Application of a new approach for modeling the oil field formation damage due to mineral scaling
Alireza Rostami, Amin Shokrollahi, Khalil Shahbazi, Mohammad Hossein Ghazanfari

To cite this version:
Alireza Rostami, Amin Shokrollahi, Khalil Shahbazi, Mohammad Hossein Ghazanfari. Application of a new approach for modeling the oil field formation damage due to mineral scaling. Oil & Gas Science and Technology - Revue d’IFP Energies nouvelles, Institut Français du Pétrole, 2019, 74, pp.62. 10.2516/ogst/2019032. hal-02171462

HAL Id: hal-02171462
https://hal.archives-ouvertes.fr/hal-02171462
Submitted on 2 Jul 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Application of a new approach for modeling the oil field formation damage due to mineral scaling

Alireza Rostami1,*, Amin Shokrollahi2, Khalil Shahbazi1,*, and Mohammad Hossein Ghazanfari2

1 Department of Petroleum Engineering, Petroleum University of Technology (PUT), P.O. Box 6198144471, Ahwaz, Iran
2 Department of Chemical and Petroleum Engineering, Sharif University of Technology (SUT), P.O. Box 113659465, Tehran, Iran

Received: 2 March 2019 / Accepted: 14 May 2019

Abstract. Mineral scaling has been considered a great concern for developing the oil production from the underground petroleum reservoirs. One of the main causes of this phenomenon is known as the chemical incompatibility of injected brine, frequently sea water, with the reservoir brine leading to the deposition of various supersaturated salts such as calcium carbonate, calcium sulfate and barium sulfate. In present communication, an evolutionary approach namely, Gene Expression Programming (GEP), was employed for rigorous modeling of formation damage by mineral scaling of mixed sulfate salt deposition. At first, a large databank of damaged permeability datapoints as a function of injected volume, injection flowrate, temperature, differential pressure and ionic concentrations of the existing chemical species in the porous media was employed. In this regard, a user-friendly correlation was extended for the first time by the aforementioned technique in the literature. Professional evaluation of the suggested GEP-based model was implemented by different statistical parameters and appealing visualization tools. Having proposed the GEP-based correlation, statistical parameters of the Average Absolute Relative Deviation Percent (AARD%) of 0.640% and determination coefficient ($R^2$) of 0.984 was calculated. Accordingly, it is demonstrated that the proposed model has a superior performance and great potential for efficient prediction of damaged permeability due to the mixed sulfate salt scaling. Moreover, the implemented outlier diagnosis technique verified the validity of the databank used for modeling, as well as the high robustness of the suggested model was confirmed. In conclusion, the developed correlation in this work can be of enormous practical value for skillful engineers and scientists in any academic study and industrial applications dealing with mixed salt deposition.

1 Introduction

In water injection scheme, deposition of mineral scales has been considered a great challenge for production development of subterranean petroleum reservoirs [1]. Success of a water injection project can be jeopardized by such scale, and even worse it can be terminated in the worst operational conditions. Therefore, the efficiency of this process, depending upon the degree of pressure maintenance and the oil production level, can be reduced. The amount of formation damage caused by the scaling is chiefly quantified by the well-known terms of permeability and porosity reduction [2, 3].

The mechanism of permeability reduction leading to the formation damage in the porous media by precipitated minerals is the deposition of the supersaturated minerals on the pore walls due to the attraction forces between the pore surface area and scale solid particles. In such condition, a number of bridges made of the scale particles across the pore throats will be created, as well as pore throats blockage by a single particle will strongly happen. The degree of formation damage is affected by the features of minerals precipitate. The morphology and quantity of the growing crystals on the pore walls are monitored by means of several factors including existence of impurities, mixing rate, temperature change and supersaturation owing to the variation in physical conditions (e.g., chemical incompatibility) [4–7].

Chemical incompatibility is classified as the main reason of the severe oilfield scaling when incompatible waters are mixed. While water is injected into the reservoir, a reaction between the formation minerals and formation water with the injected fluid will take place resulting in the establishment of scale formation through the porous media. In other words, based on the so-called phenomenon of the chemical incompatibility, the reservoir brine and injection water will undergo the strong chemical interactions together leading

* Corresponding authors: alireza.rostami.put2014@gmail.com; alireza.rostami@afp.put.ac.ir; shahbazi@put.ac.ir

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
to the precipitation of the supersaturated minerals. Precipitation of inorganic scale is dependent upon the ionic composition of fluid and the nature of scale due to their influence on the molecular or atomic interaction amongst the involved particles of commingling brines [8–11].

Precipitation instigates as long as the nucleation process have turned on. Nucleation is defined as the joining of atoms attracted for establishing submicron nuclei. The presence of impurities in fluid makes lower energy needed for construction of such nuclei than the required energy for a pure fluid. Because the impure components perform as proper sites for nucleation, in which it is termed as heterogeneous nucleation. The process of heterogeneous nucleation extensively takes place where suspended inorganic particles exist in the formation water coming into contact with the injection water [4, 12–16].

Generally, sea water is used as the injected fluid in both secondary and tertiary oil recovery processes, which is rich in sulfate and bicarbonate anionic chemical species. Despite sea water as the injection fluid, high contents of cationic species like calcium, barium and strontium exist in the formation water which can form diverse types of sulfate scales (i.e., CaSO4, BaSO4, SrSO4) and carbonate scales (i.e., CaCO3) in the oilfield. Hence sulfate and carbonate scales are recognized as the two key classes of scaling minerals in the oilfields [17, 18]. Notwithstanding carbonate scaling minerals in which they are extremely reliant on pressure alterations and pH changes, sulfate scales commonly happen on account of incompatible waters mixing and temperature ups and downs [19–21]. In addition to the sea water, disposal water, which is produced in association with oilfield production, can also be recycled and re-injected into the reservoir, even though the problem of chemical incompatibility is still present in this case. It is noteworthy that injection of compatible water resources with the connate and formation water is something impossible to implement [22, 23].

Relatively good solubility of carbonate scaling minerals makes the easier inhibition of them by conventional techniques such as soaking with suitable acid dissolver than the sulfate scales. Thus, the deposition of such scales downstream of wellhead equipment can be resolved through nonstop injection of an appropriate inhibitor into the transportation lines. Nevertheless, sulfate scales (e.g., barium sulfate) have very poor solubility, low tendency for reacting with most acids, great hardness and exposures of very little surface area at the time of deposition. In consequence, restricted number of removal methods is existing in order to preclude their deposition in sensitive regions. For this reason, squeeze treatments of scale inhibitor are usually applied so that sulfate mineral scaling will be inhibited in downstream or upstream of first completion and any mixing location [18, 20, 24, 25].

Aiming for evaluating the impact of different parameters on scale deposition such as incompatibility of mixing fluids, temperature, pressure, concentration of chemical species present in commingling fluids, innumerable investigations have been implemented [17, 18, 26, 27]. The impact of incompatible waters on the formation of carbonate and calcium sulfate scales in the synthetic porous media was extensively studied by Moghadasi et al. [2]. In continuum study, more experimental and theoretical investigations were conducted on the degree of permeability reduction as a measure of formation damage caused by calcium carbonate and calcium sulfate scaling through mixing carbonate/sulfate rich and calcium rich solutions by Moghadasi et al. [28].

Along with the experimental studies, a number of theoretical studies have been focused on the prediction of mineral scale deposition in porous media [29–32]. In view of the hydrodynamic and kinetics of gypsum deposition, Jamialahmadi et al. [33] have instituted a model with outstanding performance for mathematical modeling of deposition and removal of gypsum scaling integrating the impact of salt supersaturation, injection flowrate and temperature. This model may work ineffectively in order to predict the value of permeability reduction when mixed salt precipitation happens in solution. In more recent years, Safari and Jamialahmadi [19] initiated a highly nonlinear simulator on the basis of hydrodynamic, kinetics and thermodynamic laws for modeling deposition of both single salt (i.e., barium sulfate), and mixed salt (i.e., strontium sulfate in connection with barium sulfate) during fluid flow through porous media. Then, the authors optimized the kinetic coefficients by means of a hybrid approach namely, Pattern Search (PS) algorithm in cooperation with Particle Swarm Optimization (PSO) technique. Finally, they concluded that their model has an acceptable agreement with the experimental data with deviations less than 10% [19]. Even though the authors carried a great job out, their model is highly complex with lots of coefficients to be tuned. So, there is a vital requisite for constructing a universal model in order to have a rapid and precise estimation of permeability impairment for mixed sulfate salts scaling during water flooding process in the porous media.

Genetic based calculations have been commonly used in petroleum industry as a promising approach in estimating several parameters [34–36]. In recent years, Gene Expression Programming (GEP) [37] as an evolutionary algorithm, has been increasingly applied in different disciplines of petroleum and chemical engineering. The application of GEP [37] algorithm resulted in development of accurate models in wide varieties of petroleum industry which generally gives more precise estimates than the pre-existing models. In accordance with GEP [37] mathematical strategy, the optimum correlation format will be initiated without any assumption about the form of equation. Successful examples of GEP scheme applied in the literature can be found in the work of several researchers in the open literature [38, 39].

In current study, potential application of GEP algorithm as a powerful technique is presented so as to prepare a possible solution to the all disadvantages earlier mentioned in the field of mixed sulfate salt scaling in the porous media. For this reason, a large database was adopted from the open literature [40–43]. Afterwards, the database is divided into two sets of training (about 341 datapoints) and testing (about 85 datapoints). According to the
training dataset, the GEP-based empirically derived equation is developed. Throughout various statistical parameters and visualization tools the performance of the proposed model is exhibited. To the best of authors knowledge, there is no report on modeling permeability impairment as a representative of formation damage caused by mixed sulfate salt scaling in the open literature. To this end, the validity of the databank used for modeling is assessed by means of outlier analysis.

### 2 Data gathering

Based on the previous modeling studies in the field of soft computation, it has been demonstrated that development of a comprehensive model is crucially in the need of a large database. A database with the feature of all-inclusiveness makes the constructed model to be applied for a wider ranges of operational conditions. In other words, the proposed model is not limited to a specific condition and can be employed for mathematical description of the interest phenomenon at different conditions [35, 44–62]. In present research, 431 datapoints of permeability reduction values as a function of temperature, differential pressure, volume of injected water, initial permeability, flowrate, and ionic concentrations of cationic species (i.e., strontium, calcium and barium) and anionic species (i.e., sulfate) is taken from the open literature (see Supplementary Material) [40–43]. This database is used for developing and testing the capability of the proposed model. For model development and examining its capability, nearly 80% and 20% of the entire database are employed, respectively. This data division is carried out by a random computational process defined in GEP modeling. Table 1 shows the specification of the employed database for GEP modeling.

### 3 Gene expression programming

In an attempt to develop the genetic-based calculation, the most recent version of genetic computational models, namely GEP, in which the shortcomings of the preceding genetic models like Genetic Algorithm (GA) and Genetic Programming (GP) were modified in GEP computation strategy [63]. Unlike GP approach working with one element of Expression Trees (ETs), GEP scheme deals with two components, including ETs and chromosomes. Symbolic ETs are defined as the population individuals, and the chromosomes are responsible for encoding and translating the candidate solution into a real candidate solution [64]. In this regard, a typical chromosome is categorized into functions and variable/constant terminals. The constants are determined by the model program, however, the variables and functions are set as the inputs of the model. For each gene, the inputs and terminals are corresponded to, respectively, gene’s head and gene’s tail which are related as follows [65]:

\[
t = h(n - 1) + 1
\]

where, the symbols \( n \), \( h \) and \( t \) denote the largest function arity, the magnitude of gene’s head and the length of gene’s tail, respectively. Setting parameters of the used GEP strategy for modeling damaged permeability in this study are reported in Table 2. The similar translation procedure is observed in biological genes encoded in DNAs which are constantly transformed into proteins. Owing to the structural features of the chromosomes and reproduction processes accomplished to this technique, unlimited modifications of programs are obtained leading to effective solution to the problem [65]. It is confirmed that the convergence speed of the GEP mathematical strategy is two to four orders of magnitude larger than that of the

### Table 1. Statistical specifications of the database utilized for developing the correlation.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Unit</th>
<th>Minimum</th>
<th>Average</th>
<th>Maximum</th>
<th>SD*</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \Delta P )</td>
<td>psi</td>
<td>100.00</td>
<td>150.12</td>
<td>200.00</td>
<td>40.85</td>
</tr>
<tr>
<td>( T )</td>
<td>°C</td>
<td>50.00</td>
<td>66.71</td>
<td>80.00</td>
<td>12.48</td>
</tr>
<tr>
<td>( Q )</td>
<td>cc/min</td>
<td>8.55</td>
<td>17.78</td>
<td>31.33</td>
<td>5.37</td>
</tr>
<tr>
<td>( K_i )</td>
<td>md</td>
<td>12.30</td>
<td>12.99</td>
<td>13.87</td>
<td>0.52</td>
</tr>
<tr>
<td>( V_{inj} )</td>
<td>PV</td>
<td>1.47</td>
<td>23.72</td>
<td>83.80</td>
<td>15.55</td>
</tr>
<tr>
<td>( C_{Ca^{2+}} )</td>
<td>ppm</td>
<td>780</td>
<td>9592.76</td>
<td>30000</td>
<td>11999.84</td>
</tr>
<tr>
<td>( C_{Ba^{2+}} )</td>
<td>ppm</td>
<td>10</td>
<td>618.91</td>
<td>2200</td>
<td>919.49</td>
</tr>
<tr>
<td>( C_{Sr^{2+}} )</td>
<td>ppm</td>
<td>370</td>
<td>583.81</td>
<td>1100</td>
<td>301.36</td>
</tr>
<tr>
<td>( C_{SO_4^{2-}} )</td>
<td>ppm</td>
<td>2750</td>
<td>2854.76</td>
<td>2960</td>
<td>105</td>
</tr>
<tr>
<td>( K_d )</td>
<td>md</td>
<td>9.81</td>
<td>12.05</td>
<td>13.81</td>
<td>0.77</td>
</tr>
</tbody>
</table>

*SD* refers to the standard deviation which is calculated as follows:

\[
SD = \left( \frac{1}{N-1} \sum_{i=1}^{N} (S_i - \bar{S})^2 \right)^{\frac{1}{2}}.
\]
Table 2. Setting parameters of the used GEP strategy for modeling damaged permeability in this study.

<table>
<thead>
<tr>
<th>GEP algorithm parameters</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of chromosomes</td>
<td>30</td>
</tr>
<tr>
<td>No. of genes</td>
<td>3</td>
</tr>
<tr>
<td>Head size</td>
<td>7</td>
</tr>
<tr>
<td>Linking function</td>
<td>+</td>
</tr>
<tr>
<td>Generations without change</td>
<td>2000</td>
</tr>
<tr>
<td>Fitness function</td>
<td>Root Mean Square Error</td>
</tr>
<tr>
<td>Inversion</td>
<td>0.00546</td>
</tr>
<tr>
<td>Mutation</td>
<td>0.00138</td>
</tr>
<tr>
<td>IS transposition</td>
<td>0.00546</td>
</tr>
<tr>
<td>RIS transposition</td>
<td>0.00546</td>
</tr>
<tr>
<td>One-point recombination</td>
<td>0.00277</td>
</tr>
<tr>
<td>Two-point recombination</td>
<td>0.00277</td>
</tr>
<tr>
<td>Gene transposition</td>
<td>0.00277</td>
</tr>
<tr>
<td>Gene recombination</td>
<td>0.00277</td>
</tr>
<tr>
<td>Permutation</td>
<td>0.00546</td>
</tr>
<tr>
<td>Constants per gene</td>
<td>10</td>
</tr>
<tr>
<td>Random chromosomes</td>
<td>0.0026</td>
</tr>
<tr>
<td>Type of data</td>
<td>Floating point</td>
</tr>
<tr>
<td>Random cloning</td>
<td>0.00102</td>
</tr>
<tr>
<td>Operators used</td>
<td>+, −, ×, /, √, EXP, INV, LN, LOG, X², POW</td>
</tr>
</tbody>
</table>

Developing the correlation

Based on the existing literature concentrated on the mineral scale formation in porous media, it is fully understood that the amount of permeability reduction as a measure of formation damage is under the influence of several independent variables. These variables include ionic concentrations of sulfate anion and divalent cations (i.e., calcium, strontium, barium), differential pressure, temperature, injected volume and flowrate [19, 67–70]. Therefore, the proposed GEP-based model is extended, as follows:

\[
K_d = f\left( \Delta P, T, Q, V_{inj}, K_i, C_{Ca^{2+}}, C_{Ba^{2+}}, C_{Sr^{2+}}, C_{SO_4^{2-}} \right),
\]

\[\text{(2)}\]

where, the symbols \(K_d\), \(K_i\), \(\Delta P\), \(T\), \(Q\), \(V_{inj}\), \(C_{Ca^{2+}}\), \(C_{Sr^{2+}}\), \(C_{Ba^{2+}}\) and \(C_{SO_4^{2-}}\) indicate the damaged permeability, initial permeability, differential pressure, temperature, flowrate, injected volume, concentration of calcium ion, concentration of strontium ion, concentration of barium ion and concentration of sulfate ion, respectively. When the decision variables are defined, the subsequent mathematical strategy will be applied to find the optimal equation format described as below:

1. **Population preparation.** Randomly selected individual chromosomal structures throughout checking various algebraic operators (e.g., ×, ÷, √, tanh) with its decoded ET and corresponding algebraic expression (correlation) with the Karva language illustration. The authors used a well-known and optimized programming code for GEP modeling to simulate the interested parameter in this study.

2. **Predicting fitness value.** For each individual, the Objective Function (OF) is predicted by the subsequent formulation:

\[
OF(i) = \frac{100}{N} \sum_{i=1}^{N} \frac{|K_{exp,i} - K_{pred,i}|}{K_{exp,i}}.
\]

In equation (3), \(N\) denotes the number of datapoints, and the superscripts “exp and pred” are, in turn, representatives for experimental and predicted values of permeability reduction [71].

3. **Individuals selection.** For replacement goals, the OF value gives an indication to select the appropriate individuals indicating suitable parents. For this reason, the so-called approach of tournament is utilized to prepare the adequate variety of dataset during each generation process [64, 72].

4. **Genetic operations.** Several operators including replication, mutation and inversion are applied for the goals of genes modification and reproduction. In replication stage, the chosen chromosomes used in step 3, are accurately duplicated [71]. Moreover, through
the application of the mutation operator, the effective adaptation of individuals’ population will be resulted by selecting randomly engaged nodes, and replacing saved information with the random primitive from the similar arity. The mutation will be applied via alteration in the magnitudes of gene’s head and gene’s tail. For this, the mutation operation can be occurred everywhere in chromosomal structure by introducing the known quantity of mutation rate \( p_m \). By dint of modifying the randomly chosen gene’s head, new individuals are developed in inversion process of GEP modeling. Having defined the well-known term of inversion rate \( p_i \), the efficiency of inversion operation can be evaluated simply [71].

5. Insertion and transposition sequence components. These transposable components can be shifted easily from one location to another one in a chromosomal structure [71]. Ferreira [37] introduced three types of such elements in his work, as follows: short fragments with first position function that move to the gene’s root (RIS components); short fragments with either first position terminal or first position function moving to the gene’s head; and whole genes that move to start of chromosomes.

6. Recombination. In this process, three different types of recombination, including one-point, two-point and gene, randomly select two chromosomes for exchanging certain materials together resulting in the extension of two new chromosomes. As a result, a new generation will be established. Considering a user-defined stopping condition, the aforementioned procedure will be repeated until the interested with satisfactory precision will be achieved. For more information, the interested readers are suggested to refer the extensive instances explained in the work of Ferreira [66].

4 Results and discussions

4.1 Benchmarks for evaluation of the proposed GEP model

In this section, a GEP-based empirically derived equation will be presented subsequently as a result of applying GEP mathematical strategy for the first time in this field of study. For better assessment of the proposed GEP-based correlation, various statistical quality measures are utilized including the Average Absolute Relative Deviation Percent (AARD\(\%\)), Root Mean Square Error (RMSE), determination coefficient \( R^2 \) and Average Relative Deviation Percent (ARD\(\%\)). One of the most important statistical criterion applied in a wide variety of mathematical and numerical modeling in chemical and petroleum engineering is calculation of the AARD\(\%\) value. The AARD\(\%\) which is defined as the degree of model precision, directly indicates the total magnitude of the estimation error relative to the target experimental data. The higher value of AARD\(\%\) shows the lower model accuracy. The quantity of RMSE illustrates the amount of inaccuracy happened in the process of modeling. In other words, it shows the magnitude of deviation between the real and simulated data. The other important and widely used statistical parameter is \( R^2 \) which shows the goodness of fit, or it displays how well the model estimates are matched with the actual data. When the value of \( R^2 \) approaches to unity, the most satisfactory agreement will be achieved. The final benchmark is the ARD\(\%\) value, which shows the quality of deviation distribution in the vicinity of zero deviation. The lower ARD\(\%\) value near to zero confirms the more compacted concentration of error distribution around the zero deviation. In addition to the parametric evaluation of the proposed model, diverse graphical illustrations are utilized to confirm the superiority and large capability of the GEP-based empirically derived correlation. The most significant of all applied diagrams are crossplot, index plot, and error distribution plot which will be represented subsequently in this study.

4.2 Assessment of the proposed GEP model

Based on the GEP processing, a user-friendly equation is developed which can be used for fast estimation of damaged permeability by this novel approach for the first time in the bulk of research have been paying attention to the formation damage caused by mineral scaling. The proposed empirically-derived GEP equation is given, as below:

\[
\log (K_d) = A_1 + A_2 + A_3, \tag{4}
\]

\[
A_1 = \frac{(T - V_{\text{inj}}) \times (Q - K_1)}{Q \times \Delta P \times \left\{ \log \left( C_{\text{inj}} + C_{\text{Sr}^{2+}} + C_{\text{Ca}^{2+}} + C_{\text{SO}_2^-} \right) - 2.77550702075275 \right\}^2}, \tag{5}
\]

\[
A_2 = \frac{40.792864950713 \times Q^2}{(\Delta P - 2.43561403505515)^2 \times \{0.0656999407983034 \times T - 1.46554080471316\}^2}, \tag{6}
\]

\[
A_3 = 0.0266417196782303 \times \ln \left( 302.487101618176 - \frac{Q}{V_{\text{inj}}} \right)^2, \tag{7}
\]

where, the units of \( K_d \), \( K_1 \), \( \Delta P \), \( T \), \( Q \), \( V_{\text{inj}} \), \( C_{\text{Ca}^{2+}} \), \( C_{\text{Sr}^{2+}} \), \( C_{\text{Ba}^{2+}} \), and \( C_{\text{SO}_2^-} \) in the above equations are md, md, psi, \( ^{\circ} \text{C} \), cc/min, PV, ppm, ppm, ppm, and ppm, respectively. The statistical details of the proposed GEP-derived correlation for the individual sets of training, test and total database are shown in Table 3. Based on this table, the values of \( R^2 \), ARD\(\%\), AARD\(\%\) and RMSE for the total database are 0.9843, 0.0355\%, 0.6409\% and 0.0967\%, respectively. It means that the AARD\(\%\) < 1\% and \( R^2 > 0.98 \) which show the satisfactory performance of the proposed correlation. Moreover, this table confirms the successful testing of the proposed model because of the better performance of the test set than the training set.

The results of GEP predictions/calculations against the experimental damaged permeability are indicated in
As can be seen in this crossplot, a compressed cloud of datapoints can be observed around the unit slope line or 45° line which shows the better agreement of the GEP model estimates in comparison with the corresponding measured data of damaged permeability. Because the most ideal performance of a model will be achieved when the predictions and actual data are the same leading to settling of the datapoints on unit slope line. For GEP model the values of AARD%, RMSE and ARD% are acceptably nearby the zero error value, and the $R^2$ value is close to unity. The other characteristic diagram for justified judgment of the GEP model is represented in Figure 3 exhibiting the distribution of GEP predictions error against the actual data of damaged permeability. In Figure 3, the relative deviation percent mainly changes in an acceptable range of −2 to 2%. Furthermore, the total value of ARD% which is equal to 0.0355%, demonstrates that the central focus of relative error distribution is sufficiently near to zero value. Figure 4 shows the schematic diagram for the error distribution of the proposed GEP-based equation versus the data.

Table 3. The statistical parameters of the developed GEP model for prediction of damaged permeability.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Training set</th>
<th>Test set</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>0.9841</td>
<td>0.9858</td>
<td>0.9843</td>
</tr>
<tr>
<td>Average relative deviation, %b</td>
<td>0.0629</td>
<td>−0.0746</td>
<td>0.0355</td>
</tr>
<tr>
<td>Average absolute relative deviation, %c</td>
<td>0.6565</td>
<td>0.5782</td>
<td>0.6409</td>
</tr>
<tr>
<td>Root mean square errord</td>
<td>0.0990</td>
<td>0.0872</td>
<td>0.0967</td>
</tr>
<tr>
<td>Number of data samples</td>
<td>345</td>
<td>86</td>
<td>431</td>
</tr>
</tbody>
</table>

$a$ Determination coefficient:

$$R^2 = 1 - \frac{\sum_{i=1}^{N} \left( (K_d)_{i}^{\exp} - (K_d)_{i}^{\text{pred}} \right)^2}{\sum_{i} \left( (K_d)_{i}^{\exp} - (K_d)_{i}^{\text{exp}} \right)^2}. $$

$b$ Average relative deviation percent (ARD):

$$\text{ARD}\% = \frac{100}{N} \sum_{i=1}^{N} \left( \frac{(K_d)_{i}^{\exp} - (K_d)_{i}^{\text{pred}}}{(K_d)_{i}^{\exp}} \right).$$

c Average absolute relative deviation percent (AARD):

$$\text{AARD}\% = \frac{100}{N} \sum_{i=1}^{N} \left( \frac{|(K_d)_{i}^{\exp} - (K_d)_{i}^{\text{pred}}|}{(K_d)_{i}^{\exp}} \right).$$

d Root mean square error (RMSE):

$$\text{RMSE} = \left( \frac{\sum_{i=1}^{N} \left( (K_d)_{i}^{\exp} - (K_d)_{i}^{\text{pred}} \right)^2}{N} \right)^{\frac{1}{2}}.$$
frequency. Having focused on this figure, a normal distribution can be easily observed for both training and test subsets indicating the symmetry in the outcomes from current paper.

4.3 Outlier diagnosis

It has been known that the reliability of a utilized database directly affects the accuracy and validity of the constructed model. Nonetheless, proper data measurement is often not practicable, and diverse types of undesirable experimental errors originated from human and equipment mistakes may diffuse into the measurements. Such unwanted deviations in experimental work are accounted as a menace to the success of modeling. Hence, detection of these suspected measurements from data is vital for any modeling study [67].

In an attempt to distinguish the invalid data, the so-called technique of Leverage Value Statistics (LVS) was conducted in current study. The statistical technique of LVS is commonly applied for describing the outlier data detecting the existing arrangement between the independent and dependent parameters. As it was previously explained, the suggested GEP model has a robust capability for predicting impaired permeability as a result of mineral scaling. In LVS processing, mathematical strategies contain the computation of Hat matrix and residual values for whole database. For calculating the residuals, the deviation between the target and the corresponding GEP predicted value have to be calculated for each datapoint. Besides, Hat matrix includes the GEP model estimates and the target experimental data, as below [73–76]:

\[ H = X(X'X)^{-1}X'. \] (8)

In equation (8), \( X \) is defined as the two-dimensional Hat matrix in which the possible area of the problem lay on the diagonal of this matrix, and the symbol \( t \) denotes the transpose operator. The total numbers of the model parameters and used data determine the number of the columns (\( n \)) and rows (\( m \)) of the Hat matrix, respectively. The well-known diagram of William is broadly applied to recognize the suspected data according to the calculated residuals and the values of Hat matrix estimated by the equation (8). The relationship between the standardized residuals (\( R \)) and \( H \) indices are shown in William’s plot. A cautionary leverage limit (\( H^* \)) is calculated as 3 \((n + 1)/m\). The criterion for measurement acceptability is

---

**Fig. 2.** Comparison between experimental damaged permeability and GEP predictions/calculations.

**Fig. 3.** Illustration of relative error distribution versus the damaged permeability.

**Fig. 4.** Distribution of relative deviation in the experimental dataset including train set, and test set.
the presence of data in the cut-off limit of residuals that is equal to ±3 (indicated by two horizontal lines in William’s plot). In Figure 5, the result of outlier analysis is represented in which the most of data exist in the range of \(-3 \leq R \leq 3\) and \(0 \leq H \leq H^*\) verifying the high truthfulness and robustness of the GEP-based model in this study. With respect to the \(H\) and \(R\) ranges, three classes of outliers can be suggested including Regression, Bad High Leverage and Good High Leverage outliers [73–76].

When the conditions \(-3 \leq R \leq 3\) and \(H^* \leq H\) are fulfilled, the outlier is called “Good High Leverage”. In this outlier, the measurements do not adequately affect the determination coefficient and accommodate nearly on the regression line which passes through the measured data, even though they have large values of leverage. The measurements with \(R\) values larger than 3 or less than –3 are under the class of “Bad High Leverage” outlier which is a serious intimidation for strong modeling. The intercept and the slope of the regression line are extremely affected by this outlier. The final type of outlier is named as “Regression” not disturbing the valid range and it has no impact on the regression line in spite of having large values of residual [73–76].

In this study, for recognizing the more likely invalid data, the Hat values were calculated via equation (8), then the William’s plot was drawn in Figure 5. Accordingly, it is clear that just 1 datapoint of 427 datapoints is recognized as the off-range (outlier) data. As a consequence, the suggested GEP-derived model in this study is reliable and efficiently precise owing to the fact that the datapoints are widely held in the interior region of \(-3 < R < 3\) and \(0 < H < 0.0493\).

5 Conclusion

In current study, GEP as an evolutionary mathematical strategy was applied in order to estimate the damaged permeability as a result of mixed salt scaling. For this purpose, an extensive database was taken from the open literature to develop a user-friendly and empirically-derived equation by this novel approach. The processed database was grouped into two subsets of training and test. The training group includes 80% of the entire database used for model development, as well as testing group includes 20% of the whole database applied for examining the model. Moreover, various schematic diagrams including crossplot, index plot and relative distribution plot were employed for better model evaluation. Calculation of different statistical quality measures for the proposed model results in the AARD% of 0.640%, \(R^2\) of 0.984, RMSE of 0.097 and the ARD% of 0.036%. Therefore, the GEP-based model developed in this study has proven to have an excellent precision and superior performance in estimating the impaired permeability in comparison with the experimentally measured datapoints. Finally, the proposed tool in this study is of tremendous practical value for quick and cheap prediction of impaired permeability in water flooding schemes during co-precipitation of mixed sulfate salts.

Supplementary Material

Supplementary Material is available at https://ogst.ifpenergiesnouvelles.fr/10.2516/ogst/2019032/olm.

References


40. Merdhah A. (2007) The study of scale formation in oil reservoir during water injection at high-barium and high-salinity formation water, in: Chemical and Natural Resources Engineering, Universiti Teknologi, Malaysia.


