WILEY | Hindawi

Research Article

CNN-LSTM Model Optimized by Bayesian Optimization for Predicting Single-Well Production in Water Flooding Reservoir

Lei Zhang [ID],[1,2] **Hongen Dou** [ID],[1,2] **Kun Zhang**,[3] **Ruijie Huang** [ID],[1] **Xia Lin**,[1,2] **Shuhong Wu**,[1,2] **Rui Zhang**,[4] **Chenjun Zhang**,[1] **and Shaojing Zheng**[1]

[1]*Research Institute of Petroleum Exploration and Development, Beijing 100083, China*
[2]*Artificial Intelligence Technology R&D Center for Exploration and Development, CNPC, Beijing 100083, China*
[3]*The Twelfth Oil Production Plant, Changqing Oilfield Company, Qingyang 745099, China*
[4]*Changqing Engineering Design Co., Ltd., PCOC, Xi'an 710021, China*

Correspondence should be addressed to Hongen Dou; douhongen@petrochina.com.cn
and Ruijie Huang; huangruijie@petrochina.com.cn

Geared toward the problems of predicting the unsteadily changing single oil well production in water flooding reservoir, a machine learning model based on CNN (convolutional neural network) and LSTM (long short-term memory) is established which realizes precise predictions of monthly single-well production. This study is considering more than 60 dynamic and static factors that affect the changes of oil well production, introduce water injection parameters into data set, select 11 main control factors, and then, build a CNN-LSTM model optimized by Bayesian optimization. The effectiveness of the proposed model is verified in a realistic reservoir. The experiment results show that the prediction accuracy of the proposed model is over 90%, which suggests a penitential application in an extensive range of applications. Production forecasting by the developed model is simple, efficient, and accurate, which can provide a guidance for the dynamic analysis of a water flooding reservoir, and work as a good reference of the development and production of other types of reservoirs.

1. Introduction

Oil production prediction runs through the entire development course of a water-driven oilfield; it is the foundation of the well stimulation and plays an important role in making investment decisions. For a long time, the main production prediction methods are reservoir numerical simulation, mechanism model, and decline curve analysis, which have their respective advantages and disadvantages. Since the 1990s, fuzzy comprehensive evaluation (FE), BP neural network (BPNN), grey model (GM), and other methods have been applied to oilfield production prediction [1, 2]. Recently, with the maturity and wide application of big data, artificial intelligence, and other technical theories and methods, more and more intelligent algorithms have been introduced into the petroleum industry, providing a new way to solve complex engineering problems [3, 4].

As a nonlinear fitting method, machine learning can learn rules from data and make a prediction. In the past 20 years, there has been extensive research on production prediction based on machine learning. Common methods include random forest (RF), support vector machine (SVM), fuzzy comprehensive evaluation (FE), artificial neural network (ANN), and autoregressive integrated moving average model (ARIMA) [5–14]. These classical machine learning methods have been fully applied in the field of petroleum industry and show strong vitality.

Y. Duana et al. use RTS (Rauch Tung Striebel) to smooth the gas production series, and then, an ARIMA model was established to predict the gas production [6]. J. Gu et al. proposed an oil well production prediction model which combines ARIMA and Kalman filter to eliminate the influence of nonsynchronicity and hysteresis [7]. Viet Nguyen-Le et al. propose 3 ANN models to predict the
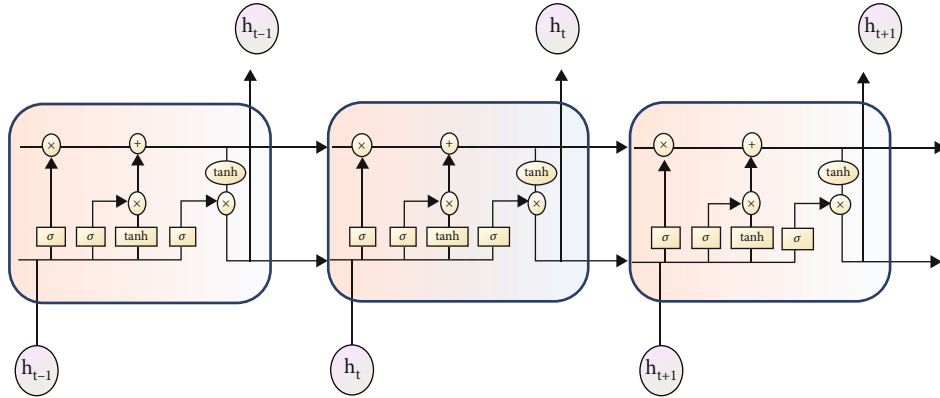
FIGURE 1: The basic structure of LSTM cell. In each cell unit, long-term information and short-term information are conveyed, respectively.

parameters in Arps's hyperbolic decline and then reconstructs the production profile [13]. Y. Zhu et al. established an "ε-SVM" production prediction model based on sequential minimal optimization (SMO) algorithm using production records and bottom hole pressure data [14].

Since 2015, the recurrent neural networks represented by long short-term memory (LSTM) and gated recurrent unit (GRU) become a new hot spot in the field of production prediction [15–22].

L. Zhang et al. proposed a GRU-FNN model to predict single-well production in water drive oilfield [23]. X. Song et al. adopted particle swarm optimization (PSO) to optimize the basic structure of LSTM model and achieved better model performance [24]. Weiss applied fuzzy ranking and neural network to establish correlations to predict oil production [25]. H. Wang et al. established a production prediction model for high-water-cut oil fields considering both timing and engineering factors based on LSTM and increased the input features to 17 items, with a relative error of only 1% [26]. Sagheer et al. compared the application effects of simple RNN and LSTM in oil and gas production predictions, verified that LSTM had better performance than simple RNN [27].

All the predictions made in the above studies are pretty accurate, but there is still room for further improvements. First, most of the works are aimed at predicting the total production of a reservoir or an oilfield, few research focused on the single-well production prediction. In fact, accurately predicting the production of single well is not easy because the fluctuation of single-well production is more irregular and more drastic than that of the block/reservoir production. Also, there are too many factors affecting single-well production, and we can hardly analyze them clearly. Second, in most studies, the production series is divided into 3 segments based on time, the earlier one for training, the middle one for validating, and the latest one for testing. However, as we all know, the production changes of oilfield/oil well have significant stages, and the early stage, middle stage, and late stage have different dominant control factors or rules of change. Learning the rules in early stage and predicting the production changes in late stage are bound to cause errors. Finally, there are few studies that consider both dynamic and static characteristics in time series prediction. Based

on these considerations, to achieve accurate single-well production prediction in water flooding oilfield, the data set was built considering over 60 geological and development factors that affect oil well production. The influence of the water injection on oil production is also quantified and added to the data set as a feature. The missing data are filled in accordance with the industry knowledge and the distribution pattern of data. Secondly, the feature correlation analyses tailored to time series prediction are carried out, through which 11 dominant control factors of oil well production are selected to build the samples. The production prediction model is established based on a one-dimensional convolution neural network and a long short-term memory neural network, and the model hyperparameters are optimized by Bayesian optimization. Finally, an example numerical test is carried out in a practical reservoir. Results show that the CNN-LSTM model has better performance compared with the Bi-LSTM model, Attention-LSTM model, or any other models. Production forecasting by the developed model is simple, efficient, and accurate, which can provide a guidance for the dynamic analysis of a water flooding reservoir and work as a good reference of the development and production of other types of reservoirs.

2. Methodology

2.1. LSTM and Conv1D.
LSTM (long short-term memory) can be the most successful recurrent neural network (RNN) nowadays. Its unique design of gate structure and cell state makes it possible to capture long-term dependencies, which is also the key to its great success. As shown in Figure 1, different from the other simple RNN units, there are three gates in the LSTM cell unit: forgetting gate, input gate, and output gate. In each time step, LSTM cells will partially pass the information to the next step and retain some information in the cell as the cell state, which is the key for LSTM to achieve long-term memory. LSTM neural network performs well in the field of time series prediction [28].

Conv1D (1-dimensional convolutional neural network) is famous in the areas of image recognition, but in recent years, people find that 1-dimensional CNN (Conv1D) can well accomplish the task of time series prediction and even outperforms LSTM and other time series prediction models
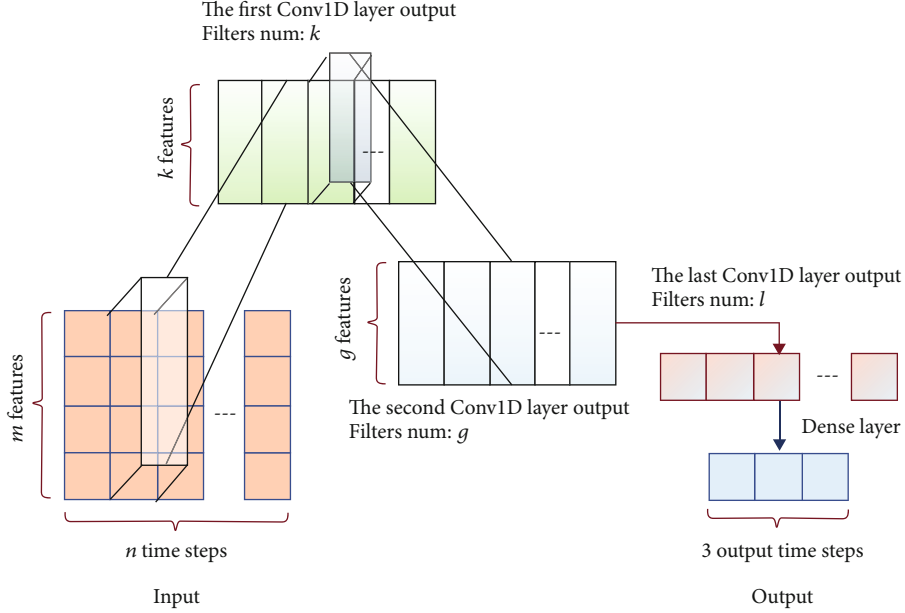
FIGURE 2: Diagram of Conv1Ds. For the inputs sized $m \times n$ and a Conv1D layer with k filters, the outputs' size is $k \times n$. Using multiple Conv1Ds, the final output size is $1 \times n$. The model illustrated in this figure wants to predict production for the next three months, that is, it wants three outputs, which we achieve using a simple dense layer containing three neurons.

TABLE 1: Initial data set. The italicized items are category features which need to be one-hot encoding.

Well basic data						
Name	X_coordinate	Y_coordinate	Depth		Reservoir data of well	
Unit	*Layer*	Start date	Flooding date	Position	Top depth (MD)	Bottom depth (MD)
Inject/production	*Well type*	Initial temperature	Initial pressure	Top depth (TVD)	Bottom depth (TVD)	Thickness of sand bodies
Initial pressure coefficient	Saturation pressure	Initial solution gas-oil ratio	Viscosity	Effective thickness	Porosity	Permeability
Monthly producing well production data						
Time	*Producing scheme*	Production days	Nozzle	Pump diameter	Pump depth	Stoke
Jig frequency	Monthly liquid	Monthly oil	Monthly water	Monthly gas	Gas-oil ratio	Water cut
Cumulative oil	Cumulative water	Cumulative liquid	Recovery	Static pressure	FBHP (bottom hole flowing pressure)	Working fluid level
Monthly water-injecting well injection data						
Injection methods	Injecting days	Pumping pressure	Casing pressure	Oil pressure	Injecting volume monthly	Cumulative injecting volume

in certain cases. When Conv1D is used for time series prediction, the inputs, n temporal steps with m features, can be viewed as an image of size of $n \times m$, and Conv1D can extract a new image from the input data by scanning and calculating in one direction. Changing the number of filters and the size of convolutional kernels, we can easily control the size of outputs (see Figure 2). It should be noted that the convolution kernel of Conv1D is also 2-dimensional, but it can only slide windows in one direction, and that is the difference between Conv1D and 2-dimensional convolutional neural network.

2.2. Spearman Correlation Coefficient. Spearman correlation coefficient, also known as the rank correlation coefficient, can measure the degree of nonlinear correlation between two features and is a method to analyze the correlation between two variables. Spearman correlation coefficient is calculated as in

$$\rho = \frac{\sum_{i=1}^{n}\left(x_j - \bar{x}\right)\left(y_j - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_j - \bar{x}\right)^2 \sum_{i=1}^{n}\left(y_j - \bar{y}\right)^2}}. \tag{1}$$
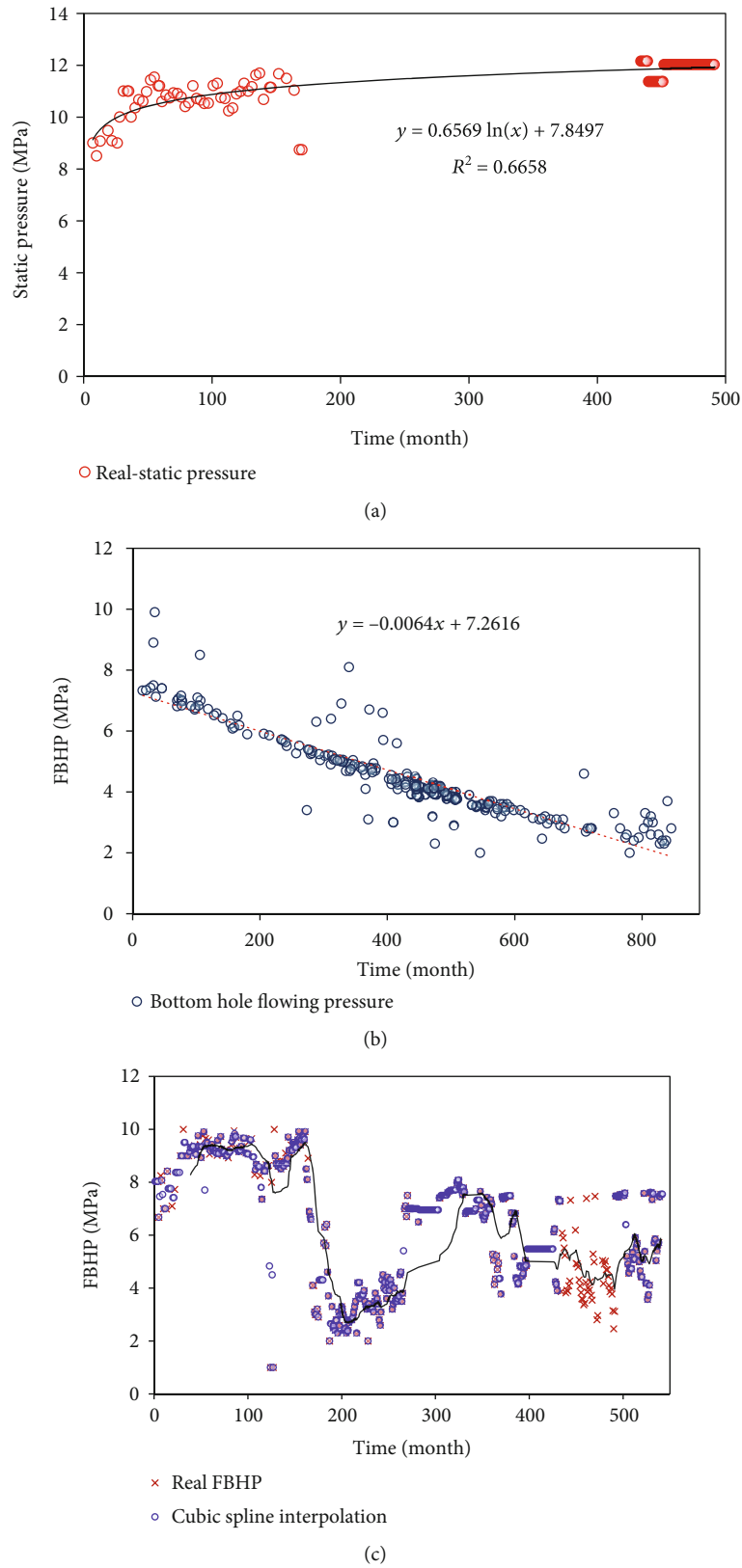
(a)

(b)

(c)

FIGURE 3: Static pressure, working fluid level and FBHP in well A. (a) The relationship between time and static pressure. (b) Comparison between real values and predicted values of FBHP. (c) The relationship between working fluid level and FBHP.

$\rho \in [-1, 1]$. When the ρ is negative, it means that the two features have a negative correlation, one feature increases means the other decreases. Similarly, if ρ is positive, the two features are positively associated. The larger the absolute value is, the stronger the correlation between two features is.

2.3. Bayesian Optimization. Acclaimed as one of the best hyperparameter tuning algorithm at present stage, Bayesian optimization is a general gradient-free global optimization strategy, which can identify a good set of hyperparameters with few iterations. It is suitable in two cases: (1) the objective function is extremely complex and time-consuming to evaluate; (2) the target function is difficult to differentiate in respect to the independent variable.

In each iteration, Bayesian optimization decomposes the optimization problem into multiple small optimization problems. It samples the original function curve by a certain method and builds an alternative curve by fitting these points. Bayesian optimization builds a model to describe the parameter-distribution of the objective function using the Gaussian process model and then solves the minimum value of the alternative model and uses this minimum value as the optimal solution of the original function in this iteration. This process is called surrogate optimization. By gradually increasing the number of sampling points, the model will gradually approach the original objective function, and an optimal combination of hyperparameters will be obtained.

In this paper, Bayesian optimization is used to optimize network model hyperparameters, including network depth and width, initial learning rate, and network activation functions. For all hyperparameters to be optimized, a maximum value and a minimum value are first specified for them, respectively. Based on the Bayesian optimization method in the package *Keras_tuner*, appropriate maximum iteration and optimization objective are set and then executed to obtain the optimal combination of hyperparameters.

3. Data Processing and Feature Engineering

3.1. Sources of Data. Data used in this paper were collected from 426 oil wells and 94 water wells in a Chinese reservoir named A with an average production time of 406 months. The oldest wells in the data set can trace back to 1956. During more than 60 years of development, the reservoir has experienced a variety of production methods such as pumping production, water flooding production, and fracturing production. At present, the reservoir has entered the ultrahigh water cut stage, with an average water cut of more than 95%.

The initial data include monthly production data of 426 oil wells, monthly injection data of 94 water-injecting wells, and single-well reservoir information. The specific items are shown in Table 1.

3.2. Data Preprocessing

3.2.1. Filling. In this paper, most of the missing data are bottom hole flowing pressure, reservoir static pressure, or dynamic liquid level height. According to their own charac-
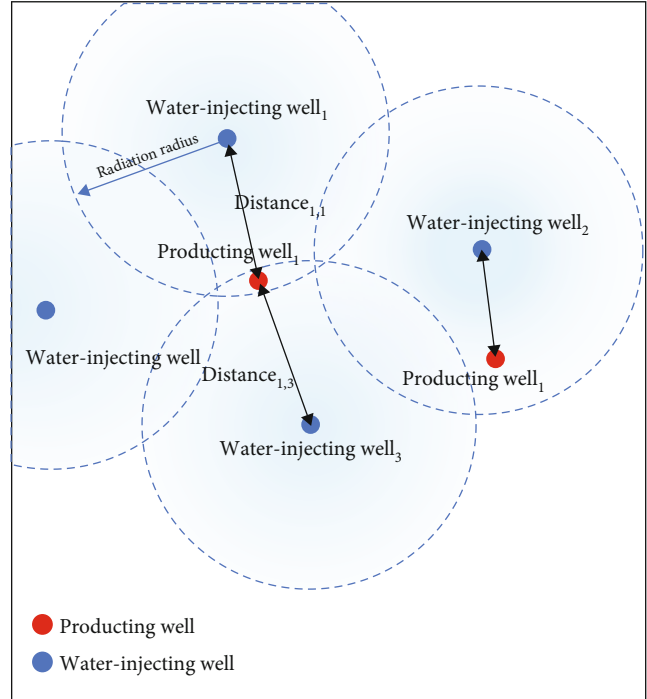

FIGURE 4: Diagram of injection-production relationship. Producing wells always be influenced by several water-injecting wells.

teristics, we use different filling methods, taking well A as an example to illustrate.

(1) Reservoir static pressure

According to the oil field development experience, the change of static pressure during the oilfield development process will not be significant and may have an obvious trend; therefore, the regression fitting method can be useful. Through the regression analysis using existing data, the relationship between the static pressure and time can be obtained, and the missing static pressure data will be calculated, as shown in Figure 3(a).

(2) Bottom hole flowing pressure (FBHP)

Figure 3(b) shows that changes of bottom hole flowing pressure during a well's production life has stages, and the range of FBHP in different stages varies greatly, which limits the application of fitting regression method. In this case, we use the cubic spline interpolation, and the filling results are shown in Figure 3(b), which is basically consistent with the actual value distribution.

(3) Working fluid level

As the working fluid level and FBHP were highly correlated (see Figure 3(c)), we can easily obtain a linear relationship between the two features, and one can be calculated from another. If either the working fluid level height or bottom hole flowing pressure is available, cubic spline interpolation is used.
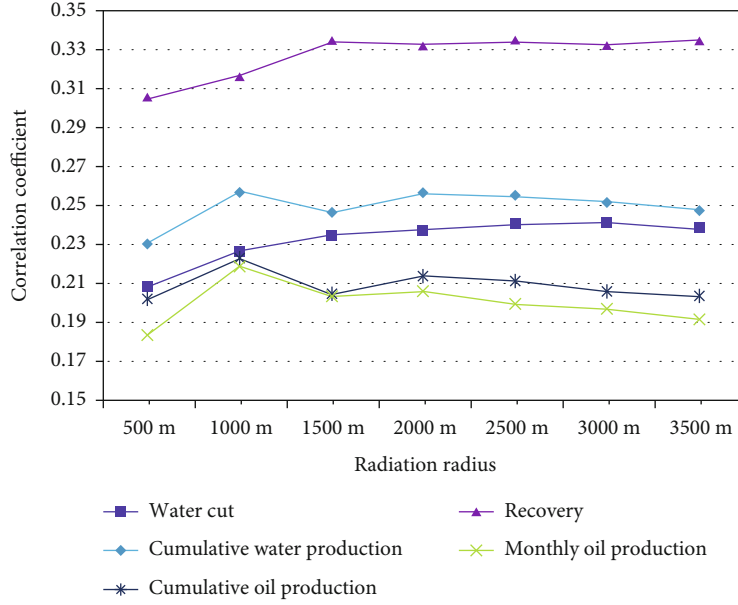
FIGURE 5: The correlation between producing well injected volume and production features. Seven correlation coefficients are used to compare; they are ρ (radiation radius, water cut), ρ (radiation radius, cumulative oil production), ρ (radiation radius, cumulative water production), ρ (radiation radius, monthly oil production), and ρ (radiation radius, recovery).

3.2.2. Encoding. Since a large part of static features in the data set are category features, the one-hot encoding process for category features should be carried out to concordance numerical features and category features. The features that are treated with this operation include units, layers, well types, and producing schemes.

3.2.3. Measurement of Injection-Production Relationship. A major factor that affects the production of oil well in water-driven reservoir is waterflooding measure. In this paper, we propose the parameter injection$_i^t$, producing well i injected volume in month t, and add it into the data set as a feature "inject."

As shown in Figure 4, an oil/producing well is often controlled by n water/water-injecting wells. We assume that a water-injecting well has a certain radiation radius, and in the radiation radius around a water well, the oil well within the radius is affected by that water-injecting well, and the smaller the distance is, the greater the impact of inject well on producing well is, and the larger producing well injected volume is. The total producing well injected volume equals the superposition of all "affected water-injecting wells."

According to this hypothesis, we can quantify the producing well injected volume, as shown in

$$\text{distance}_{i,j} = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2}, \quad (2)$$

$$\text{injection}_i^t = \sum_{j=1}^{m} \frac{\alpha_j^t}{\text{distance}_{i,j}}. \quad (3)$$

Here, i denotes the producing well i, j denotes the water-injecting well j, m denotes the number of water-injecting
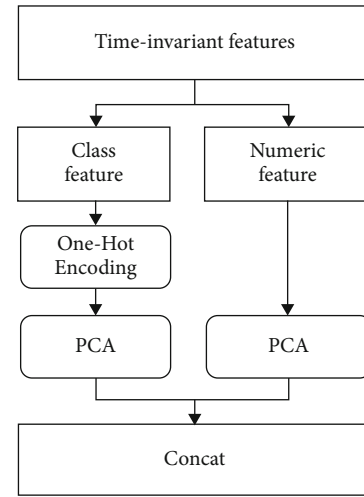

FIGURE 6: The data processing working flow. Static features and dynamic features need to be treated separately, also the discrete features and continuous features do.

wells who influences producing well i. And the distance between producing well i and water-injecting well j is distance$_{i,j}$. In tth month, α_j^t units of water were injected to reservoir by water-injecting well j.

The radiation radius is determined by correlation analysis: the closer the radiation radius to the actual radiation radius, the higher the correlation between producing well injected volume and the well production. By establishing the relationship between radiation radius and correlation coefficient, we can find the optimal radiation radius. As shown in Figure 5, 1000 m is the optimal radiation radius. Under this radius, producing well injected volume in a
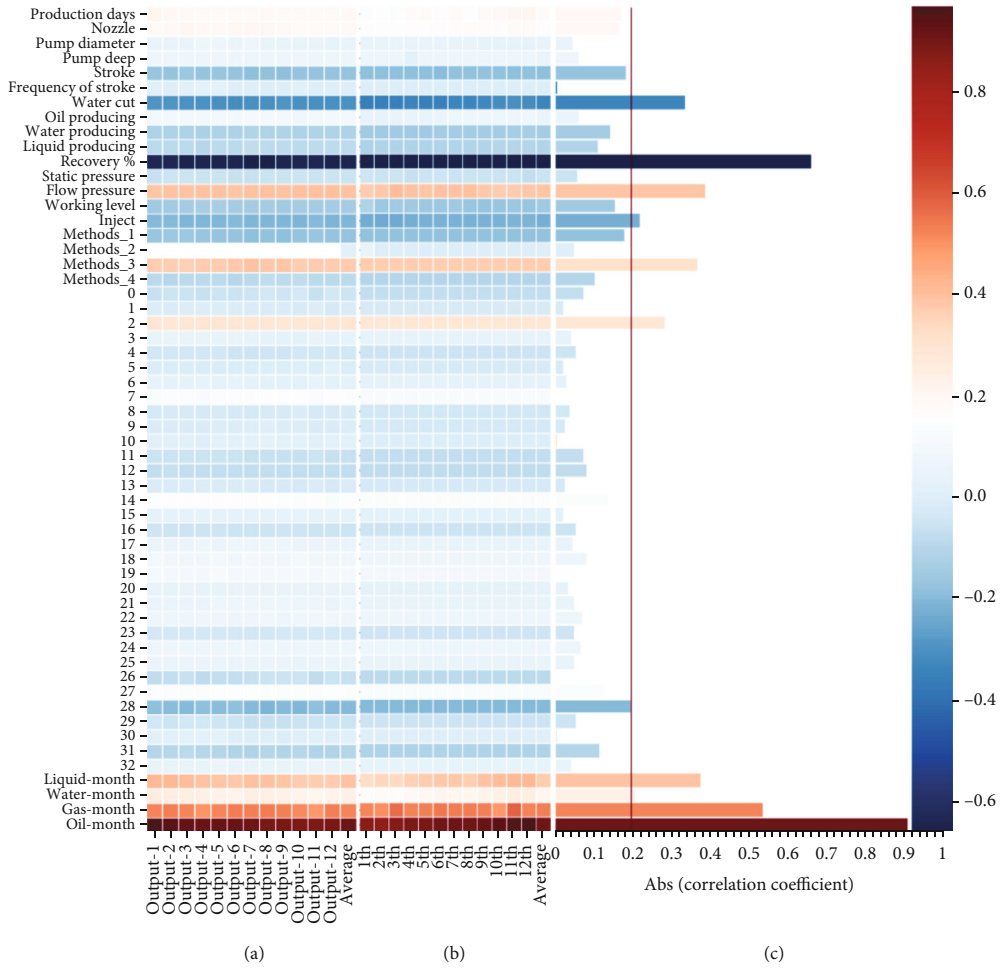
FIGURE 7: Heatmap of correlation coefficient between inputs and outputs. (a) The average of 12 correlation coefficients between the production in the next nth month and the features in the last 12 months. (b) The correlation coefficient between the production in the next 1th month and the features in the last mth months. (c) The average of 12 correlation coefficients between the production in the next 1th month and the features in the last 12 months.

certain time and certain producing well is calculated and added to the data set as the feature "inject."

3.3. Feature Engineering

3.3.1. Feature Dimension Reduction. For better prediction accuracy, feature compression is needed to improve the quality of features and limit the number of features. In the process of feature compression, static features and dynamic features need to be treated separately, so are the discrete features and continuous features. Otherwise, the static features representing the characteristics of individual wells will submerge in the continuously changing dynamic features. Similarly, the highly sparse 0-1 features created by one-hot encoding will also submerge in the continuously changing numerical features. In the data processing working flow of this paper (Figure 6), we first separate the dynamic features (features that change over time), as shown in Table 1. There are in total 11 dynamic features without dimension reduction. Secondly, we separate the 0-1 features from the static features for principal component analysis (PCA) dimensionality reduction (with a confidence of 95%); then,
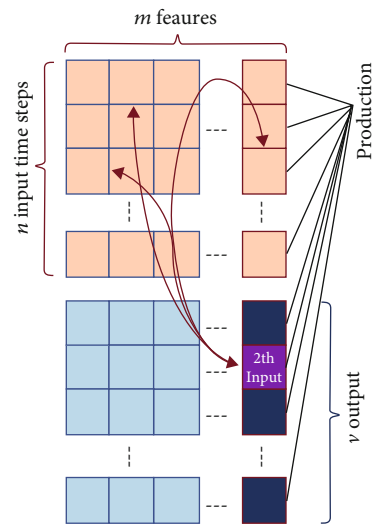

FIGURE 8: Correlation analysis method in this paper. The correlation is between inputs' features and outputs, not simply between the different features.
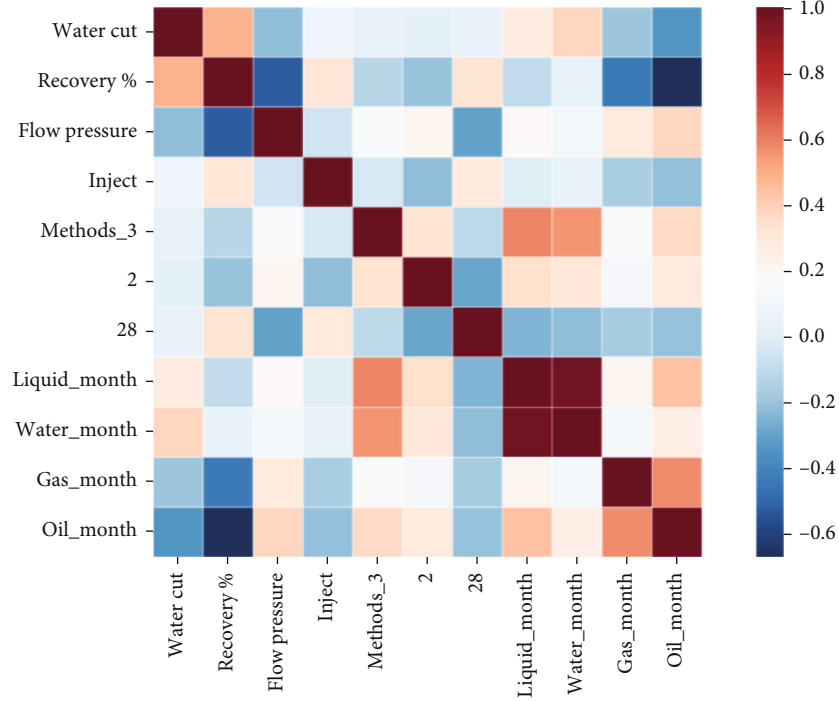
Figure 9: Heatmap of correlation coefficients between 11 features.

Table 2: Model evaluation metrics.

MAE	$MAE = \dfrac{1}{n}\sum_{i=1}^{N}\lvert y_i - \widehat{y_i}\rvert$
MAPE	$MAPE = \left(\dfrac{1}{n}\sum_{i=1}^{N}\dfrac{\lvert y_i - \widehat{y_i}\rvert}{y_i}\right)\times 100$
RMSE	$RMSE = \sqrt{\dfrac{\sum_{i=1}^{N}\lvert y_i - \widehat{y_i}\rvert}{N}}$
R^2	$R^2 = \dfrac{\sum_{i=1}^{N}(y_i - \bar{y})^2 - \sum_{i=1}^{N}(y_i - \widehat{y_i})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$
Accuracy	$accurate = 1 - \left(\dfrac{\sum_{i=1}^{N}((\lvert y_i - \widehat{y_i}\rvert)/y_i)}{N}\right)$

y_i: the measured value at time i; $\widehat{y_i}$: the predicted value at time i; $\bar{y}$: the mean of y_i; N: sample size.

the remaining numerical features are compressed by another PCA model (with a confidence of 95%), and two sets of compressed features are spliced to form the features numbered no. 0-no. 32 (see Figure 7) finally.

3.3.2. Feature Selection. In previous studies, the analyses of correlation with time series are generally carried out in the same time step in terms of production prediction; this correlation is the relationship between the current month's output and the current month's flowing pressure (or other features). This correlation can reflect the relationship between targets and variables and between variables and variables. However, there are still some loose places: the data used in prediction is not the data of the current month, but the data of the pre-

vious n months, which means that if we want to get a more accurate correlation coefficient, the calculation should also be carried out between the target month's output and the input data used to predict it (see Figure 8).

The study takes the input length of 12 and the output length of 12 (using the previous 12 months' data to predict the next 12 months' production) as an example to illustrate the problem.

Figure 7 shows the visualization results of Spearman correlation coefficient between the inputs and outputs. Figure 7(a) shows the average of 12 correlation coefficients between the production in the next nth month and the feature^k in the last m months:

$$\rho\left(output_n, feature^k\right) = \frac{\sum_{m=1}^{12}\rho\left(output_n, fature_m^k\right)}{12}. \quad (4)$$

Here, $\rho(output_n, feature^k)$ denotes the Spearman correlation coefficient between $output_n$ and $feature^k$; $output_n$ denotes the production in the nth month; $feature^k$ denotes the kth feature; $fature_m^k$ denotes the $feature^k$ in the last m months.

Figure 7(b) shows the correlation coefficient between the production in the next 1th month and the features in the last mth months. When a certain feature is fixed, from left to right, the color of the grids in Figure 7(a) gradually becomes lighter, and the color of the grids in Figure 7(b) gradually becomes darker, which means that the larger the time interval between the inputs and outputs, the lower the correlation between them, and the greater the difficulty of prediction.

TABLE 3: Model structure (CNN-LSTM).

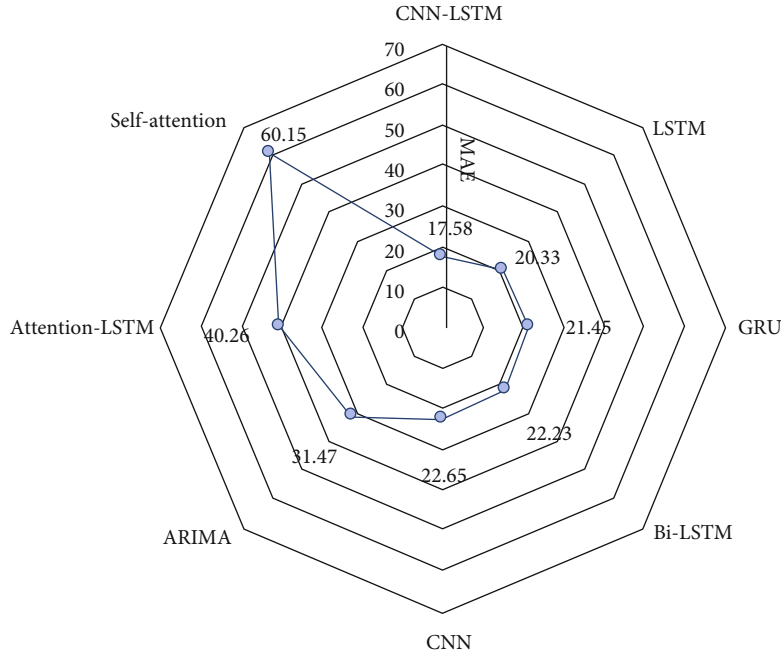| Layer (type)	Hyperparameters		
	Filter	Kernel size	Dilation rate
Conv1D_1	128	1	1
Conv1D_2	241	1	1
	Unit	Activation	
LSTM_1	156	Softsign	
LSTM_2	136	Softsign	
		Dropout rate	
Dropout		0.2	
	Unit	Activation	
LSTM_3	57	None	
Dense	Prediction time steps	None	

FIGURE 10: The predicting performance of different models.

Observing row by row, we can easily find that dynamic features have stronger correlation with future productions than static features, which can be easily explained: the correlation coefficient depends to a certain extent on the relative variation of the value of features in the sample. Compared with the production which remains changed over time, static features only vary from well to well, so the overall correlation is weak. We also note that the forecast target (production) has a strong autocorrelation—future production is far more correlated with past production than any other characteristic (up to 0.9), which means that the forecast value will be largely determined by the past production data.

To reduce the degree of difficulty in model fitting, accelerate the convergence, and eliminate invalid features, features are selected, as shown in Figure 7(c): with ±0.2 as the threshold, the features of correlation (accumulative correlation score over all outputs) ≥ 0.2 or ≤−0.2 will be retained, and 11 features are eventually entered into the data set. The correlation between them is shown in Figure 9. We can find that the correlation between monthly water production and monthly liquid production is much greater than that between monthly oil production and monthly liquid production, which confirms that the block has entered the ultrahigh water cut stage, and most of the produced liquid is water.

3.4. Sample Generation. Before generating the input samples, we need to standardize the data, so that different features have the same scale, and they will have a fair chance to be learned by the model. As mentioned in the introduction, in order to enable the model to learn the rule of production variation at each production stage, we divided the data set by wells: 341 wells constitute the training set for model training, 43 wells constitute the validation set for hyperparameter optimization, and 42 wells constitute the test set for model

TABLE 4: Performance of different models (84-3).

Model	MAE	MAPE	R^2	RMSE	Accurate	Training time/min
CNN-LSTM	17.58	6.54	0.99	37.26	0.93	6.33
LSTM	20.33	7.02	0.99	44.36	0.9	5.18
GRU	21.45	7.15	0.99	47.14	0.9	3.04
Bi-LSTM	22.23	7.69	0.99	48.12	0.88	6.44
CNN	22.65	7.73	0.99	48.87	0.88	3.16
ARIMA	31.47	15.68	0.98	37.17	0.83	2.11
Attention-LSTM	40.26	18.26	0.97	60.68	0.82	8.26
Self-attention	60.15	25.36	0.96	80.67	0.66	5.01

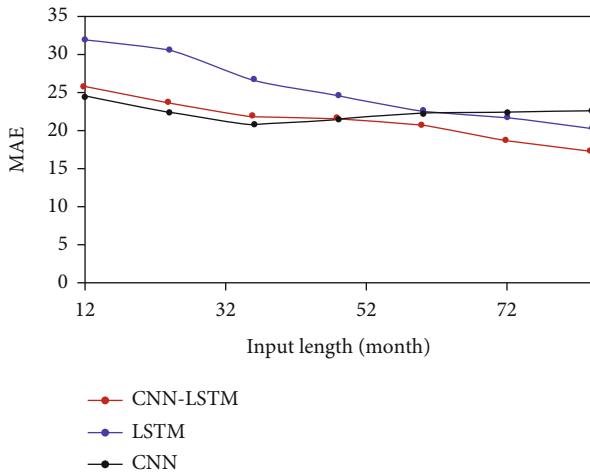84-3: using the data from the previous 84 months to predict oil production of the next three months.

FIGURE 11: Different input length vs. model prediction error.

testing. The samples were manufactured by sliding time window and rearrange order to achieve better training effect.

4. Model Design and Evaluation

4.1. Evaluation Metrics. In this paper, in order to evaluate the prediction performance of the model comprehensively, five evaluation metrics are used to evaluate the model, as shown in Table 2. It is worth noting that, to reduce the uncertainty caused by the potential randomness of the model, the model evaluation results in this paper are all from the average of 3 experiments under the same setting.

4.2. Model Structure Design. After extensive investigations and comparative experiments, LSTM and CNN are selected to build a hybrid model. CNN is stacked in the first place, its excellent ability of feature extraction enables the model to extract as much hidden knowledge as possible. After the CNN layers, we stacked the LSTM layers in the hope that the model could learn the changes of timing sequence better. In addition, the layer normalization is used between the LSTM layers, which prevents possible gradient extinction and gradient explosion.

In this paper, Bayesian optimization is used to optimize the hyperparameters and the structure of the model. Since the limited space, only the structure and hyperparameter optimization results of the optimal model (CNN-LSTM) are presented here, as shown in Table 3.

4.3. Comparative Model. There were 8 models built for the optimal model selection, as shown below:

CNN-LSTM: Conv1D for feature extraction. LSTM for timing sequence capture [29, 30].

LSTM: classical LSTM model.

Bi-LSTM: the normal LSTM model can only learn the information from front to back, but cannot catch the information from back to front. Bidirectional LSTM is an improvement in this aspect. It combines forward LSTM and backward LSTM to capture forward and backward information at the same time [31, 32].

GRU: GRU is one of the varieties of LSTM and maybe the most successful one. It simplifies the three gate structures of LSTM cells into two and can achieve almost the same prediction accuracy as LSTM while greatly accelerating convergence [23].

CNN: one-dimensional convolutional neural network model for time series prediction [33].

Attention-LSTM: the LSTM model supplemented with an attention mechanism in hope that the addition of attention mechanism can help the model better capture the critical time steps which may contain the key information about production changes. This paper adopts Luong Attention mechanism and "General" score function. To focus attention on multiple time steps instead of one, we modify the weight activation function to "sigmoid" [34–36].

Self-Attention: do not use RNNs or CNNs, and the multihead self-attention mechanism is used to realize the prediction of time series. The structure of the model refers to BERT [37], but word embedding and location embedding are dropped, and the activation function adopts "sigmoid" [38].

ARIMA: one of the most common time series prediction models.

It is worth noting that the hyperparameters of all the above models are also the most combined ones obtained by Bayesian optimization.

4.4. The Training Set. The input of the neural network is selected features as well as production records in a certain

TABLE 5: Comparison of prediction accuracy of different models under different feature selection strategies (84-3).

Models (84-3)		MAE			MAPE/%			Accurate/%	R^2
		MAE	Max	Min	MAPE	Max	Min		
A	CNN-LSTM	17.60	123.1	0.24	6.55	34.2	0.04	93.30	0.99
	CNN	22.66	125.31	0.21	7.19	38.1	0.03	88.37	0.99
	LSTM	23.55	120.13	0.19	7.32	35.1	0.03	86.11	0.99
	GRU	22.45	121.45	0.23	7.31	37.21	0.04	85.34	0.99
B	CNN-LSTM	17.58	120.03	0.23	6.54	35.4	0.03	93.23	0.99
	CNN	22.65	125.47	0.22	7.73	37.9	0.03	88.36	0.99
	LSTM	20.33	120.15	0.2	7.02	34.1	0.03	90.01	0.99
	GRU	21.45	121.63	0.24	7.15	34.2	0.04	90.24	0.99

A: use all features; B: use selected features.

duration, and the model is trained in a "supervised" fashion with future well production being the outputs. In this case, the inputs at each step of the model include historic production/features data for 82 consecutive months; the output is the production in the next 3 months.

In the training process, the model based on the encoder-decoder uses the "Teacher Forcing" hybrid training strategy [35], in which 60% of the input data of each time step in the decoder stage is real data, and the other 40% is the predicted value of the previous model output. This hybrid strategy can prevent overfitting caused by rapid convergence while ensuring the prediction accuracy of the model. Through the experiment, "*Adagrad*" was chosen as the model optimizer. The size of batch and the initial learning rate were set at 36 and 0.05, respectively. Callback function *ReduceLROnPlateau* was used to adapt the learning rate. The maximum epoch is 150; training process is controlled by the callback function *EarlyStop*. And model loss is MAE of predicted production.

5. Results and Discussion

5.1. Discussion of Different Models. Among the 8 models proposed, CNN-LSTM model achieved the best performance, and the predicting performance of other models are shown in Figure 10 and Table 4.

Results show that the prediction precision of attention mechanism model (including self-attention mechanism) is the worst. In many papers, attention mechanism improves the score of time series prediction [32, 34–36]. However, it is not the case in this paper. In fact, the production of a well in a certain month does not heavily depend on one or more certain previous months, so adding weights to different time steps is not very helpful in forecasting. Besides, the addition of attention mechanism greatly aggravates the training burden of the model, which is another possible reason for the low prediction accuracy of the model.

Bi-LSTM model also failed to reach the expected score, and its performance was slightly worse than that of classical LSTM model. Unlike the semantic recognition task such as machine translation, changing the order of oil well production series makes no significant difference in predicting production, and the use of Bi-LSTM brings more complex model structure undoubtedly, making model training more difficult.

GRU and CNN models gain similar results, with LSTM model having slightly higher accuracy than LSTM, but it should be noted that the training time of GRU and CNN was almost half of LSTM. GRU or CNN models may be more appropriate in some less-desirable cases.

5.2. Discussion of Model Input Length. In this paper, CNN, LSTM, and CNN-LSTM model were used to conduct contrast experiments, respectively, and to verify the influence of different input length on model prediction error, as shown in Figure 11.

It is easy to see that the performance of CNN model is better than that of LSTM and CNN-LSTM when the input length is short. However, with the increase of the input length, the prediction error of CNN model also increases significantly, while that of LSTM and CNN-LSTM is decreasing. Compared with CNN and LSTM, CNN-LSTM model seems to inherit advantages from both, better than LSTM in short case and better than the other two in long case. Considering data characteristics, case requirements, and model performance, the following example uses 84 months' data to predict monthly well production over the next three months.

5.3. Discussion of Feature Selection. To verify the effectiveness of feature selection, two feature selection plan were used for feature selection experiment. Plan A: input all features; plan B: input select features. The prediction accuracy of each model using different features is shown in Table 5.

It is found that features and algorithms have cross influence on the prediction accuracy of the model. The RNN model has strong sensitivity to the number of features, and the selection of the number of features can reduce its prediction error. The model with convolutional layer is insensitive to the changes in the number of features. It can be seen from Table 5 that there is not much difference in model prediction accuracy before and after feature selection. The analysis shows that the RNN (LSTM/GRU) model is more inclined to capture the connection of samples in time series, while the convolution structure model is more inclined to analyze and extract high-dimensional features. Therefore, prior feature selection for the LSTM model means that part of the
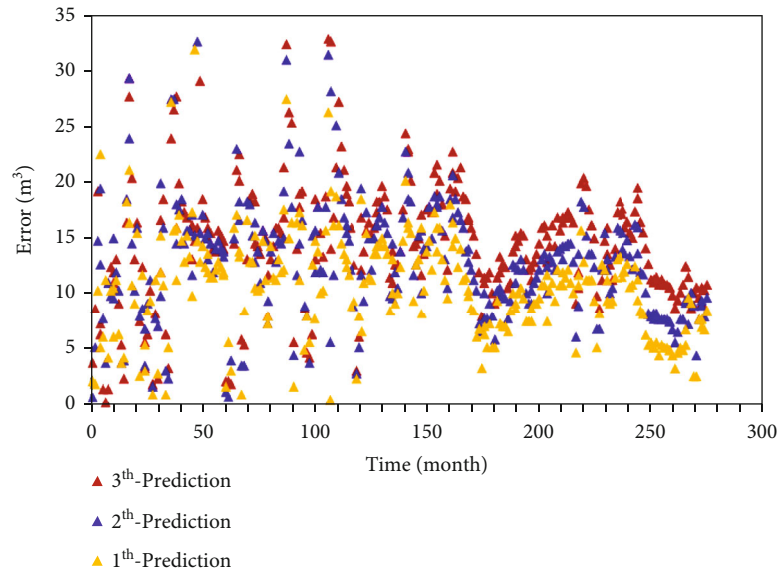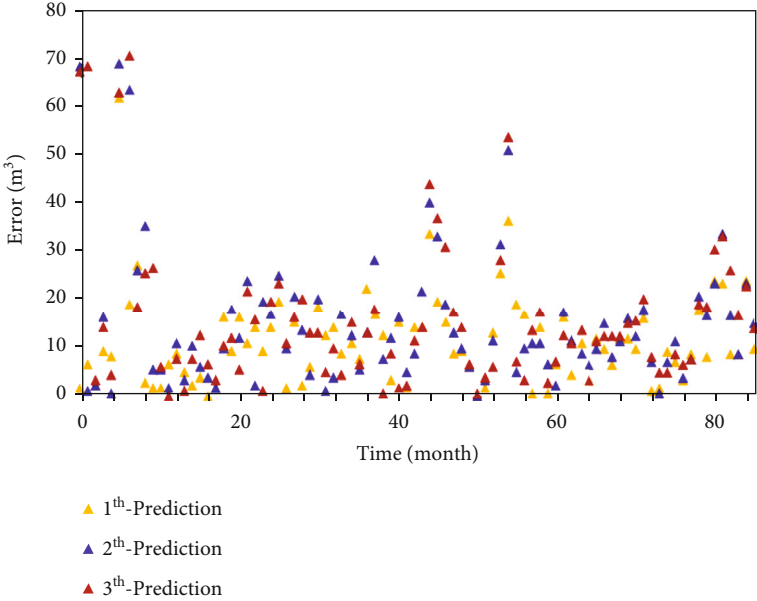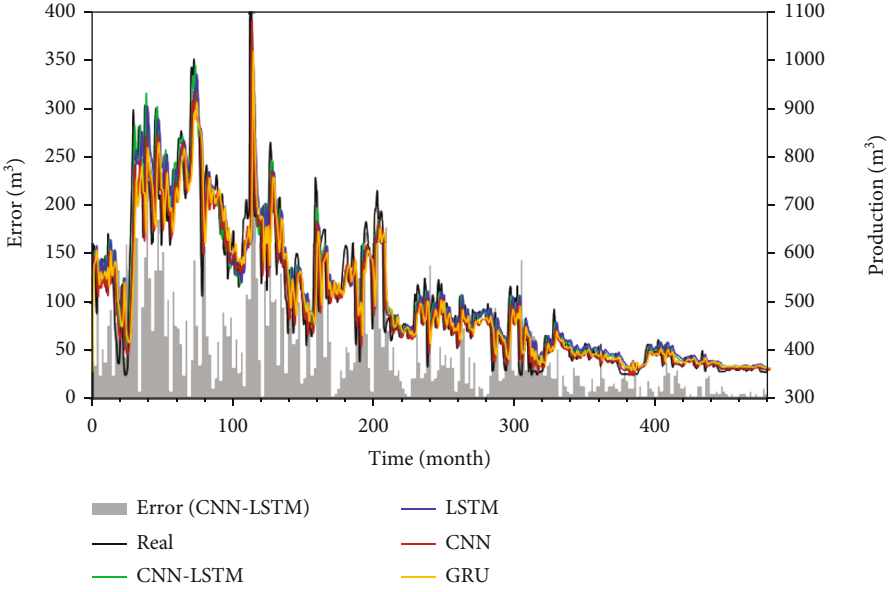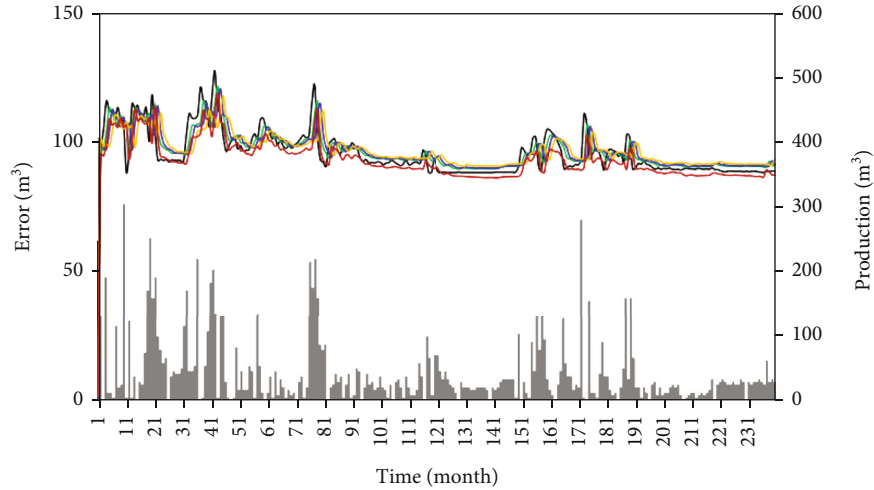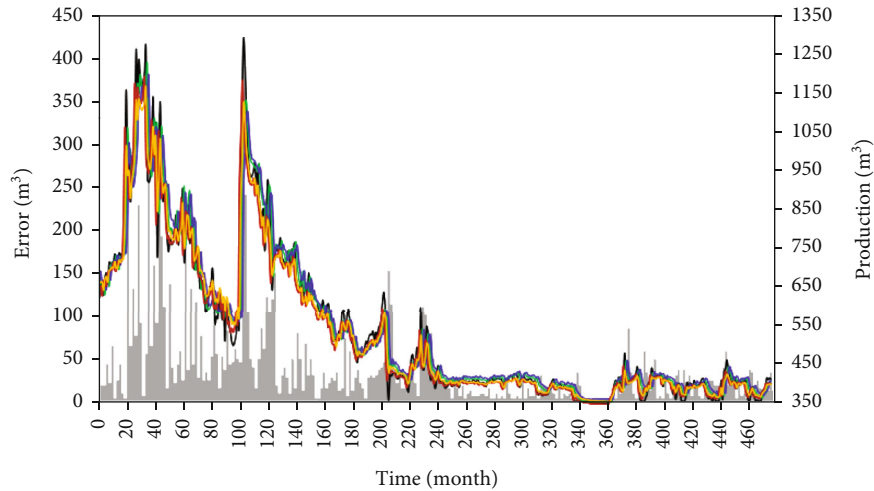
(a)
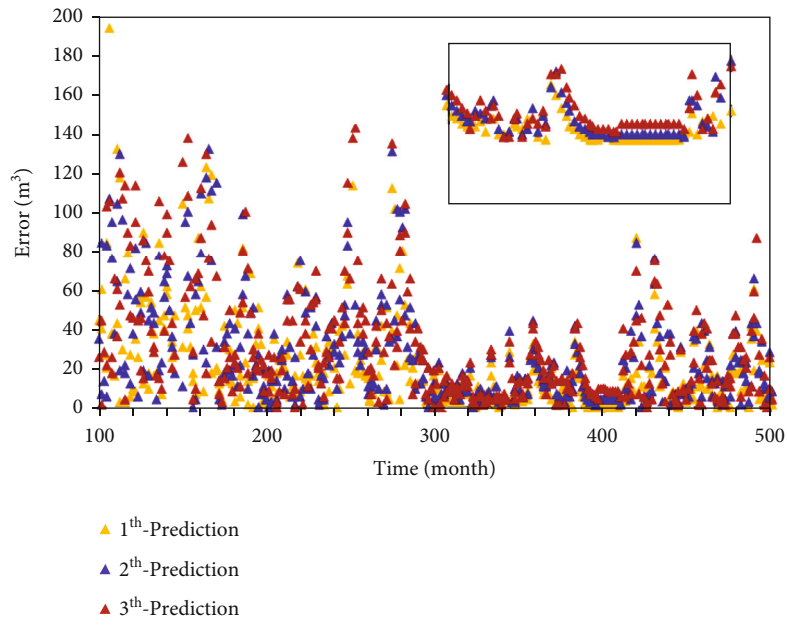
(b)

Figure 12: Continued.

(c)

(d)

FIGURE 12: Continued.

(e)

(f)

Figure 12: The prediction results of 6 wells in reservoir A. (a–c) The error distribution of the 3 predicted production of well A\B\C by CNN-LSTM model. "1th-prediction" denotes the next month's predicted production. Similarly, "2th-prediction" denotes the predicted production of the month after next. And "3th-prediction" means the predicted production of the third month in the future. (e–f) The comparison of real production curves and the 1th predicted production curves of well E/F/G by CNN-LSTM model, LSTM model, GRU model, and CNN model. The absolute error between real production and production predicted by CNN-LSTM is also shown.

feature extraction work has been completed before model training, which reduces the difficulty of LSTM training, so the LSTM model has higher prediction accuracy for data sets with fewer features. CNN itself can well realize feature extraction, and feature selection only accelerates its training speed, but has little impact on the final prediction accuracy.

5.4. Case Verification. The proposed models are used to predict 43 wells' production in the test set, and the results of 6 wells were selected randomly. The results are shown in Figure 12.

Figures 12(a)–12(c) compare the prediction error of the three outputs of the model with wells B, C, and D as examples. Obviously, the 1th outputs have the minimum error consistent with the correlation analysis.

Figures 12(d)–12(f) compared different models' prediction results of wells E, F, and G. Similar to the results shown in Table 4, CNN-LSTM model's prediction is closer to the

real tendency, but other models also can fit the curve closely. Another obvious phenomenon is that the predictions at earlier time stages are worse compared to latter ones; in other words, the error at high production is greater than the error at low production. This is because production changes are more dramatic in the early stages of the well development.

Extensive experiments show that proposed scheme significantly improves production prediction accuracy and enhances predict efficiency.

6. Conclusion

This paper proposes a machine learning model for predicting single-well production in water flooding reservoir. The specific conclusions are as follows:

(1) More than 60 factors of geology and development that affect the changes of oil well production was comprehensively considered to build the data set. Data filling and feature extraction were carried out, respectively, according to the characteristics of data. Features were analyzed and selected from the perspective of time sequence. The data set is divided by wells to make the sample distribution more practical

(2) Eight models with good performance in the time series prediction are constructed. By comparison, CNN-LSTM model gains the best score, while the improvement of attention mechanism and Bi-LSTM model is limited. It also illustrates that complex models that are doing well in other tasks may not be suited to the well production prediction

(3) The Bayesian optimization is used to optimize the hyperparameter of the models, which can greatly improve the efficiency of hyperparameter optimization and improve the prediction accuracy of the models

(4) An experiment case is carried out in reservoir A, and the results prove that the model proposed in this paper can accomplish the prediction task of single-well production successfully and provides a good reference and guidance for development and production in water flooding reservoir

Data Availability

The manuscript is a self-contained data article; the entire data used to support the findings of this study are included within the article. If any additional information is required, this is available from the corresponding author upon request to zhanglesley@foxmail.com.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] J. Huang and J. Zhang, "Overview of oil and gas production forecasting by machine learning," *Petroleum Reservoir Evaluation and Development*, vol. 11, pp. 613–620, 2021.

[2] K. Wang, B. Jiang, H. Li et al., "Rapid and accurate evaluation of reserves in different types of shale-gas wells: production-decline analysis," *International Journal of Coal Geology*, vol. 218, p. 103359, 2020.

[3] G. Luo, L. Xiao, and Y. ShiI, "Machine learning for reservoir fluid identification with logs," *Petroleum Science Bulletin*, vol. 1, pp. 24–33, 2022.

[4] G. Liao, Y. Li, and L. Xiao, "Prediction of microscopic pore structure of tight reservoirs using convolutional neural network model," *Petroleum Science Bulletin*, vol. 1, pp. 26–38, 2020.

[5] R. R. A. Mendes, A. P. Paiva, R. S. Peruchi, P. P. Balestrassi, R. C. Leme, and M. B. Silva, "Multiobjective portfolio optimization of ARMA-GARCH time series based on experimental designs," *Computers and Operations Research*, vol. 66, pp. 434–444, 2016.

[6] Y. Duana, H. Wang, and M. Wei, "Application of ARIMA-RTS optimal smoothing algorithm in gas well production prediction," *Petroleum*, vol. 8, no. 2, pp. 270–277, 2021.

[7] J. Gu, G. Sui, and Z. Li, "Oil well production forecasting method based on ARIMA-Kalman filter data mining model," *Journal of Shenzhen University Science and Engineering*, vol. 35, no. 6, pp. 575–581, 2018.

[8] Z. Hou, K. Song, and G. Ma, "Neural units with higher-order synaptic operations for robotic image processing applications," *Soft Computing*, vol. 11, no. 3, pp. 221–228, 2007.

[9] X. Song, Y. Liu, and J. Ma, "Productivity forecast based on support vector machine optimized by grey wolf optimizer," *Lithologic Reservoirs*, vol. 32, no. 2, pp. 134–140, 2020.

[10] L. Gao, C. Tan, and X. Cheng, "Oilfield output forecast based on the chaotic sequence time forecasting method," *Peak Data Science*, vol. 2, p. 5, 2016.

[11] A. D. Morozov, D. O. Popko, and V. M. Duplyakov, *Data-Driven Model for Hydraulic Fracturing Design Optimization: Focus on Building Digital Database and Production Forecast*, 2019.

[12] R. Zhang and H. Jia, "Production performance forecasting method based on multivariate time series and vector autoregressive machine learning model for waterflooding reservoirs," *Petroleum Exploration and Development*, vol. 48, no. 1, pp. 201–211, 2021.

[13] V. Nguyen-Le and H. Shin, "Artificial neural network prediction models for Montney shale gas production profile based on reservoir and fracture network parameters," *Energy*, vol. 244, article 123150, 2022.

[14] Y. Zhu, Y. Xian, and Q. Li, "Shale gas productivity forecast based on big data," *Well Testing*, vol. 28, no. 1, pp. 1–6, 2019.

[15] G. Hui, S. Chen, Y. He, H. Wang, and F. Gu, "Machine learning-based production forecast for shale gas in unconventional reservoirs via integration of geological and operational

factors," *Journal of Natural Gas Science and Engineering*, vol. 94, p. 104045, 2021.

[16] J. Zhang, Z. Yan, and X. Zhang, "Developing a long short-term memory (LSTM) based model for predicting water table depth in agricultural areas," *Journal of Hydrology*, vol. 561, pp. 918–929, 2018.

[17] W. Liu, W. D. Liu, and J. Gu, "Forecasting oil production using ensemble empirical model decomposition based long short-term memory neural network," *Journal of Petroleum Science and Engineering*, vol. 189, article 107013, 2020.

[18] Y. Ren, J. Gu, and W. Cui, "Oilfield production prediction model based on improved fruit fly algorithm and long-short term memory neural network," *Science Technology and Engineering*, vol. 20, no. 18, pp. 7–21, 2020.

[19] Q. Zhang, Z. Chen, and Y. Zeng, "Data-driven approaches for time series prediction of daily production in the Sulige tight gas field, China," *Artificial Intelligence in Geosciences*, vol. 2, pp. 165–170, 2021.

[20] Z. Guo, J. Zhao, and Z. You, "Prediction of coalbed methane production based on deep learning," *Energy*, vol. 230, article 1208471, 2021.

[21] N. C. Chakra, K. Y. Song, and M. M. Gupta, "An innovative neural forecast of cumulative oil production from a petroleum reservoir employing higher-order neural networks (HONNs)," *Journal of Petroleum Science & Engineering*, vol. 106, no. 6, pp. 18–33, 2013.

[22] R. Huang, C. Wei, B. Wang et al., "Well performance prediction based on long short-term memory (LSTM) neural network," *Journal of Petroleum Science and Engineering*, vol. 208, article 109686, 2022.

[23] L. Zhang, H. Dou, and H. Wang, "Neural network optimized by genetic algorithm for predicting single well production in high water cut reservoir," in *2021 3rd International Conference on Intelligent Control, Measurement and Signal Processing and Intelligent Oil Field (ICMSP)*, pp. 297–306, Xi'an China, 2021.

[24] X. Song, Y. Liu, L. Xue et al., "Time-series well performance prediction based on long short-term memory (LSTM) neural network model," *Journal of Petroleum Science and Engineering*, vol. 186, article 106682, 2020.

[25] W. W. Weiss, R. S. Balch, and B. A. Stubbs, *How Artificial Intelligence Methods Can Forecast Oil Production*, no. article 75143, 2002SPE, 2002.

[26] H. Wang, L. Mu, and F. Shi, "Production prediction at ultra-high water cut stage via recurrent neural network," *Petroleum Exploration and Development*, vol. 47, no. 5, pp. 1009–1015, 2020.

[27] A. Sagheer and M. Kotb, "Time series forecasting of petroleum production using deep LSTM recurrent networks," *Neurocomputing*, vol. 323, pp. 203–213, 2019.

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

[29] A. Agga, A. Abboua, and M. Labbadi, "CNN-LSTM: an efficient hybrid deep learning architecture for predicting short-term photovoltaic power production," *Electric Power Systems Research*, vol. 208, article 107908, 2022.

[30] A. Agga, A. Abbou, M. Labbadi, and Y. el Houm, "Short-term self consumption PV plant power production forecasts based on hybrid CNN-LSTM, ConvLSTM models," *Renewable Energy*, vol. 177, pp. 101–112, 2021.

[31] X. Huang, Q. Li, and Y. Tai, "Time series forecasting for hourly photovoltaic power using conditional generative adversarial network and Bi-LSTM," *Energy*, vol. 246, article 1234031, 2022.

[32] N. Mughees, S. AliMohsin, and A. Mughees, "Deep sequence to sequence Bi-LSTM neural networks for day-ahead peak load forecasting," *Expert Systems with Applications*, vol. 175, article 114844, 2021.

[33] Z. Yuan, H. Huang, Y. Jiang, and J. Li, "Hybrid deep neural networks for reservoir production prediction," *Journal of Petroleum Science and Engineering*, vol. 197, article 108111, 2021.

[34] Y. Liu, L. Shan, and D. Yu, "An echo state network with attention mechanism for production prediction in reservoirs," *Journal of Petroleum Science and Engineering*, vol. 209, article 109920, 2022.

[35] M. T. Luong, H. Pha, and C. D. Manning, *Effective Approaches to Attention-Based Neural Machine Translation*, Computer ence, 2015.

[36] J. Qin, Y. Zhang, and S. Fan, "Multi-task short-term reactive and active load forecasting method based on attention-LSTM model," *International Journal of Electrical Power & Energy Systems*, vol. 135, article 108111, 2021.

[37] J. Devlin, M. W. Chang, and K. Lee, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, 2018.

[38] A. Vaswani, N. Shazeer, and N. Parmar, "Attention Is All You Need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.