Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 5547288, 13 pages
https://doi.org/10.1155/2021/5547288

Research Article

Prediction of Heavy Oil Steam Stimulation Based on Data-Driven and Mechanism Model

Chaochao Zhao,[1,2] Chao Min,[1,2] Chuanfei Wang,[3] Yanfeng Lin,[1] and Mengshu Long[1]

[1]*School of Science, Southwest Petroleum University, Chengdu 610500, Sichuan, China*
[2]*Institute for Artificial Intelligence, Southwest Petroleum University, Chengdu 610500, Sichuan, China*
[3]*Research Institute of Exploration and Development, Shengli Oilfield Company, SINOPEC, Dongying 257015, Shandong, China*

Correspondence should be addressed to Chao Min; minchao@swpu.edu.cn

Received 11 February 2021; Revised 25 March 2021; Accepted 19 April 2021; Published 5 May 2021

Academic Editor: Xin Ma

In the middle and late stages of heavy oil development, formulating a scientific and reasonable mining plan is the key to improving oilfield efficiency. At present, steam stimulation is still the main development method of heavy oil. The determination of its production is not only limited by boiler conditions, surface pipelines, and wellbore conditions but also by the steam absorption capacity of the formation. Therefore, local analysis cannot achieve the best effect in the whole process of steam stimulation. The mechanism model is the most commonly used method to predict heavy oil production, but too many idealized assumptions make the prediction results quite different from the actual production situation. With the rapid development of machine learning, people can achieve rapid prediction of production through field data. However, when the range of the actual parameter is small, the generalization ability of the model is weak and overfitting occurs. Based on the above background, this paper conducts a coupling study on surface steam pipeline flow, steam injection wellbore flow, and formation flow from the perspective of data-driven. Firstly, based on the correlation coefficient and the feature selection of Random Forest, the importance of the characteristics affecting liquid production and water content was ranked. Secondly, through the comparison of five typical machine learning algorithms, we select the optimal prediction model and optimal characteristics suitable for the sample of this paper. Finally, because of the poor generalization ability of the prediction model, we sampled the mechanism model and increased the diversity of steam dryness samples. We find that the accuracy of the optimal prediction model is improved and the generalization ability of the model is improved after the training of new samples. This paper provides a new idea for the production prediction of heavy oil steam stimulation reservoirs, which is helpful for the efficient development of heavy oil reservoirs.

1. Introduction

As a rich mineral resource, heavy oil has important practical significance for its efficiency and economic development. However, due to the high viscosity and poor flowability of heavy oil, it is difficult to achieve ideal results with conventional technology. Therefore, steam stimulation is still the main development method of heavy oil. The local analysis theory of steam stimulation technology in surface pipelines, wellbores, and formations has been relatively mature and applied to the actual production of oilfields [1, 2]. For a given heavy oil block, the mining effect of steam stimulation depends on the injection and production parameters and the degree of thermal energy utilization of the injected steam. However, the steam injection parameters are only designed through this local software, which cannot make the whole steam stimulation process the best.

The dynamic prediction of steam stimulation wells is the basis of injection parameter design and production design optimization. To improve the mining effect of steam stimulation, researchers have conducted a lot of research on the index prediction of steam stimulation wells. Marx and Langenheim used the energy balance to calculate the heating area of the oil layer [3]. Boberg proposed a steam stimulation production prediction model, which can reflect the mechanism of heating viscosity reduction and oil increase in the

process of steam stimulation, but there are many limitations [4]. Hou and Chen proposed an improved steam stimulation productivity prediction model based on previous studies and introduced the shape coefficient to correct the influence of the overlap phenomenon in the steam injection process [5]. Zheng et al. established a new analytical model for steam stimulation productivity prediction based on the Marx–Langenheim model [6]. The model shows an exponential change in the temperature field in the hot oil area, which is more in line with the actual reservoir. When the temperature is lower than a certain temperature, the heavy oil presents a non-Newtonian fluid state. Yang et al. considered the non-Newtonian steam stimulation productivity prediction model of heavy oil [7].

From the perspective of percolation mechanics and cybernetics, the reservoir system belongs to the distributed parameter system. The basic physical quantities describing the reservoir state are water saturation field and reservoir pressure field. Different parameters represent different underground conditions. The mechanism model reflects our induction and summary of real phenomena and is a reliable and prior cognition of the flow law of underground fluid. Although the mechanism model developed more and more perfectly but compared with the reservoir numerical simulation method, the parameters considered are much less. In 1953, Bruce et al. simulated the one-dimensional gas-phase unstable radial and linear flow [8]. Although limited by the computer level and solving algorithm at that time, it was a milestone in the history of reservoir numerical simulation. With the breakthrough of the numerical solution of linear equations, in 1968, Stone introduced the first numerical solver SIP [9]. In 1974, Coats et al. developed a three-dimensional three-phase steam injection thermal oil recovery model [10]. On this basis, several reservoir numerical simulation software such as CMG series and Eclipse series have been developed.

So far, reservoir numerical simulation software has made a great breakthrough in the integration of functions. For different types of oil and gas reservoirs, different mining methods can almost be used to deal with reservoir numerical simulation software [11–14]. We sample different underground conditions by reservoir numerical simulation and then describe the reservoir mining state by partial differential equations, but its accuracy is based on accurate geological models. Therefore, some idealized assumptions are needed. Since the production law is affected by many unquantifiable main control factors, this may lead to a large difference between the predicted results and the actual production data.

In recent years, artificial intelligence methods have been widely used in the field of petroleum engineering [15–19], which are mainly used for production control and optimization, information prediction, and model simulation in petroleum engineering [20–24]. However, limited by the actual conditions, there is little difference in the data of stratigraphic conditions and production systems between steam stimulation wells in the same block so that when applied to actual oilfield data, the generalization ability of the model is weak and overfitting occurs. Therefore, it is difficult to simply reflect the relationship between some key variables and output indicators from data analysis. This is because the basis of the approximate function space is uncertain and directionless when the simulation is carried out directly by the black-box method. The parameters can only be used blindly for fitting, and its stability cannot be guaranteed.

The innovations of this paper are as follows. (1) Based on previous studies, we conduct a coupled study on surface steam pipeline flow, steam injection wellbore flow, and formation flow based on data-driven. (2) Based on the correlation coefficient and Random Forest feature selection, this paper ranks the features that affect liquid production and water content in importance. (3) For a heavy oil field in eastern China, we used five typical machine learning algorithms to model and compare its field data. It is found that the six characteristics of produced degree, dynamic liquid surface, soaking time, stroke, stroke times, and well pattern mode have little effect on liquid production and water content, which are eliminated. At the same time, the prediction models of liquid production and water content based on Random Forest have the highest accuracy of 86% and 83%, respectively, but the generalization ability of the prediction models is poor. (4) We sampled the mechanism model, increased the diversity of steam dryness samples, and trained the new samples again. It is found that the accuracy of the optimal prediction model obtained previously was improved, making the prediction results more accurate and reliable, and the generalization ability of the model was improved.

The content of this paper is arranged as follows. The second part introduces the data source and data preprocessing. The third part is the establishment and verification of the input and output model of the reservoir system based on data-driven. The fourth part is the establishment and verification of the input and output model of the oil reservoir system based on hybrid data-driven. The fifth part is the conclusion.

2. Data Source and Preprocessing

2.1. Data Source. The data used in this paper are collected from the dynamic and static information, steam injection data, and production data of 109 heavy oil blocks in a heavy oil field in eastern China. Among them, the static information includes oil area, produced reserves, porosity, permeability, and other information data. Dynamic indicators include cumulative oil and cumulative water production. Steam injection data include steam quantity at the boiler outlet, steam pressure at the boiler outlet, and so on. Production data include liquid production and water content.

2.2. Data Preprocessing. Data preprocessing is also an important part of data-driven index prediction, which greatly affects the accuracy of prediction. There are many missing or abnormal values in the actual production data, which cannot be directly trained. Therefore, data cleaning and other operations must be carried out first to obtain higher prediction accuracy.

2.2.1. Outlier Processing. We remove outliers according to the PauTa criterion (3σ criterion). Assuming that the measured variables are measured with equal accuracy, x_i is obtained. If the residual error $v_b\,(1 \le b \le n)$ of a measurement value x_b satisfies $|v_b| = |x_b - x| > 3\sigma$, then x_b is considered to be a bad value with a gross error value, and it is deleted. The formula for standard error σ is as follows:

$$\sigma = \left[\frac{1}{n-1}\sum_{i=1}^{n} v_i^2\right]^{(1/2)} = \left\{\frac{\left[\sum_{i=1}^{n} x_i^2 - \left(\left(\sum_{i=1}^{n} x_i\right)^2/n\right)\right]}{(n-1)}\right\}^{(1/2)}, \tag{1}$$

where $i = 1, 2, \ldots, n$, x is the arithmetic mean, and the residual error is $v_i = x_i - x$.

2.2.2. Missing Value Filling. For the collected samples, if there is too much missing data for a certain group of samples or the sample is missing the two important data of liquid production and water content, the sample is deleted. For the missing values of other parameters such as steam temperature and steam pressure, the *K*-nearest neighbor algorithm is used for filling [25]. We compare the original dataset with the corresponding features in the new dataset and calculate the distance between the new data and each sample in the original dataset. Then, the category of the new data is voted by *K* samples with the smallest distance. The sample distance calculation formula is as follows:

$$\mathrm{d}(p,q) = \mathrm{d}(q,p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}, \tag{2}$$

where d is the relative distance between two fault feature samples and p_i and q_i are the corresponding point data of different fault feature samples, respectively.

After outlier processing and missing value filling, we finally sorted out 97 heavy oil blocks from 109 heavy oil blocks, a total of 780 groups of samples.

2.2.3. Feature Selection. Feature selection is also called feature subset selection or attribute selection. It is a data preprocessing operation that selects from the original features to reduce the data dimension and improve the generalization ability of the model. In practice applications, although more parameters can be used to integrate more information, too many parameters will reduce learning efficiency and even affect prediction accuracy.

Since many factors affecting the development index of heavy oil steam stimulation, it is necessary to go through a systematic index analysis process to find the development index more accurately. Based on the basic theory of reservoir engineering and combined with related research [2, 13, 26–28], we obtained the factors affecting the production of heavy oil steam stimulation, which can be divided into the following five categories:

(1) Reservoir characteristic: reservoir type, surface crude oil viscosity, initial formation temperature, reservoir buried depth, edge-bottom water, oil area, dynamic reserve, primitive oil-bearing saturability, reservoir effective thickness, porosity, net total ratio of oil layers, permeability, original formation pressure, and dynamic liquid surface, in turn with x_1~x_{14}

(2) Productive regulation: soaking time, well distance, well spacing density, well pattern mode, startup well number, stroke, stroke times, production time, and annual turnover, in turn with x_{15}~x_{23}

(3) Characteristics of historical production: cumulative oil production, cumulative water production, and produced degree, in turn with x_{24}~x_{26}

(4) Control variable: steam quantity at the boiler outlet, steam flow rate at the boiler outlet, steam pressure at the bottom of the steam injection well, and steam dryness at the bottom of the steam injection well

(5) Output variable: liquid production and water content, represented by y_1 and y_2, respectively

In the data-driven process, considering that the interaction between data may have a negative impact on the final result, appropriate choices are therefore needed. The four control variables directly affect the final mining effect so as the input of the model, and this paper only selects the remaining 26 variables.

The correlation coefficient is a type of statistical analysis index, which is usually used to determine the direction and degree of linear correlation of variables. The formula is as follows:

$$r = \frac{\sigma_{xy}^2}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n} (y_i - \overline{y})^2}}. \tag{3}$$

We get the correlation coefficient between 26 independent variables and 2 dependent variables, as shown in Table 1.

Feature screening based on Random Forest refers to how much contribution each feature makes on each tree in the Random Forest [29, 30], and then, take the average and compare the contribution of different features. The Gini

Table 1: Correlation coefficient of partial variables.

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}
y_1	0.574	0.718	0.671	0.622	0.419	0.698	0.697	0.785	0.727	0.772	0.640	0.485	0.386
y_2	0.685	0.526	0.532	0.399	0.640	0.778	0.715	0.688	0.641	0.675	0.583	0.640	0.665
	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{20}	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}
y_1	0.562	0.078	0.760	0.746	0.109	0.646	0.353	0.029	0.509	0.575	0.876	0.745	0.237
y_2	0.082	0.056	0.399	0.558	0.117	0.622	0.165	0.336	0.656	0.684	0.796	0.787	0.348

index is usually used as an evaluation index to measure; its calculation formula is as follows:

$$GI_m = 1 - \sum_{k=1}^{|K|} p_{mk}^2, \tag{4}$$

where K represents the category and p_{mk} represents the proportion of category k in node m.

Then, the importance of feature x_j at node m is as follows:

$$\mathrm{VIM}_{jm}^{(\mathrm{Gini})} = GI_m - GI_l - GI_r, \tag{5}$$

where GI_l and GI_r, respectively, represent the Gini index of the two new nodes after branching.

If the node of feature x_j in the decision tree κ is set M, then the importance of feature x_j in the tree$\kappa$is as follows:

$$VIM_{\kappa j}^{(\mathrm{Gini})} = \sum_{m \in M} VIM_{jm}^{(\mathrm{Gini})}. \tag{6}$$

Assuming that there are J trees in the Random Forest, the importance of feature x_j throughout the Random Forest is as follows:

$$\mathrm{VIM}_{j}^{(Gini)} = \frac{1}{J} \sum_{j=1}^{J} \mathrm{VIM}_{\kappa j}^{(Gini)}. \tag{7}$$

We get the importance of 26 characteristics that affect liquid production and water content, as shown in Table 2.

We obtained the correlation coefficient and the importance ranking based on Random Forest feature selection and then added them to make a comprehensive comparison and to obtain the importance ranking of variables affecting liquid production and water content. The results are shown in Table 3.

3. Establishment and Verification of the Input and Output Model of the Reservoir System Based on Data-Driven

The steam stimulation oil recovery process is composed of a steam injection system, reservoir system, and lifting system. They perform the steam injection, soaking, and production, as shown in Figure 1. The reservoir system is the hub of the entire oil production system, which directly affects the energy consumption and system efficiency of steam injection and lifting systems. At the same time, due to the complexity of heavy oil formation conditions, it is very difficult to study the reservoir system from the perspective of the mechanism. Therefore, this paper explores the flow law of steam in the formation through data-driven to further improve the mining effect of steam stimulation. We convert the steam injection data from the boiler outlet to the bottom of the well through a simplified mechanism model, as shown in Figure 2. This paper assumes that only the steam dryness and steam pressure change during the steam flow process, while steam quantity and steam flow rate remain unchanged.

3.1. Calculation of Steam Pressure and Steam Dryness at the Bottom of the Steam Injection Well. This paper uses the steam injection wellhead and bottom hole as nodes to couple surface steam pipeline flow, steam injection wellbore flow, and formation flow. To explore the complex formation flow law, firstly, we convert the field data from the boiler outlet to the bottom of the well through a simplified mechanism, as shown in Figure 1. Secondly, we explore the formation flow law through data-driven, to predict the heavy oil steam stimulation production. This paper assumes that only the steam dryness and steam pressure change during the steam flow process, and the other injection and production parameters remain unchanged.

3.1.1. Steam Dryness Change of the Steam Pipeline. We make the following assumptions [2]:

(1) The pressure loss when steam flows in the pipeline is not considered

(2) The steam temperature and atmospheric temperature are fixed

(3) There is an insulating layer outside the steam pipeline

Since reaching the wellhead is still saturated steam and we ignore the change of pressure, its temperature is constant. At the same time, we do not consider the change of kinetic energy and potential energy, but only consider the change of steam internal energy. Then, the wellhead dryness can be calculated by the energy balance principle. We have

$$q_l \cdot L = i_s \left\{ \left[x_g \cdot h_s + \left(1 - x_g\right) h_\omega \right] - \left[x_\omega h_s + (1 - x_\omega) h_\omega \right] \right\}. \tag{8}$$

The dryness loss of the steam pipeline is as follows:

$$\Delta x_{gx} = \frac{q_l \cdot L}{i_s \cdot \left(h_s - h_w\right)}. \tag{9}$$

3.1.2. Steam Pressure Change in the Steam Injection Wellbore. We make the following assumptions [2]:

Table 2: Characteristic importance based on the Random Forest Gini index.

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}
y_1	0.126	0.135	0.186	0.094	0.167	0.008	0.113	0.006	0.096	0.094	0.126	0.178	0.167
y_2	0.169	0.087	0.167	0.172	0.005	0.135	0.124	0.107	0.085	0.167	0.091	0.134	0.142
	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{20}	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}
y_1	0.001	0.008	0.124	0.148	0.001	0.007	0.001	0.001	0.183	0.139	0.009	0.191	0.002
y_2	0.007	0.002	0.098	0.109	0.002	0.136	0.003	0.014	0.009	0.008	0.126	0.135	0.016

Table 3: Ranking of characteristic variables affecting liquid production and water content.

Ranking	Parameter	Ranking	Parameter
1	Cumulative water production	2	Cumulative oil production
3	Porosity	4	Initial formation temperature
5	Well spacing density	6	Reservoir type
7	Dynamic reserve	8	Primitive oil-bearing saturability
9	Oil area	10	Original formation pressure
11	Well distance	12	Surface crude oil viscosity
13	Reservoir effective thickness	14	Permeability
15	Production time	16	Annual turnover
17	Startup well number	18	Reservoir buried depth
19	Net total ratio of oil layers	20	Edge-bottom water
21	Produced degree	22	Dynamic liquid surface
23	Soaking time	24	Stroke times
25	Stroke	26	Well pattern mode

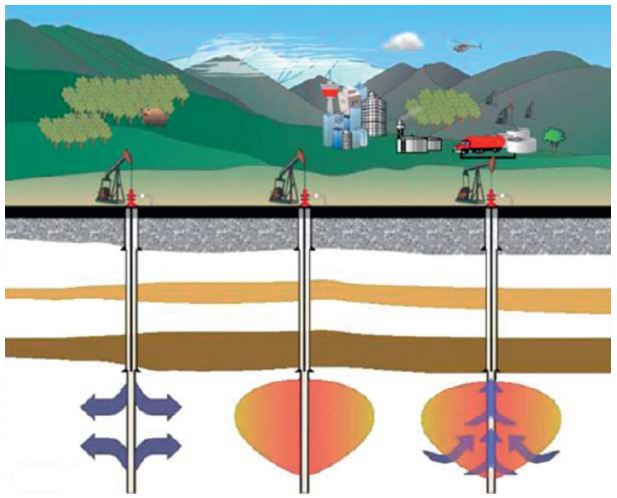

Figure 1: Steam stimulation oil recovery process [31].

(1) The steam injection rate, steam pressure, and steam quality of the wellhead remain unchanged

(2) We assume that the heat transfer from the oil well to the cement ring is one-dimensional stable, and the heat transfer from the cement ring to the formation is one-dimensional unstable heat transfer and ignores the heat transfer along the well depth direction

(3) We consider pressure changes in the wellbore

(4) We assume that the thermal conductivity of the formation is constant

This paper only considers the case of vertical injection wells. Since saturated steam is injected into the well, it becomes a two-phase flow of water and vapor. Therefore, according to the pressure balance equation, the pressure drop formula is expressed as follows:

$$\frac{dP}{dZ} = \rho_m g - \tau_f - \frac{\rho_m v dv}{dZ}. \tag{10}$$

We obtain the steam pressure change of the steam injection wellbore as follows:

$$\Delta P_{jt} = \frac{\rho_m g - \tau_f}{1 - \left(\left(i_s q_g\right)/\left(A_P^2 \cdot P\right)\right)} \cdot \Delta Z. \tag{11}$$

Considering the limitation of the article content, the proof process is shown in Appendix A.

3.1.3. Steam Dryness Change in the Steam Injection Wellbore. In unit time, the heat loss on the length dZ of the wellbore is dQ. According to the assumptions in Section 3.1.2, we have

$$\frac{dQ}{dZ} = 2\pi r_2 U_2 \left(T_s - T_h\right). \tag{12}$$

The heat loss of the wellbore will inevitably lead to a decrease in saturated steam energy, which will result in a decrease in steam dryness. We have

$$\frac{dQ}{dZ} = -i_s \frac{dh_m}{dZ} - i_s \frac{d}{dZ}\left(\frac{v^2}{2}\right) + i_s g. \tag{13}$$

Furthermore,

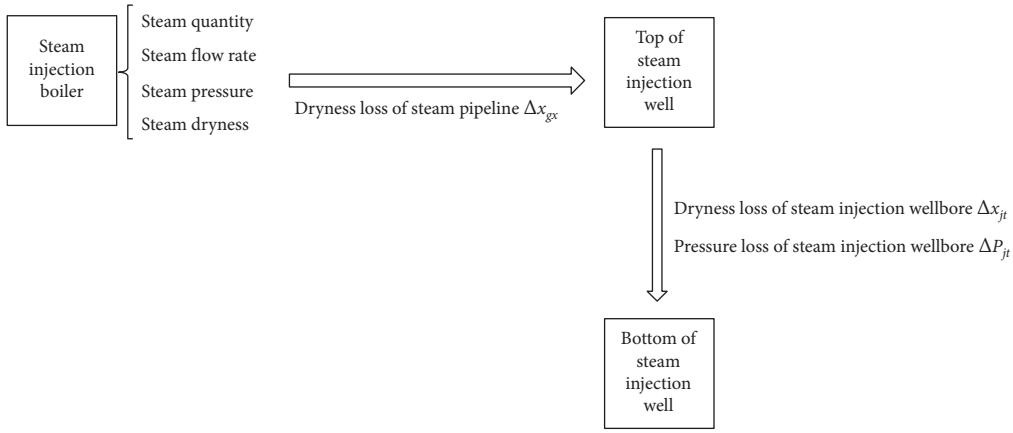

Figure 2: Steam flow process.

$$i_s(h_s - h_w)\frac{dx}{dZ} + i_s\left[\left(\frac{dh_s}{dP} - \frac{dh_w}{dP}\right)\frac{dP}{dZ}\right]x + \frac{dQ}{dZ} + i_s\frac{dh_w}{dP}\frac{dP}{dZ} + \frac{i_s^3}{A^2\rho}\frac{d}{dZ}\left(\frac{1}{\rho}\right) - i_s g = 0. \tag{14}$$

We make the following transformation:

$$\begin{aligned} C_1 &= i_s(h_s - h_w), \\ C_2 &= i_s\left[\left(\frac{dh_s}{dP} - \frac{dh_w}{dP}\right)\frac{dP}{dZ}\right], \\ C_3 &= \frac{dQ}{dZ} + i_s\frac{dh_w}{dP}\frac{dP}{dZ} + \frac{i_s^3}{A^2\rho}\frac{d}{dZ}\left(\frac{1}{\rho}\right) - i_s g. \end{aligned} \tag{15}$$

Therefore, the solution of equation (14) is as follows:

$$x = e^{-(C_2/C_1)Z}\left(-\frac{C_3}{C_2}e^{-(C_2/C_1)Z} + x_w + \frac{C_3}{C_2}\right). \tag{16}$$

We obtain the following dryness loss of the steam injection wellbore:

$$\Delta x_{jt} = x_w - x. \tag{17}$$

Considering the limitation of the article content, the proof process is shown in Appendix B. See Appendix C for parameter description.

3.2. Introduction and Evaluation of the Data-Driven Model. According to the importance ranking results of the characteristics affecting the liquid production and water content in Section 2.2, this section is based on five typical machine learning algorithms of *N*-Neighbours, Linear Regression, Random Forest, AdaBoost, and Support Vector Regression to predict the liquid production and water content of heavy oil steam stimulation, and select the optimal prediction model and the optimal number of features suitable for the problem samples in this paper. In order to evaluate the prediction effect of the model, we use the R^2 (determination coefficient) of the model on the liquid production and water content as the measurement standard. The larger the R^2, the better the model accuracy. The formula for R^2 is as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{j}\left(x_{j,c} - x_{j,p}\right)^2}{\sum_{i=1}^{j}\left(x_{j,p} - x_{j,a}\right)^2}, \tag{18}$$

where $x_{j,c}$ is the actual observation value, $x_{j,p}$ is the predicted value, and $x_{j,a}$ is the average value of the actual observation value.

For the 780 groups of samples sorted out in Section 2.2, we used the above five typical machine learning algorithms to conduct ten-fold cross validation on liquid production and water content, and the average value of R^2 of cross-validation results was used as the estimation of algorithm accuracy. The effects of the feature number on the determination coefficients of liquid production and water content are shown in Figures 3 and 4.

It can be seen that when the number of features is 24, the prediction accuracy of liquid production and water content based on the Random Forest algorithm is the highest, which are 86% and 83%, respectively. At this time, the determination coefficients of the five algorithms for liquid production and water content are shown in Table 4.

3.3. Model Validation. In order to further verify the accuracy of the model after adding dryness samples, we randomly selected two blocks (A and B) from 97 heavy oil blocks and used the established model to simulate the influence of steam quantity, bottom-hole steam pressure, and bottom-hole steam dryness on oil production and liquid production by the control variable method. The results are shown in Figures 5–7.

According to Figure 5, we can see that the oil production and liquid production increase with the increase of steam quantity, but the rising range gradually decreases, which is consistent with the actual change law.

According to Figure 6, we can see that the oil production and liquid production first increase with the increase of bottom-hole steam pressure and then gradually decrease after a "peak" appears.

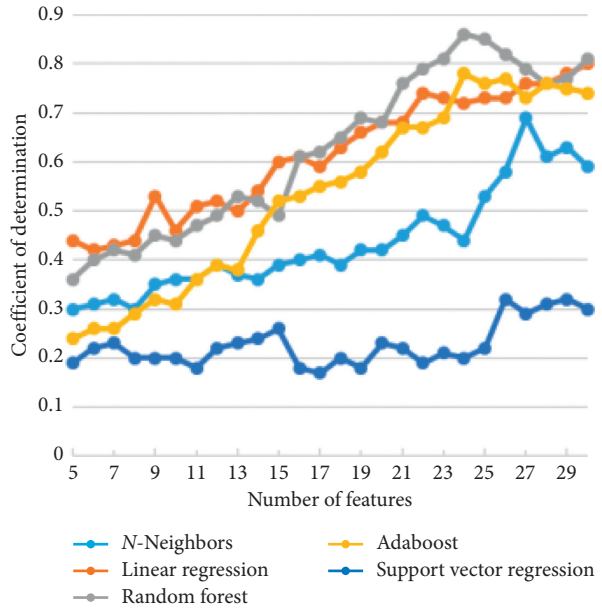

Figure 3: The effect of the feature number on liquid production.

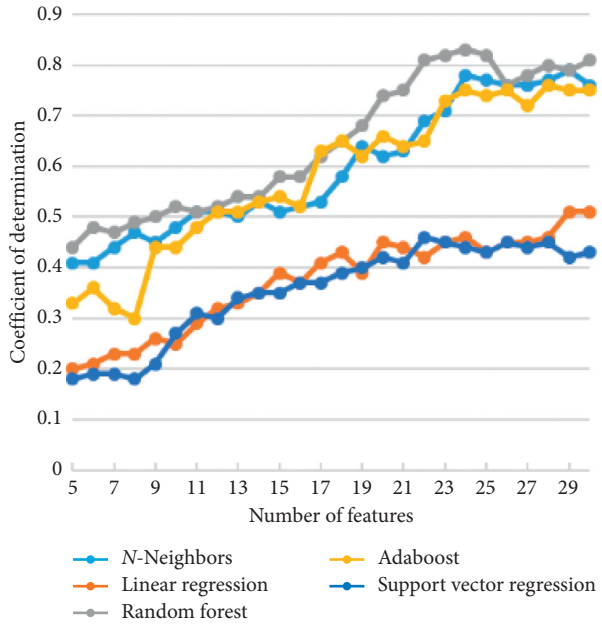

Figure 4: The effect of the feature number on water content.

According to Figure 7, it can be seen that, with the increase of bottom-hole steam dryness, the oil production and liquid production are gradually reduced, which is inconsistent with the actual changes. The reason for the poor consistency is that the actual data indicators fluctuate slightly, which leads to insufficient sample diversity and weak generalization ability after training.

4. Establishment and Verification of the Input and Output Model of the Oil Reservoir System Based on Hybrid Data-Driven

The essence of training the model through field data is function fitting, and the fitting function has no clear direction, as shown in Figure 8(a). If the variation range of

Table 4: Coefficient of determination of liquid production and water content.

Algorithm	R^2 of liquid production	R^2 of water content
N-Neighbors	0.44	0.78
Linear Regression	0.72	0.46
Random Forest	0.86	0.83
AdaBoost	0.78	0.77
Support Vector Regression	0.20	0.44

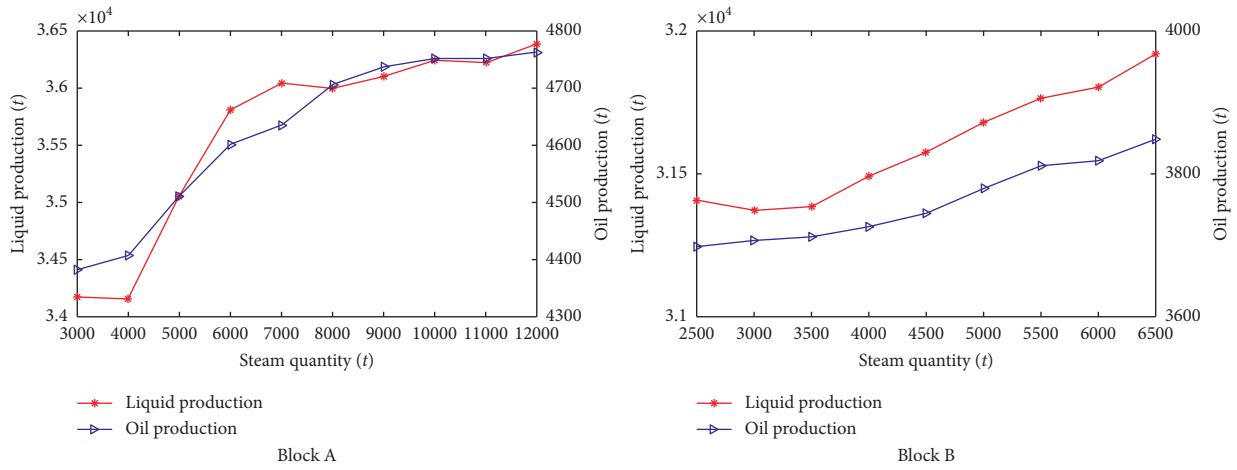

Figure 5: Effect of steam quantity on oil and liquid production.

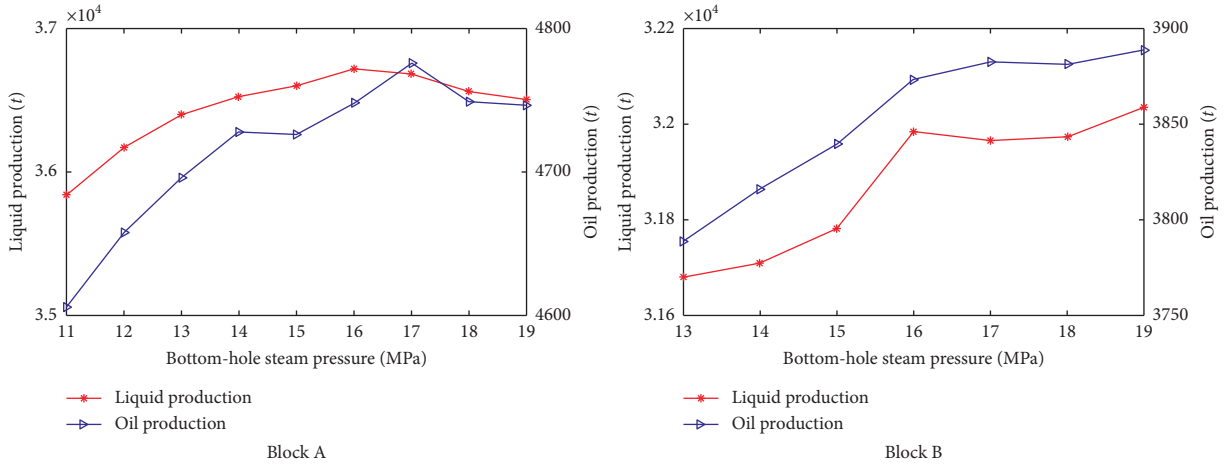

Figure 6: Effect of bottom-hole steam pressure on oil and liquid production. (a) Block *A*. (b) Block *B*.

parameters is small, the generalization ability of the model is weak and there may be overfitting. When predicting simply based on the mechanism model, it is essentially an abstract description of physical laws, as shown in Figure 8(b). Although the generalization ability of the model is strong, because the theoretical basis is the ideal model, the results are not necessarily consistent with the actual situation. Therefore, this paper samples the mechanism model and combines it with the field data to train the model. In this way, it can implicitly and automatically realize the parameter adjustment and fitting work that originally required a large amount of manual operation during the machine learning training process and improve the fitting accuracy. It can also artificially adjust the parameters of mechanism simulation to increase the data diversity and improve the generalization ability of the training model, which is conducive to the reliability of the established prediction model, as shown in Figure 8(c).

4.1. Introduction and Evaluation of the Hybrid Data-Driven Model. In Section 3.3, the effect of steam dryness on liquid production and water content is inconsistent with the actual change. Therefore, in this section, we sample the mechanism

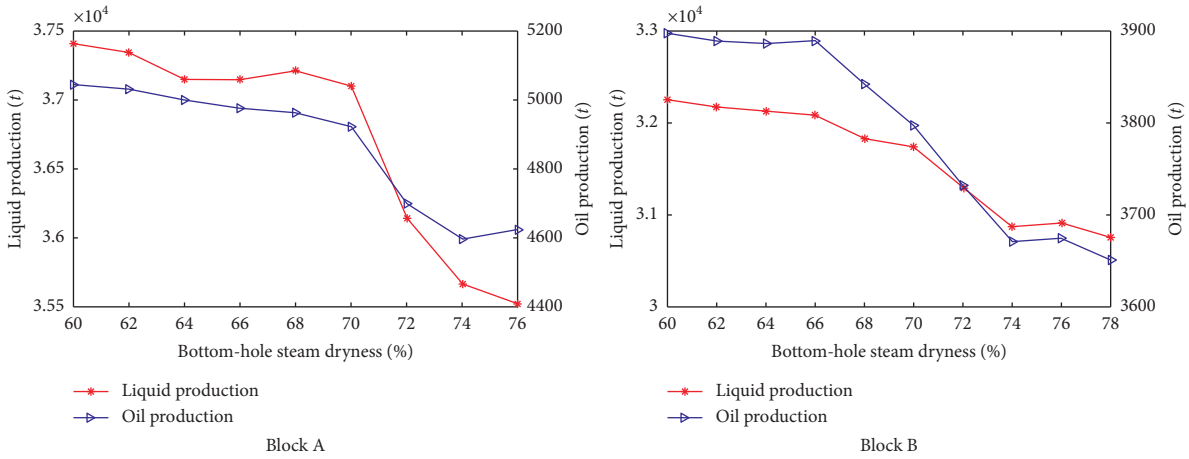

Figure 7: Effect of bottom-hole steam dryness on oil and liquid production. (a) Block *A*. (b) Block *B*.

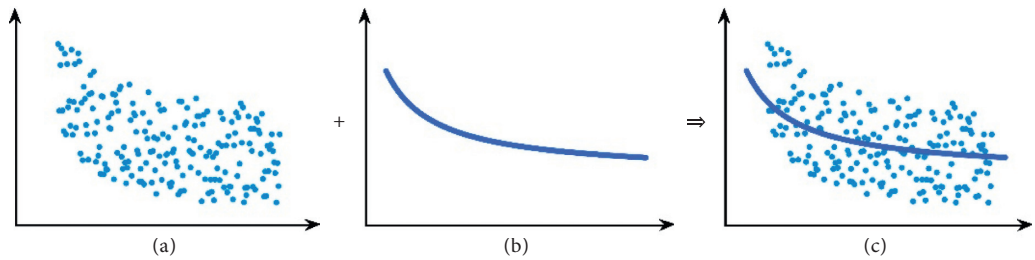

Figure 8: Hybrid data-driven process.

model by reservoir numerical simulation to increase the diversity of steam dryness samples and add them to the field data samples. To verify whether the model accuracy is improved after increasing the number of samples, we select the number of features as 24 and then use the above five typical machine learning algorithms to re-predict the liquid production and water content. The determination coefficients of the five algorithms for liquid production and water content are shown in Table 5.

According to Table 5, we can see that the prediction accuracy of liquid production and water content based on the Random Forest algorithm is the highest, which are 88% and 85%, respectively. At the same time, compared with Tables 4 and 5, we found that, after sampling the mechanism model and combining it with the field data, only the fitting effect of the water content prediction model based on AdaBoost and the liquid production prediction model based on Support Vector Regression did not change, while the fitting effect of the other models was improved.

4.2. Model Validation. In order to further verify the accuracy of the model after adding dryness samples, we used a new prediction model for blocks *A* and *B* to simulate the influence of steam quantity, bottom-hole steam pressure, and bottom-hole steam dryness on oil production and liquid production by the control variable method. The results are shown in Figures 9–11.

Table 5: Coefficient of determination of liquid production and water content.

Algorithm	R^2 of liquid production	R^2 of water content
N-Neighbors	0.49	0.80
Linear Regression	0.76	0.47
Random Forest	0.88	0.85
AdaBoost	0.80	0.77
Support Vector Regression	0.20	0.45

According to Figure 9, we can see that the oil production and liquid production increase with the increase of steam quantity, but the rising range gradually decreases. Eventually, it tends to be flat, which is consistent with the actual change law.

According to Figure 10, we can see that the oil production and liquid production first increase and then decrease with the increase of bottom hole pressure, which is consistent with the actual change law.

Figure 11 shows that the oil production and liquid production increase with the increase of steam dryness, but the rising range gradually decreases. It is consistent with the actual change law. At the same time, compared with Figure 7, we can see that the generalization ability of the algorithm has been improved, which lays the foundation for further exploring the deep learning algorithm based on field data and surrogate model.

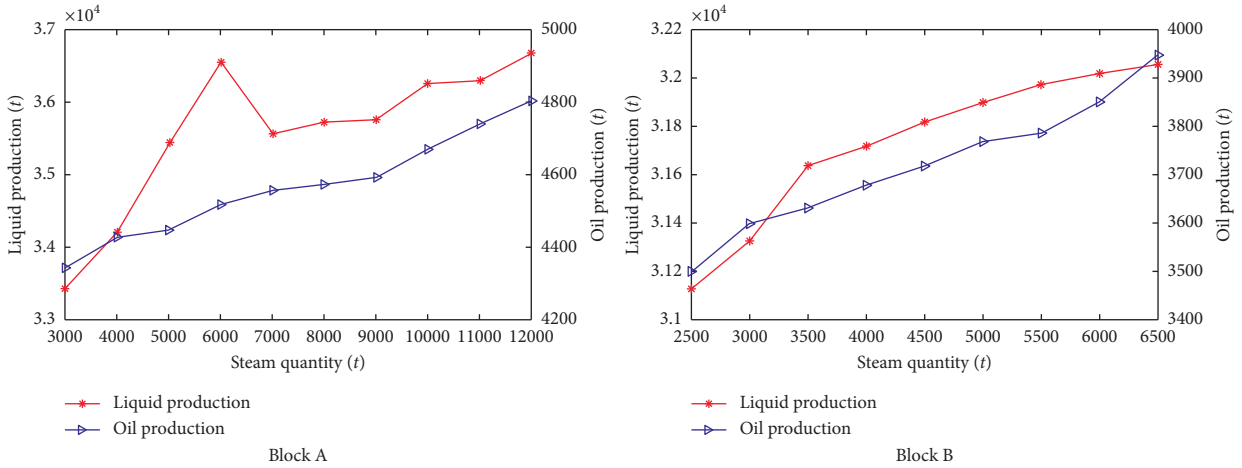

Figure 9: Effect of steam quantity on oil and liquid production. (a) Block *A*. (b) Block *B*.

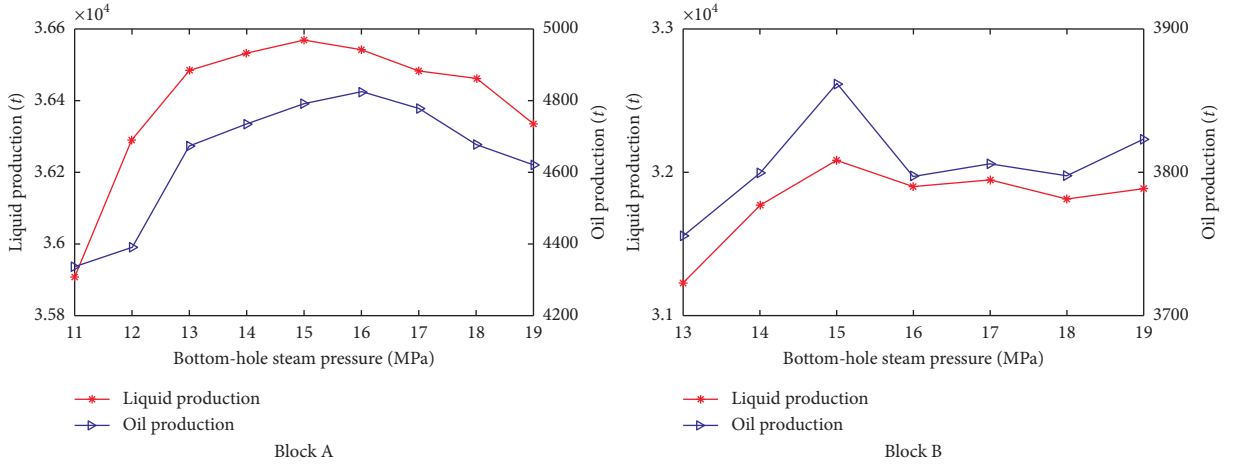

Figure 10: Effect of bottom-hole steam pressure on oil and liquid production. (a) Block *A*. (b) Block *B*.

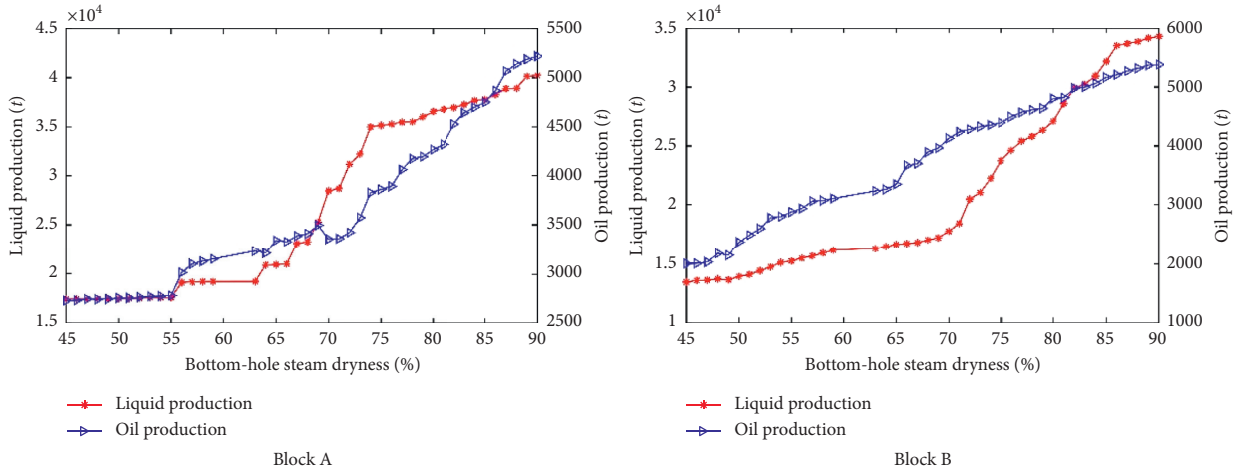

Figure 11: Effect of bottom-hole steam dryness on oil and liquid production. (a) Block *A*. (b) Block *B*.

Table 6: Parameter description.

Letter	Physical quantity	Unit
h_s	The enthalpy of saturated steam at a certain pressure	kcal/kg
h_w	The enthalpy of saturated water under a certain pressure	kcal/kg
x_g	Steam dryness of boiler outlet	Decimal
x_w	Steam dryness of steam injection wellhead	Decimal
L	Length of pipeline	m
i_s	Saturated steam mass flow rate	kg/h
q_l	Heat loss per unit time and unit length of steam pipeline	kcal/(h·m)
P	The pressure at a point in the wellbore	kgf/m^2
Z	Well depth	m
ρ_m	The density of the saturated steam mixture	kg/m^3
G	Acceleration of gravity	m/s^2
τ_f	Friction loss gradient	(Kgf/m^2)/m
V	The flow rate of the saturated steam mixture	m/h
r_2	The outer radius of the inner pipe	m
U_2	Overall heat transfer coefficient	kcal/(h·m^2·°C)
T_s	Steam temperature	°C
T_h	Outer-edge temperature of cement ring	°C
q_g	The volume flow of steam	m^3/h
A_p	Pipe cross-sectional area	m^2

5. Conclusions

(1) Based on previous studies, this paper conducts a coupled study on surface steam pipeline flow, steam injection wellbore flow, and formation flow based on data-driven. This provides a new idea for the prediction of heavy oil steam stimulation production and a theoretical basis for further formulating scientific and reasonable development plans.

(2) Based on the correlation coefficient and Random Forest feature selection, this paper ranks the features that affect liquid production and water content in importance.

(3) For a heavy oil field in eastern China, we compared the field data through five typical machine learning algorithms and selected the optimal prediction model and the optimal number of features suitable for the sample problem in this article, but the generalization ability of the prediction model is poor. Therefore, we sampled the mechanism model, increased the diversity of steam dryness samples, and trained the new samples again. It is found that the previously obtained optimal prediction model not only improved the accuracy but also the generalization ability of the model.

It is feasible to study the steam stimulation production of heavy oil from the perspective of mechanism model and field data in this paper. However, this paper still has some limitations. Firstly, there is a certain error in the collection of field data, which may affect our results. Secondly, the lack of samples leads to weak generalization ability after training. Thirdly, the content of steam stimulation is complex, and many factors are affecting the production of steam stimulation. In the selection of features, this paper did not consider the influence of heavy oil lifting methods and viscosity reduction technology.

Appendix

A. Calculation of the Bottom-Hole Steam Pressure Based on the Mechanism Model

Based on the assumptions in Section 3.1.2, we know that wellbore pressure drop is the sum of friction energy loss, potential energy change, and kinetic energy change. According to the pressure balance equation, the pressure drop formula of vertical injection wells can be expressed as

$$\frac{\mathrm{d}P}{\mathrm{d}Z} = \rho_m g - \tau_f - \frac{\rho_m v \mathrm{d}v}{\mathrm{d}Z}. \tag{A.1}$$

The change of kinetic energy has obvious significance only in the case of the fog flow. For the fog flow, the gas volume flow is much larger than the liquid volume flow. Therefore, according to the law of ideal gas, we have

$$\begin{aligned} v &= \frac{i_s}{\rho_m A_P}, \\ \rho_m v \mathrm{d}v &= \rho_m \cdot \frac{i_s}{\rho_m A_P} \cdot \mathrm{d}\left(\frac{i_s}{\rho_m A_P}\right) = \frac{i_s^2}{A_P^2}\mathrm{d}\left(\frac{1}{\rho_m}\right). \end{aligned} \tag{A.2}$$

At the same time,

$$PV = RT,$$

$$\rho = \frac{M}{V},$$

$$\frac{1}{\rho} = \frac{RT}{PM},$$

$$\mathrm{d}\left(\frac{1}{\rho_m}\right) = \frac{RT}{M}\mathrm{d}\left(\frac{1}{P}\right) = -\frac{RT}{MP^2}\mathrm{d}P = -\frac{1}{\rho_m P}\mathrm{d}P. \tag{A.3}$$

So,

$$\frac{\rho_m v \mathrm{d}v}{\mathrm{d}Z} = -\frac{i_s q_g}{A_P^2 \cdot P}\frac{\mathrm{d}P}{\mathrm{d}Z}. \tag{A.4}$$

We replace equation (A.4) with equation (A.1) and obtain the following changes in steam pressure in the steam injection wellbore:

$$\Delta P_{jt} = \frac{\rho_m g - \tau_f}{1 - \left(\left(i_s q_g\right)/A_P^2 \cdot P\right)} \cdot \Delta Z. \tag{A.5}$$

B. Calculation of the Bottom-Hole Steam Dryness Based on the Mechanism Model

In unit time, the heat loss on the length dZ of the wellbore is dQ. Under the assumptions in Section 3.1.2, we have

$$\frac{\mathrm{d}Q}{\mathrm{d}Z} = 2\pi r_2 U_2 \left(T_s - T_h\right). \tag{B.1}$$

The heat loss of the wellbore will inevitably lead to a decrease in saturated steam energy, which will result in a decrease in steam dryness. We have

$$\frac{\mathrm{d}Q}{\mathrm{d}Z} = -i_s\frac{\mathrm{d}h_m}{\mathrm{d}Z} - i_s\frac{\mathrm{d}}{\mathrm{d}Z}\left(\frac{v^2}{2}\right) + i_s g. \tag{B.2}$$

Among them,

$$h_m = (1-x)h_w + xh_s,$$

$$\frac{\mathrm{d}h_m}{\mathrm{d}Z} = (1-x)\frac{\mathrm{d}h_w}{\mathrm{d}Z} + x\frac{\mathrm{d}h_s}{\mathrm{d}Z} + h_w\frac{\mathrm{d}(1-x)}{\mathrm{d}Z} + h_s\frac{\mathrm{d}x}{\mathrm{d}Z}. \tag{B.3}$$

Here,

$$\frac{\mathrm{d}h_w}{\mathrm{d}Z} = \frac{\mathrm{d}h_w}{\mathrm{d}P}\cdot\frac{\mathrm{d}P}{\mathrm{d}Z},$$

$$\frac{\mathrm{d}h_s}{\mathrm{d}Z} = \frac{\mathrm{d}h_s}{\mathrm{d}P}\cdot\frac{\mathrm{d}P}{\mathrm{d}Z}. \tag{B.4}$$

At the same time,

$$\frac{\mathrm{d}}{\mathrm{d}Z}\left(\frac{v^2}{2}\right) = \frac{\mathrm{d}}{\mathrm{d}Z}\left(\frac{i_s^2}{2\rho^2 A^2}\right) = \frac{i_s^2}{\rho^2}\frac{1}{\rho}\frac{\mathrm{d}}{\mathrm{d}Z}\left(\frac{1}{\rho}\right). \tag{B.5}$$

So,

$$i_s\left(h_s - h_w\right)\frac{\mathrm{d}x}{\mathrm{d}Z} + i_s\left[\left(\frac{\mathrm{d}h_s}{\mathrm{d}P} - \frac{\mathrm{d}h_w}{\mathrm{d}P}\right)\frac{\mathrm{d}P}{\mathrm{d}Z}\right]x + \frac{\mathrm{d}Q}{\mathrm{d}Z} + i_s\frac{\mathrm{d}h_w}{\mathrm{d}P}\frac{\mathrm{d}P}{\mathrm{d}Z} + \frac{i_s^3}{A^2\rho}\frac{\mathrm{d}}{\mathrm{d}Z}\left(\frac{1}{\rho}\right) - i_s g = 0. \tag{B.6}$$

We make the following transformation:

$$C_1 = i_s\left(h_s - h_w\right),$$

$$C_2 = i_s\left[\left(\frac{\mathrm{d}h_s}{\mathrm{d}P} - \frac{\mathrm{d}h_w}{\mathrm{d}P}\right)\frac{\mathrm{d}P}{\mathrm{d}Z}\right],$$

$$C_3 = \frac{\mathrm{d}Q}{\mathrm{d}Z} + i_s\frac{\mathrm{d}h_w}{\mathrm{d}P}\frac{\mathrm{d}P}{\mathrm{d}Z} + \frac{i_s^3}{A^2\rho}\frac{\mathrm{d}}{\mathrm{d}Z}\left(\frac{1}{\rho}\right) - i_s g. \tag{B.7}$$

Therefore, the solution of equation (12) is as follows:

$$x = e^{-\left(C_2/C_1\right)Z}\left(-\frac{C_3}{C_2}e^{-\left(C_2/C_1\right)Z} + x_w + \frac{C_3}{C_2}\right). \tag{B.8}$$

We obtain the following dryness loss of the steam injection wellbore:

$$\Delta x_{jt} = x_w - x. \tag{B.9}$$

C. Parameter Description

Table 6

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors are grateful to all of the anonymous reviewers for their careful reading and valuable comments on how to improve this work. This work was supported by the National Natural Science Foundation of China (no. 11601451), the

International Cooperation Program of Chengdu City (no. 2020-GH02-00023-HZ) and the Scientific Research Project of Sinopec Corporation "Heavy oil steam stimulation low-consumption and high-efficiency development of overall optimization technology" (no: P19018-5).

References

[1] M. Zhang, J. Zhang, F. Ye, Y. Liu, and H. B. Huang, "Prediction model of economic oil steam ratio limit for heavy-oil stimulation," *Special Oil & Gas Reservoirs*, vol. 27, no. 3, pp. 121–124, 2020.

[2] Y. M. Chen, *Steam Injection Thermal Oil Recovery*, Petroleum University Press, Chengdu, China, 1996.

[3] J. W. Marx and R. H. Langenheim, "Reservoir Heating by Hot Fluid Injection Petroleum Transactions," *Transactions of the AIME*, vol. 216, pp. 312–315, 1959.

[4] T. C. Boberg, *Thermal Oil Recovery Engineering Method*, Petroleum Industry Press, Chengdu, China, 1980.

[5] J. Hou and Y. M. Chen, "Improved steam soak predictive model," *Petroleum Exploration and Development*, vol. 24, no. 3, pp. 53–56, 1997.

[6] J. Zheng, G. X. Chen, and P. C. Liu, "A new analytical model for productivity prediction in steam soak," *Journal of Oil and Gas Technology*, vol. 33, no. 5, pp. 111–114, 2011.

[7] J. Yang, X. F. Li, Z. X. Chen, J. Tian, L. Huang, and X. G. Liu, "A productivity prediction model for cyclic steam stimulation in consideration of non-Newtonian characteristics of heavy oil," *Acta Petrolei Sinica*, vol. 38, no. 1, pp. 84–90, 2017.

[8] G. H. Bruce, D. W. Peaceman, H. H. Rachford, and J. D. Rice, "Calculations of unsteady-state gas flow through porous media," *Journal of Petroleum Technology*, vol. 5, no. 3, pp. 79–92, 1953.

[9] H. L. Stone, "Iterative solution of implicit approximations of multidimensional partial differential equations," *SIAM Journal on Numerical Analysis*, vol. 5, no. 3, pp. 530–558, 1968.

[10] K. H. Coats, W. D. George, C. Chu, and B. E. Marcum, "Three-dimension simulation of steam flooding," *Society of Petroleum Engineers Journal*, vol. 14, no. 6, pp. 573–592, 1974.

[11] J. Hou, K. Zhou, H. Zhao, X. Kang, S. Wang, and X. Zhang, "Hybrid optimization technique for cyclic steam stimulation by horizontal wells in heavy oil reservoir," *Computers & Chemical Engineering*, vol. 84, pp. 363–370, 2016.

[12] Y. Wang, S. Ren, and L. Zhang, "Mechanistic simulation study of air injection assisted cyclic steam stimulation through horizontal wells for ultra heavy oil reservoirs," *Journal of Petroleum Science and Engineering*, vol. 172, pp. 209–216, 2019.

[13] E. H. Luo, Z. F. Fan, Y. L. Hu, L. Zhao, and B. Bo, "An efficient optimization framework of cyclic steam stimulation with experimental design in extra heavy oil reservoirs," *Energy*, vol. 192, pp. 1–19, 2020.

[14] W. Liu, W. Liu, and J. W. Gu, "Oil production prediction based on a machine learning method," *Oil Drilling & Production Technology*, vol. 42, no. 1, pp. 70–75, 2020.

[15] X. Ma, Y.-s. Hu, and Z.-b. Liu, "A novel kernel regularized nonhomogeneous grey model and its applications," *Communications in Nonlinear Science and Numerical Simulation*, vol. 48, pp. 51–62, 2017.

[16] X. Ma and Z.-B. Liu, "The kernel-based nonlinear multivariate grey model," *Applied Mathematical Modelling*, vol. 56, pp. 217–238, 2018.

[17] D. Y. Fan, H. Sun, J. Yao, K. Zhang, X. Yan, and Z. X. Sun, "Well production forecasting based on ARIMA-LSTM model considering manual operations," *Energy*, vol. 220, Article ID 119708, 2021.

[18] O. Akbilgic, D. Zhu, I. D. Gates, and J. A. Bergerson, "Prediction of steam-assisted gravity drainage steam to oil ratio from reservoir characteristics," *Energy*, vol. 93, pp. 1663–1670, 2015.

[19] S. Wang, Z. Chen, and S. Chen, "Applicability of deep neural networks on production forecasting in Bakken shale reservoirs," *Journal of Petroleum Science and Engineering*, vol. 179, pp. 112–125, 2019.

[20] B. Zeng, X. Ma, and M. Zhou, "A new-structure grey Verhulst model for China's tight gas production forecasting," *Applied Soft Computing*, vol. 96, Article ID 106600, 2020.

[21] D. Jia, H. Liu, J. Zhang et al., "Data-driven optimization for fine water injection in a mature oil field," *Petroleum Exploration and Development*, vol. 47, no. 3, pp. 674–682, 2020.

[22] Z. Zhong, Y. S. Alexander, Y. Y. Wang, and R. Bo, "Predicting field production rates for waterflooding using a machine learning-based proxy model," *Journal of Petroleum Science and Engineering*, vol. 194, Article ID 107574, 2020.

[23] L. Yu, X. Ma, W. Q. Wu, Y. Wang, and B. Zeng, "A novel elastic net-based NGBMC (1, model with multi-objective optimization for nonlinear time series forecasting," *Communications in Nonlinear Science and Numerical Simulation*, vol. 96, Article ID 105696, 2021.

[24] L. Yu, X. Ma, W. Q. Wu, X. W. Xiang, Y. Wang, and B. Zeng, "Application of a novel time-delayed power-driven grey model to forecast photovoltaic power generation in the Asia-Pacific region," *Sustainable Energy Technologies and Assessments*, vol. 44, Article ID 100968, 2021.

[25] F. Wang, "Research on K nearest neighbor algorithm based on class division and neighbor selection," M.Sc. Thesis, Xi'an University of Technology, Xi'an, China, 2020.

[26] W. Liu, W. D. Liu, J. Gu, and X. Shen, "Predictive model for water absorption in sublayers using a machine learning method," *Journal of Petroleum Science and Engineering*, vol. 182, Article ID 106367, 2019.

[27] Y. K. Chen, H. Zhao, Q. Zhang et al., "Development and application of a coupling method for well pattern and production optimization in unconventional reservoirs," *Journal of Circuits, Systems & Computers*, vol. 29, no. 7, Article ID 2050105, 2020.

[28] J. W. Gu, M. Zhou, Z. T. Li, X. J. Jia, and Y. Liang, "Oil well production forecast with long- short term memory network model based on data mining," *Special Oil & Gas Reservoirs*, vol. 26, no. 2, pp. 77–81, 2019.

[29] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, "Variable selection using random forests," *Pattern Recognition Letters*, vol. 31, no. 14, pp. 2225–2236, 2010.

[30] S. Y. Zhang, K. Yang, C. M. Xia, C. L. Jin, Y. L. Wang, and H. X. Yan, "Research on feature reduction and classification of pulse signal based on random forest," *Modernization of Traditional Chinese Medicine and Materia Medica-World Science and Technology*, vol. 22, no. 7, pp. 2418–2426, 2020.

[31] C. T. Frenette, M. Saeedi, and J. L. Henke, "Integrated economic model for evaluation and optimization of cyclic-steam-stimulation projects," *SPE Economics & Management*, vol. 8, pp. 11–22, 2016.