Hindawi
International Journal of Energy Research
Volume 2024, Article ID 6704100, 15 pages
https://doi.org/10.1155/2024/6704100

Research Article

An Interwell Connectivity Assessment Model for Polymer Flooding Short-Term Development Data Based on A-LSTM and EFAST Methods

Ming Li,[1] Jihong Zhang,[1] Xinjian Tan,[1] and Ruixue Zhang[2]

[1]*Laboratory of Enhanced Oil Recovery of Education Ministry, Northeast Petroleum University, No. 99, Xuefu Street, Daqing 163318, China*
[2]*First Oil Production Plant, Daqing Oilfield Co. Ltd., No. 34, Zhongqi Road, Daqing 163000, China*

Correspondence should be addressed to Jihong Zhang; dqzhjh@126.com

Received 2 March 2024; Revised 9 April 2024; Accepted 12 April 2024; Published 7 May 2024

Academic Editor: Mohamed Louzazni

Interwell connectivity assessment in polymer-driven reservoirs is critical for setting appropriate injection rates and improving oil recovery. Traditional deep learning techniques often lack accuracy and reliability when applied to short-term oilfield production data. In response, the A-LSTM algorithm is proposed, which integrates the attention mechanism with a long- and short-term memory network (LSTM). The predictive accuracy of A-LSTM is assessed and juxtaposed with LSTM and support vector regression (SVR) algorithms for short-term single-well daily oil production analysis. The Huber loss function was utilized to quantify the difference between predicted and actual results, resulting in a dynamic production prediction model. An interwell connectivity (IWC) assessment model is then obtained by fusing the dynamic production prediction model with the EFAST method, thus demonstrating the superior prediction accuracy of A-LSTM in oil production prediction and connectivity assessment. Moreover, the credibility of the assessment is further corroborated through numerical simulations and interwell tracer tests. The study results showed that the interwell connectivity evaluation model based on the A-LSTM algorithm and EFAST method is not only capable of accurately predicting the single-well daily oil production using a small sample dataset but also a highly reliable method for interwell connectivity evaluation, and the application of the interwell connectivity assessment model can further guide polymer flooding work in oilfields.

1. Introduction

Interwell connectivity (IWC) represents a crucial measure to determine the equilibrium between reservoir injection and recovery and serves as a primary reference for guiding the development plans of oilfields. Study on IWC in oil reservoirs encompasses predicting the daily oil production of producing wells, quantitative description of IWC, and analysis of dynamic changes [1]. Traditional IWC study methods typically involve tracer testing, pressure testing, well testing, and other complex and costly means [2]. In order to overcome these drawbacks, scholars have established mathematical-physical models for IWC analysis, utilizing static and dynamic parameters and production measures derived from oilfield production data. The models include the Spearman rank correlation analysis model [3], multiple linear regression model [4], capacitance and resistance model [5], and multiwell production index model [6]. However, these methods suffer from lower accuracy of calculation and insufficient consideration of production parameters in the model calculation equation. Furthermore, the diversity in reservoir geological conditions and development methods introduces uniqueness, thereby diminishing the adaptability and generalizability of early-established IWC analysis models. This limitation restricts field application to only a subset of fields.

In recent years, machine learning has been widely used in linguistics, clinical medicine, computer science, and other fields of data processing and analysis work. It is an important means of data classification. Regression, using computer programs to simulate human learning, can be used to

analyze and mine the actual oilfield production data to obtain the hidden relationships between the data and achieve the learning objectives. Scholars in the field of oil and gas exploration and development have been inspired by machine learning algorithms to carry out a series of IWC research work based on the combination of static and dynamic oilfield data and machine learning algorithms. Panda and Chopra [7] first applied the artificial neural network (ANN) algorithm to fluid flow simulation and interwell interaction prediction in inhomogeneous permeable media in 1998. Demiryurek et al. [8] proposed a sensitivity analysis method based on the partial derivatives of the output variables with respect to the input variables to quantitatively describe the production rate of production wells. The method compensated for the inability to quantify the connectivity between injection and extraction wells in a single direction in earlier IWC studies using neural network algorithms, but the study did not consider the complex interwell interference effects between multiple injection wells. For some time afterward, fewer scholars utilized the ANN algorithm for interwell connectivity studies due to its poor fitting to time series data. Instead, most scholars focused their research efforts on improving the capacitance model and its application process [9–11]. In 2016, Elons et al. [12] first introduced the long short-term memory (LSTM) algorithm for dynamic prediction of daily oil production in oilfields using time series data. The LSTM algorithm demonstrated superior applicability for time series prediction tasks, leading to its gradual replacement of the ANN algorithm in this domain. Then, Cheng et al. [13] used the Extended Fourier Amplitude Sensitivity Test (EFAST) method to perform global sensitivity analysis on the production prediction model based on the LSTM algorithm, which fully considered the nonlinearity of the injection and extraction relationship, the coupling effect between multiple injection wells and a single production well, and computed the yield of the LSTM algorithm. Jiang et al. [14] combined the material balance equation in CRM with ANN algorithm to propose a physical knowledge interaction neural network for daily oil production prediction work, and this method increased the interpretability of the model. Data quality is also particularly important in the IWC research process, which determines the accuracy and reliability of the model analysis results. Albertoni and Lake [4] constructed a nonlinear filter for daily oil production data. This filter takes into account the time lag and decay of the flow and propagation process of injected water in the formation. It is based on the principle of pressure drop superposition. However, this method is not applicable to special reservoir environments such as low and ultralow permeability, tight reservoirs, and other reservoir types where the internal seepage pattern is non-Darcy seepage. Liu et al. [15] used the integrated empirical mode decomposition (EEMD) method which was used to preprocess the daily oil production time series data to obtain the intrinsic mode function (IMF), and the DTW algorithm was used to select the IMF as the input of the LSTM to predict the daily oil production. The method decomposes the wave function of daily oil production data over time from the perspective of improving model data quality, so that the data signal, which is inherently nonlinear and non-smooth, is transformed into multiple smooth wave functions. Wang et al. [16] preprocess the raw data into a custom form so that each sample contains additional local wave information and historical residual energy information, and in predicting long-term production data of bottomhole pressure (BHP), data performed better. However, it is difficult to obtain production data continuously and for a long period of time in the actual production process of oilfields, and there is an urgent need to further optimize and improve the traditional LSTM algorithm so as to adapt it to the prediction work of short-term production data. One promising approach to address these study gaps is to incorporate the attention mechanism from natural language processing (NLP) into the LSTM [17] to enhance the screening and focusing of key information in features. This approach could potentially improve the prediction accuracy of the model by accounting for factors that were previously ignored.

To address the issues encountered in the aforementioned study process, this study proposes a novel approach that integrates the A-LSTM algorithm and EFAST to evaluate interwell connectivity. The proposed method leverages actual reservoir production data as the basis for feature extraction and establishes a production dynamic time dataset by screening and cleaning the data using various data preprocessing techniques. The attention mechanism is then incorporated into the LSTM algorithm by modifying the weight coefficient search method within the LSTM gating unit and utilizing the additive attention score function to strategically search for weights. The resulting A-LSTM algorithm, along with LSTM and SVR algorithms, is employed to construct a single-well daily oil production prediction model, with the Huber loss function serving as the error metric to quantify the differences between predicted and actual values. Finally, the superiority of the A-LSTM algorithm in interwell connectivity assessment is verified through numerical simulation and inter-well tracer testing.

2. Methodologies

2.1. Data Preprocessing

2.1.1. Feature Selection. Table 1 presents a range of basic dynamic and static characteristic parameters of the oilfield production process, which are available for consideration in our IWC study. However, the selection of characteristic parameters should follow certain principles to ensure the validity of our IWC study.

(1) Static feature parameters are not considered during the model learning process, as the data should remain dynamic

(2) Dynamic characteristic parameters that do not directly affect the connectivity between injectors and producers, such as production time and well-working mode, will not be studied

(3) The selected dynamic characteristic parameters should reflect the internal energy changes of the

Table 1: The basic characteristic parameters of oilfield production data.

Dynamic feature parameters	
Single producer daily production time	h
Wellhead tubing pressure	MPa
Annulus pressure	MPa
Casing pressure	MPa
Hydrostatic pressure	MPa
Daily liquid production	m^3/d
Daily oil production	m^3/d
Water cut	%
Daily polymer injection	m^3/d
Static feature parameters	
Mean temperature	°C
Mean effective rock thickness	m
Mean porosity	%
Mean permeability	cp
Mean oil saturation	%
Pump discharge capacity	m^3/d
Stroke	m
Okinawa	once/min

reservoir between injectors and producers during the polymer flooding process, thereby enabling the accurate assessment of IWC

To ensure consistent production well capacity in oilfield production, it is necessary to adjust the daily injection volume of injectors. This adjustment is based on the bottomhole pressure of injectors, which is considered the dependent variable of the daily injection volume. However, the change in the daily injection volume is generally minimal due to factors such as well depth, well diameter, and physical parameters of the injection polymer. On the other hand, the bottomhole pressure of producers and the recovery rate typically exhibit a linear and exponential relationship. Therefore, this paper utilizes only the time series data of daily injection volume for each injector as input data. The daily oil production of the extraction well is used as the prediction target, enabling the construction of a production dynamic time series dataset for the daily oil production prediction of producers. This approach reduces the feature dimensionality of the input dataset and accelerates the model's operation.

2.1.2. Data Cleaning and Transformation. After data feature extraction, data cleaning, transformation, and statutes are usually required to improve the quality of the time series dataset for oil production. Data cleaning involves handling missing and outlier values. Interpolation methods are commonly used to handle missing values, such as empirical interpolation, multiple imputation by chained equation (MICE) interpolation [18], K-nearest neighbor (KNN) interpolation [19], and random forest (RF) interpolation [20]. The handling of outlier values usually involves both supervised and unsupervised methods [21]. For datasets with small sample sizes and few features, unsupervised detection methods are usually preferred. In this paper, statistical methods such as boxplots and clustering-based outlier detection are used, along with professional experience to analyze the causes of outliers and determine the appropriate outlier handling method. The changes in oilfield production dynamic data over a short period of time are often not significant, so empirical interpolation can be used to estimate missing data by using data from adjacent time nodes. If missing data cannot be estimated using empirical interpolation, the KNN interpolation method can be used to estimate the value of missing data points by identifying K similar or nearby samples in the dataset. This method is simple and more suitable for continuous data types. Deletion of missing records is generally not preferred due to the limitations in the sample size. To handle outlier values, we can intuitively use boxplots to detect outliers, which are simple tools for outlier detection. Outliers are more likely to appear in the daily polymer injection feature parameters, and the treatment method is usually to consider them as missing values or not to handle them. Data transformation is a data normalization technique that aims to eliminate differences in feature dimensions. For oilfield production data, such as daily polymer injection of injectors and daily oil productions of producers, features exhibit nonuniformity in dimensionality and nonlinear and nonsmooth variation over time. Therefore, it is necessary to normalize production dynamic data using data transformation methods to make the data dimensionless, improve data quality, and accelerate model training and prediction speeds. In this paper, the minimum-maximum normalization method in Equation (1) is used to scale production dynamic time series to the range of [0,1] by linear transformation. The purpose of this approach is to address potential numerical issues that may arise during LSTM operation, satisfy the requirements of the tanh activation function, and accelerate model computation.

$$x_{\text{std}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}. \tag{1}$$

The aim of this study is to enhance efficiency by reducing the dimensionality of the data and identifying the minimal input data subset through feature extraction. The initial step involves data screening based on this principle, which leads to low-dimensional dynamic time series data with small sample sizes, eliminating the need for data statute processing. Following this, dataset partitioning is required for data preprocessing to segregate the dataset into training and testing subsets to facilitate model training and testing.

2.2. Prediction Model Based on A-LSTM

2.2.1. The Structure of LSTM. The recurrent neural network (RNN) algorithm, which effectively addresses the issues of gradient disappearance and explosion in ANN algorithms, as well as long-term time dependence in RNN algorithms, serves as the predecessor of the LSTM [22–24]. Compared to the common RNN recurrent network, the LSTM possesses a more complex hidden layer structure, and Figure 1

depicts the hidden layer cell structure of the LSTM. In this figure, x_t refers to the input of the current node's LSTM unit; h_{t-1} refers to the hidden state of the LSTM hidden layer unit of the previous node; c_{t-1} refers to the unit state of the previous node; the σ and tanh functions denote the sigmoid and inverse tangent activation functions, respectively. Furthermore, f_t denotes the output vector of the forgetting gate, while i_t denotes the output vector of the input gate. The current node input state $\bar{c}_t$ is leveraged to extract valid information from the current input, while c_t denotes the current node unit state, which is composed of the previous node unit state and the current node input state. Finally, o_t denotes the output vector of the output gate, which is employed to regulate the impact of long-term memory on the current output, and h_t denotes the hidden state of the current node.

A representation of the forward propagation process in the LSTM hidden layer can be expressed through Equations (2)–(7), which calculates the hidden state of the t-node within the hidden layer.

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right), \tag{2}$$

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right), \tag{3}$$

$$o_t = \sigma\left(W_o \cdot [h_{t-1}, x_t] + b_o\right), \tag{4}$$

$$\bar{c}_t = \tanh\left(W_c \cdot [h_{t-1}, x_t] + b_c\right), \tag{5}$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \bar{c}_t, \tag{6}$$

$$h_t = o_t \cdot \tanh\left(c_t\right). \tag{7}$$

W_f, W_i, W_o, and W_c denote the weight coefficient matrices associated with the oblivion gate, input gate, output gate, and cell state, respectively. The parameters b_f, b_i, b_o, and b_c are the biases connecting the corresponding gates and cell states. Sigmoid activation function and tanh activation function are shown in the following equations:

$$\begin{aligned} \sigma(x) &= \text{sigmoid}(x) = \frac{1}{1+e^{-x}}, \\ \tanh(x) &= \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}. \end{aligned} \tag{8}$$

In terms of activation function selection, the sigmoid function used is a linear function that takes values in the range of (0,1) which can control the gate opening well; the tanh function is a nonlinear function that is used to control the cell state c_t and the hidden layer hidden state h_t.

2.2.2. Proposed Structure of A-LSTM. The A-LSTM structure, as shown in Figure 2, incorporates an attention mechanism within the hidden layer of the LSTM. This integration allows for the redistribution of weight coefficients in the LSTM's hidden layer, enabling strategic weight searching. Consequently, this enhances the speed of operations and augments the predictive accuracy of the model.

$x_1, x_2, x_3, \ldots, x_{t-1}$ denote the history input sequence, x_{ti} denotes the set of history input sequences, and x_t denotes the t-node input sequence; $h_1, h_2, h_3, \ldots, h_{t-1}$ denote the history hidden state obtained after the history input sequence is input to the LSTM hidden layer unit, h_{ti} denotes the set of history hidden states, and h_t' denotes the set of t-node LSTM hidden states with the attention mechanism added; $s_1, s_2, s_3, \ldots, s_{t-1}$ denote the attention scoring function of historical hidden states, and s_{ti} denotes the set of attention scoring functions of historical hidden states; $\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_{t-1}$ denote attention probability weights of historical input hidden states to current input, and C denotes the state of the LSTM hidden layer input unit with the attention mechanism added at node t. The process comprises four distinct components.

(1) Calculating Attention Scores. The process begins with utilizing the LSTM hidden layer to compute the historical hidden states. Next, an attention scoring function is employed to allocate weights and biases to each of the historical hidden states, thereby obtaining the attention score for each of them. Various attention scoring functions are available for selection, with different scoring functions categorized according to the attention aggregation method. This selection process is akin to the selection of activation functions in the LSTM hidden layer cell. Equation (9) provides the scoring function, which utilizes an additive attention mechanism and demonstrates suitability for processing data of varying dimensionalities. The additive attention mechanism is demonstrated to exhibit good adaptability to low- and high-dimensional data [25].

$$s_{ti} = \nu^T \cdot \tanh\left(W_{ti} \cdot h_{ti} + b_{ti}\right). \tag{9}$$

The scoring function, s_{ti}, is a component of a forward neural network that consists of a single hidden layer. The weight coefficient matrix after activation by the hidden layer is denoted by ν, and the transposition of this weight coefficient matrix is denoted by ν^T. The weight coefficient matrix of h_{ti} before activation is denoted as W_{ti}, while b_{ti} denotes the bias of h_{ti} before activation.

(2) Calculating Historical Attention Probability Weights. For single-objective probability weight calculation problems, such as in LSTM gating units, the sigmoid function can be used to calculate the gate opening. In the context of computing probabilistic weights for multiple objectives, the application of the softmax function, as shown in Equation (10), becomes indispensable, which calculates the probability of multiple variables between (0,1), and the sum of these probabilities is number one.

$$\alpha_{ti} = \text{softmax}(s_{ti}) = \frac{\exp\left(s_{t-1}\right)}{\sum_{j=1}^{T-1}\exp\left(s_{tj}\right)}. \tag{10}$$

The notation used in this context includes $T-1$, which denotes the number of hidden states from historical time steps. Additionally, s_{tj} denotes the jth weight of attention probability for the s_{ti} function, where j ranges from 1 to $T-1$. Finally, α_{ti} denotes the set of attention probability

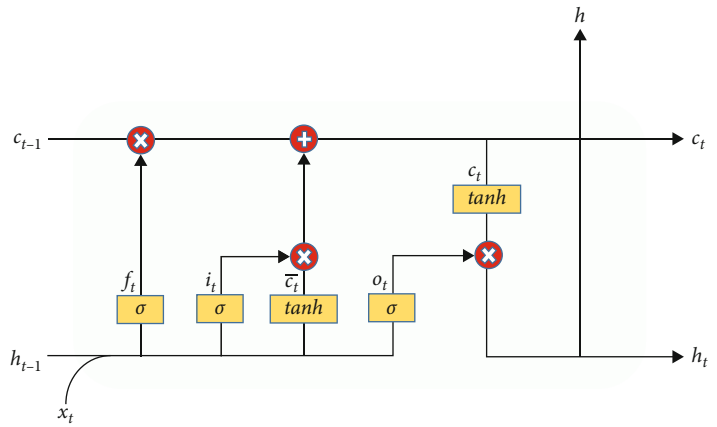


Figure 1: The structure of the LSTM hidden layer.

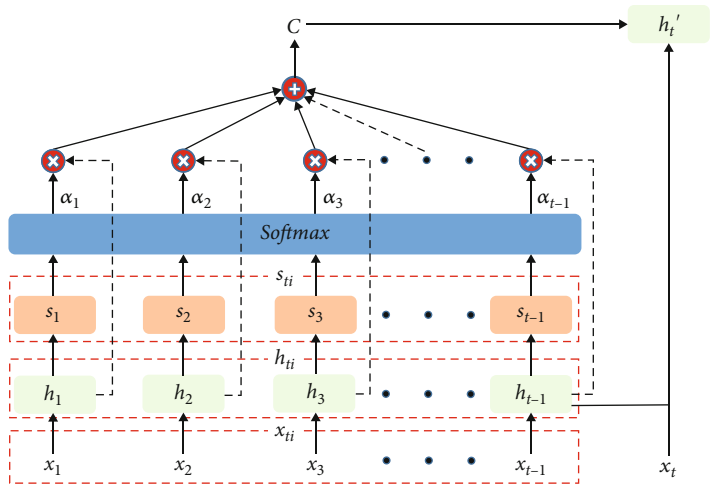


Figure 2: The structure of A-LSTM.

weights assigned to the historical input hidden states in relation to the current input.

(3) Update the Cell State of the t-Node LSTM Hidden Layer Input. The cell state of the new t-node LSTM hidden layer input is obtained by weighting and summing all the attention probability weights α_{ti} with the corresponding historical hidden states h_{ti}. This unit state reflects the process of redistributing the attention probability weights of the historical node hidden states to the t-node hidden states, i.e., state C in the following equation:

$$C = \sum_{i=1}^{T-1} \alpha_{ti} \cdot h_{ti}. \tag{11}$$

(4) Update t-Node Hidden Layer State. Equation (12) illustrates that within the A-LSTM hidden layer, the hidden state of node t undergoes an update process that incorporates the cell state C, the previous time step's hidden state h_{t-1}, and the present input x_t. This process generates a feature vector that contains information about the weights of the historical input nodes. In order to integrate the attention mechanism into the LSTM hidden layer cell structure, the feature vector needs to be modified to include the weights of the historical input nodes.

$$h_t' = H(C, h_{t-1}, , x_t). \tag{12}$$

2.3. The EFAST Analysis. The EFAST analysis technique employs variance analysis for global sensitivity analysis to assess the sensitivity of daily oil production to daily polymer injection. Specifically, this study utilizes a variance decomposition approach to determine the contribution of different daily polymer injections from injectors to the total variance of a trained daily oil production dynamic prediction model, which yields the IWC coefficient. The dynamic prediction model for daily oil productions is denoted as $Y = f(x_1, x_2, \cdots, x_k)$, where $x_1, x_2, \cdots, x_k$ refer to the first, second, and kth input factors, each comprising multiple time nodes. The model variance D quantifies the uncertainty associated with the impact of daily polymer injection from injectors on the daily oil production of extraction wells.

The process of mapping a multidimensional input time series onto a one-dimensional search space $s \in (-\infty, +\infty)$, denoted as $Y = f(x_1, x_2, \cdots, x_k)$, is transformed into a one-dimensional representation $Y = f(s)$, where x_k denotes the kth dimensional input factor in the multidimensional input time series. Equation (13) illustrates that each input factor x_k can be expressed as a specific frequency ω_k.

$$x_k(s) = G_k(\sin \omega_k \bullet s),\ \forall k = 1, 2, \cdots, n. \tag{13}$$

The search function G_k is determined by the probability density function of the model input factor x_k, as stated in reference [26], where ω_k denotes linearly uncorrelated positive integer frequencies. Equation (14) provides the Fourier transform process of $f(s)$.

$$f(s) = \sum_{j=-\infty}^{+\infty} \left(A_j \cos \omega_j s + B_j \sin \omega_j s\right). \tag{14}$$

If the function s is sampled at equal intervals n times in the interval $[-\pi, \pi]$, resulting in the sampling points s_0, s_1, …, s_{n-1}, which are then inputted to the model, the corresponding Fourier coefficients A_j and B_j for $j \in Z = (-\infty, +\infty)$ can be approximated as shown in

$$\begin{aligned} A_j &= \frac{1}{N_s} \sum_{n=1}^{N_s} f(s_n) \cos\left(\omega_j s_n\right), \\ B_j &= \frac{1}{N_s} \sum_{n=1}^{N_s} f(s_n) \sin\left(\omega_j s_n\right). \end{aligned} \tag{15}$$

The variable N_s denotes the sample size which can be denoted by

$$N_s = 2M\omega_{\max} + 1,\ M \in N^*. \tag{16}$$

M denotes the maximum number of harmonics, which is usually taken as either 4 or 6. $\omega_{\max}$ denotes the maximum value in the set of frequencies ω_k.

The expected variance D_k of the input factor x_k can be obtained using Parseval's theorem. The expected variance D_k of the input factor x_k shown in Equation (17) is calculated using Perceval's theorem.

$$D_k = 2 \sum_{m=1}^{M} \left(A_{m\omega_k}^2 + B_{m\omega_k}^2\right),\ m = 1, 2, \cdots, n. \tag{17}$$

The overall variance of the model is further obtained as D shown in

$$D = 2 \sum_{m=1}^{M} \left(A_m^2 + B_m^2\right). \tag{18}$$

m denotes the number of harmonics, and the parameters $A_{m\omega_k}$ and $B_{m\omega_k}$ denote the two Fourier coefficients corresponding to the mth harmonic.

The first-order sensitivity index S_{Fk} shown in Equation (19), which disregards the coupling effect of other input factors with x_k, can be obtained using the expected variance of x_k and the overall variance of the model. This sensitivity evaluation result can be referred to as the local sensitivity analysis result. To obtain the global sensitivity analysis result, the contribution of the coupling effect between the input factor x_k and other input factors $x_{\sim k}$ to the overall variance of the model must be considered. Finally, the global sensitivity index S_{Tk} shown in Equation (20) of the input factor x_k can be obtained.

$$S_{Fk} = \frac{D_k}{D}, \tag{19}$$

$$S_{Tk} = 1 - \frac{D_{\sim k}}{D}. \tag{20}$$

D_k denotes the variance of the input factor x_k, and $\sim k$ denotes the values of all input factors except x_k. Thus, $D_{\sim k}$ denotes the sum of the total variance of all other input factors.

To visually compare the connectivity status between injectors and producers, this paper maps the global sensitivity indices of all input factors to the [0, 1] interval. This mapping is done to obtain the normalized global sensitivity index, which is referred to as the IWC coefficient S_{Tk}^* shown in

$$S_{Tk}^* = \frac{S_{Tk}}{\sum_{k=1}^{n} S_{Tk}},\ k = 1, 2, \cdots, n. \tag{21}$$

3. Experimental

3.1. Data Preprocessing

3.1.1. Data Acquisition and Feature Extraction. This paper focuses on a test well group located in the *S*-well area of the *D* field, which is characterized by medium to high permeability. The well group, featuring a typical five-point method well network deployment, comprises of producers labeled as *P* and surrounding injectors labeled as *W*1, *W*2, …, and *W*4. Historical production data from the polymer flooding phase of the well group were collected between October 26, 2020, and June 10, 2022. The multidimensional time series dataset comprises 5 dimensions and 593 time nodes, arranged in chronological order, which includes injection rate data from four injectors and recovery rate data from one extraction well. These data provide insights into the dynamic behavior of the oilfield's production.

3.1.2. Data Cleaning and Normalization

(1) Data Cleaning.

(1) Missing Value Processing

The KNN interpolation method stands for K-nearest neighbor interpolation. It is a simple and effective method for filling in missing values in a dataset. The basic idea of KNN interpolation is to find the K-nearest neighbors to

the missing value and calculate the average of their values to fill in the missing value. The distance between the missing value and other data points is calculated using a distance metric, such as Euclidean distance or Manhattan distance. In the context of the study mentioned in the question, the KNN interpolation method was used to fill the missing daily polymer injection data, which was fluctuating and could not be accurately filled using the empirical interpolation method. The KNNImputer function from the scikit-learn in Python 3.9 can be used to fill in the missing values in the dataset. The n_neighbors parameter determines the number of neighboring data points to use in the estimation process, and the weight parameter can be set to "distance" to give more weight to closer neighbors. Based on prior knowledge and validation, setting n_neighbors = 3 and weights = "distance" is a reasonable choice for filling in missing values.

(2) Outlier Handling

The boxplot method was utilized to detect outliers in the daily polymer injection of injectors shown in Figure 3 and the daily oil production of producers shown in Figure 4. Further analysis was conducted to determine the reason behind the identified outliers. Based on the investigation results, a decision was made to either treat the outliers as missing values or retain them in the dataset due to the influence of the injector production system.

Based on the box line diagram of daily polymer injection, it is apparent that there are four anomalous values in the daily polymer injection of injector $W2$. Upon analyzing the dynamic data of daily polymer injection, it is possible to determine the specific time period where these anomalous values occurred. Further analysis reveals that during this time period, the daily polymer injection of $W2$ well as a whole was adjusted to over 95 m^3/day. Additionally, field production measure records indicate that hydraulic fracturing measures were conducted on the new production level of the $W2$ well during this time period, leading to an increase in its daily polymer injection. Therefore, there is no need to address the anomalous values as they can be attributed to the aforementioned hydraulic fracturing measures.

(2) Normalization. The normalization function was developed in Python 3.9 within the Spyder integrated environment. Once construction was completed, the cleaned production dynamic data was normalized to mitigate potential numerical issues, low model accuracy, and difficulty in model convergence. However, when making predictions, the predicted values must be reverse normalized in order to compare and analyze the results with the predicted target.

(3) Dataset Division. Following the normalization of the input dataset, it is necessary to split the data into a training set and a test set at an 8 : 2 ratio. Specifically, the first 475 timestamps of data will comprise the training set data, while the remaining 118 timestamps of data will serve as the test set data.

3.2. A-LSTM Model Training and Optimization. During the model training process, two key areas require attention:

(1) Model learning rate setting: this involves defining the initial learning rate of the model, as well as selecting a loss function that allows for dynamic adjustment of the learning rate

(2) Model hyperparameter and structure optimization: this encompasses selecting appropriate methods for optimizing both the model's hyperparameters and its overall structure

3.2.1. A-LSTM Model Learning Rate Adjustment. The Adam [27] adaptive optimizer is utilized to adjust the learning rate of the training set during iterative updates of the A-LSTM network weights, thereby minimizing the loss function and increasing the model's convergence speed. To ensure optimal convergence, it is crucial to select an appropriate initial learning rate, as values that are too large may cause the model to fail to converge or skip optimal/suboptimal solutions, while values that are too small can result in slow convergence and increased training time. The selection of a suitable loss function is also critical, as it serves as the learning rate adjustment evaluation function to estimate the deviation of the model's daily oil production prediction from the actual value. In this study, we adopt the smoothed mean absolute error (Huber loss) [28] as the loss function L_δ shown in Equation (22), which combines the advantages of mean square error (MSE) and mean absolute error (MAE).

$$L_\delta(y, f(x)) = \begin{cases} \frac{1}{2}\left(y_i^{\text{obs}} - y_i^{\text{pred}}\right)^2, \text{if } \left|y_i^{\text{obs}} - y_i^{\text{pred}}\right| \le \delta, \\ \delta\left|y_i^{\text{obs}} - y_i^{\text{pred}}\right| - \frac{1}{2}\delta^2, \text{if } \left|y_i^{\text{obs}} - y_i^{\text{pred}}\right| > \delta, \end{cases} \tag{22}$$

where y_i^{obs} denotes the actual value of daily oil production, y_i^{pred} denotes the predicted value of daily oil production, and δ denotes the parameter obtained by cross-validation of the Huber function; when $\delta \sim 0$, the Huber loss will tend to MAE; when $\delta \sim \infty$, the Huber loss will tend to MSE.

3.2.2. A-LSTM Model Hyperparameter Optimization. The Keras platform is used for hyperparameter optimization of the IWC evaluation model in this study. The hyperparameters that are optimized in the model include the number of hidden layers of the A-LSTM and the number of nodes within the hidden layers. The optimization of these hyperparameters has a direct impact on the model's ability to reflect the complexity of IWC, as well as its accuracy and generalization ability for prediction. To optimize these hyperparameters, we use a genetic algorithm (GA) [29], which is known for its ease of implementation, strong robustness, and ability to find globally optimal solutions compared to gradient search-based hyperparameter optimization algorithms.

3.3. Evaluation of Model Prediction Effectiveness. The effectiveness of model prediction is evaluated through three metrics, the MAE, the root mean square error (RMSE), and the

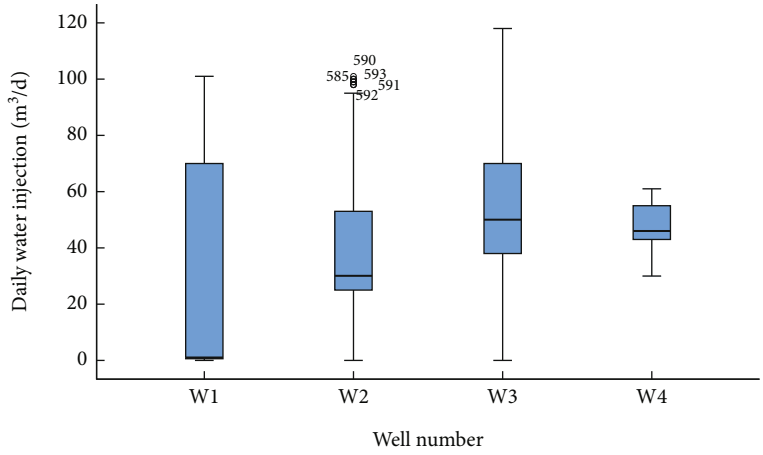


Figure 3: Box line diagram of daily polymer injection of four injectors.

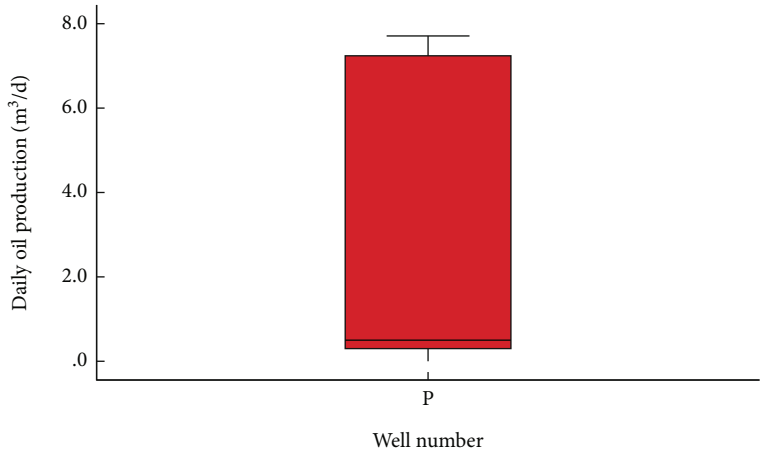


Figure 4: Box line diagram of daily oil production of a producer.

adjusted coefficient of determination ($\bar{R}^2$), as shown in the following equations:

$$\begin{aligned} \mathrm{MAE} &= \frac{1}{n}\sum_{i=1}^{n}\left|y_i^{\mathrm{obs}} - y_i^{\mathrm{pred}}\right|, \\ \mathrm{RMSE} &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i^{\mathrm{obs}} - y_i^{\mathrm{pred}}\right)^2}, \\ R^2 &= 1 - \frac{\sum_{i=1}^{n}\left(y_i^{\mathrm{obs}} - y_i^{\mathrm{pred}}\right)^2}{\sum_{i=1}^{n}\left(y_i^{\mathrm{obs}} - \bar{y}_i^{\mathrm{obs}}\right)^2}, \\ \bar{R}^2 &= 1 - \frac{\left(1-R^2\right)\left(N_s - 1\right)}{N_s - p - 1}. \end{aligned} \tag{23}$$

The actual daily oil productions are denoted as y_i^{obs}, the predicted daily oil productions are denoted as y_i^{pred}, and the number of independent variables is denoted as p.

3.4. Calculation of IWC Factor. The EFAST sensitivity analysis technique on a global scale necessitates the establishment of diverse parameters, with preeminent emphasis accorded to the identification of the search function and the interference factor, as shown in Equations (24) and (25). The search function implemented in the analysis is predicated upon the G function adopted in Sobol's sensitivity analysis technique, which utilizes the Monte Carlo sampling methodology [30].

$$G(X_1, X_2, \cdots, X_k, a_1, a_2, \cdots, a_k) = \prod_{i=1}^{k} g_i, \tag{24}$$

$$g_i = \frac{|4X_i - 2| + a_i}{1 + a_i},\ i = 1, 2, 3 \cdots k. \tag{25}$$

a_i denotes the minimum value of the ith input factor.

Based on the equations for determining the number of sampling times and input factors ($N = 2q + 1$) and given that there are 4 input factors, the number of sampling times was set to 9. In consideration of the number of samples and data dimension, an interference factor of $M = 4$ was employed. First-order and global sensitivity indices were subsequently calculated and mapped onto the interval [0,1] to provide a

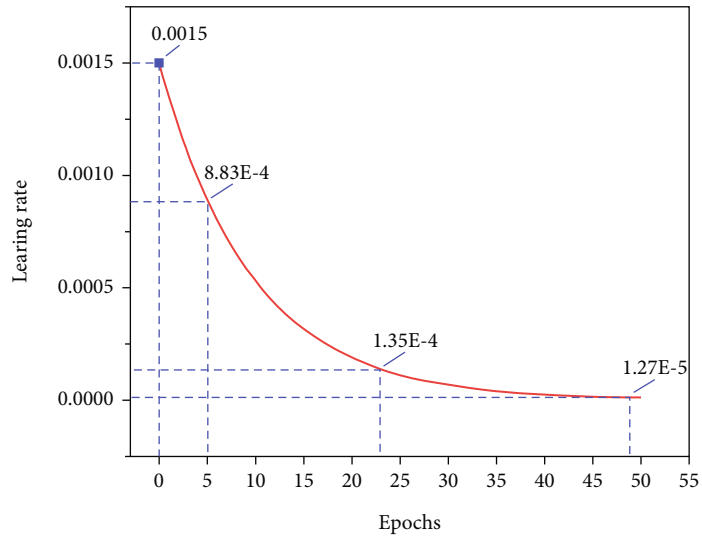


Figure 5: The learning rate curve of the training set.

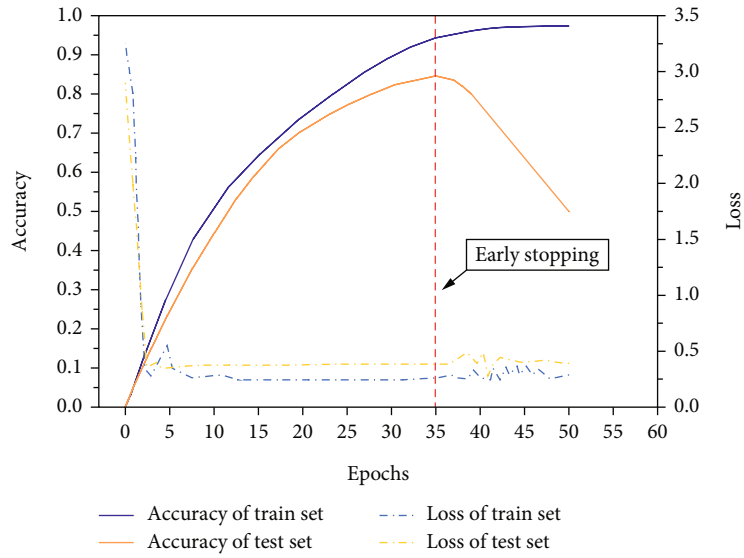


Figure 6: The accuracy and Huber loss rate curves of both the training and test sets.

visual representation of the connectivity between injectors and producers. This information can then be used to adjust the development plan between injectors and producers in a timely manner.

4. Results and Discussion

4.1. Results of A-LSTM Model Training and Optimization

4.1.1. Learning Rate Adjustment Results. The optimal value for the hyperparameter δ is determined to be 0.3. Furthermore, the initial learning rate of 0.0015 is employed along with exponential decay rates of 0.85 and 0.999 for the first-order and second-order moment estimations, respectively. The optimal values of the hyperparameters epochs and batch_size are set to 35 and 16, respectively, via a combination of accuracy, Huber loss rate change curves, and results obtained using an early stopping mechanism applied to both the training and test sets. Figure 5 shows the learning rate curve of the training set, while Figure 6 shows the accuracy and Huber loss rate curves of both the training and test sets.

4.1.2. Model Hyperparameter Optimization Results. In order to fit nonlinear data, the number of LSTM layers and dense layers was explored in the range of 1-3. It was determined that a 3-layer network was sufficient for this purpose. Using a GA for hyperparameter optimization, the optimal number of LSTM layers was determined to be 2, with a hidden layer of 16 neurons. The input_shape required three input parameters, namely, sample, time steps, and feature, with values of 475, 1, and 4, respectively. In order to evaluate the accuracy of the A-LSTM algorithm, this paper also optimizes the LSTM and SVR algorithms. The optimized hyperparameters of LSTM are as follows: the LSTM layer is still 2 layers, the

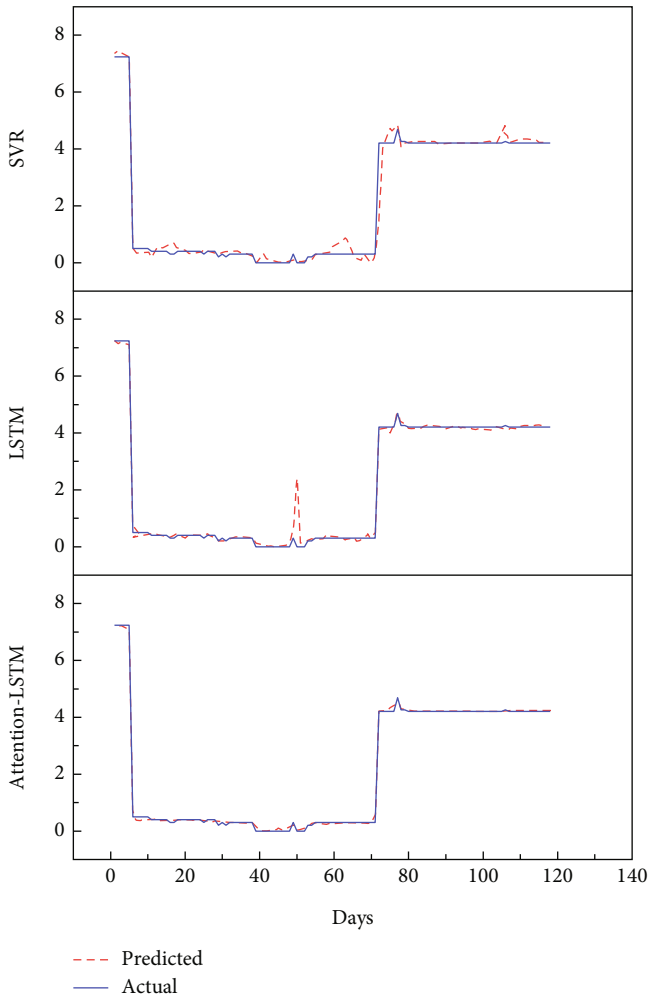


Figure 7: Performance of the test set of three algorithmic models of A-LSTM, LSTM, and SVR in predicting daily oil production.

optimal values of epochs and batch_size are set to 51 and 8, respectively, and the change of the dynamic learning rate is the same as that of the A-LSTM setup method, while the SVR model chooses the commonly used Gaussian radial basis function as the nonlinear kernel function. For SVR model, Gaussian radial basis function is chosen as the nonlinear kernel function, and the optimized kernel function parameter is 0.022 and the error boundary is 0.251, and the penalty factor C and the window length are set to 3 and 12, respectively.

4.2. Prediction Results of Three Prediction Models, A-LSTM, LSTM, and SVR. Figure 7 shows the A-LSTM model and compares its performance to that of the LSTM model and the SVR algorithm model in predicting the daily oil production for the test set. The results clearly demonstrate that the A-LSTM model outperforms both the LSTM model and the SVR algorithm model in predicting the daily oil production of a single well.

The predicted daily oil production curve generated by the A-LSTM model is smoother and flatter compared to the curves produced by the other two models. However, it was observed that the LSTM model generated a high anomaly in the test data at day 51, which did not correspond to any significant abrupt changes in the actual daily oil production time curve. Upon further investigation, it was found that there was a significant increase in the daily polymer injection of an injector connected to the tested well before this time point. This suggests that the accuracy of the model's predictions can be affected when the input data produces abrupt changes. However, such points should not be treated as anomalies in the actual production process, as they may be the result of expanding polymer injection or production measures. For such problems, data cleaning should be performed according to the magnitude of the data signal-to-noise ratio. The presence of a certain degree of data noise can make the data more robust as a whole. The A-LSTM model applied in this study accurately excluded the outlier data as the object of concern, maintaining high prediction accuracy when predicting other untrained production dynamic data daily oil production time series, i.e., the test set. The performance of the three models was further evaluated using three evaluation functions, MAE, RMSE, and $\bar{R}^2$, and the results are shown in Table 2.

The best performance of the A-LSTM model can be seen from the performance evaluation results of the model.

Table 2: Performance evaluation results of three algorithm models of A-LSTM, LSTM, and SVR.

Model	Train set			Test set		
	MAE	RMSE	$\bar{R}^2$	MAE	RMSE	$\bar{R}^2$
A-LSTM	0.058	0.293	0.992	0.089	0.351	0.989
LSTM	0.263	0.481	0.961	0.598	0.626	0.935
SVR	0.966	1.695	0.912	1.356	2.215	0.869

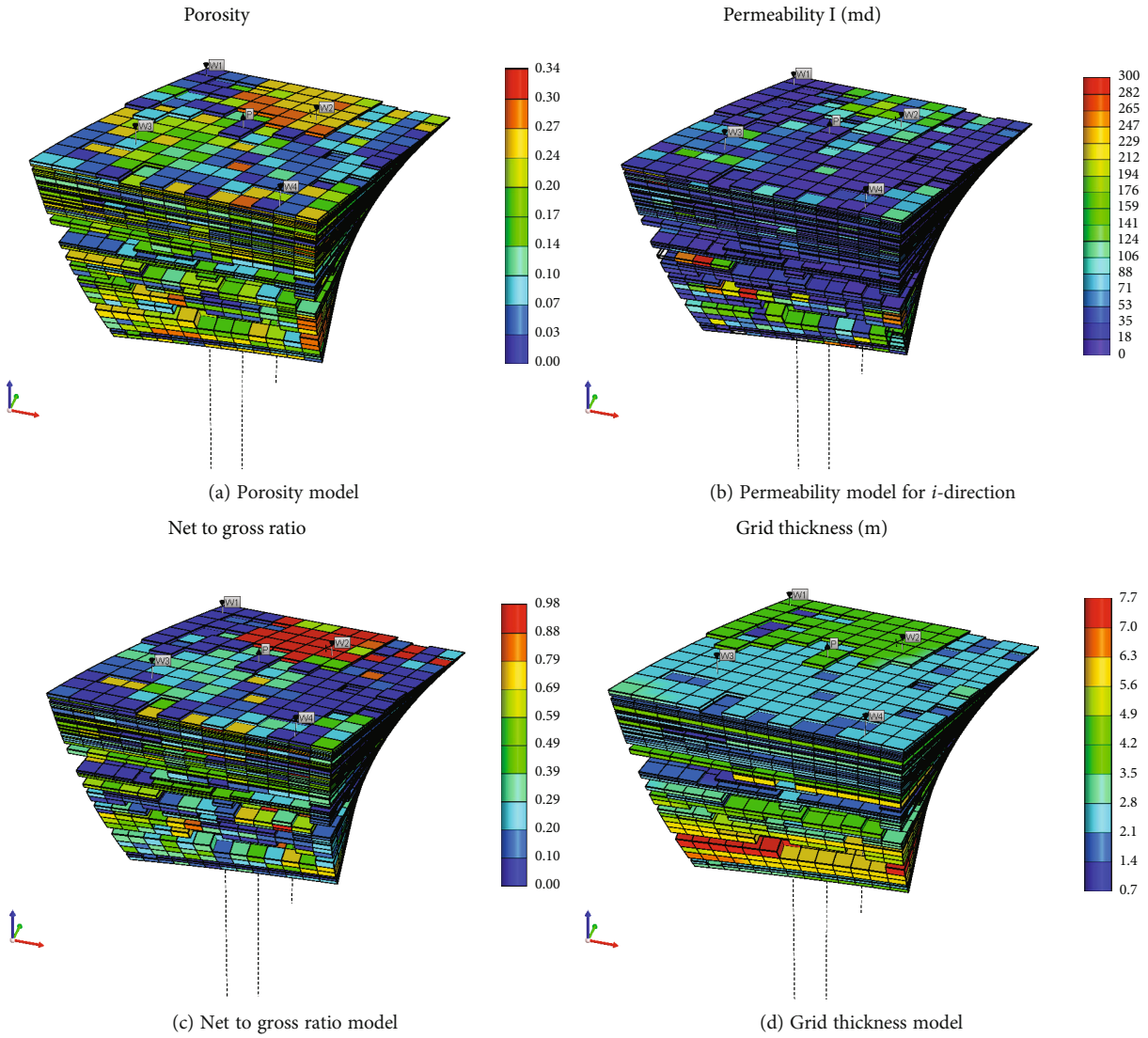
Table 3: Global sensitivity index of the test well area.

Model	Producer	Injectors			
		$W1$	$W2$	$W3$	$W4$
A-LSTM	P	0.89	0.35	0.78	0.16
LSTM		0.85	0.41	0.66	0.18
SVR		0.79	0.31	0.89	0.11

Table 4: IWC factor in the test well area.

Model	Producer	Injectors			
		$W1$	$W2$	$W3$	$W4$
A-LSTM	P	0.41	0.16	0.36	0.07
LSTM		0.40	0.20	0.31	0.09
SVR		0.38	0.15	0.42	0.05

4.3. Evaluation Results of IWC. Normalized global sensitivity indices in Table 3 are utilized to derive IWC coefficients in Table 4. Using the results of IWC calculations in Tables 3 and 4, it can be analyzed that the connectivity between the $W4$ wells and the production P is poor, which is 0.1 or less,



(a) Porosity model

(b) Permeability model for *i*-direction

(c) Net to gross ratio model

(d) Grid thickness model

Figure 8: The constructed three-dimensional numerical simulation model.

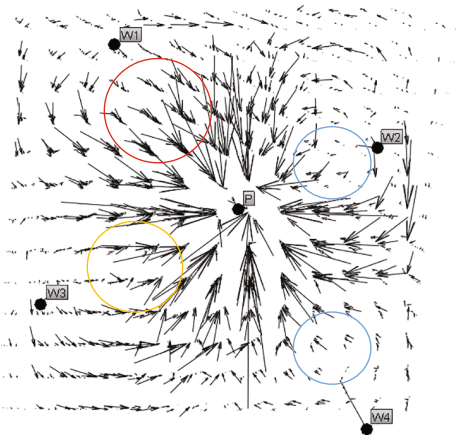


Figure 9: The three-dimensional polymer phase flow field of the test well area in June 2022.

while the *W*1-*W*3 wells have a good connectivity with the production well.

4.4. Reliability Verification

4.4.1. The Numerical Simulation Method. We construct a numerical simulation model of polymer flooding in nonhomogeneous reservoirs based on the dynamic and static physical parameters of the actual production process in the test well area, so as to further verify the reliability of the method used in this paper, and the constructed three-dimensional numerical simulation model is shown in Figure 8. The oil-bearing area of the test area is 0.38 km^2, with a geological reserve of 23.62×10^4 t and a pore volume of 85.7×10^4 m^3. During the model construction process, Petrol and CMG software were used for the establishment of the geological model and the numerical simulation model, with a grid step length of 25 m, vertically including 21 simulation layers, containing the numerical simulation model (Figure 8) grid $13 \times 13 \times 21$ with 3549 grids.

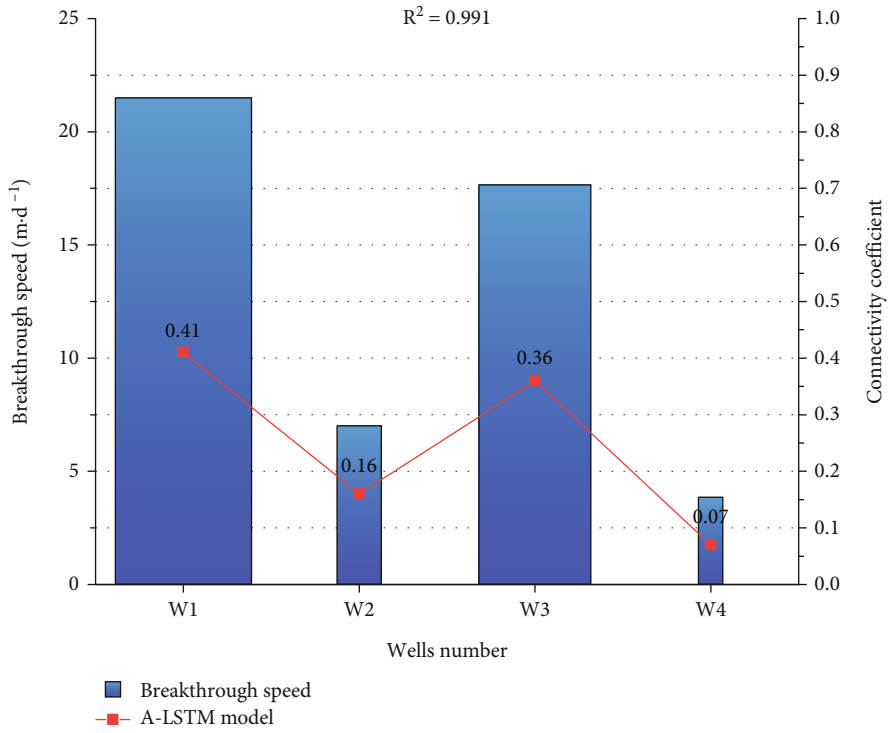


Figure 10: Comparison of tracer breakthrough velocity test and A-LSTM model IWC coefficient prediction results.

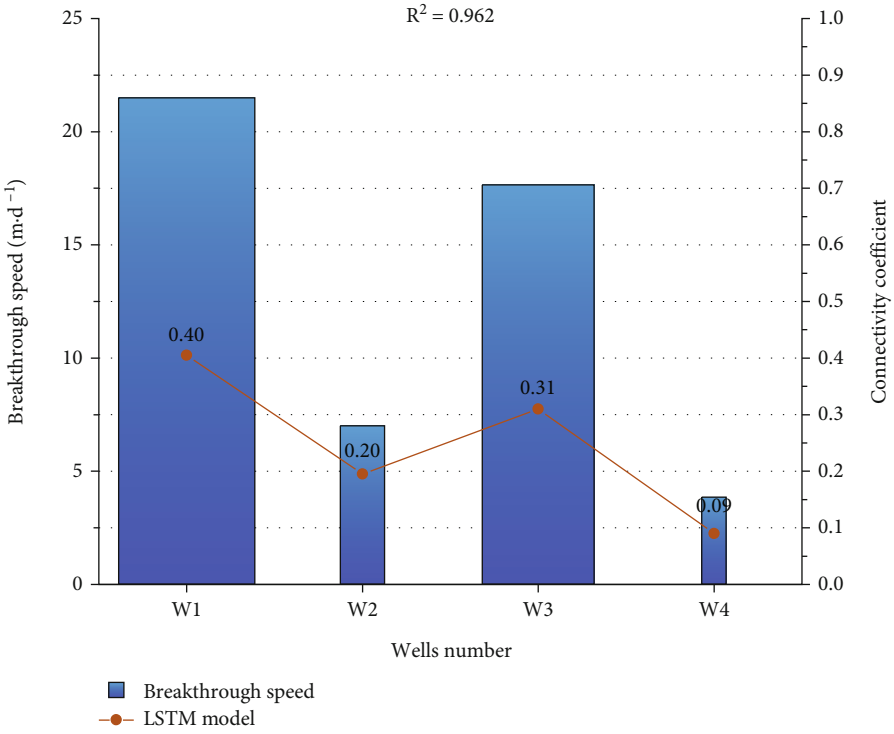


Figure 11: Comparison of tracer breakthrough velocity test and LSTM model IWC coefficient prediction results.

The model was employed to validate the methodology presented in this paper. The results of the simulation include the three-dimensional distribution of the polymer phase flow field in June 2022, as shown in Figure 9, and the distribution of the three-dimensional oil and polymer phase flow, as shown in Figure 10.

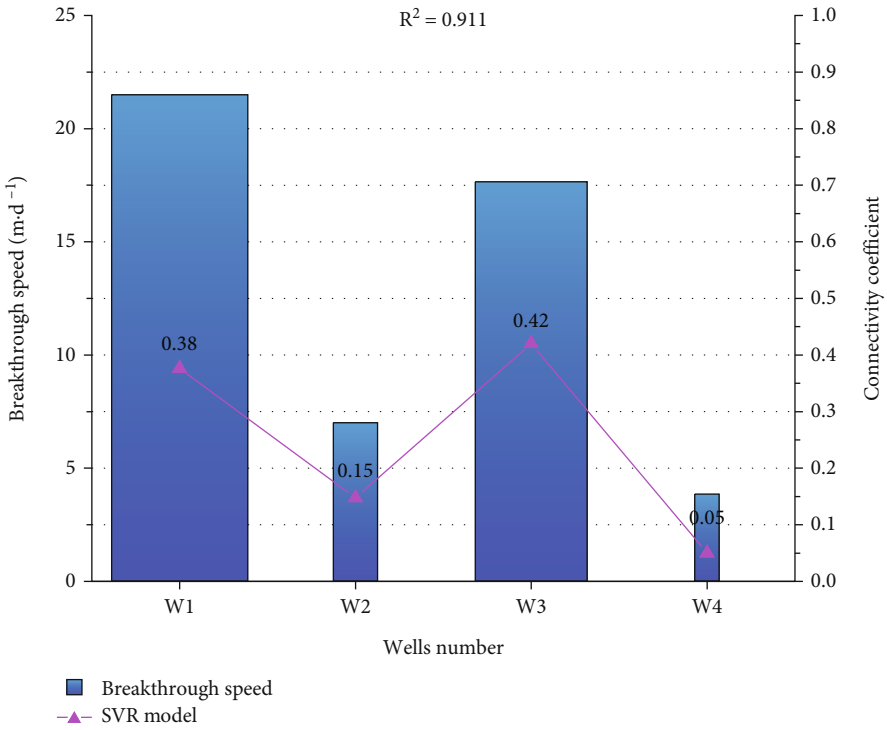


Figure 12: Comparison of tracer breakthrough velocity test and SVR model IWC coefficient prediction results.

The three-dimensional polymer phase flow field diagram provides a visual representation of the flow lines and their densities between the wells, enabling us to assess the IWC. The observation of dense flow lines between well $W1$ and well P indicates strong connectivity between them, followed by well $W3$, while the flow lines between wells $W2$ and $W4$ and well P are sparse, indicating weaker connectivity between them. Overall, the three-dimensional polymer phase flow field provides a valuable tool for qualitatively evaluating IWC, while the model developed in this paper enables a more quantitative assessment.

4.4.2. The Interwell Tracer Testing. To assess the accuracy of the calculated results from three algorithm models, namely, A-LSTM, LSTM, and SVR, the interwell tracer test results were employed for evaluation. In September 2022, a new tracer was used in the testing process that is more environmentally friendly. Compared with the traditional tracer, the new type of tracer has the characteristics of nonwater solubility, insoluble in water, and distributed in the form of spherical droplets, so that it can not spread the concentration and is less contaminated. The new tracer consists of a variety of non-water-soluble liquid mixture, specific gravity of 0.8 g/cm^3~1.6 g/cm^3.The breakthrough velocity of the tracer can serve as an approximation of the influence coefficient (i.e., connectivity coefficient) of the producer by the surrounding injection wells. The R^2 coefficient of determination was used to calculate the variability of the model-predicted connectivity coefficients with respect to the tracer test results, leading to the results shown in Figures 10–12. The tracer test results indicate that well P is affected by four injection wells in the dominant connection direction of wells $W1$ and $W3$, with breakthrough velocities of 23.0 and 13.8 m/d, respectively. The coefficient of determination between the predicted IWC coefficient of the A-LSTM model and the tracer breakthrough velocity test results was found to be 0.991, as shown in Figure 10. The coefficients of determination for the LSTM and SVR models were found to be 0.962 and 0.911, as shown in Figures 11 and 12.

5. Conclusions and Future

This paper introduces a novel methodology that integrates A-LSTM with EFAST to enhance the accuracy of predicting daily oil production time series during the polymer flooding stage and to assess interwell connectivity (IWC) in real reservoirs. The proposed approach consists of three key stages. Initially, data preprocessing techniques are employed to enhance data quality. In the subsequent stage, the attention mechanism is incorporated into the LSTM algorithm to develop the A-LSTM algorithm, which is then compared with LSTM and SVR algorithms in terms of multiple performance evaluation metrics for predicting single-well daily oil production. Utilizing the Huber loss function as the error function enhances the model's resilience and reduces susceptibility to outliers, resulting in superior performance of the A-LSTM algorithm in accurately forecasting daily oil production. In the final stage, EFAST global sensitivity analysis is utilized to estimate IWC coefficients between producers and injectors using the dynamic prediction model of daily oil production. The proposed method offers several advantages, including maximizing data quality through various preprocessing techniques and capturing essential time series features while filtering out irrelevant information via

A-LSTM. Additionally, the EFAST analysis method effectively evaluates connectivity and polymer injection effects in multiple directions, a capability unmatched by other local sensitivity analysis methods. To validate the proposed approach, numerical simulations are conducted to generate three-dimensional flow field maps of the polymer phase in the test area. Furthermore, tracer test results are employed to assess and compare connectivity strength and weakness between polymer injection and oil recovery wells. The findings demonstrate close alignment between the outcomes of the IWC assessment model and results obtained from numerical simulations and interwell tracer tests, indicating the robustness of the proposed model for guiding injection and production operations in the field.

In the future, the proposed method can be further optimized by parallelizing the historical data of the polymer flooding phase in multiple well groups, improving the interpretation of IWC coefficients, and exploring the quantitative assessment method of interstratigraphic connectivity status.

Data Availability

The datasets generated and analyzed during this study need to be obtained from the corresponding authors using ∗.csv format data with the consent of the relevant departments of the oilfield, which can be obtained by contacting the corresponding author.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

We are extremely grateful for the support of the National Natural Science Foundation of China and the field data provided by the DQ Oilfield as the basis for this study. This work was supported by the National Natural Science Foundation of China (51874096).

References

[1] F. Liu, J. M. Mendel, and A. M. Nejad, "Forecasting injector/producer relationships from production and injection rates using an extended Kalman filter," *SPE Journal*, vol. 14, no. 4, pp. 653–664, 2009.

[2] S. H. Yousefi, F. Rashidi, M. Sharifi, M. Soroush, and A. J. Ghahfarokhi, "Interwell connectivity identification in immiscible gas-oil systems using statistical method and modified capacitance-resistance model: a comparative study," *Journal of Petroleum Science and Engineering*, vol. 198, article 108175, 2021.

[3] K. J. Heffer, R. J. Fox, C. A. McGill, and N. C. Koutsabeloulis, "Novel techniques show links between reservoir flow directionality, earth stress, fault structure and geomechanical changes in mature waterfloods," *SPE Journal*, vol. 2, no. 2, pp. 91–98, 1997.

[4] A. Albertoni and L. W. Lake, "Inferring Interwell connectivity only from well-rate fluctuations in waterfloods," *SPE Reservoir Evaluation & Engineering*, vol. 6, no. 1, pp. 6–16, 2003.

[5] A. A. Yousef, P. Gentil, J. L. Jensen, and L. W. A. Lake, "A capacitance model to infer interwell connectivity from production- and injection-rate fluctuations," *SPE Reservoir Evaluation & Engineering*, vol. 9, no. 6, pp. 630–646, 2006.

[6] D. Kaviani and P. P. Valkó, "Inferring interwell connectivity using multiwell productivity index (MPI)," *Journal of Petroleum Science and Engineering*, vol. 73, no. 1-2, pp. 48–58, 2010.

[7] M. N. Panda and A. K. Chopra, "An integrated approach to estimate well interactions," in *SPE India Oil and Gas Conference and Exhibition*, Mumbai, India, 2010.

[8] U. Demiryurek, F. Banaei-Kashani, C. Shahabi, and F. Wilkinson, "Neural-network based sensitivity analysis for injector-producer relationship identification," in *SPE Intelligent Energy International Conference and Exhibition*, Amsterdam, The Netherlands, 2008.

[9] M. Sayarpour, E. Zuluaga, C. S. Kabir, and L. W. Lake, "The use of capacitance–resistance models for rapid estimation of waterflood performance and optimization," *Journal of Petroleum Science and Engineering*, vol. 69, no. 3, pp. 227–238, 2009.

[10] M. Sayarpour, C. S. Kabir, and L. W. Lake, "Field applications of capacitance-resistance models in waterfloods," *SPE Reservoir Evaluation & Engineering*, vol. 12, no. 6, pp. 853–864, 2009.

[11] Z. Zhang, H. Li, and D. Zhang, "Water flooding performance prediction by multi-layer capacitance-resistive models combined with the ensemble Kalman filter," *Journal of Petroleum Science and Engineering*, vol. 127, pp. 1–19, 2015.

[12] A. S. Elons, M. Y. Elgendy, and D. A. Magdy, "A real time dynamic model for optimizing hydrocarbons' production based on deep recurrent network," in *2016 11th International Conference on Computer Engineering & Systems (ICCES)*, pp. 483–488, Cairo, Egypt, 2016.

[13] H. Cheng, V. Vyatkin, E. Osipov, P. Zeng, and H. Yu, "LSTM based EFAST global sensitivity analysis for interwell connectivity evaluation using injection and production fluctuation data," *IEEE Access*, vol. 8, pp. 67289–67299, 2020.

[14] Y. Jiang, H. Zhang, K. Zhang et al., "Reservoir characterization and productivity forecast based on knowledge interaction neural network," *Mathematics*, vol. 10, no. 9, 2022.

[15] 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 106367, 2019.

[16] H. Wang, J. Han, K. Zhang et al., "An interpretable interflow simulated graph neural network for reservoir connectivity analysis," *SPE Journal*, vol. 26, no. 4, pp. 1636–1651, 2021.

[17] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, http://arxiv.org/abs/1409.0473.

[18] S. Van Buuren and K. Groothuis-Oudshoorn, "Mice: multivariate imputation by chained equations inR," *Journal of Statistical Software*, vol. 45, no. 3, pp. 1–67, 2011.

[19] O. Troyanskaya, M. Cantor, G. Sherlock et al., "Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001.

[20] D. R. Cutler, T. C. Edwards Jr., K. H. Beard et al., "Random forests for classification in ecology," *Ecology*, vol. 88, no. 11, pp. 2783–2792, 2007.

[21] A. Nowak-Brzezińska and I. Gaibei, "How the outliers influence the quality of clustering?," *Entropy*, vol. 24, no. 7, p. 917, 2022.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[23] M. M. Salamattalab, M. Hasani Zonoozi, and M. Molavi-Arabshahi, "Innovative approach for predicting biogas production from large-scale anaerobic digester using long-short term memory (LSTM) coupled with genetic algorithm (GA)," *Waste Management*, vol. 175, pp. 30–41, 2024.

[24] Y. Zhang, C. Li, H. Duan, K. Yan, J. Wang, and W. Wang, "Deep learning based data-driven model for detecting time-delay water quality indicators of wastewater treatment plant influent," *Chemical Engineering Journal*, vol. 467, article 143483, 2023.

[25] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[26] A. Saltelli, S. Tarantola, and K. P.-S. Chan, "A quantitative model-independent method for global sensitivity analysis of model output," *Technometrics*, vol. 41, no. 1, pp. 39–56, 1999.

[27] D. Kinga and J. B. Adam, "A method for stochastic optimization," *International Conference on Learning Representations (ICLR)*, vol. 5, 2015.

[28] P. J. Huber, "A robust version of the probability ratio test," *The Annals of Mathematical Statistics*, vol. 36, no. 6, pp. 1753–1758, 1965.

[29] Y. Chen, Z. Dong, Y. Wang et al., "Short-term wind speed predicting framework based on EEMD-GA-LSTM method under large scaled wind history," *Energy Conversion and Management*, vol. 227, article 113559, 2021.

[30] S. Burhenne, D. Jacob, and G. P. Henze, "Sampling based on Sobol'sequences for Monte Carlo techniques applied to building simulations," in *Proceedings of Building Simulation, 12th Conference of International Building Performance Simulation Association*, pp. 1816–1823, Sydney, 2011.