Pattern Recognition for Steam Flooding Field Applications Based on Hierarchical Clustering and Principal Component Analysis

Na Zhang,* Mingzhen Wei,* Baojun Bai, Xiaopeng Wang, Jian Hao, and Shun Jia

ABSTRACT: Steam flooding is a complex process that has been considered as an effective enhanced oil recovery technique in both heavy oil and light oil reservoirs. Many studies have been conducted on different sets of steam flooding projects using the conventional data analysis methods, while the implementation of machine learning algorithms to find the hidden patterns is rarely found. In this study, a hierarchical clustering algorithm (HCA) coupled with principal component analysis is used to analyze the steam flooding projects worldwide. The goal of this research is to group similar steam flooding projects into the same cluster so that valuable operational design experiences and production performance from the analogue cases can be referenced for decision-making. Besides, hidden patterns embedded in steam flooding applications can be revealed based on data characteristics of each cluster for different reservoir/fluid conditions. In this research, principal component analysis is applied to project original data to a new feature space, which finds two principal components to represent the eight reservoir/fluid parameters (8D) but still retain about 90% of the variance. HCA is implemented with the optimized design of five clusters, Euclidean distance, and Ward’s linkage method. The results of the hierarchical clustering depict that each cluster detects a unique range of each property, and the analogue cases present that fields under similar reservoir/fluid conditions could share similar operational design and production performance.

INTRODUCTION

Steam flooding is the oldest and most successful commercial enhanced oil recovery (EOR) technique for oil reservoirs that have been used since the 1960s, and it is recognized as one of the most efficient oil recovery techniques for depleting the oil in various types of reservoirs because steam flooding results in higher ultimate oil recovery compared to other EOR techniques.1-4

The mechanisms of steam flooding EOR are intimately related with the thermal effects on the reservoir rock and fluid properties. Mechanisms that benefit the ultimate oil recovery include (1) increase of rock and fluid temperature from heat convection and conduction, (2) reduction in reservoir fluid (e.g., oil and water) viscosities, (3) increase in reservoir rock and fluid volumes that serve as a depletion drive energy, (4) vaporization of the light fraction of crude oils (often called distillation), (5) reduction of interfacial tensions and change in the relative permeability to oil and water, (6) gravity segregation, (7) solution gas drive, and (8) emulsion drive. These thermal effects are typically not applied uniformly to the whole reservoir, usually resulting in several temperature-fluid flow regions.5-7 When steam flooding is applied to reservoirs with different characteristics, the relative importance of the EOR mechanisms changes.8 It is evident that oil viscosity reduction is a dominating factor for heavy oil recovery, which significantly increases the oil mobility in the improved reservoir conditions;9-10 in contrast, for light oil reservoirs, thermal expansion and distillation are of greater importance than other EOR mechanisms.11

Steam flooding has been widely used for the production of heavy crude oil in shallow, thick sandstone formations.12,13 Most of the steam flooding projects have been implemented in sandstone formations because most of the EOR techniques have been tested at the pilot and commercial scales in this type of lithology.14,15 On the other hand, steam flooding is also one of the EOR techniques that can be applied to various reservoir and fluid conditions with improved operational techniques. There is an increasing number of steam flooding projects in carbonate reservoirs,16-19 light oil reservoirs,20,21 thin heavy oil reservoirs,22-24 and offshore developments.25-27 Before the implementation of steam flooding at a full field scale, a series of detailed preliminary studies, including laboratory tests, reservoir characterization, simulation, and pilot tests, are preformed to reduce the uncertainties and to minimize the risks.28 However, these evaluation studies are expensive and time-consuming. The reservoir/fluid properties change under different sets of steam flooding conditions.
different conditions, which brings the challenge of decision-making to operational design and production performance prediction.

In recent years, artificial intelligence (AI) has become a hot topic, with more and more AI techniques being implemented in the oil industry for advanced data analysis. Both supervised and unsupervised learning algorithms have been employed in the literature to assist the decision-making of EOR techniques. Implementation of AI in the oil industry could be classified as the prediction of the efficiencies/parameters and pattern recognition.29−32 In machine learning, the predictions are normally treated as regression problems (supervised learning), where objective functions, loss functions, or activation functions are required in the establishment of models. In contrast, pattern recognition methods in the oil industry usually refer to an unsupervised learning approach, which is used to find the hidden patterns behind the data set itself and no objective function is needed. Artificial neural network (ANN), particle swarm optimization (PSO), and support vector machine (SVM) have been widely used in prediction models in the oil industry. For example, Zhang et al. proposed the implementation of SVM and multiple regression methods to predict the recovery factor and the CO2 injection efficiency for CO2 immiscible flooding.33 Shafiei et al. developed models for the prediction of the steam flooding reaction factor and the cumulative steam–oil ratio by the implementation of ANN-PSO.34 However, only limited research has been conducted for pattern recognition in the oil industry, which mainly focuses on the EOR method selection process. Siena et al. applied Bayesian clustering and principal component analysis to build a model for EOR selection based on six reservoir/fluid properties, where the EOR selection result is revealed by the analogy projects.28 Alvarado et al. present a 2D graphical expert map to visualize the percentage of each EOR method included in the clusters for inspection,35 where the EOR method recommendation depends on the cluster that the new project merged with. The main idea of the pattern recognition is to detect hidden patterns in the data set so that a recommendation can be provided after the characterization of the proposed patterns.
In this paper, we implement the hierarchical clustering algorithm (HCA) to worldwide steam flooding projects that were collected from existing EOR surveys and publications to find the hidden patterns within steam flooding projects because the steam flooding techniques have been conducted under various conditions. Based on the patterns revealed from the HCA, the analogue assessment for new candidate steam flooding projects enables us to find the most similar cases from the existing projects, which assists in the decision-making process of risk reduction by providing recommendations for operational design and production performance.

This paper is organized as follows. The Data Preparation section describes the establishment of the worldwide steam flooding data set that we used for pattern recognition. The Methodologies section details the approaches included in this work followed by the results received from each method with the presentation of three field case applications to examine the effectiveness of the proposed methodologies for which the operational design and production performance have been well documented. Final remarks are then summarized in the Conclusions section.

## DATA PREPARATION

Figure 1 illustrates a graphical workflow of the steam flooding pattern recognition process. Four steps are integrated into this work: (1) data preparation of the worldwide steam flooding data set; (2) data cleansing and pre-processing; (3) design and implementation of hierarchical clustering for pattern recognition; and (4) data analysis for clustering results. The first step relies on extensive review and examination of successful pilot/field steam flooding projects that were published in *Oil and Gas Journal* biannual EOR surveys, SPE publications, DOE reports, and AAPG databases. Eight main reservoir/flow parameters are selected and extracted as these parameters are commonly available and used for EOR project data analysis.36 These parameters include porosity, average permeability (matrix permeability and fracture permeability), depth, net thickness, oil viscosity, oil gravity, temperature, and oil saturation before steam flooding started.

The second step ensures that the data quality and clustering analysis meet requirements. In this study, all projects with missing values are deleted to avoid biased results if the missing properties could not be found from supplemental publications or reports. A severe duplicate data problem is revealed in the steam flooding data set as the data set is formed with the integration of various sources, so identical projects are removed. Senseless or incorrect data are detected from boxplots and scatterplots as explained in previous studies.37−40 Where incorrect data are erased or corrected based on the literature. After cleansing for data quality enhancement, 384 projects are retained in the data set. Figure 2 presents the location of oil fields and the number of projects in each country that applies steam flooding technology. The United States, Venezuela, Canada, and Trinidad are the leaders for the conduction of steam flooding, which makes up 93% of the total projects.

## METHODOLOGIES

After we finalize the data set, a series of robust data transformation techniques and data analysis methodologies are applied to assist the steam flooding pattern recognition process. Permeability and oil viscosity were first transformed into the logarithmic scale because the values of these two parameters vary over several orders of magnitude, which are about $10^2$ to $10^4$ times greater than other properties and have large ranges for steam flooding projects. Standardization and principal component analysis were employed to pre-process the data and to improve the interpretation results of the proposed methodology. HCA and reservoir/flow property classification techniques were applied to detect the hidden patterns of steam flooding projects, which are detailed in the following subsections.

### Standardization

To ensure that the selected or extracted eight reservoir/flow properties have the same importance and to avoid the features with high variances from "illegitimately" dominating the principal component analysis and hierarchical clustering results, we applied z-score standardization on our data set to force the properties into the same scale, which is defined as

$$X' = \frac{X - \bar{X}}{\sigma}$$

where $X'$ is transformed value, $X$ is original value, $\bar{X}$ is average value of a reservoir/flow property, and $\sigma$ is standard deviation of a reservoir/flow property.

The main advantage of using the z-score standardization method is not only bringing all properties into the scale but also preserving all relationships among properties.41

### Principal Component Analysis

In the implementation of machine learning techniques, a phenomenon called “the curse of dimensionality” has been widely observed. Machine learning techniques perform well with a low dimensionality of data; however, with the increase of dimensionality of the analyzed data, most algorithms have poor performance on similarity measurements due to the high computational complexity on similarity measurements.42 In statistics, a method called principal component analysis (PCA) has been commonly used to solve the high-dimensional data problem by reducing the dimensionalities of the data set while retaining the main variances. The goal of PCA is to find directions/vectors that project the data set with minimized projection errors. The primary mechanism of PCA is a series of orthogonal transformations that are applied to convert a set of observations into linear uncorrelated principal components (PC), where each principal component represents a combination of all input variables and reveals the associations between reservoir/flow properties. After the data is transformed by PCA, the original data set with high dimensions can be effectively reduced to two dimensions (2D) or three dimensions (3D) without losing much information. Typically, a good PCA result should retain more than 90% variance from the original data set.

### Hierarchical Clustering Algorithm

Clustering is considered one of the most crucial unsupervised learning algorithms that deals with finding a structure in a collection of unlabeled data. The goal of clustering is to determine the intrinsic grouping or hidden pattern in a data set by computing the pairwise distance. In this study, we apply the agglomerative hierarchical clustering technique to the steam flooding data set because this method is straightforward and has a stable structure, which allows for a fully customized design in the algorithm (e.g., number of clusters cut off, distance, and linkage) and prevents the “black-box” processing information from being stored as in other algorithms (like ANN).43−45 The structure and outcomes of HCA can be presented in a
dendrogram and scatterplot, which depicts the closeness among all projects, detects special cases, and reveals the hidden pattern in the data set. The framework of the implementation of HCA in this work is made up of six main steps:

1. Perform data preprocessing.
2. Define distance function.
3. Determine the linkage method by the computation of linkage coefficient.
4. Find the optimized value of the number of clusters.
5. Use HCA with the defined distance function, linkage method, and number of clusters.
6. Analyze clustering result.

Figure 3 presents the process for the implementation of an agglomerative HCA with a bottom-up structure. The agglomerative HCA starts with each data point (project) being a single cluster, and then merges the data points that are closest (smallest distance). The merging process ends when all objects are forced in one superior cluster. The root node represents the whole dataset, and each leaf of the tree represents a sample. The intermediate nodes describe the clusters at that level, and the height of the dendrogram usually displays the distance between each paired cluster.

In mathematics, numerous methods exist to define the distances between objectives. As the HCA is a distance-based algorithm, the definition of distance is critical for the design of HCA because by using different methods, the computation results will be different, which determines how clusters/projects are merged together. In this research, the Euclidean distance is used to determine the closeness between projects and clusters because this method has been most commonly utilized for numerical features. Average, single, complete, and Ward linkage methods have been considered in the design of HCA to define how clusters are merged to a higher level. The number of clusters is another required parameter in the implementation of HCA. Thirty indices are employed to find the optimal number of clusters. This was proposed by Charrad et al. because they present a comprehensive evaluation and combination of the majority of existing methodologies in the literature, including Silhouette, elbow, gap statistic, and so forth.55

Reservoir/Fluid Property Classification. The classification of reservoir/fluid properties is essential as they are closely associated with the driving mechanisms that affect the performance of steam flooding. Based on the oil viscosity at reservoir condition, oil has been commonly classified as viscous oil ($\mu < 100$ cp), heavy oil ($100 \text{ cp} \leq \mu \leq 10,000$ cp), and extra heavy oil ($\mu > 10,000$ cp).5,34-47 A similar oil classification based on oil gravity has also been well accepted by the oil industry.5,46,47 The reservoir depth is also an important parameter for steam flooding applications, which are generally classified as either a shallow reservoir or a deep reservoir. In the oil field, the criterion for the classification of depth is ambiguous. For example, most steam flooding projects were conducted in shallow reservoirs because deep reservoirs have more heat loss and have higher requirements for the insulating tubing, leading to higher costs. However, no specific value was given in the literature to define the specific depth of a shallow or deep reservoir. In this study, we classify the reservoir depth for steam flooding based on the collection of numerous publications that mentioned “shallow reservoir” or “deep reservoir”. We find that the 3000 ft burial depth is the critical value for steam flooding applications. Many studies have showed that heat loss is the main reason why the steam flooding technique is applied mostly in shallow reservoirs, however, it is essential to point out that the temperature of injected steam is more important. Based on the pressure-enthalpy phase diagram of water, the lower reservoir pressure requires a lower steam temperature to provide the same amount of energy (enthalpy).50 For a naturally pressured reservoir with a burial depth of 3000 ft, the reservoir pressure is about 1350 psi, which requires the steam temperature to reach about 600 °F.51,52 Uniquely designed downhole equipment is needed to meet the requirements of high temperatures. Therefore, 3000 ft is a threshold depth for steam flooding projects, where reservoirs with a depth greater than 3000 ft are considered deep reservoirs. Otherwise, they are shallow reservoirs.

For the rest of the reservoir/fluid properties (porosity, permeability, start oil saturation, temperature, and net thickness), we employ a statistical method by using boxplots to classify the properties, as shown in Figure 4. The goal of applying a boxplot is to display the range and distribution of each property for the existing projects, which not only
facilitates the classification of properties but also presents the feasibility of steam flooding applications. Minimum, Q1 (25th percentile), median (50th percentile), mean (average), Q3 (75th percentile), and maximum values are illustrated in the boxplot. A property is classified in a low category when the value is smaller than Q1 (25th percentile), which means that more than 75% of the existing steam flooding projects were conducted with a higher value. Similarly, the high category is defined as when the property value is greater than Q3 (75th percentile), which indicates that only less than 25% of the existing projects are greater than the given property value. The range from Q1 to Q3 is categorized as a medium category because this range represents most projects.

RESULTS AND DISCUSSION

Dimensionality Reduction. Because eight reservoir/fluid parameters are selected for pattern recognition (clustering), PCA transforms the eight-dimensional data into eight PCs. The column chart in Figure 5 illustrates the variance expressed by each PC based on the input data, and the red dotted line denotes the cumulative variance explained by the first several PCs. The results depict that the first two PCs retained about 90% of the variance, which proves that the PCA could be effectively used in the steam flooding data set for dimensionality reduction. Therefore, a two-dimensional PCA results in a high variance explained from the original steam flooding data set, which demonstrates that the clustering results with two PCs could be clearly visualized and evaluated. A visualized comparison of the clustering results with and without PCA pre-processing process will be presented in the following section.

Hierarchical Clustering. Agglomerative coefficient has been commonly used in R language for the evaluation of different linkage methods based on clustering structure. Table 1 presents the Ward linkage method, which is based on the optimization of error sum of squares (minimum variance) and is selected as a criterion to choose the paired clusters in each step. Figure 6 presents the frequency distributions of 24 out of 30 indices that recommend having less than 10 clusters. The other six indices elucidate that the data set should be split into more than 10 clusters. The horizontal axis in Figure 6 shows the number of clusters, while the vertical axis illustrates the total number of indices/methods that recommend each value of the number of clusters. For example, five indices suggest splitting the original data into two clusters/groups, while three indices agree to divide the data into three clusters based on (1) the maximum/minimum value of the index, (2) the maximum/minimum difference between hierarchy levels, (3) the maximum/minimum second differences between hierarchy levels, (4) critical values such as in the gap statistic, and (5) the significant local change in the measurement.53,54 The results recommend that five clusters with seven supporting indices are the optimal value in the steam flooding data set. Figure 7 illustrates the visualization of PC1 and PC2 with five clusters by retaining about 90% of the variance from the original steam flooding data set, which demonstrates that the clustering results with two PCs could be clearly visualized and evaluated. The number of steam flooding projects in each cluster is shown in Table 2, where cluster 1 (C1) is the biggest group containing 126 projects, followed by C2 (105 projects), C4 (82 projects), C3 (47 projects), and C5 (24 projects). The results in Figure 7 elucidate the clear boundaries between clusters, which means that five clusters are distinguished from each other by including significantly different reservoir/fluid
properties. In contrast, Figure 8 shows a messy distribution with the same HCA design (distances, linkages, and number of clusters) where PCA did not pre-process the original data set. The main reason for the unclear boundaries between clusters is the high dimensionality of data, where the dimensions of the original data with 8 reservoir/fluid parameters are difficult to be clearly visualized in a 2D plot. Therefore, data transformation with PCA is essential for steam flooding projects.

**Characterization of Clusters.** As the purpose of HCA is to recognize the hidden patterns in steam flooding data sets that cannot be seen from direct observations, the characterization of clusters is critical for studying reasons why the clusters are distinguished from each other. Blue boxplots in Figure 9 demonstrate comparisons between five clusters for each reservoir/fluid property.

![Figure 7. Hierarchical clustering visualization with PCA in data pre-processing.](image1)

### Table 2. Number of Steam Flooding Projects in Each Cluster

| cluster number | number of projects |
|----------------|--------------------|
| C1             | 126                |
| C2             | 105                |
| C3             | 47                 |
| C4             | 82                 |
| C5             | 24                 |

![Figure 8. Hierarchical clustering visualization without PCA in data pre-processing.](image2)
Figure 9 indicates that C3 and C5 include the special reservoirs (2 projects) that have high porosity (up to 65%), which is caused by the lithologies in the reservoirs. Most of the steam flooding projects have been implemented in sandstone formations because most of the EOR techniques have been tested at the pilot and commercial scales in this type of lithology. The normal porosity for the sandstone reservoirs is less than 35%. The projects with extremely high porosities are found in the Midway-Sunset field and South Belridge field and in diatomite formations, where the diatomite reservoirs generally have low matrix permeability (less than 1 mD) with a high porosity (40−70%).

Figure 9b illustrates the average permeability ranges based on the matrix and fracture permeabilities. Although C1 is the biggest cluster, most of the projects in C1 fall into a well-concentrated range of permeability from 2000 to 3000 mD, which reveals that C1 had been effectively grouped with projects with similar permeability. Also, permeability boxplots for C4 and C5 show that the permeability of these two clusters is condensed from 2000 to 3500 mD and from 2000 to 3000 mD, respectively.

Figure 9c displays that most steam flooding projects were applied in reservoirs with a depth less than 2500 ft, which is shallower than other EOR methods. The deepest project being conducted in a reservoir is 5740 ft. The ranges of net thickness of the formation are presented in Figure 9d. Most of the steam flooding projects were applied with a thickness less than 200 ft. However, C4 detected most of the projects that the reservoir is thicker than 200 ft. Normally, steam flooding could not be applied in thick reservoirs so as to avoid the steam overriding problem, which reduces the sweep efficiency.

Figure 9e,f displays the ranges for reservoir temperature and the oil saturation before the application of steam flooding, respectively. C1 contains most of the projects with reservoir...
temperatures less than 90 °F and an average oil saturation of 65%, while C5 contains a broader range of temperatures (>90 °F) with a small range of oil saturation. The boxplots demonstrate that most projects are conducted in lower reservoir temperatures compared to other EOR techniques.39,58,59 The shallower burial depth is one of the reasons for the lower temperature, where the temperature is positively related to the depth with an average geothermal gradient of 2 °F/100 ft.60 Another reason is that lower reservoir temperatures may cause a greater temperature difference when the same amount of steam is injected with the same temperature, which results in a more significant reduction of oil viscosity, especially in heavy oil reservoirs. Also, boxplots in C2, C3, and C4 elucidate similar ranges for both temperature and oil saturation, which means that other reservoir/fluid properties may have significant differences between C2, C3, and C4 (e.g., porosity, permeability).

Figure 9g,h summarizes the ranges for both oil gravity and oil viscosity. In Figure 9g, only cluster 3 detected the light oil projects from the steam flooding data set, which includes the projects with oil gravity greater than 25 °API. The projects in other clusters illustrate a condensed range from 12 to 14 °API, which means most of the projects in C1, C2, C4, and C5 are heavy oil reservoirs. Figure 9h shows that C1 captured the projects with extremely heavy oil (μ > 100,000 cp), and that C4 grouped the projects with high oil viscosity ranging from 4000 to 10,000 cp, which is higher than the ranges in C2, C3, and C5.

**Analogue Reasoning.** The goal of analogue reasoning is to examine the effectiveness of the established PCA/HCA method and to find the most similar project to the new candidate steam flooding project. The analogue case shares the similar reservoir/fluid properties with the new candidate project so that valuable operational design experiences and production performance from the analogue cases can be referenced for decision-making. The analogue process is carried out by the computation of Euclidean distances that were embedded in the hierarchical clustering process between new candidate steam flooding project(s) and the existing steam flooding data set. A project with minimal distance to the candidate field is considered as being the closest case to the new project. Figure 10 illustrates the visualized analogue results of three new candidate projects. As shown in Figure 10, three cases fall into different patterns/clusters that were

![Figure 10. Analogue visualization with three new steam flooding testing projects in a scatterplot.](image)

| Table 3. Reservoir/Fluid Properties of Testing Cases and Analogue Results |
|-----------------------------------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| case # | field name | country | cluster | net thickness, ft | porosity, % | permeability, mD | depth, ft | oil gravity, °API | oil viscosity, cp | temperature, °F | start oil saturation, % | references |
|--------|------------|---------|---------|------------------|-------------|-----------------|--------|-----------------|-----------------|-----------------|---------------------|------------|
| case 1 | Shanjiasi Forest Reserve | China | cluster 2 | 82               | 30           | 5000            | 3983   | 19              | 9200            | 131             | 60                  | 61         |
| analog 1 | Wolf Lake | Canada | cluster 1 | 75               | 31           | 205             | 3000   | 10              | 32              | 10,000          | 60                  | 62         |
| case 2 | Wolf Lake | Canada | cluster 1 | 75               | 33           | 3000            | 1398   | 10              | 10,000          | 10              | 65                  | 63         |
| analog 2 | Ruhelertwist Emlichheim | Germany | cluster 3 | 49               | 28           | 5000            | 1400   | 50              | 45,000          | 45              | 78                  | 63         |
| case 3 | Ruehlertwist | Germany | cluster 3 | 79               | 30           | 6000            | 2650   | 25              | 175             | 25              | 51                  | 64−66      |
| analog 3 | Emlichheim | Germany | cluster 3 | 79               | 30           | 6000            | 2400   | 24.5            | 175             | 24.5            | 62                  | 65−67      |
revealed by the PCA and HCA. The first case is allocated to C2, and case 2 is merged with C1, while case 3 is integrated with C3. Each case represents a scenario of the analogue result, case 1 analog to a foreign oil field, case 2 analog to the same field, and case 3 analog to an adjacent field. Table 3 depicts the comparison of reservoir/liquid properties between the testing cases and the analogue cases.

The analogue reasoning of case 1 from the established PCA and HCA methods reveals that the reservoir/liquid properties of the Forest Reserve field from Trinidad are the most similar project to the Shanjiashi field from China. Both fields applied the cyclic steam flooding technology with an averaged soaking period of 3−4 days in each cycle. Although the well schemas are so different in the two fields, where 280 production wells were drilled in the Shanjiashi field compared to 70 production wells in the Forest Reserve field, the averaged enhanced oil production after the implementation of steam flooding for each well is similar, which are 32.6 (Shanjiashi field) and 30.9 bbl/d/well (Forest Reserve field).65 Besides, the conclusion of steam flooding was the first attempt to enhance the oil recovery for both fields. Thinner insulating tubing than that in the Shanjiashi field was installed in the Forest Reserve field with a more insulated cement sheath to reduce the drilling cost and to ensure the steam quality.65 The analogue results from case 1 imply that the design and performance are similar when the reservoir/liquid properties are close. Therefore, the analogue assessment could assist in predicting the effectiveness of steam flooding for new candidate steam flooding projects based on existing experiences from a similar field, especially when the field data is limited.

The second case is the Wolf Lake field from Canada, which implemented steam flooding in 1985 and consisted of 187 production wells. The analogue result presents that the most similar existing project is located in the same field, which applied steam flooding in 1982 with one production well. In fact, case 2 is the expansion of the analogue project, so the reservoir/liquid properties are almost the same except the viscosity.65 Because oil viscosity reduction is the main mechanism for steam flooding, especially in bitumen reservoirs, the oil viscosity decreased significantly after steam injection from the pilot test, which caused viscosity reduction compared with the analogue case. The analogue result proves that the proposed PCA/HCA methodology is still capable of detecting the similar cases from the same field because all the reservoir/liquid properties were normalized before the implementation of HCA.

Case 3 represents a scenario that finds an adjacent oil field. Oil fields with close geographical locations normally share similar reservoir/liquid properties because the depositional environments are the same, which results in smaller distances between the analogue project and the candidate case. The third case is selected from the Ruehlertwist field from Germany in Lower Saxony, and the analogue results show that the nearby Emlichheim field is the most similar and is only 13.4 miles away from the Ruehlertwist field.

Classification of Reservoir/Fluid Properties. Table 4 shows the classification results of all reservoir/liquid properties for steam flooding. As described in the previous section, the classification results for porosity, permeability, net thickness, reservoir temperature, and start oil saturation are based on the yellow boxplots illustrated in Figure 9, where a property value less than Q1 is considered a low value, a value greater than Q3 is in the high category, and a value between Q1 and Q3 is in the medium category. The domain knowledge is applied for the classification of oil gravity and oil viscosity based on previous studies from experts.8,34,46

| Table 4. Classification Results of Reservoir/Fluid Properties Based on Worldwide Steam Flooding Projects and Domain Knowledge |
|-----------------|-----------------|-----------------|-----------------|
| property        | category        | Value Range     | references      |
| porosity, %     | low             | <30             | based on worldwide steam data |
|                 | medium          | [30,35]         |                 |
|                 | high            | >35             |                 |
| permeability, mD| low             | <1000           | based on worldwide steam data |
|                 | medium          | [1000,3000]     |                 |
|                 | high            | >3000           |                 |
| depth, ft       | shallow         | ≤3000           | 61, 62, 71 |
|                 | deep            | >3000           |                 |
| net Thickness, ft| thin            | <98.4           | based on worldwide steam data |
|                 | thick           | >205            |                 |
| temperature, °F | low             | <90             | based on worldwide steam data |
|                 | medium          | [90,110]        |                 |
|                 | high            | >110            |                 |
| oil gravity, °API| light oil       | >25             | 8, 46, 47 |
|                 | medium oil      | [20,25]         |                 |
|                 | heavy oil       | >10             |                 |
|                 | extra heavy oil | <10             |                 |
| oil Viscosity, cp| viscous oil     | <100            | 8, 34, 46, 47 |
|                 | heavy oil       | [100,10,000]    |                 |
|                 | extra heavy oil | >10,000         |                 |
| start oil saturation, % | low | <57 | based on worldwide steam data |
|                 | medium          | 57−80           |                 |
|                 | high            | >80             |                 |

CONCLUSIONS

In this paper, a combination of principal component analysis and HCAs is applied to identify the hidden patterns in worldwide steam flooding projects and to examine the effectiveness of the proposed method via the analogue reasoning process. Based on the computation of 30 indices and the clustering structure, we detected that the optimum number of clusters is 5, which indicates five stabilized cluster patterns among all steam flooding projects. We further characterized the clusters to study the patterns revealed by the HCA. We found the reservoir/liquid properties C1, C4, and C5 have small concentrated ranges, while the projects in C2 and C3 contain special cases for porosity, permeability, depth, and oil gravity. The comparison with/without PCA before the implementation of HCA illustrates that the HCA associated with PCA transformation provides clear clustering boundaries and reduces the dimensionalities from 8D to 2D while still retaining about 90% of the variance. In addition, the reservoir/liquid properties are classified based on domain knowledge from literature, and the values of Q1 and Q3 as revealed by the boxplots. The threshold depth for the implementation of steam flooding is 3000 ft due to the limitation of infrastructure. Most of the steam flooding projects were applied with the burial
depth less than 3000 ft and are classified as the shallow reservoir.

A blind test of the proposed method was performed by considering three field cases. The analogue results demonstrate that the established method is capable of providing assistance for capturing the most similar existing steam flooding projects that share similar reservoir/ﬂuid properties. In addition, the analogue cases indicate that the operational designs and performance of steam flooding are close even though the candidate case and the analogue field are from different countries (case 1). Therefore, the analogy based on the PCA/HCA not only provides assistance for operational design decision-making in new steam flooding candidate fields but may also provide a prediction for the future performance based on existing projects.

# ASSOCIATED CONTENT

## Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.2c01693.

Worldwide steam flooding data set (XLSX)

# AUTHOR INFORMATION

## Corresponding Authors

Na Zhang — Shandong University of Science and Technology, Qingdao 266590, China; Missouri University of Science and Technology, Rolla 65401, United States; Email: nzrdf219@163.com, nzrdf@most.edu

Mingzhen Wei — Missouri University of Science and Technology, Rolla 65401, United States; Email: weim@most.edu

## Authors

Baojun Bai — Missouri University of Science and Technology, Rolla 65401, United States

Xiaopeng Wang — Changing Oilfield Branch of PetroChina, Xi’an, Shaanxi 710200, China

Jian Hao — Shandong University of Science and Technology, Qingdao 266590, China

Shun Jia — Shandong University of Science and Technology, Qingdao 266590, China

Complete contact information is available at https://pubs.acs.org/10.1021/acsomega.2c01693

## Notes

The authors declare no competing financial interest.

# ACKNOWLEDGMENTS

This research was supported by the Natural Science Foundation of Shandong Province (grant ZR2021QF076), Project of Shandong Province Higher Educational “Youth Innovation Science and Technology Plan” (grant 2021JKJ060), and National Natural Science Foundation of China (grants S2174121 and 71971130).

# NOMENCLATURE

| Abbreviation | Description |
|--------------|-------------|
| EOR          | enhanced oil recovery |
| PC1          | first principal component |
| ANN          | artifical neural network |
| PC2          | second principal component |
| SVM          | support vector machine |

HCA hierarchical clustering algorithm
PSO particle swarm optimization
ANN artificial neural network
PCA principal component analysis
Q1 first percentile
PC(s) principal component(s)
Q3 third percentile

# REFERENCES

1. Santos, R. G.; Loh, W.; Bannwart, A. C.; Trevisan, O. V. An overview of heavy oil properties and its recovery and transportation methods. Braz. J. Chem. Eng. 2014, 31, 571−590.

2. Shafiei, A.; Ahmadi, M.; Dusseaum, M.; Elkanem, A.; Zendehboudi, S.; Chatzis, I. Data Analytics Techniques for Performance Prediction of Steamflooding in Naturally Fractured Carbonate Reservoirs. Energies 2018, 11, 292.

3. Dong, X.; Liu, H.; Chen, Z.; Wu, K.; Lu, N.; Zhang, Q. Enhanced oil recovery techniques for heavy oil and oilsands reservoirs after steam injection. Appl. Energy 2019, 239, 1190−1211.

4. Pang, Z.; Wang, L.; Yin, F.; Lyu, X. Steam chamber expanding processes and bottom water invading characteristics during steam flooding in heavy oil reservoirs. Energy 2021, 234, 121214.

5. Venkatramani, A. V.; Okuno, R. Mechanistic simulation study of expanding-solvent steam-assisted gravity drainage under reservoir heterogeneity. J. Petrol. Sci. Eng. 2018, 169, 146−156.

6. Alajmi, A. F. Heat loss effect on oil bank formation during steam flood. J. Petrol. Sci. Eng. 2021, 199, 108262.

7. Wu, C. H. A Critical Review of Steamflood Mechanisms. SPE California Regional Meeting; OnePetro, 1977.

8. Blevins, T.; Duerksen, J.; Ault, J. W. Light-oil steamflooding an emerging technology. J. Petrol. Technol. 1984, 36, 1115−1122.

9. Kumar, A.; Hassanzadeh, H. Impact of shale barriers on performance of SAGD and ES-SAGD—A review. Fuel 2021, 289, 119850.

10. Liu, P.; Zhang, Y.; Liu, P.; Zhou, Y.; Qi, Z.; Shi, L.; Xi, C.; Zhang, Z.; Wang, C.; Hua, D. Experimental and numerical investigation on extra-heavy oil recovery by steam injection using vertical injector-horizontal producer. J. Petrol. Sci. Eng. 2021, 205, 108945.

11. Yang, S.; Nie, Z.; Wu, S.; Li, Z.; Wang, B.; Wu, W.; Chen, Z. A Critical Review of Reservoir Simulation Applications in Key Thermal Recovery Processes: Lessons, Opportunities, and Challenges. Energy Fuels 2021, 35, 7387−7405.

12. Feder, J. Heavy-Oil Steamflood Validates Machine-Learning-Assisted Model. J. Petrol. Technol. 2020, 72, 53−55.

13. Osterloh, W. T.; Jones, J. Novel Thermal Process for Recovery of Extremely Shallow Heavy Oil. SPE Reservoir Eval. Eng. 2003, 6, 127−134.

14. Manrique, E. J.; Thomas, C. P.; Ravikiran, R.; Izadi Kamouei, M.; Lantz, M.; Romero, J. L.; Alvarez, V. EOR: Current Status and Opportunities. SPE Improved Oil Recovery Symposium; Society of Petroleum Engineers: Tulsa, Oklahoma, USA, 2010; p 21.

15. Al-Saedi, H. N.; Flori, R. E.; Mortadha, A. Coupling Low Salinity Water Flooding and Steam Flooding for Heavy Oil in Sandstone Reservoirs; Low Salinity-Alternating-Steam Flooding LSASF: A Novel EOR Technique. Abu Dhabi International Petroleum Exhibition & Conference; OnePetro, 2018.

16. Alkafeef, S. F.; Zaid, A. M. Review of and outlook for enhanced oil recovery techniques in Kuwait oil reservoirs, International Petroleum Technology Conference; OnePetro, 2007.

17. Choudhary, M. A.; Wani, M. R.; Mahmeed, A.; Al-Rasheed, H. R.; Redha, M. Recovering Heavy Oil from a Middle East Carbonate Reservoir: EOR Potential, Screening, Challenges, Uncertainties and Risk Management Measures. SPE Enhanced Oil Recovery Conference; Society of Petroleum Engineers: Kuala Lumpur, Malaysia, 2011; p 11.

18. Pratama, R. A.; Babadagli, T. Effect of Temperature, Phase Change, and Chemical Additives on Wettability Alteration During...
Steam Applications in Sands and Carbonates. SPE Reservor Eval. Eng. 2020, 23, 292−310.

(19) Taraskin, E. N.; Zakharian, A. Z.; Zakharian, A. Z.; Uregov, S. O. Adaptive option of steam injection technological efficiency evaluation for carbonate high-viscosity oil reservoir conditions (Russian). Neft. khozyaystvo-Oil Ind. 2018, 2018, 102−107.

(20) Wu, S.; Han, M.; Ma, D.; Wu, Y.; Yu, Q.; Chunyi, Y.; Dehuang, S. A Case Study: A Successful Steam Flooding Project to Enhance Oil Recovery of Low-permeability, Light-oil Waterflooding Reservoir. SPE Enhanced Oil Recovery Conference; OnePetro, 2011.

(21) Zhang, N.; Wei, M.; Fan, J.; Aldhaiberti, M.; Zhang, Y.; Bai, B. Development of a hybrid scoring system for EOR screening by combining conventional screening guidelines and random forest algorithm. Fuel 2019, 256, 115915.

(22) Tian, Y.; Ji, B.; Li, G.; Liu, N.; Dong, Y.; Ma, S. The comprehensive model for solvent assisted steam flooding in thin heavy oil reservoirs considering asphaltene deposition. J. Petrol. Sci. Eng. 2020, 185, 106676.

(23) Zhao, D. W.; Wang, J.; Gates, I. D. Thermal recovery strategies for thin heavy oil reservoirs. Fuel 2014, 117, 431−441.

(24) Huang, S.; Chen, X.; Liu, H.; Jiang, J.; Cao, M.; Xia, Y. Experimental and numerical study of solvent optimization during horizontal-well solvent-enhanced steam flooding in thin heavy-oil reservoirs. Fuel 2018, 228, 379−389.

(25) Wang, S.; Li, Y.; Ma, K.; Yang, J.; Zhang, G. Evaluation of the First Cyclic Steam Pilot in Offshore Oilfield of China. SPE Asia Pacific Enhanced Oil Recovery Conference; OnePetro, 2015.

(26) Han, X.; Liu, Y.; Liu, H.; Wang, Q.; Zou, J.; Zhang, H.; Wang, H.; Wu, X. Case Study: Realization and Evaluation of Cyclic Steam Stimulation Pilot for Offshore Oilfield, China. SPE EOR Conference at Oil and Gas West Asia; OnePetro, 2018.

(27) Han, X.; Zhong, L.; Liu, Y.; Zou, J.; Wang, Q. Study and Pilot Test of Multiple-Thermal-Fluid Stimulation in Offshore Nanpu Oilfield. SPE Prod. Oper. 2020, 35, 592−603.

(28) Siena, M.; Guadagnini, A.; Rossa, E. D.; Lamberti, A.; Masserano, F.; Rotondi, M. A Novel Enhanced-Oil-Recovery Screening Approach Based on Bayesian Clustering and Principal-Component Analysis. SPE Reservoir Eval. Eng. 2016, 19, 382−390.

(29) Berry, M. W.; Mohamed, A.; Yap, B. W. Supervised and Unsupervised Learning for Data Science; Springer, 2019.

(30) Lam, D.; Wunsch, D. C. Academic Press Library in Signal Processing; Elsevier, 2014; Vol. 1, p 1115−1149.

(31) Li, B.; Billiter, T. C.; Tokar, T. Rescaling Method for Improved Machine-Learning. Decline Curve Analysis for Unconventional Reservoirs. J. SPE 2021, 26, 1759−1772.

(32) Kashra, A. A.; Sakhhaee-Pour, A.; Hussein, I. A. SPE Reservoir Evaluation & Engineering; OnePetro, 2021; pp 1−20.

(33) Zhang, N.; Wei, M.; Bai, B. Applicability of Worldwide CO2 Worldwide Immiscible Flooding and Prediction. Carbon Management Technology Conference; OnePetro: Houston, Texas, USA; 2017; p 10.

(34) Shafiei, A.; Dusseault, M. B.; Zendehboudi, S.; Chatzis, I. A new screening tool for evaluation of steamflooding performance in Naturally Fractured Carbonate Reservoirs. Fuel 2013, 108, 502−514.

(35) Alvarado, V.; Ranson, A.; Hernandez, K.; Manrique, E.; Matheus, J.; Liscano, T.; Prosperi, N. Selection of EOR/IOR Opportunities Based on Machine Learning. European Petroleum Conference; OnePetro: United Kingdom, 2002; p 11.

(36) Taber, J. J.; Martin, F. D.; Serigraph, R. S. EOR screening criteria revisited-Part 1: Introduction to screening criteria and enhanced recovery field projects. SPE Reservoir Eval. Eng. 1997, 12, 189−198.

(37) Saleh, L. D.; Wei, M.; Bai, B. Data analysis and updated screening criteria for polymer flooding based on oilfield data. SPE Reservoir Eval. Eval. 2014, 17, 15−25.

(38) Hama, M. Q.; Wei, M.; Saleh, L. D.; Bai, B. Updated Screening Criteria for Steam Flooding Based on Oil Field Projects Data. SPE Heavy Oil Conference-Canada; OnePetro: Calgary, Alberta, 2014; p 19.

(39) Zhang, N.; Yin, M.; Wei, M.; Bai, B. Identification of CO2 sequestration opportunities: CO2 miscible flooding guidelines. Fuel 2019, 241, 459−467.
(62) Ramkhalawan, C. D.; Khan, J.; Bainey, K. R. Thirty 30 Years of Steamflooding: Reservoir Management and Operational Experiences. SPE Annual Technical Conference and Exhibition; OnePetro: Dallas, Texas, 1995; p 8.

(63) Hallam, R. J.; Hajdo, L. E.; Donnelly, J. K.; Baron, P. R. Thermal Recovery of Bitumen at Wolf Lake. SPE Reservoir Eng. 1989, 4, 178–186.

(64) Proyer, G.; Chaziteodorou, G.; Muss, H.; Rosskamp, M. Results of a Steamdrive Pilot Project in the Ruehlertwist Field, Federal Republic of Germany. J. Petrol. Technol. 1985, 37, 284–294.

(65) Leonard, J. Increased rate of EOR brightens outlook. Oil Gas J. 1986, 84, 71–89.

(66) Leonard, J. Steam dominates enhanced oil recovery. Oil Gas J. 1986, 80, 139–159.

(67) Robertson, E. P. Selection of Analytical Steam Stimulation Models Based on Common Reservoir. SPE/DOE Improved Oil Recovery Symposium; OnePetro: Tulsa, Oklahoma, 1998; p 9.

(68) Moritis, G. More US EOR projects start but EOR production continues decline. Oil Gas J. 2008, 106, 41.

(69) Koottungal, L. Worldwide EOR Survey. Oil and Gas Journal, 2008.

(70) Matheny, S. L., Jr. EOR methods help ultimate recovery. Oil Gas J. 1980, 78 (13).

(71) Dehghani, K.; Ehrlich, R. Evaluation of Steam Injection Process in Light Oil Reservoirs. SPE Annual Technical Conference and Exhibition; OnePetro: New Orleans, 1998; p 13.