a function of the three lithotypes based on their ranges of permeability: sand, shaly sand, and shale, as depicted in Figure 6.

4.3 | History matching

The built geostatistical model has grids of 210, 202, and 45 in I, J, and K directions was then upscaled to grids of 69, 66, and 12 to simulate the GAGD process. To preserve the reservoir heterogeneity, arithmetic and harmonic mean formula were adopted for the porosity and permeability upscaling, respectively. The upscaled model was the base to build the compositional reservoir flow simulation by the use...
of the CMG-GEM package. The reservoir model should then be history matched to be used for future field development and planning. In the history matching, the trial and error technique considers frequently adjusting the reservoir model parameters and calculating the flow responses to check the matching with the observed measurements. The procedure continues until we reach the minimum mismatch error (satisfactory) match between the calculated and observed flow responses. In this paper, excellent curve matches of the cumulative oil production, field oil rate, cumulative water injection, and field injection rate were achieved in the early steps after adjusting the aquifers parameters of permeability and radius along with considering the Carter-Tracy approach to simulate the infinite-active aquifer. The acquired matching is an excellent indicator of the model's performance as it replicates water cuts and saturation distribution. The production and injection history of this field covered 56 years of production activities until the first quarter of 2010. Consequently, history matching was obtained between 1954 and 2010 as shown in Figures 7 and 8.

4.4 | PVT fluid behavior

A detailed pressure-volume-temperature (PVT) model was built through the WinProp package to spoor the interaction of primary and secondary fluids in terms of their compositions. WinProp is a Computer Modeling Group (CMG) compositional equation-of-state package used for phase behavior and property calculations. In WinProp, the fluids are characterized based on their components to tolerate the compositional interaction through the reservoir. The following steps of calculation were conducted to build the full PVT model to simulate the GAGD process in the South Rumaila oil field:

1. Fluid component properties and plus fraction characterization.
2. Multiple-contact miscibility calculations and ternary diagram generation.
3. Flash, saturation, and envelope calculations.
4. Recombination and constant composition expansion calculations.
5. Differential liberation and constant volume depletion.
6. Separator and swelling tests.

The compositions of the primary fluids (available oil) and the secondary (injected gas) were depicted in Figure 9. The general PVT data of solution gas-oil ratio (GOR), oil and gas formation volume factor (FVF), and oil and gas viscosity, all as a function of pressure, were illustrated in Table 1.

4.5 | The GAGD process simulation setting

After achieving a satisfactory history matching, the reservoir model was considered for the GAGD process evaluation within 10 years of the future production
The development strategy focuses on placing vertical and horizontal wells for simultaneous gas injection and oil production, respectively. The wells were placed at the early stage of development, all at the same time to provide the fast and efficient injection and production. A total of 33 wells have been used to implement the key concept of the GAGD process. These wells included 22 vertical injection wells and 11

**FIGURE 8** History matching of entire field injection of South Rumaila oil field

**FIGURE 9** The primary and secondary fluid compositions for the South Rumaila oil field
horizontal producers with 3000 m lateral length that are placed in the reservoir layers where the lithology is sand and shaly sand. First, CO₂ injection is commenced by the vertical injection wells at the shallower two layers. Simultaneously, the following three layers are utilized as a transition region to provide vertical space for gas gravity drainage. The next steps involve setting up the horizontal producers through layers 6–8, which contain the highest oil saturation in the reservoir. Eventually, the remaining four layers did not involve injection or production processes since the water saturation in these layers is 100% from the infinite water aquifer. Figure 10 illustrates the 3-D reservoir model of facies distribution with the positions of injection and production wells, which were used for the CO₂ injection in the GAGD method. In Figure 10, the reservoir body is represented by the red color that refers to the shale zones. Sand zones and shaly-sand zones (indicated in 1 and 2, respectively, in Figure 10) were perforated in production and injection wells since they are considered to be high permeable zones.

5 | OPTIMIZATION APPROACHES

DoE is a methodical numerical technique that generates a suitable group of experiments to be the base of the simulation runs. DoE is mainly utilized for determining the highest critical parameters that impact the response during the sensitivity analysis practice. The DOE tool provides a way to acquire the most likely case, which accomplishes the optimum response from a certain procedure. 36 DOE technique was demonstrated as an efficient tool for carrying out several tasks such as system optimization, variable screening, risk evaluation, and  

### TABLE 1 General PVT data used for the compositional simulation of the GAGD process

| Pressure (Psia) | Solution GOR (SCF/STB) | Oil FVF (bbl/STB) | Gas FVF (CF/SCF) | Oil vis (cp) | Gas vis (cp) |
|----------------|------------------------|-----------------|----------------|-------------|-------------|
| 14.70          | 0.0000                 | 1.0469          | 0.21852        | 1.3624      | 0.00855     |
| 100.00         | 36.9525                | 1.0748          | 0.03127        | 1.0614      | 0.01017     |
| 200.00         | 89.6405                | 1.1096          | 0.01540        | 0.9292      | 0.01075     |
| 300.00         | 128.3230               | 1.1333          | 0.01018        | 0.8457      | 0.01111     |
| 400.00         | 161.8011               | 1.1529          | 0.00758        | 0.7805      | 0.01136     |
| 500.00         | 193.0021               | 1.1705          | 0.00603        | 0.7252      | 0.01155     |
| 600.00         | 223.1102               | 1.1872          | 0.00499        | 0.6767      | 0.01170     |
| 700.00         | 252.7181               | 1.2033          | 0.00426        | 0.6333      | 0.01183     |
| 739.46         | 264.2473               | 1.2095          | 0.00402        | 0.6175      | 0.01188     |
| 1074.80        | 366.4098               | 1.2641          | 0.00272        | 0.5008      | 0.01227     |
| 1644.93        | 567.8513               | 1.3717          | 0.00175        | 0.3561      | 0.01298     |
| 2313.55        | 868.1831               | 1.5330          | 0.00123        | 0.2438      | 0.01415     |
| 4577.55        | 3889.9479              | 3.2476          | 0.00069        | 0.0646      | 0.02749     |

Abbreviations: FVF, formation volume factor; GAGD, gas-assisted gravity drainage; GOR, gas-oil-ratio.

**FIGURE 10** Production and injection well locations in sand and shaly-sand zones
robust design.\textsuperscript{37} It is essential to accomplish the most precise experimental design model that imitates the physical model or process for faster, cheaper, and more flexible implementation. To accomplish this task, the necessary group of elements and interactions have to be investigated to enable the analysis and implementation of results to be accurate and trustworthy.\textsuperscript{2}

The key terms of DoE are response parameter that symbolizes the result from a specific experiment and factor. The factors are described as a variable that impacts the response parameter and can be classified as primary and secondary based on the different levels of sensitivity. The overall count of designed experiments is defined by an exponential relationship. For illustration, the experiments count with \( k \) variables and 4 levels will be equal to \( 4^k \). The Latin Hypercube Sampling was implemented in this study along with the proxy modeling to identify the optimal values of the operational production decision factors for the GAGD method optimization.

Latin Hypercube Sampling (LHS) is described as a numerical sampling method that is employed to produce random samples out of given input factors to create several computer experiments and trials from a distribution with several levels.\textsuperscript{38} To represent various stages of variation for every single factor with the lowest number of trials, the sampling procedures deliver restricted data points in the design field with a uniform distribution by the space-filling strategy.\textsuperscript{39} LHS is an efficient design that generates uniform and low discrepancy observations.\textsuperscript{38} Figure 11 illustrates the space-filling design by LHS for two variables. In this figure, the sampled data are allocated within the entire space randomly, but all the points are uniformly distributed to capture the entire variation of the process being studied.

LHS creates further effective experiments for \( K \) parameters compared with the basic Monte Carlo sampling technique. More explicitly, LHS delivers a steady points design since it retains the highest space between each design point compared with the other points.\textsuperscript{40} In LHS, \( K \) variables sampling is conducted by splitting every parameter into several equivalent parts. Furthermore, LHS is a thorough practice that randomly produces a new group of trials or experiments in case the given data set does not characterize the problem. However, there is no explicit approach to define the required number of trials or experiments that can be generated.\textsuperscript{41} The computer experiments, which were generated for optimization, were accomplished in R statistical language by [lhs] package.\textsuperscript{42}

6 | PROXY MODELING

The proxy approach deals with building a simplified model alternative to the complex model (metamodel). This proxy model empowers users to achieve results equivalent to the results acquired by the sophisticated models with a significantly lower computational time that might be a few seconds for millions of simulation runs. Nonetheless, the complex model consumes several days to obtain the results of hundreds of runs. The proxy model is performed by fitting the operational parameters
training data to the response factor, which is represented by the following relationship:

\[ y = f(X_1, X_2, ..., X_k) + \varepsilon, \] (1)

where \( X_1, X_2, ..., X_k \) represent the input variables, \( y \) refers to the anticipated response factor, and \( \varepsilon \) refers to the regression error.

Cross-validation is necessary to maximize the opportunity of achieving global optima and to enhance the forecast precision of the proxy model. The random sampling cross-validation was implemented on the entire experiment through sampling and dividing the given data set into two groups: 30% testing set for forecast and prediction while 70% training set for building the model. More explicitly, the training subset was used as the base for cumulative oil production modeling, obtained by the reservoir simulator, as a function of operational parameters. The prediction is then utilized for testing subset data using the simulator as well as the proxy model.

In this paper, the comparison between the proxy models was conducted for the missing match of the cumulative produced oil that was estimated through the reservoir simulation model and by the proxy model depending on the testing subset, not the same training data set. More specifically, this cross-validation approach provides an indication of the model’s generalization ability and confirms making exterior forecasts on the new data set that was unseen in the training subset-based modeling (the results can be trusted when implemented on external datasets). The mismatch was computed by estimating the root mean square prediction error (RMSE) and the updated \( R^2 \) adj. RMSE quantifies the predictable squared variance of the reservoir simulator and proxy model response factors, for instance, the cumulative produced oil. While, the updated \( R^2 \) adj is a modified version of \( R^2 \), which shows how much variance can be described by a model. However, the updated \( R^2 \) adj is adjusted for several predictors in the model and it grows only if the new term enhances the model.

\[
RMSE = \frac{1}{n} \sqrt{\frac{1}{n} \sum_{j=1}^{n} (\hat{y}_j(x_i) - f_j(x_i))^2}, \] (2)

\[
R^2_{adj} = 1 - \frac{(1-R^2)(n-1)}{n-k-1}, \] (3)

where \( n \) is the count of experiments and \( k \) is the count of predictors (operational decision factors).

Figure 12 illustrates the comprehensive flowchart for DoE, Proxy model, cross-validation, and application domain of the optimal proxy model.

From the aforementioned DoE and proxy modeling, 643 simulation jobs (experiments of the operational parameters) were generated for the training and validation runs. These runs were then adopted for the comparison of four proxy models: Polynomial (Quadratic) Regression (QM), Multivariate Additive Regression Splines (MARS), Fuzzy log-GEnetic Algorithm, and Generalized Boosted Regression Model (GBM).

### 6.1 Polynomial regression

The second-degree polynomial regression has been employed to generate the RSM by constructing a nonlinear function that relates the input parameters with the response factor. The common model for the polynomial RSM is the second-degree quadratic model. Models with more than two parameters, are characterized by the following formula:

\[ y = \alpha_0 + \sum_{j=1}^{k} \alpha_j X_j + \sum_{j=1}^{k} \alpha_{jj} X_j^2 + \sum_{i<j}^{k} \alpha_{ij} X_i X_j + \varepsilon, \] (4)

where \( \alpha_j \) is the linear coefficient while \( \alpha_{jj} \) is the coefficient of the quadratic term, and \( \alpha_{ij} \) is the interaction coefficient of every two factors. The RSM was entirely implemented via R-statistical programming language through rsm package.

### 6.2 Multivariate additive regression splines

MARS is described as an improved technique of linear regression that represents a nonparametric regression approach that spontaneously formulates a function between variables considering the nonlinearity by utilizing piecewise linear slices, which noun as splines. The MARS technique employs a group of coefficients and functions that define a formula to relate the response parameters and the forecasted variables. MARS technique is proper for multidimensional predictors since the fundamental functions divide the given input data into sections where every section contains its coefficients group to eliminate the potential outliers that might be available in the data set.

Modeling data by MARS is applied by two key stages: a forward phase that explores possible knots to advance the performance of modeling, and a backward procedure that removes the unimportant predictors. MARS model is characterized by the below function that applies the forward step:
where:

\[ y = c_0 + \sum_{i=1}^{N} c_i \prod_{j=1}^{K_i} b_j(X_{v(j,i)}) \]  \hspace{1cm} (5)

\( y \) is the response or output variable, \( c_0 \) the constant, \( c_i \) the vector of coefficients of the irregular fundamental functions, \( v(j,i) \) the index of independent parameters utilized in the \( i \)th term of the \( j \)th product, \( b_j(X_{v(j,i)}) \) the truncated power fundamental function with \( v(j,i) \), \( K_i \) the parameter that limits the order of interactions, \( b_j \) the spline function.

Non-influential predictors are eliminated through the backward step depending on the global cross-validation (GCV) principle, which adaptively deals with the diverse trends of data.\textsuperscript{45} The entire implementation of MARS approach was performed through earth package in the R-statistical programming language.\textsuperscript{49}
6.3  Fuzzy logic-genetic algorithm

Fuzzy logic is a method of knowledge illustration proper for notions that cannot be identified accurately, but their contexts govern it. Fuzzy Logic is a convenient way to build a fuzzy input and output data model. The fuzzy logic system comprises three phases: fuzzifier, fuzzy inference system, and defuzzifier.\(^\text{50}\) In particular, the mechanism of the fuzzy logic system can be described as follows: in the fuzzifier stage, the raw inputs to the system form fuzzy inputs. Later, these fuzzy inputs are to be populated into the inference environment or system in which the real calculations are accomplished. The next step includes incorporating the rule base with fuzzy inputs and the inference engine to generate fuzzy results for every rule. The rule base is described as a confinement of expert understanding. The fuzzy results create a fuzzy group and this group is converted into a crisp value through the defuzzifier stage.\(^\text{51}\) However, a Genetic Algorithm is a random search tool to generate potential answers that compete with each other to define the most suitable solution by applying operators of recombination, transformation, and selection that mimic the genetic regeneration in a biological environment comparable to the Natural Section theory that was proposed by Darwin.\(^\text{52}\)

FUuzzy logic-GEnetic algorithm (FUzzy-GEnetic) is an evolutionary algorithm of fuzzy systems population, which is randomly generated by Genetic Algorithm, to be used as a prediction model by fitting the given training data as labels. The full procedure of FUzzy-GEnetic proxy modeling was implemented through fugeR R-package.\(^\text{53}\) In fugeR, the given data is used to verify the entire fuzzy system. The algorithm forecast is then compared with the labels and each system has a quantified performance. The population of chromosomes is then utilized to create a 20% population for the subsequent creation based on the crossover and mutation. The final creation identifies the fuzzy system that achieved the optimum performance.

6.4  Generalized boosted regression model

GBM is an influential technique that has various implementations in machine learning, GBM was developed by\(^\text{54,55}\) to imitate the complicated dependencies of a nonlinear relationship. Specifically, GBM can be described as an application of augmentation to Freund and Schapire’s AdaBoost algorithm and J. Friedman’s gradient boosting machine.\(^\text{56}\) In the literature, GBM has been proven to have wide implementation within machine learning and data science subjects that achieves high accuracy of forecast and modeling of the response parameters. In GBM modeling, fitting new models repeatedly will lead to achieving precise modeling as it helps to minimize the difference between the observed and the forecasted responses. A key concept of GBM is to train the data to attain the highest formulation with the negative gradient of the loss function.\(^\text{57}\)

The concept beyond the GBM loss function is to eliminate high deviations from the objective results and to ignore the insignificant residuals.\(^\text{57}\) The GBM technique starts with allocating a differentiable loss function and begins with a base model \(F\). Afterward, iteration is executed to compute the next negative gradient and the process finishes when convergence is obtained.

\[
-g_i(x_i) = -\frac{\partial L (y_i, F(x_i))}{\partial F(x_i)}.
\]

Then, the regression tree \(h\) is matched to the negative gradients \(-g_i(x_i)\). The full practice of GBM-based proxy modeling has been applied by \textit{gbm} library in R language.\(^\text{58}\)

7  GAGD production optimization

The five operational decision parameters investigated for the immiscible GAGD production optimization are maximum oil production rates (MAX_STO), lowest bottom hole pressure (MIN_BHP), and water-cut (MAX_WCUT) in producers, along with the maximum gas rate (MAX_BHG) and minimum bottom hole injection pressure (MAX_BHP) in injectors. These five parameters represent the main controls on the production and injection activities in the field. Specifically, these parameters govern the amount of produced oil from the production wells by specifying the maximum limits of oil production and water cut along with determining the minimum limit of bottom hole pressure. However, the maximum injection rate and maximum injection pressure control the total amount of injected water into the reservoir, which improves reservoir pressure and oil production. Table 2 illustrates the default parameters of the base simulation scenario of the GAGD technique in addition to the ranges of each parameter (minimum and maximum levels) within the optimization procedure.

In Table 2, the main predefined constraint is the maximum water cut value, which was specified according to the capabilities of surface facilities that separate water from oil in the field. Moreover, the minimum
bottom hole pressure in the production well represents the second constraint that governs the amount of oil production and keeps the pressure difference reasonable to improve oil production. The over-predicting field performance happens when we specify the minimum bottom hole pressure to the lowest levels, which causes a large pressure difference and over-prediction of oil production. The levels of each factor in Table 2 were combined by the Latin Hypercube Sampling to produce hundreds of simulation jobs (experiments). Then, the designed experiments were assessed by the compositional flow model to estimate the cumulative produced oil during 10-year prediction period (January 1, 2026).

The optimal solution referred to the simulation job that results to achieve the maximum cumulative oil production, as depicted in Figure 13. It also demonstrates that field cumulative produced oil through the base simulation scenario of the GAGD technique in addition to the general solutions that represent the nonoptimal cases. The total number of the generated simulation jobs including the optimal solution was approximately 625 runs.

The optimal solution was identified and visualized relative to the field cumulative produced oil, as outlined in Figure 14. The optimum solution represents the maximum field cumulative produced oil during 10 years of prediction time. The general solutions in Figure 14, represented by the green curves, refer to the least flow response that combines low and/or poor combinations of the factors' levels. Hence, they lead to low levels of field cumulative oil production.

The cumulative produced oil during the prediction time through the base simulation scenario of the GAGD technique was 4.39 bn STB. Nonetheless, the optimal solution, obtained from LHS-based proxy optimization (Optimal Case), managed to enhance the cumulative produced oil production to 4.6 bn STB. The incremental oil recovery is 215.2 million STB, as illustrated in Figure 15, which compares the base simulation case and optimal GAGD process performance along with the primary production case of no injection. The optimum

| TABLE 2 The operational decision parameters of the GAGD production optimization |
|------------------------------------------|----------------|----------------|----------------|
| Response CumOilProd, STB Factor          | Min            | Base case      | Max            |
| MAX_STO, STB/DAY                         | 350,000        | 750,000        | 625,000        |
| MIN_BHP, Psia                            | 2000           | 2660           | 3500           |
| MAX_WCUT                                 | 0.70           | 0.95           | 0.90           |
| MAX_BHG, ft³/DAY                         | 7.5E06         | 10E06          | 15E06          |
| MAX_BHP, Psia                            | 2000           | 3000           | 3500           |

Abbreviation: GAGD, gas-assisted gravity drainage.
cumulative oil production was acquired through achieving the optimal levels of the production control factors, which are illustrated in Table 3.

In Figure 15, a substantial increment in oil recovery was acquired from the optimal solution in comparison with the base simulation scenario of the original conditions of the operational parameters. The field cumulative produced oil acquired through the base scenario in 10 years can be produced after only 18 months within the prediction period.
7.1 Evaluation and validation of the proxy models

The entire procedure of proxy-based optimization was illustrated considering the iterative procedure of constructing a polynomial model. The 643 simulation jobs were used for a comparison of building new proxy models through four various approaches in a different procedure. The new four proxy metamodels are Polynomial Regression, Multivariate Additive Regression Splines (MARS), Fuzzy Logic-Genetic Algorithm, and Generalized Boosted Model (GBM). This method

| Response | Base case | Optimal |
|----------|-----------|---------|
| CumOilProd, STB | 4.3887E09 | 4.6039E09 |
| MAX_STO, STB/DAY | 750,000 | 500,000 |
| MIN_BHP, Psia | 2660 | 2000 |
| MAX_WCUT | 0.95 | 0.9 |
| MAX_BHG, ft³/DAY | 10E06 | 15E06 |
| MAX_BHP, Psia | 3000 | 3500 |

**TABLE 3** Optimal levels of the operational decision factors in comparison with the base case

**FIGURE 16** Comparison of proxy-predicted and simulator calculated-QM given the test subset

**FIGURE 17** Comparison of proxy-predicted and simulator calculated-QM given the full data set
combines multiple techniques such as cross-validation, Root Mean Square Error and adjusted $R^2_{adj}$ aiming to identify the optimum proxy model that is possible to be assumed as an ideal metamodel of the nonlinear CO$_2$ injection through the GAGD technique.

After sampling and subdividing the data set of 643 designed experiments into 75% training and 25% testing subsets, the modeling was implemented based on 450 simulation jobs (training subset) through the four aforementioned proxy approaches. The prediction from each of the four proxy models was then adopted based on 193 runs (testing subset). The RMSE along with the adjusted $R^2_{adj}$ have been computed to differentiate the estimated cumulative produced oil from the simulation model and the forecasted through the proxy model. Figures 16, 18, 20, and 22 illuminate the matching of the estimated cumulative produced oil from the observed, which is from the simulation model, and the forecasted from the proxy models (Predicted) concerning the test subset jobs for the QM, MARS, FUzzy-GENetic, and GBM, respectively. Figures 17, 19, 21, and 23 depict the matching of the estimated cumulative produced oil from the simulation model (Observed) along with the forecasted based on the proxy models (Predicted) concerning the full data set jobs for the QM, MARS, FUzzy-GENetic, and GBM, respectively. In the (Left) figures, the blue

![Figure 18](image1.png)

**Figure 18** Comparison of proxy-predicted and simulator calculated-fuzzy-genetic given the test subset

![Figure 19](image2.png)

**Figure 19** Comparison of proxy-predicted and simulator calculated-fuzzy-genetic given the full data set
balls refer to the observed response values, cumulative oil production estimated from the compositional reservoir simulator. However, the red balls represent the predicted cumulative oil production from the four proxy models. In addition, the (Right) figures represent the scatter plots of matching of simulation and proxy models cumulative produced oil (Figures 16–23).

The comparison between the four proxy metamodels: QM, FUzzy-GEnetic, GBM, and MARS was justified based on the RMSE and $R_{adj}^2$ with respect to the test subset and full datasets, as illustrated in Table 4.

The GBM proxy model was much better than the QM, FUzzy-GEnetic, and MARS models as GBM had the least RMSE and highest $R_{adj}^2$. Moreover, the scatter matching of the simulation along with proxy models cumulative produced oil from the Gradient Boosting Regression model was more matched than QM, FUzzy-GEnetic, and
MARS models as most of the points in GBM fit the 45° degree line.

8 | APPLICABLE DOMAIN OF GBM MODEL

To conclude, a statistical analysis of outlier identification was carried out to assess whether the GBM model and the data set upon which it was built were statistically valid. In this study, outliers were identified using the leverage strategy, which was then shown using a Williams plot because of its high accuracy in detecting outliers. The standardized residual (R) values resulting from model predictions are shown on a graph in relation to hat (H) values, which reflect the diagonal elements of the hat matrix and are characterized as follows:

$$H = X(X'X)^{-1}X',$$  

(7)
where $X$ represents a matrix with a size equal to $n \times D$, $n$ stands for the number of data points, $D$ represents the number of parameters, and $X^t$ denotes the transpose matrix of $X$. The applicability domain represents a leverage limit value ($H^*$) calculated as $3(D + 1)/n$ in the Williams plot. Importantly, eligible data points must have $R$ values in the range of $-3$ to $3$. Consequently, acceptable data points are placed within the following ranges: $0 \leq H \leq H^*$ and $-3 \leq R \leq 3$. A high proportion of valid data points implies a model that is extremely consistent in its predictions.

Figure 24 depicts the respective Williams plots of cumulative oil production by the GBM model. As can be observed, 96.11% of cumulative oil production of data points fell within the following valid range: $0 \leq H \leq 0.032$ and $-3 \leq R \leq 3$. These findings provide further evidence of the statistical validity of the GBM model and the data set employed in this investigation.

### SUMMARY AND CONCLUSIONS

The summary and conclusions of this study were outlined as follows:

1. The compositional reservoir simulator was conducted for the GAGD process evaluation within a nonhomogeneous main pay reservoir in the South Rumaila field.
2. After achieving the history matching, 10 years of future production was set as a prediction period to acquire the optimum recovery ratio through changing the operational decision parameters, which govern the injection and production processes.
3. The Latin Hypercube Sampling was implemented based on a low-discrepancy Experimental Design technique to generate several hundreds of trials and experiments to be assessed by the simulation model to identify the optimum recovery ratio and to construct the proxy model. That DoE and proxy modeling approach include determining an ideal group of operational decision parameters through the immiscible GAGD technique.
4. The parameters are CO$_2$ injection rate and highest BHP in the injectors, along with the highest oil production rate, lowest BHP, and maximum water cut in the horizontal producers. This optimal case led to obtaining 4.6039 million STB with an increment of 212.5 million STB of oil over the

| Algorithm      | Test Subset $R^2_{adj}$ | RMSE     | Full Data set $R^2_{adj}$ | RMSE     |
|----------------|-------------------------|----------|---------------------------|----------|
| QM             | 0.9417                  | 16.476E+6| 0.9346                    | 18.254E+6|
| FUzzy-GEnetic  | 0.9167                  | 22.466E+6| 0.9171                    | 20.557E+6|
| GBM            | 0.9955                  | 3.999E+6 | 0.9973                    | 3.704E+6 |
| MARS           | 0.9441                  | 14.186E+6| 0.9528                    | 15.509E+6|

FIGURE 24 William’s plot is demonstrated for applicability domain of the GBM model
5. The first proxy modeling workflow includes generating simulation jobs as training runs to build the proxy model, which was iteratively validated through four sets of validation tests (verification runs).

6. To create an accurate proxy model that truly models the compositional reservoir simulator (metamodel), the polynomial regression in addition to three more approaches were adopted to build proxy models. The four constructed models are Polynomial Regression, Fuzzy Logic-Genetic Algorithm (Fuzzy GEnetic), Generalized Boosted Model (GBM), and Multivariate Additive Regression Splines (MARS). The accuracy comparison between the four proxy models was conducted concerning the $R^2_{adj}$ and RMSE for the prediction of test subsets and full data sets.

7. It was observed that the GBM model was the most accurate metamodel for the GAGD process as it attains the minimum RMSE and the maximum adjusted $R^2_{adj}$ that both reflect the lowest mismatch of the cumulative produced oil estimated through the simulation model and forecasted by the GBM-proxy models.

8. In addition, the MARS proxy model was the second-best matching model, followed by the polynomial and FUzzy-GEnetic proxy models. Additionally, the cumulative produced oil of both the simulation and proxy models from the GBM has better scatter points matching than the MARS, QM, and FUzzy-GEnetic.

9. Consequently, the GBM can be implemented as a simplified substitute metamodel instead of the high-resolution compositional reservoir model by the GAGD technique assessment and forecast.

10. Also, the Williams plots of estimated cumulative oil production corroborated the valid results of the database and the findings provided by the GBM model, with below than 3.88% of data points highlighted as probable outliers, as shown by the results.

11. In the end, one of the most valuable outcomes of this study is the implementation of an effective GBM-based proxy model for forecasting and evaluating the viability of CO$_2$-EOR initiatives. This robust proxy model may be utilized as a reliable template for predicting the EOR performance in mature fields in real-world environments due to the extensive use of actual field data in its field development. Machine learning’s modeling approach may be simply adapted to different oil fields with similar reservoir characteristics.

**NOMENCLATURE**

- $X_i$ the input variables in the regression model
- $y$ the response or output variable
- $\varepsilon_i$ the regression error
- $f_i(x_i)$ the predicted response variable $y$
- $f(x_i)$ the measured response variable $y$
- $R^2$ coefficient of correlation
- $R^2_{adj}$ adjusted coefficient of correlation
- $n$ total number of measurements
- $k$ total number of predictors in the regression model
- $\alpha_0$ the intercept in the regression model
- $\alpha_j$ the linear coefficient
- $\alpha_{jj}$ the coefficient of the quadratic term
- $\alpha_{ij}$ the interaction coefficient of every two factors
- $\beta$ vector of coefficients of the irregular fundamental functions
- $v(j, i)$ the index of independent parameters utilized in the $i$th term of the $j$th product
- $b_{ji}(X_{v(j,i)})$ the truncated power fundamental function
- $K_i$ a parameter that limits the order of interactions
- $b_{ji}$ the spline function
- $h$ the regression tree
- $-g(x_i)$ the negative gradients
- $X$ the matrix with size equal to $n \times D$
- $n$ the number of data points
- $D$ the number of parameters
- $X^t$ the transpose matrix of $X$

**ACKNOWLEDGMENTS**

The authors thank the Computer Modeling Group for the CMG software support.

**ORCID**

Watheq J. Al-Mudhafar  http://orcid.org/0000-0002-5327-8300

Hung Vo Thanh  http://orcid.org/0000-0002-7094-9380

**REFERENCES**

1. Rao DN. Gas-assisted gravity drainage process for improved oil recovery. United States patent 8,215,392 B2 (2012).
2. White CD, Royer SA. Experimental design as a framework for reservoir studies. SPE Reservoir Simulation Symposium, Houston, TX; 2003. doi:10.2118/79676-MS
3. Lee TH, Lee CS, Jung J, Kim H-W, Hong S, Choi J. Prediction of the motion of tracked vehicle on soft soil using kriging metamodel. Fifth ISOPE Ocean Mining Symposium, Tsukuba, Japan; 2003.
4. Badru O, Kabir CS. Well placement optimization in field development. SPE Annual Technical Conference and
15. Zubarev DI. Pros and cons of applying a proxy model as a substitute for full reservoir simulations. SPE Annual Technical Conference and Exhibition, New Orleans, LA; 2009. doi:10.2118/124815-MS

16. Goodwin N. Bridging the gap between deterministic and probabilistic uncertainty quantification using advanced proxy based methods. SPE Reservoir Simulation Symposium, Houston, TX, 23-25 February 2015. doi:10.2118/173301-MS

17. He Q, Mohaghegh SD, Liu Z. Reservoir simulation using smart proxy in SACROC unit-case study. SPE Eastern Regional Meeting, Canton, OH, 13-15 September, 2016. doi:10.2118/184069-MS

18. Yeten B, Castellini A, Guyaguler B, Chen WH. A comparison study on experimental design and response surface methodologies. SPE Reservoir Simulation Symposium, The Woodlands, 31 January-2 February, 2005. doi:10.2118/93347-MS

19. Jaber AK, Awang MB, Stephen KD. Proxy modelling for rapid optimisation of miscible CO2-WAG injection in heterogeneous clastic reservoirs: a case study from Southern Iraqi oil fields. Int J Oil Gas Coal Technol. 2021;26(3):263-280. doi:10.1504/IJOGCT.2021.113134

20. Avansi GD. Use of proxy models in the selection of production strategy and economic evaluation of petroleum fields. SPE Annual Technical Conference and Exhibition, New Orleans, 4-7 October, 2009. doi:10.2118/129512-STU

21. Hassani H, Sarkheil H, Foroud T, Karimpooli S. A Proxy Modeling Approach to Optimization Horizontal Well Placement. American Rock Mechanics Association; 2011.

22. Osterloh WT. Use of multiple-response optimization to assist reservoir simulation probabilistic forecasting and history matching. SPE Annual Technical Conference and Exhibition, Denver, CO, 21-24 September, 2008. doi:10.2118/116196-MS

23. Vo Thanh H, Sugai Y, Nguete R, Sasaki K. Robust optimization of CO2 sequestration through a water alternating gas process under geological uncertainties in Coo Long Basin, Vietnam. J Nat Gas Sci Eng. 2020;76:1-15. doi:10.1016/j.jngse.2020.103208

24. Rao DN, Ayirala SC, Kulkarni MM. Development of gas-assisted gravity drainage (GAGD) process for improved light oil recovery. SPE/DOE Symposium on Improved Oil Recovery, Tulsa, OK, 17-21 April, 2004. doi:10.2118/89357-MS

25. Al-Obaidi DA, Al-Mudhafar WJ, Al-Jawad MS. Experimental evaluation of carbon dioxide-assisted gravity drainage process (CO2-AGD) to improve oil recovery in reservoirs with strong water drive. Fuel, 2022;324:1-15. doi:10.1016/j.fuel.2022.124409

26. Al-Ansari R. The petroleum geology of the upper sandstone member of the Zubair formation in the Rumaila South. Geological Study. Ministry of Oil; 1993.

27. Al-Obaidi RY. Identification of palynozones and age evaluation of Zubair formation, Southern Iraq. J Al-Nahrain Univ. 2009;12(3):16-22.

28. Mohammed WJ, Al Jawad MS, Al-Shamaa DA. Reservoir flow simulation study for a sector in main pay South Rumaila oil field. SPE Oil and Gas India Conference and Exhibition. Mumbai, India; 2010. doi:10.2118/126427-MS

29. Al-Mudhafar WJ, Al Jawad MS, Al-Shamaa D. A. Using optimization techniques for determining optimal locations of additional oil wells in South Rumaila oil field. CPS/SPE International Oil and Gas Conference and Exhibition. Beijing, China; 2010. doi:10.2118/130054-MS

30. Kabir CS, Mohammed NI, Choudhary MK. Lessons learned from energy models: Iraq’s South Rumaila case study. SPE Middle East Oil and Gas Show and Conference. Manama, Bahrain; 2007. doi:10.2118/105311-MS

31. Al-Mudhafar W. Multiple-point geostatistical lithofacies simulation of fluvial sand-rich depositional environment: a case study from Zubair formation/South Rumaila oil field. Offshore Technology Conference, Houston, TX, USA, 2-5 May 2016. doi:10.4043/27273-MS
32. Al-Mudhafar W. How is multiple-point geostatistics of lithofacies modeling assisting for fast history matching? A case study from a sand-rich fluvial depositional environment of Zubair formation in South Rumaila oil field. OTC-28662-MS, the Offshore Technology Conference. Houston, TX, USA 2000. doi:10.4043/28662-MS

33. Hagoort J. Oil recovery by gravity drainage. SPE J. 1980;20(03):139-150. doi:10.2118/7424-PA

34. CMG. The Computer Modeling Group’s Reservoir Simulation Manual; 2015.

35. Thomas, LK, Hellums LJ, Reheis GM. A nonlinear automatic history matching technique for reservoir simulation models. SPE J. 12:508-514. doi:10.2118/3475-PA

36. Lazic ZR. Design of Experiments in Chemical Engineering. Wiley-Vch; 2006.

37. Amudo C, Graf T, Dendekar RR, Randle JM. The pains and gains of experimental design and response surface applications in reservoir simulation studies. SPE Reservoir Simulation Symposium, Houston, TX, 2-4 February 2009. doi:10.2118/118709-MS

38. McKay MD, Beckman RJ, Conover WJ. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics. 1979;21(2): 239-245. doi:10.2307/1268522

39. Bhat CR. Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. Transp Res B Methodol. 2001;35(7):677-693. doi:10.1016/S0167-9473(00)00014-X

40. Stocki R. A method to improve design reliability using optimal latin hypercube sampling. Comput Assist Mech Eng Sci. 2005;12:87-105.

41. Stein M. Large sample properties of simulations using latin hypercube sampling. Technometrics. 1987;29:143-151.

42. Carnell R. Package ‘lhs’. 2015. Accessed March 22, 2022. https://cran.r-project.org/web/packages/lhs/lhs.pdf

43. Al-Mudhafar W. Incorporation of bootstrapping and cross-validation for efficient multivariate facies and petrophysical modeling. SPE Low Perm Symposium, Denver, CO, 5-6 May 2016. doi:10.2118/180277-MS

44. Lenth RV. Response-surface methods in R, using RSM. J Stat Softw. 2009;32(7):1-17. doi:10.18637/jss.v032.i07

45. Friedman JH. Multivariate adaptive regression splines. Ann Stat. 1991;19:1-67.

46. Kooperberg C. Multivariate adaptive regression splines. In: El-Shaarawi AH, Piegsch RW, eds. Encyclopedia of Environmetrics. Vol 4. John Wiley & Sons, Ltd. 2006.

47. Adoko AC, Jiao YY. Predicting tunnel convergence in weak rocks: multivariate adaptive regression spline and rough set theory approaches. ISRM International Symposium-8th Asian Rock Mechanics Symposium, Sapporo, Japan, 14-16 October 2014.

48. Samui P. Multivariate adaptive regression spline (Mars) for prediction of elastic modulus of jointed rock mass. Geotech Geol Eng. 2013;31:249. doi:10.1007/s10706-012-9584-4

49. Milborrow, S. Package ‘earth’. 2016. Accessed July 20, 2021. https://cran.r-project.org/web/packages/earth/earth.pdf

50. Al-Mudhafer WJM, Alabbas M. Application of a hybrid system of genetic algorithm & fuzzy logic as optimization techniques for improving oil recovery in a sandstone reservoir in Iraq. SPE-149982-MS, presented at the SPE Latin America and Caribbean Petroleum Engineering Conference, Mexico City, Mexico, 16-18 April 2012. doi:10.2118/149982-MS

51. Hinterding R, Michalewicz Z, Eiben AE. Adaptation in evolutionary computation: a survey. IEEE/CEP: Proceeding of the IEEE Conference on EC, IEEE World Congress on Computational Intelligence; 1997. doi:10.1109/ICEC.1997.592270

52. Goldberg DE. Genetic Algorithm in Search, Optimization and Machine Learning. Addison-Wesley; 1989.

53. Bujard A. Package ‘fugeR’: FUzzy GENetic, a machine learning algorithm to construct prediction model based on fuzzy logic. Accessed August 8, 2012. https://cran.r-project.org/src/contrib/Archive/fugeR/fugeR_0.1.2.tar.gz

54. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001;29(5):1189-1232. http://www.jstor.org/stable/2699886

55. Friedman JH. Stochastic gradient boosting. Comput Stat Data Anal. 2002;38(4):367-378. doi:10.1016/S0167-9473(01)00065-2

56. Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Comput Neurosci. 2013;7:21. doi:10.3389/fncom.2013.00021

57. Ridgeway G. Generalized boosted models: a guide to the ‘gbm’ package. 2007. Accessed August 3, 2007. http://www.saedsayad.com/docs/gbm2.pdf

58. Rousseeuw PJ, Leroy AM. Robust Regression and Outlier Detection. Vol. 589. John Wiley & Sons; 2005.

59. Gramatica P. Principles of QSAR models validation: internal and external. QSAR Comb Sci. 2007;26:694-701. doi:10.1002/qsar.200610151

60. KISI ö. Generalized regression neural networks for evapotranspiration modelling. Hydrolog Sci J. 2006;51:1092-105.

61. Leroy AM, Rousseeuw PJ. Robust Regression and Outlier Detection, J. Wiley & Sons, New York; 1987.

How to cite this article: Al-Mudhafar WJ, Rao DN, Srinivasan S, Vo Thanh H, Al Lawe EM. Rapid evaluation and optimization of carbon dioxide-enhanced oil recovery using reduced-physics proxy models. Energy Sci Eng. 2022;10:4112-4135. doi:10.1002/ese3.1276