ABSTRACT: Recently, production optimization has gained increasing interest in the petroleum industry. The most computationally intensive and critical part of the production optimization process is the evaluation of the production function performed by the numerical reservoir simulator. Employing proxy models as a substitute for the reservoir simulator is proposed for alleviating this high computational cost. In this study, a new approach to construct adaptive proxy models for production optimization problems is proposed. An adaptive difference evolution algorithm (SaDE) optimized least-squares support vector machine (LSSVM) is used as an approximation function, while training is performed using a self-adaptive response surface experimental design (SaRSE). SaDE selects the optimal hyperparameters of LSSVM during the training process to improve the prediction accuracy of the proxy model. Cross-validation methods are used in the recursive training and network evaluation phases. The developed method is used to optimize the production of block gas reservoir models. Computational results confirm that the developed adaptive proxy model outperforms traditional regression methods. It is further verified that when the experimental data are updated, the alternative model still has high prediction accuracy when performing the objective function evaluation. The results show that the proposed proxy modeling approach enhances the entire optimization process by providing a fast approximation of the actual reservoir simulation model with better accuracy.

1. INTRODUCTION

The development of onshore oil and gas has made rapid progress over the past decade, primarily due to the development of unconventional resources, particularly shale gas.\textsuperscript{1-7} Although shale gas extraction has broken through the industrial capacity barrier, it is at the critical point of marginal benefits, making it difficult to achieve beneficial development. Shale gas production optimization is getting increasing attention in the oil industry. Production evaluation is one of the most critical and difficult processes in production optimization under complex geological and engineering conditions. Reasonable and accurate production evaluation is essential for production optimization.

Numerical reservoir simulation technology, developed by reservoir geologists, reservoir engineers, and reservoir numerical simulators, is a physics-driven modeling approach recognized as the standard decision-making approach in the petroleum industry.\textsuperscript{8} Numerical reservoir simulations can effectively simulate discrete fracture networks, fluid seepage, and key physical effects that could be important in shale reservoirs. Zhang et al.\textsuperscript{9} applied a porous system with bedrock grid to describe the flow of unsteady gas from bedrock to fracture interval and used numerical simulation software to simulate the effects of parameters such as porosity, permeability, bedrock-fracture coupling parameters, and fracture half-length spacing on shale gas production capacity. Hu et al.\textsuperscript{10} simulated the sensitivity of production capacity to parameters such as the number of fracture bars, fracture conductivity, and fracture spacing, which provided guidance for designing fracture construction plans for shale gas wells. Numerical simulations with higher accuracy have been frequently used as a powerful tool for engineering design and optimization. However, hundreds or thousands of reservoir simulations need to be run to perform optimization tasks. A single run of a simulated reservoir model consisting of thousands or even millions of grid blocks can take several hours. In addition, the large number of control parameters exacerbates this numerical simulation-based optimization design problem.

Essentially, the proxy model is a data mining tool. Advances in data mining, especially in machine learning, made it possible to create machine learning models that outperformed many...
existing computer models. Hassani et al.11 used the traditional quadratic, multiplicative, and radial basis function (RBF) models to approximate the cumulative outflow from the reservoir when drilling a new horizontal well, which significantly reduced the computational time. Schuetter et al.12 employed simple regression and other advanced methods such as random forests (RFs), support vector regression (SVR), gradient propulsion machines (GBM), and multi-dimensional kriging to build prediction models for production indicators. Zhang et al.13 proposed a new multipoint adaptive Gaussian process agent model for design domain (MAGPSM-IDD) and applied it to an integrated field and an actual reservoir, showing superior computational performance. Wang et al.14 employed a deep neural network (DNN) and Sobol global sensitivity analysis to analyze 2919 wells (2780 multistage hydraulically fractured horizontal wells and 139 straight wells) in the Bakken Formation. A thermal coding method was also employed to process the classified data. A reliable DNN model was built by evaluating Xavier initialization, exit techniques, and batch normalization. The results show that the proposed DNN model can integrate directly into existing hydraulic fracturing design programs. Brantson et al.15 applied back-propagation artificial neural network (BPANN), radial basis function neural network (RBNN), and generalized regression neural network (GRNN) as proxy models to predict the historical production decreasing trend of extralow porosity tight gas reservoirs and further validated the effectiveness of the proxy models.

Although many studies have shown the broad application of data mining techniques in various engineering disciplines, its application in petroleum engineering is still in its infancy. Moreover, the performance of production forecasting in unconventional shale reservoirs is not impressive.16−18 An important reason is that reservoir properties, including permeability, porosity, gas adsorption, gas saturation distribution, and reservoir pressure, as well as fracture stimulation design, will have a severe impact on shale gas production. More significantly, the prediction performance of machine learning algorithms is influenced a lot by their hyperparameters, such as learning rate, initial weights and thresholds, activation function, number of neurons, and the number of layers of a multilayer neural network, as well as the characteristics of the training samples. Hyperparameter tuning or optimization is the significant process of automatically testing different configurations for training a machine learning model. Prabusankarlal et al.19 combined the adaptive difference evolution algorithm (SaDE) with extreme learning machine (ELM), where the SaDE algorithm optimizes the hidden node learning parameters of ELM to improve the model prediction accuracy and generalization performance significantly.

We used a self-adaptive response surface experimental design as the data set. Response surface design is a statistical method that uses reasonable experimental design methods and obtains specific data through experiments. Using multiple quadratic regression equations to fit the functional relationship between factors and response values, we seek optimal process parameters by analyzing the regression equations to solve multivariate problems.20−22 If the training sample is limited, then the proxy model will be challenging to extract discriminative features automatically and guarantee regression accuracy. As the number of training samples increases, all of these estimates approach the true regression accuracy, which is the accuracy of a proxy designed with the full knowledge of the actual situation. This study will use a self-adaptive response surface design scheme, which generates training samples with specified accuracy. This adaptive accuracy model ensures that the training samples have actual reservoir characteristics, thus avoiding unsatisfactory learning results due to errors in the training samples. In addition, we introduce the least-squares support vector machine (LSSVM) as a proxy model. The LSSVM is functionalized by SaDE to optimally select the unknown parameters of the LSSVM and compare the results with other optimization algorithms (genetic algorithm (GA), particle swarm algorithm (PSO), gray wolf algorithm (GWO)) to generate a reasonable production proxy model. Our approach efficiently (in a short time) automatically searches for the best machine learning algorithm and hyperparameter values for a given machine learning problem.

2. CONSTRUCTING THE PROXY MODEL

2.1. Methodology of Least-Squares Support Vector Machines. The support vector machine (SVM) is a typical machine learning method based on statistical theory. It is effective in solving problems with a few samples, nonlinearity, and high dimensionality. With the advancement of the research, SVM is being widely used for complex nonlinear modeling problems in various fields. The main disadvantage of SVMs is the dimensional disaster problem. LSSVM is an SVM-based improvement type proposed by Suykens23 in 1999. The higher computational load of SVM is overcome by LSSVM, which uses a set of linear equations to solve the problem.

The basic principle of LSSVM is to classify the hyperplane that can correctly partition the data set and has the largest geometric interval for a given data set. As shown in Figure 1, \(\omega \cdot x + b = 0\) is the classification hyperplane. For a linearly divisible data set, there are infinitely many such hyperplanes, but the classification hyperplane with the largest geometric interval is the only one.

This study utilizes the regression form of LSSVM as the proxy model. LSSVM can transform linearly inseparable data into linearly separable data using kernel functions to map them to a high-dimensional space. The function of the kernel function in LSSVM is shown in Figure 2.

For the training sample set \(S = \{(x_i, y_i)\}_{i=1}^{n}\) with size \(n\), consisting of the input value \(x_i \in R\) and the corresponding output values \(y_i \in R\), its LSSVM regression model is determined as follows

\[
y = f(x) = \omega^T \varphi(x) + b
\]  

(1)
where $\omega$ is the weight vector, $b$ is the offset, and the nonlinear transformation $\varphi(x)$ is the mapping from the low-dimensional space to the high-dimensional space. According to the principle of structural risk minimization, in the LSSVM, the squared error loss function in the optimization targets is selected so that the regression problem is transformed into a quadratic optimization problem as follows

$$
\min \frac{1}{2} \omega^T \omega + \frac{1}{2} \gamma \sum_{i=1}^{n} e_i^2
$$

(2)

where $e_i$ is the slack variable and $\gamma$ is the penalty. The constraint condition is as follows

$$
y_i = \omega^T \varphi(x_i) + b + e_i
$$

(3)

To solve the optimization problem, the Lagrange function was introduced

$$
L = \frac{1}{2} \| \omega \|^2 + \frac{1}{2} \gamma \sum_{i=1}^{n} e_i^2 - \sum_{i=1}^{n} \alpha_i [(\omega^T \varphi(x_i)) + b + e_i]
$$

(4)

where $\alpha_i$ represents the Lagrange multiplier. The following formula could be obtained based on the Karush–Kuhn–Tucker (KKT) optimization conditions

$$
y_i = \sum_{k=1}^{n} (\alpha_k K(x_i, x_k)) + b + \frac{\alpha}{\gamma}\n$$

(5)

where $K(x_i, x_k)$ refers to the kernel function. In this study, the radial basis function was selected as the kernel function of the LSSVM regression model for its excellent generalization ability and wide-range convergence domain. Its expression is as follows

$$
K(x_i, x_k) = \exp\left(-\frac{\|x_i - x_k\|^2}{2\sigma^2}\right)
$$

(6)

where $\sigma$ refers to the kernel function width; $\gamma$ and $\sigma$ have a significant degree of impact on the learning and generalization ability of LSSVM models. Therefore, optimization algorithms (e.g., PSO, GA, SaDE, GWO) are generally applied to select the hyperparameters $\gamma$ and $\sigma$ of the LSSVM model comprehensively and optimally.

The quadratic optimization problem can be transformed into the problem of linear equation

$$
\begin{bmatrix}
0 & 1 & 0 & \cdots & 1 \\
1 & K(x_1, x_1) + \frac{1}{\gamma} & K(x_1, x_2) & \cdots & K(x_1, x_n) \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & K(x_n, x_1) & K(x_n, x_2) & \cdots & K(x_n, x_n) + \frac{1}{\gamma}
\end{bmatrix}
$$

(7)

The ultimate LSSVM regression function was built as follows

$$
y = f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x_k) + b
$$

(8)

2.2. Methodology of Parameter Optimization of the LSSVM. Differential evolutionary (DE) algorithm is an excellent swarm intelligence optimization algorithm after genetic algorithm (GA) and particle swarm algorithm (PSO). The DE algorithm has been extensively applied because of its simple structure, few control parameters, easy implementation using accurate number coding, and fast convergence speed.

2.2.1. Differential Evolution Algorithm. The basic procedures of the DE algorithm include initialization, mutation, crossover, and selection. Individuals are randomly generated in the search space by initialization; then, mutation and crossover generate new individuals and eventually select which individual will enter the next generation. This process will continue to repeat until its termination condition is reached.

2.2.1.1. Initialization. For minimizing the objective function

$$
\min_{x \in \mathbb{R}^D} f(x)
$$

concerning the parameter vector $x \in \mathbb{R}^D$, the population containing NP parameter vectors evolves toward the global minimum. The ith individual is as follows

$$
x_{i,G} = [x_{i,G}^1, x_{i,G}^2, \ldots, x_{i,G}^D]
$$

(9)

where $i = 1, 2, \ldots, NP$. $G$ refers to the number of evolutionary generations.

Initialize to generate a set of NP parameter vectors $x_{i,G}$ to cover the parameter space as much as possible

$$
x_{i,G} = x_{\min} + \text{rand}(0, 1) \cdot (x_{\max} - x_{\min})
$$

(10)

where $x_{\min} = [x_{\min}^1, x_{\min}^2, \ldots, x_{\min}^D]$ and $x_{\max} = [x_{\max}^1, x_{\max}^2, \ldots, x_{\max}^D]$ refer to the maximum and minimum parameters boundary, respectively.

2.2.1.2. Mutation. The existing vectors are converted into new mutant vectors by a self-organization scheme that takes the difference vectors of randomly selected population vectors. For each parameter vector $x_{i,G}$ of the current generation, a mutation vector $V_{i,G}$ is generated by some mutation strategy. Several popular mutation strategies (MT) are listed as follows $MT_1$: DE/rand/1.

$$
V_{i,G} = x_{i,G}^m + F \cdot (x_{r1,G}^m - x_{r2,G}^m)
$$

(11)
vector-based strategies are often not effective for multipeaked issues, leading to premature strategy, but they require a higher computational cost. The perturbation parameter vectors, a crossover approach was usually exhibits a slow convergence rate. The two-di

above, MT1 has a strong exploration capability and is suitable fi

selection procedure is performed using the from the target vector and are randomly generated within the range: \[1, 2, 3\].

will be generated according to the following crossover equation

\[ x_{G+1,G} = x_{G,G} + F \cdot (x_{G,best,G} - x_{G,G}) + F \cdot (x_{G,rand,G} - x_{G,G}) \]

where \(F\) refers to the positive amplification factor, which controls the scaling of the difference vector and is usually chosen in the range \(0 \leq F \leq 2\). \(K\) refers to the control parameter, which is randomly generated in the range \(0 \leq K \leq 1\). Indices \(r\) are \(r_1, r_2, r_3, r_4\) and \(r_5\) mutually exclusive integers and are randomly generated within the range: \([1, 2, \ldots, NP]\).

Among the different vector generation strategies mentioned above, MT1 has a strong exploration capability and is suitable for solving multimodal problems. However, this strategy usually exhibits a slow convergence rate. The two-difference vector-based strategies “MT2” and “MT3” have better perturbation effects than the one-difference vector-based strategy, but they require a higher computational cost. The MT3 strategy relies on the best solution found so far, has a fast convergence rate, and performs well for single-peaked problems. However, this strategy will likely fall into a local optimum for multi-peaked issues, leading to premature convergence. The “MT4” strategy is rotationally invariant. The results show that the MT4 strategy is effective for solving multiobjective optimization problems.

2.2.1.3. Crossover. To increase the diversity of the perturbation parameter vectors, a crossover approach was used. At generation \(G\), concerning each mutation vector \(V_{G} = [V_{1,G}, V_{2,G}, \ldots, V_{NP,G}]\), the trial vector \(U_{G} = [U_{1,G}, U_{2,G}, \ldots, U_{NP,G}]\) will be generated according to the following crossover equation

\[ U_{G} = \begin{cases} V_{G} & \text{if } (\text{rand}_{j} \leq \text{CR}) \text{ or } (j = j_{\text{rand}}) \\ x_{G} & \text{otherwise} \end{cases} \]

where CR refers to the crossover rate, which controls the fraction of parameter values obtained from the mutation vector with positive values between [0,1], \(\text{rand}_{j}\) refers to the \(j\)th sought value of the uniform random number generator, which takes values in \([0,1]\), and \(j_{\text{rand}}\) is a chosen integer in \([1,D]\) randomly and it ensures that one parameter in \(U_{G}\) is different from the target vector \(x_{G}\) at least.

2.2.1.4. Selection. For the minimization problem, each target vector \(x_{G}\) and its corresponding trial vector \(U_{G}\), a selection procedure is performed using the fitness function. The population with the lower fitness function value will be selected as the next-generation population.

\[ x_{G+1} = \begin{cases} U_{G+1} & \text{if } (f(U_{G+1}) < f(x_{G})) \\ x_{G} & \text{otherwise} \end{cases} \]

2.1.2. Self-Adaptive Differential Evolution Algorithm. The control parameters used by the DE algorithm, such as \(CR\), \(F\), and \(NP\), with the learning strategy, determine the algorithm’s performance. For achieving superior performance on a given problem, fine-tuning individual key control parameters is essential. Additionally, but more importantly, all available learning strategies must be tested in the mutation phase. A better solution is to use the SaDE algorithm that can adjust the control parameters and learning strategies during the evolutionary process through the self-adaptive mechanism automatically.

Among the three key control parameters of the DE algorithm, \(CR\), \(F\), and \(NP\), \(NP\) is adopted as a user-defined value to deal with different dimensions. \(F\) corresponds to the speed of convergence, which should be maintained throughout the evolutionary process for local searches with smaller values of \(F\) and global searches with larger values of \(F\), to produce influential mutant individuals. For different individuals in the current population, \(F\) can select any value in the \((0, 2)\) range of the normal distribution with mean \((0.5)\) and standard deviation \((0.3)\) randomly.

In general, \(CR\) affects the convergence speed and robustness of the search process, where choosing a good \(CR\) can achieve better learning results under different learning strategies. Therefore, the value of \(CR\) can be dynamically adjusted within a particular generation interval by accumulating previous learning experiences. After several generations, the \(CR\) values are modified several times, using all recorded \(CR\) values, and then the mean of the \(CR\) normal distribution is recalculated based on the vector of successful trials. When the process is repeated with this new mean and standard deviation of the normal distribution, the range of correct \(CR\) values for the current problem has been automatically learned.

This work introduces a self-adaptive mechanism based on statistical experience and the roulette wheel selection method to select the better mutation strategy from the pool of variant strategy candidates at different stages of evolution. MT1, MT2, MT3, and MT4 are four candidate learning strategies that can be selected during evolution. The probabilities of applying the learning strategies to each individual in the current population are \(P_1, P_2, P_3,\) and \(P_4\), respectively. The initial possibilities of all four strategies are equal \((0.25)\). For a population of size \(NP\), a probability vector of length \(NP\) can be randomly generated and uniformly distributed in the range \([0, 1]\).

If the value of the \(l\)th element of the vector is less than or equal to \(P_1\), the strategy \(MT_1\) will be applied to the \(l\)th individual in the current aggregate or moved to the next available strategy. After evaluating all trial vectors, the number of new trial vectors generated by different strategies into the next-generation \(G\) is \(NS_1, NS_2, NS_3,\) and \(NS_4\). The number of discarded trial vectors is \(NF_1, NF_2, NF_3,\) and \(NF_4\). These values are accumulated within a specific number of \(G\) that is called a learning period. Then, the probability \(p\) of strategy \(l\) is updated as follows

\[ p_l = \frac{MT_l}{\sum_{l=1}^{4} MT_l} \]

At the end of the learning period, it will automatically update the probabilities of applying the four learning strategies and reset the values obtained in the previous period for \(NS_1\)–\(NS_4\) and \(NF_1\)–\(NF_4\). This process will gradually evolve as the most suitable learning strategy for the given problem.

The performance of the LSSVM is affected to a significant degree by two key hyperparameters (eqs 5–8): \(y\) and \(\sigma\). Therefore, these hyperparameters are to be optimized using the SaDE algorithm, and the optimization process is shown in Figure 3.
The variables are chosen as the hyperparameters $\gamma$ and $\sigma$ to be optimized, with the selected fitness value function as follows

$$R(\gamma, \sigma) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y(x_i) - y^*(x_i))^2}$$

$$= \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left[ \sum_{j=1}^{n} a_i K(x_i, x_j) - b \right]^2}$$

where $y(x_i)$ and $y^*(x_i)$ are the true and proxy data at the test point $x_i$, respectively, and $R(\gamma, \sigma)$ refers to the root mean square of the prediction error (RMSE), which varies with the LSSVM parameters $\gamma$ and $\sigma$. For proxy models, any error or correlation metric is obtained by validation in a test data set. The complexity of a proxy model requires multiple methods to assess its accuracy. The indicator for evaluating the performance of a regression model is generally the regression accuracy, which is also known as the model accuracy. In addition to RMSE, relative error (RE) is also a fundamental factor to be considered.

Generally, the relative error better reflects the confidence level of the calculation. It is defined by the following equation

$$\text{RE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left( \frac{y(x_i) - y^*(x_i)}{y(x_i)} \right)^2}$$

(19)

3. CONSTRUCTING INTEGRATED OPTIMIZATION MODEL

The utilization of hydraulic fracturing and horizontal drilling has facilitated the recovery of oil reserves from shale and other tight formations, resulting in horizontal well design and fracture optimization being the most critical aspects of low-permeability reservoir development design. The characteristics of horizontal well placement and fracturing design are analyzed, combined with the proxy model in Section 2, which establishes an integrated mathematical model for optimal strategy in this section.

Most conventional horizontal well design and fracturing optimization methods focus on finding the optimal design parameters to specific metrics, such as estimated ultimate recovery (EUR), cumulative gas production over a while, or net present value (NPV). The economic analysis recommends the decision with maximum NPV (present value of net benefits, or benefits minus costs, over time) or maximum benefit-cost ratio (present value of benefits to present value of fees). The NPV is calculated to obtain the present value of future income generated by eliminating the use of standard economic techniques. NPV is the difference between the present value of cash inflows and outflows associated with a project, adjusted for inflation to its present value, the objective function most widely used in reservoir optimization. For the optimization problem of horizontal well placement and fracturing design in shale gas reservoirs, the objective function (NPV) of the optimization is calculated as follows

$$f(x_f, L_w) = \left\{ \sum_{t=1}^{N_t} \left[ \frac{1}{1 + b} \left( \frac{P_g Q_{gas}^t}{(1 + b)^t} - C_{\text{operate}}^t \right) \right] - C_{\text{tax}} Q_{gas}^t - C_{\text{ground}} - C_{\text{drill}} - C_{\text{fracture}} \right\}$$

(20)

where $x_f$ and $L_w$ denote the optimization variables fracture half-length and horizontal well length, respectively; $N_t$ and $t_s$ are the total year and the time step calculated to year $t_s$, respectively; $b$ is the discount rate; $P_g$ denotes the natural gas price; $Q_{gas}^t$ denotes the total gas production in the $t_s$-th year; $C_{\text{operate}}$ is the operating cost of natural gas in the $t_s$; $C_{\text{tax}}$ is the tax constant; $C_{\text{ground}}$ is the cost of ground works; and $C_{\text{drill}}$ and $C_{\text{fracture}}$ denote drilling costs and fracturing costs, respectively.

Figure 3. Flow diagram for LSSVM’s parameter optimization with SaDE.
Drilling costs are usually defined as a linear function of horizontal well length

\[ C_{\text{drill}} = C_{\text{d fix}} + C_{\text{d}} L_w \]  

(21)

where \( C_{\text{d fix}} \) denotes the fixed cost of drilling a new horizontal well and \( C_{\text{d}} \) is the drilling penetration cost per lateral length of the horizontal well. Table 1 provides the values for the fixed cost and penetration cost of the drilling process.

Table 1. Parameter for the NPV Calculation

| parameter                  | values | unit | symbols |
|----------------------------|--------|------|---------|
| discount rate              | 8      | %/100| b       |
| gas price                  | 1.3    | CNY/m³ | \( P_g \) |
| operating cost             | 2.64   | \( 10^6 \) CNY | \( C_{\text{operate}} \) |
| tax constant               | 0.3277 | CNY/m³ | \( t_{\text{tax}} \) |
| cost of ground works       | 5      | \( 10^6 \) CNY | \( C_{\text{ground}} \) |
| fixed cost of drilling      | 19.923 | \( 10^6 \) CNY | \( C_{\text{fix}} \) |
| drilling cost per lateral length | 2209   | CNY/m | \( C_{\text{fix}} \) |
| fixed fracturing cost       | 8.57   | \( 10^6 \) CNY | \( C_{\text{fix}} \) |

In horizontal well multistage fracturing, large volumes of fracturing fluid and proppant are pumped into the formation. The proppant is pumped into the fractured rock to help improve the permeability of the formation fluids once the injected fracturing fluid has released the fracturing pressure. Therefore, the fracturing cost is primarily determined by the fracturing fluid and proppant. The following fracturing costs are defined

\[ C_{\text{fracture}} = C_{\text{f}_{\text{fix}}} + C_{\text{f}} n_{\text{f}} + C_{\text{fix}}' \]  

(22)

where \( C_{\text{f}} \) and \( C_{\text{fix}}' \) are the fracturing fluid and proppant costs per fracturing section, respectively; \( n_{\text{f}} \) is the fracture section number, which is determined by the horizontal section length and fracture spacing; \( C_{\text{f}_{\text{fix}}} \) is the fixed fracturing cost per well; \( r_{\text{p}} \) is the proppant concentration in the fracturing fluid; and \( V_{\text{fix}} \) is the total volume of fracturing fluid. Typically, \( x_i \) and \( F_{\text{fix}} \) are nonlinearly related. This is partly because as the fracture half-length increases, the filtration loss increases due to the increased contact area between the fracturing fluid and the formation during injection and also because the injection pressure decreases as the fracture half-length increases under the influence of formation frictional resistance. This ultimately leads to the linear increase of injection volume and does not achieve the fracturing result of the linear increase of fracture half-length. In this study, according to the literature,\(^{31,32}\) we regressed the relationship between \( F_{\text{fix}} \) and \( x_i \) as a function of the following, based on the statistics of actual fracturing data

\[ F_{\text{fix}} = 0.1184 x_i^2 + 2.4152 x_i - 6.4397 \]  

(23)

It is essential to emphasize that production forecasting and optimal design should be an integrated system. With the development of computer technology, researchers utilize optimization algorithms combined with proxy models to deal with production problems such as well placement and fracturing optimization design.\(^{33-36}\) The optimization results tend to be more influenced by the optimization algorithm when the accuracy of the proxy model is sufficient.

Guyaguler et al.\(^ {37} \) proposed a hybrid genetic algorithm (HGA), which reduces the computational burden of performing many simulations by combining GA and general kriging algorithms. They obtained results consistent with observations from the Pompano field in the Gulf of Mexico and used the HGA to study the PUNQ-S3 reservoir. Ma et al.\(^ {38} \) applied gradient-based finite difference (FD), discrete simultaneous regressive stochastic approximation (DSPSA), and genetic algorithm (GA) to resolve the hydraulic fracture arrangement problem. Plaksina et al.\(^ {39} \) offer the algorithm that gives quantitative and qualitative measures of “goodness” of the optimal production plans, which can handle objectives effectively and produce the Pareto optimal solutions without requiring the user to assign weights to each aim inside an aggregate function. Li et al.\(^ {40} \) proposed a dynamic simplex interpolation-based alternating subspace (DSIAS) search method for the mixed-integer optimization problem of shale gas development projects and also presented a case study on the development of the Barnett Shale gas field and validated the optimization effect.

4. RESULTS AND DISCUSSION

In this section, we apply the SADE-LSSVM model to actual blocks (the lower Silurian Longmaxi Formation from the southern Sichuan Basin). First, a production proxy model is built based on the adaptive response surface experimental design with the historical fitting of regional production data. The applicability of the proxy model was also verified, with the proxy model remaining efficient and accurate when the data set was updated.

4.1. Model Building and Experiment Designing. The central target formation in the study area is the Longmaxi shale gas formation, which has a reservoir thickness of 30 m, a reservoir depth of about 2500–3000 m, average porosity of 4%.
4%, an average permeability of $3 \times 10^{-5}$ mD, low-pressure coefficient, and fault and microfracture developed. The reservoir is assumed to be homogeneous, with uniform fracture spacing and porosity, and permeability independent of stress. Only gas flows in the reservoir, and the effects of gas desorption are considered. Detailed reservoir information for this section is shown in Table 2.

Self-adaptive response surface experimental design (SaRSE) was applied based on six uncertain parameters with reasonable ranges of parameter values:

- porosity ($\phi$)
- matrix permeability ($p$)
- fracture half-length ($x_f$)
- horizontal section length ($l_w$)
- gas content saturation ($S_g$)

SaRSE experimental data are determined by the response surface regression accuracy, with the number of experiments increasing until the specified accuracy is reached. Experimental data thus obtained can better reflect the actual reservoir characteristics and seepage patterns. Numerical simulation software was used to calculate the 15-year gas production ($C_g$). Some of the experimental design results are shown in Supporting Information Table S1.

Figure 5. SaDE-LSSVM comparison of predicted and target $C_g$ for (a) training set, (b) validation set, (c) testing set, and (d) total set.

Figure 6. PSO-LSSVM comparison of predicted and target $C_g$ for (a) training set, (b) validation set, (c) testing set, and (d) total set.
Figure 7. GA-LSSVM comparison of predicted and target \( C_g \) for (a) training set, (b) validation set, (c) testing set, and (d) total set.

Figure 8. GWO-LSSVM comparison of predicted and target \( C_g \) for (a) training set, (b) validation set, (c) testing set, and (d) total set.

Table 3. Comparison of LSSVM Model Errors (RMSE and RE)

| proxy model  | SaDE-LSSVM | PSO-LSSVM | GA-LSSVM | GWO-LSSVM |
|--------------|------------|-----------|----------|-----------|
| training set | RMSE       | 0.002264  | 0.005254 | 0.007137  | 0.005262  |
|              | RE         | 0.001350  | 0.002044 | 0.002488  | 0.002047  |
| validation set | RMSE     | 0.002610  | 0.005651 | 0.007552  | 0.005659  |
|              | RE         | 0.001368  | 0.002677 | 0.003418  | 0.002680  |
| testing set  | RMSE       | 0.002955  | 0.007006 | 0.009437  | 0.007016  |
|              | RE         | 0.001429  | 0.002632 | 0.003381  | 0.002636  |
| total set    | RMSE       | 0.002503  | 0.005780 | 0.007808  | 0.005788  |
|              | RE         | 0.001372  | 0.002290 | 0.002856  | 0.002293  |
4.2. Proxy Model Accuracy. Based on SaRSE, the response surface regression formula will be obtained. The results of adaptive response surface regression are shown in Figure 4. The response surface formula predicts results with low accuracy. Therefore, a new method is warranted to build the agent model.

Cross-validation is introduced to improve the model’s reliability. The total data set was divided into three data sets: training set, validation set, and test set. More specifically, first, 25% of the entire data set was randomly selected as the test set and excluded from the training and validation process. The remaining data set was then randomly divided into 10 parts, and a fivefold cross-validation method was used to use 60% of
the data as the training set and 15% of the data to validate the model for each cycle. Figure 5 depicts the $C_g$ for the training, validation, test, and total data sets. The results show that the error difference between the training and test data sets is slight, indicating that the training process is reliable (no overfitting).

The predictions are in good agreement with the target values, and the accuracy is within the acceptable range. We compared the SaDE-LSSVM model with the PSO-LSSVM, GA-LSSVM, and GWO-LSSVM models, showing the high prediction accuracy of the SaDE optimized LSSVM, as shown in Figures 6–8. Table 3 shows a more comprehensive set of error metrics.

### 4.3. Sensitivity Analysis and Optimization Results

A first sensitivity analysis was carried out according to the experimental design in Supporting Information Table 1. The $F$-value of analysis of variance is a fine indicator of the degree of influence of different parameters. Figure 9 shows that all investigated parameters have different degrees of influence on $C_g$ and NPV. The results showed that $C_g$ and NPV exhibited the same behavior in response to parameter variations. $C_g$ and NPV are sensitive to geological properties ($\phi$ and $S_g$) to a high degree. The effect of $k$ on $C_g$ is relatively small. Shale gas is mainly recovered by hydraulic fracturing with high inflow capacity, where changing the matrix permeability does not bring significant changes in production, which is the reason why hydraulic fracturing is used for shale gas development.

A second sensitivity study was then performed for each influencing parameter to determine the effect of individual parameters on $C_g$ and NPV based on the experimental data in Supporting Information Table 1. The effect patterns of parameters on $C_g$ and NPV are shown in Figure 10. Obviously, the increase of $\phi$ and $S_g$ can directly increase $C_g$ with no cost constraint, which eventually makes the NPV show the same trend as $C_g$. However, an $x_f$ growth leads to a higher $C_{fracture}$ while increasing $C_g$, which results in a different trend of $C_g$ and NPV. For $l_{fracture}$ both $C_g$ and $C_{dill}$ are close to linear to $l_{fracture}$, which leads to $C_{dill}$ not being able to constrain NPV, which makes $C_g$ and NPV trend the same while the optimal $l_{fracture}$ is taken to the extreme value on the other hand.

In this study, the SaDE-LSSVM proxy model with relatively high accuracy is applied to the single-well NPV optimization with the optimization variables $x_f$ and $l_{fracture}$. The final optimization results are $x_f = 155$ m and $l_{fracture} = 2000$ m.

### 4.4. Practical Application of Proxy Model

The numerical simulation itself is a simplified mathematical model, so the LSSVM fitting is excellent. The reliability of the LSSVM method can be tested with actual field data. Therefore, to further validate the applicability of the proxy models, we developed the EUR proxy model and the absolute open flow (AOF) proxy model for the study area based on the data as the training set and 15% of the data to validate the model for each cycle.

Figure 11 depicts the $C_g$ for the training, validation, test, and total data sets. The results show that the error difference between the training and test data sets is slight, indicating that the training process is reliable (no overfitting).

The predictions are in good agreement with the target values, and the accuracy is within the acceptable range. We compared the SaDE-LSSVM model with the PSO-LSSVM, GA-LSSVM, and GWO-LSSVM models, showing the high prediction accuracy of the SaDE optimized LSSVM, as shown in Figures 6–8. Table 3 shows a more comprehensive set of error metrics.

### Table 4. Comparison of Errors in Predicting QAOF and EUR by LSSVM Models

| proxy model | SaDE-LSSVM | PSO-LSSVM |
|-------------|------------|-----------|
| QAOF        |            |           |
| training set| RMSE 0.039221 | 0.022444  |
| RE          | 0.828875   | 0.425939  |
| testing set | RMSE 0.030152 | 0.019894  |
| EUR         |            |           |
| training set| RMSE 0.013228 | 0.054516  |
| RE          | 0.032447   | 0.042602  |
| testing set | RMSE 0.098707 | 0.100341  |
| RE          | 0.079433   | 0.100720  |

The numerical simulation itself is a simplified mathematical model, so the LSSVM fitting is excellent. The reliability of the LSSVM method can be tested with actual field data. Therefore, to further validate the applicability of the proxy models, we developed the EUR proxy model and the absolute open flow (AOF) proxy model for the study area based on the data as the training set and 15% of the data to validate the model for each cycle.
actual data. We select 16 real production wells data as training samples. The characteristic variables are formation pressure, clump number, fracturing fluid volume, proppant dose, $\varnothing$, k, $S_p$, rejection rate, fracturing pressure, and pumping stop pressure. Five wells were also selected as the test set to verify the rejection rate, fracturing pressure, and pumping stop pressure. Five wells were also selected as the test set to verify the accuracy of the proxy model. Figure 11 shows the accuracy of the SADE-LSSVM model and PSO-LSSVM model, respectively. A comprehensive set of error evaluations is shown in Table 4. The proxy model errors are within acceptable limits, indicating the reliability of using actual production data to build the proxy models.

■ CONCLUSIONS

In this study, a proxy model of shale gas production, SADE-LSSVM, is developed using a swarm optimization algorithm and machine learning. Its errors are compared and evaluated, showing a high accuracy of the proxy model. The proxy model built by the swarm optimization algorithm combined with the machine learning approach still has strong applicability when the data set is updated. We obtained the following recognition points:

1. The developed adaptive proxy modeling method has high accuracy, due to the fact that a sufficient number of data are selected with the self-adaptive response surface design in the developed method, which improves the accuracy of the proxy model.
2. The developed proxy model can simulate the reservoir simulator response with acceptable accuracy, thus replacing numerical simulation in the production optimization process to satisfy the optimization process under multiple iterations and multiple control parameter variations, which is particularly important for production optimization applications.
3. The proxy models can be easily updated when any new data sets are available. These data sets include geological data, drilling data, fracturing data, or other custom design-of-experiment data. This data-driven approach has great potential for application by quickly implementing proxy models on other shale reservoir data sets.

■ ASSOCIATED CONTENT

Supporting Information
The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.1c05158.

Complete experimental design results (Table S1) (XLSX)

■ AUTHOR INFORMATION

Corresponding Authors
Shuangfang Lu — Key Laboratory of Deep Oil and Gas, China University of Petroleum, Qingdao 266580, China; School of Geosciences, China University of Petroleum (East China), Qingdao, Shandong 266580, China; orcid.org/0000-0002-0758-4075; Email: luschuangfang@upc.edu.cn
Taohua He — Key Laboratory of Deep Oil and Gas, China University of Petroleum, Qingdao 266580, China; School of Geosciences, China University of Petroleum (East China), Qingdao, Shandong 266580, China; orcid.org/0000-0002-9649-1662; Email: b17010030@s.upc.edu.cn

Authors
Lu Qiao — Key Laboratory of Deep Oil and Gas, China University of Petroleum, Qingdao 266580, China; School of Geosciences, China University of Petroleum (East China), Qingdao, Shandong 266580, China
Huijun Wang — Key Laboratory of Deep Oil and Gas, China University of Petroleum, Qingdao 266580, China; School of Geosciences, China University of Petroleum (East China), Qingdao, Shandong 266580, China
Yang Liu — Key Laboratory of Deep Oil and Gas, China University of Petroleum, Qingdao 266580, China; School of Geosciences, China University of Petroleum (East China), Qingdao, Shandong 266580, China

Complete contact information is available at: https://pubs.acs.org/doi/10.1021/acsomega.1c05158

Notes
The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors are grateful for the financial support provided by the National Major Science and Technology Project (2016ZX05061) and research project funded by SINOPEC (P19017-3).

■ REFERENCES

(1) Li, W.; Lu, S.; Li, J.; Wei, Y.; Feng, W.; Zhang, P.; Song, Z. Geochemical modeling of carbon isotope fractionation during methane transport in tight sedimentary rocks. Chem. Geol. 2021, 565, No. 120033.
(2) Chen, G.; Gang, W.; Chang, X.; Wang, N.; Zhang, P.; Cao, Q.; Xu, J. Paleoproductivity of the Chang 7 unit in the Ordos Basin (North China) and its controlling factors. Palaeoecogr., Palaeoclimatol., Palaeoecol. 2020, 551, No. 109741.
(3) He, T.; Lu, S.; Li, W.; Wang, W.; Sun, D.; Pan, W.; Zhang, B. Geochemical characteristics and effectiveness of thick, black shales in southwestern depression, Tarim Basin. J. Pet. Sci. Eng. 2020, 185, No. 106607.
(4) Hu, T.; Pang, X.; Jiang, F.; Wang, Q.; Liu, X.; Wang, Z.; Jiang, S.; Wu, G.; Li, C.; Xu, T.; et al. Movable oil content evaluation of lacustrine organic-rich shales: Methods and a novel quantitative evaluation model. Earth-Sci. Rev. 2021, 214, No. 103545.
(5) Wang, H.; Qiao, L.; Lu, S.; Chen, F.; He, X.; Zhang, J.; He, T.; et al. A novel shale gas production prediction model based on machine learning and its application in optimization of multistage fractured horizontal wells. Front. Earth Sci. 2021, 9, No. 675.
(6) Zou, C.; Zhu, R.; Chen, Z.; Ogg, J. G.; Wu, S.; Dong, D.; Qu, Z.; Wang, Y.; Wang, L.; Lin, S.; et al. Organic-matter-rich shales of China. Earth-Sci. Rev. 2019, 189, 51–78.
(7) Wang, D.; Wang, X.; Ge, H.; Sun, D.; Yu, B. Insights into the Effect of Spontaneous Fluid Imbibition on the Formation Mechanism of Fracture Networks in Brittle Shale: An Experimental Investigation. ACS Omega 2020, 5, 8847–8857.
(8) Kulga, B.; Artun, E.; Ertekin, T. Development of a data-driven forecasting tool for hydraulically fractured, horizontal wells in tight-gas sands. Comput. Geosci. 2017, 103, 99–110.
(9) Zhang, X.; Du, C.; Deimbacher, F.; Cricht, M.; Harikesevanallur, A. K. In Sensitivity Studies of Horizontal Wells with Hydraulic Fractures in Shale Gas Reservoirs, International Petroleum Technology Conference, 2009.
(10) Jia, H.; Meng, Y. Multiple fracturing of horizontal well in shale gas productivity factors numerical simulation researching. Pet. Ind. Appl. 2013, 32, 34–39.
(11) Shahani, H.; Sarkheil, H.; Forouud, T.; Kaimpoili, S. In A Proxy Modeling Approach to Optimization Horizontal Well Placement, 45th U.S. Rock Mechanics/Geomechanics Symposium, 2011; pp 26–29.
(12) Schuetter, J.; Mishra, S.; Zhong, M.; LaFollette, R. A Data-Analytics Tutorial: Building Predictive Models for Oil Production in an Unconventional Shale Reservoir. SPE J. 2018, 23, 1075–1089.
(13) Zhang, L.; Li, Z. P.; Lai, F. P.; Li, H.; Adenutsi, C. D.; Wang, K.; Yang, S.; Xu, W. Integrated optimization design for horizontal well placement and fracturing in tight oil reservoirs. J. Pet. Sci. Eng. 2019, 178, 82–96.
(14) Wang, S.; Chen, Z.; Chen, S. Applicability of deep neural networks on production forecasting in Bakken shale reservoirs. J. Pet. Sci. Eng. 2019, 179, 112–125.
(15) Brantson, E. T.; Ju, B.; Ziggah, Y. Y.; Akwensi, P.; Sun, Y.; Wu, D.; Addo, B. J. Forecasting of Horizontal Gas Well Production Decline in Unconventional Reservoirs using Productivity, Soft Computing and Swarm Intelligence Models. Nat. Resour. Res. 2019, 28, 717–756.
(16) Montgomery, J. B.; O’Sullivan, M. Spatial variability of tight oil well productivity and the impact of technology. Appl. Energy 2017, 195, 344–355.
(17) Zhou, Q.; Dilmore, R.; Kleit, A.; Wang, J. Y. Evaluating Gas Production Performances in Marcellus Using Data Mining Technologies. J. Pet. Sci. Eng. 2014, 20, 109–120.
(18) Lolon, E.; Hamidieh, K.; Weijers, L.; Mayerhofer, M.; Mechler, H.; Ouda, O. In Evaluating the Relationship between Well Parameters and Production Using Multi-Variate Statistical Models: A Middle Bakken and Three Forks Case History, SPE Hydraulic Fracturing Technology Conference, 2016.
(19) Prabuankaral, K. M.; Thirumoorthy, P.; Manavalan, R. Classification of breast masses in ultrasound images using self-adaptive differential evolution extreme learning machine and rough set feature selection. J. Med. Imaging 2017, 4, No. 024507.
(20) Yu, W.; Kamy, S. Optimization of Multiple Hydraulically Fractured Horizontal Wells in Unconventional Gas Reservoirs. J. Petrol. Sci. Eng. 2013, 2013, 1–16.
(21) Yu, W.; Sepehrnoori, K. An Efficient Reservoir-Simulation Approach to Design and Optimize Unconventional Gas Production. J. Can. Petrol. Technol. 2014, 53, 109–121.
(22) Nguyen-Le, V.; Shin, H. Development of reservoir economic indicator for Barnett Shale gas potential evaluation based on the reservoir and hydraulic fracturing parameters. J. Nat. Gas. Sci. Eng. 2019, 66, 159–167.
(23) Suykens, J. A. K.; Vandewalle, J. Least Squares Support Vector Machine Classifiers. Neural Process Lett. 1999, 9, 293–300.
(24) Fonseca, R. M.; Leeuwenburgh, O.; Rossa, E. D.; Paul, M. J.; Jansen, J. D. In Ensemble-Based Multi-Objective Optimization of On-Off Control Devices Under Geological Uncertainty, SPE Reservoir Simulation Symposium, 2015.
(25) Awotunde, A. A. In On the Joint Optimization of Well Placement and Control, Society of Petroleum Engineers SPE Saudi Arabia Section Technical Symposium and Exhibition, 2014.
(26) Oliveira, D. F.; Reynolds, A. An Adaptive Hierarchical Multiscale Algorithm for Estimation of Optimal Well Controls. SPE J. 2014, 19, 909–930.
(27) Wilson, K. C.; Durlofsky, L. J. Optimization of shale gas field development using direct search techniques and reduced-physics models. J. Pet. Sci. Eng. 2013, 108, 304–315.
(28) Xu, S.; Feng, Q.; Wang, S.; Farzam, J.; Li, Y. Optimization of multistage fractured horizontal well in tight oil based on embedded discrete fracture model. Comput. Chem. Eng. 2018, 117, 291–308.
(29) Wang, X.; Peng, X.; Zhang, S.; Ying, L.; Peng, F.; Zeng, F. Guidelines for Economic Design of Multistage Hydraulic Fracturing, Yanchang Tight Formation, Ordos Basin. Nat. Resour. Res. 2019, 29, 1413–1426.
(30) Li, J. C.; Gong, B.; Wang, H. G. Mixed integer simulation optimization for optimal hydraulic fracturing and production of shale gas fields. Eng. Optim. 2015, 1378–1400.
(31) Zhang, M. L.; Zhang, T. Y.; Fan, J. Y. Calculation and analysis of fracture extension parameters based on PKN model. Sci. Technol. Eng. 2019, 19, 116–123.