Energy Proceedings
Vol 51, 2025
ISSN 2004-2965

An Integrated Method of Data-Driven and Water-Drive Mechanism for Oil Production Forecast in High Water Cut Reservoir#

Qingshuang Jin[1], Yongchao Xue[1*], Xiangyu Ren[1], Aile Zheng[1], Xiaobiao Wang[1]

1 State Key Laboratory of Petroleum Resources and Prospecting, China University of Petroleum, Beijing 102249, China

(Corresponding Author: Yongchao Xue, Tel: 861089733218, E-mail: xyc75@cup.edu.cn)

ABSTRACT

The Q oilfield is a medium-to-high permeability heavy oil reservoir, currently in the high water cut development stage, with continuously declining production. Accurate production forecasting can provide guidance for adjusting production strategies in the oilfield. In recent years, the rise of machine learning models has offered a better alternative for predicting well production. This paper is based on data-driven daily production dynamic forecasting of oil wells, integrating data-driven models with water drive characteristic curves to enhance the accuracy of dynamic production forecasting for oil wells.

Firstly, using exclusion discriminant analysis, time, water cut, and daily oil production were selected as machine learning feature data, and the optimal water drive characteristic curve was determined. Secondly, the optimization of long short-term memory neural network structure is carried out, and the optimal results are used to carry out the recursive multi-step prediction of long short-term memory neural network. Furthermore, the optimal data-driven model was further integrated with the optimal water drive characteristic curve to establish an oil well daily oil production forecasting model that combines water drive characteristic curves and LSTM models. Finally, using data from two wells in the Q oilfield, the predictive performance of the pure data-driven model, the integrated model, and the numerical simulation model was compared and evaluated. The results showed that the integrated model had the best predictive performance and the lowest error.

This paper establishes a dynamic production forecasting model for oil wells that deeply integrates water drive characteristics with data-driven models, improving traditional daily production forecasting methods and achieving superior predictive results.

Keywords: oil production prediction, data driven, water-drive characteristics, integrated model

NONMENCLATURE

Abbreviations	
LSTM	Long Short Term Memory Network
DTW	Dynamic Time Warping
Symbols	
d_{DTW}	Cumulative distance of DTW
$d(a_i)$	The distance from a selected a_i to b
a,b	Time series data is needed to measure distances
m	Time series length
X_t	The input at t
h_{t-1}	The output of the LSTM unit at the previous time
C_{t-1}	The memory of the LSTM unit at the previous time
σ	The sigmoid activation function
b_f	The bias term
W_f	The input weight
f_t	The expression for forget gate
C_t	The cell state at time t
O_t	The value of the output gate
W_p	The cumulative water production
N_p	The cumulative oil production
L_p	The cumulative liquid production
A, B, A_2, B_2, A_3, B_3, A_4, B_4	The constant term

1. INTRODUCTION

Oil production forecasting is a multidisciplinary field that combines geology, engineering, and data analysis to estimate future oil output. Accurate oil production forecasting is crucial for the stable extraction and sustainable development of oilfields. It helps oilfield companies formulate effective development plans and

This is a paper for the 16th International Conference on Applied Energy (ICAE2024), Sep. 1-5, 2024, Niigata, Japan.

make informed decisions in resource management, investment, and energy planning.

Currently, traditional methods for production forecasting include decline curve analysis, material balance equations, empirical formulas, and reservoir numerical simulations. The decline curve analysis methods for oil and gas wells can be broadly categorized into two types: empirical methods and type curve methods. Arps [1]first introduced the Arps decline curve analysis method, which is still widely used today. Blasingame and others [2-3] introduced normalized rate, rate integral, and rate integral derivative functions, establishing the Blasingame decline curve analysis method. Subsequently, Agarwal, Agarwal-Gardner, Mattar, and others[4-5] improved these methods. Among the conventional methods for dynamic analysis of oil and gas wells, the Blasingame decline curve analysis is the most widely applied. Although decline curve analysis is simple and convenient to use, and data is easily obtainable, this method is only applicable during the production decline phase of a reservoir. Additionally, the method's simplicity can lead to lower accuracy in predictions, and it is sensitive to noise, making it easily affected by anomalous field data. The fundamental principle of the material balance equation is to consider the surface volume of all fluids in the reservoir as constant, meaning that at any point during development, the volume of produced fluids plus the remaining fluid volume in the reservoir equals the original fluid volume (all at surface conditions)[6-7]. As a zero-dimensional model for fluid flow in oil and gas reservoirs, the material balance equation is simple and easy to understand. It is widely used in dynamic geological reserve calculations, drive mechanism identification, energy evaluation, production capacity analysis, and dynamic forecasting[8-11], becoming one of the essential formulas for reservoir engineers. However, the material balance equation has many limitations, such as the assumption that oil, gas, and water phases reach equilibrium instantaneously. Important parameters (such as high-pressure fluid properties, gas cap index, oil-water and gas-oil interfaces) are difficult to obtain accurately[12-16], and the equation's strong nonlinearity makes it challenging to solve. Empirical formulas can quickly predict well production using historical data, but specific empirical formulas only provide good predictive results for a particular oilfield and cannot be applied to other types of oilfields, thus limiting their applicability. While reservoir numerical simulation methods can be applied to most oilfields, they involve significant computation and long history matching periods, often leading to extended project timelines.

In recent years, with the development of big data and artificial intelligence technologies, data-driven methods have provided better alternative solutions for dynamic oil well production prediction. In 2000, Tamhane D. et al. [17] proposed using soft computing techniques (including neural computing, fuzzy logic, and evolutionary computing) to improve reservoir description, addressing the inefficiency and inaccuracy of traditional methods. In 2004, Nguyen H.H. et al. [18] used single and multiple neural network models to predict future oil well production and experimentally validated that multiple neural network models outperformed single neural network models in long-term prediction accuracy. In 2018, Bhattacharya S. et al. [19] employed Bayesian network theory and random forest algorithms to predict lithofacies and fractures, achieving high-precision predictions for lithofacies and fractures in unconventional shale and conventional sandstone and carbonate reservoirs using common well log data. In 2019, Noshi Christine Ikram et al. [20] utilized gradient boosting trees (GBT), adaptive boosting (Adaboost), and support vector regression (SVR) algorithms to forecast future oil well production, addressing the complexity of production prediction and achieving higher accuracy than traditional analytical models. Niu W.T. et al. [21] proposed a machine learning model based on early data (including production and flowback rate data) to tackle the challenge of accurately predicting the ultimate recovery (EUR) of shale gas wells, achieving a high-precision prediction with a mean absolute percentage error of 13.41%, with support vector machines (SVM) being considered the most reliable model. In 2024, Mahlon Kida Marvin et al. [22] introduced a method using Echo State Networks (ESN) for reservoir waterflooding prediction and net present value (NPV) optimization, resolving the computational burden of traditional model development, achieving a prediction accuracy of up to 90.79% under low geological uncertainty, and realizing higher NPV in the optimized scenario compared to the base scenario. In the same year, Chen M.J. et al. [23] proposed a data-driven neural network method based on decline curves for predicting tight gas well production, addressing the low prediction accuracy of traditional methods in practical applications, and achieving high-precision prediction with a mean absolute percentage error of 14.11% and a root mean square error of 1.491.

However, these data-driven methods lack constraints from fundamental mechanisms and

principles, leading to predictions that do not align with the basic characteristics of reservoirs and often suffer from poor interpretability. This study combines data-driven methods with waterflood characteristic curves to predict oil well production, enhancing the model's interpretability and achieving better prediction results.

2. METHODOLOGY

2.1 Dynamic Time Warping (DTW)

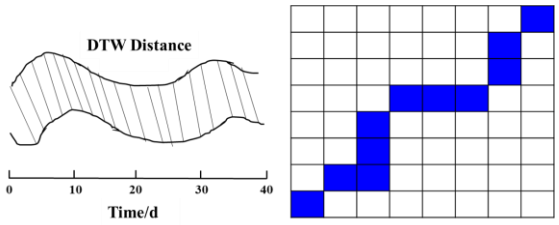


Fig. 1 DTW distance

Dynamic time warping (DTW) allows sequences to be warped in the time dimension, enabling optimal similarity matching of time series with unequal lengths through similar waveforms[24]. DTW aligns two time series, thus avoiding the issue of temporal misalignment. Given two time series, with lengths m and n(where m$\geqslant$ n), the DTW distance between the two sequences can be represented as follows:

$$d_{DTW} = \min_{a} \sum_{i=1}^{m} d(a_i) \tag{1}$$

In order to obtain the dynamic time-bending distance, a grid matrix is constructed to find an optimal regular path and finally minimize the cumulative distance.

2.2 Long Short-Term Memory (LSTM)

Traditional neural networks cannot achieve continuous memory; they can only handle the relationship between a few features and a label and cannot deal with problems related to previously input historical data. Recurrent neural networks (RNNs) essentially solve this issue. However, RNNs tend to suffer from gradient vanishing or gradient exploding problems in long-time sequence predictions. Long Short-Term Memory (LSTM) networks introduce gating mechanisms, effectively controlling the flow of information and mitigating the gradient vanishing and exploding problems, thereby demonstrating better performance in long sequence learning.

The core of LSTM model is memory cell, which cannot directly control what information needs to be remembered, so special network structures such as forget gate, input gate and output gate are needed to adjust the memory information. The forget gate is used to delete unnecessary past information, the input gate is used to store new useful information in the cell state, and the output gate determines the output information. Figure 2 shows the specific structure of the LSTM, where X_t represents the input at t, h_{t-1} represents the output of the LSTM unit at the previous time, C_{t-1} represents the memory of the LSTM unit at the previous time, and the symbol × represents element-by-element multiplication. The small square with σ represents the sigmoid activation function, and the small square with tanh represents the tanh activation function [25-27].

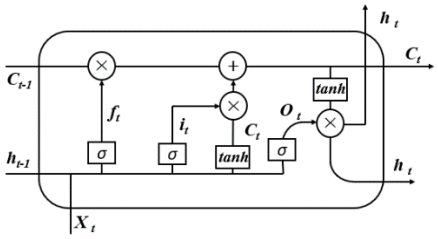


Fig. 2 Schematic diagram of long and short term memory neural network

The first step in the LSTM network is to decide what old information should be forgotten, a process known as "forget gate" and represented by ft. The output of the forget gate is a vector between 0 and 1 that determines the information that should be retained in the previous cell state, C_{t-1}. The calculation of ft is based on the input X_t and the hidden state h_{t-1} of the previous step, weighted summing using the bias term b_f and the input weight W_f. Specifically, the expression for forget gate f_t is:

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

The second step is to determine what new information the current cell state should require. First, X_t and h_{t-1} are used in the input gate it to determine information changes. Next, X_t and h_{t-1} are combined via tanh to generate new candidate cell information to update the cell's information. The bias term and input weight of the input gate it are b_i and W_i respectively, while the bias term and input weight of the candidate cell state C are b_c and W_c respectively. The specific calculation method is as follows:

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

$$\tilde{C}_t = \tanh\left(W_c \cdot \left(X_t, h_{t-1}\right) + b_c\right) \tag{4}$$

The third step is to update the current cell sta tus. The forgetting gate ft is multiplied by the prev

ious cell state C_{t-1} to determine the forgotten information. In the second step, the current candidate cell state and input gate it have been determined, and the new candidate cell information that needs to be added is multiplied to determine. According to the above calculation, the cell status update value C_t at time t can be calculated, • is the dot product, the specific calculation method is:

$$C_t = i_t \cdot \tilde{C}_t + f_t \cdot C_{t-1} \quad (5)$$

After the cell status update, the final step in the LSTM network is to calculate the value of the output gate O_t. The output gate O_t determines the final output result, which is determined by the cell state C_t and the current hidden state ht. The calculation of the output gate O_t is based on the input X_t and the hidden state h_{t-1} of the previous step, weighted summing using the offset term b_o and the input weight W_o. Specifically, the calculation formula of output gate O_t is:

$$O_t = \sigma\left(W_o \cdot \left(X_t, h_{t-1}\right) + b_o\right) \quad (6)$$

$$h_t = O_t \cdot \tanh\left(C_t\right) \quad (7)$$

2.3 Water drive characteristic curve

The empirical methods for predicting the regularity of water cut rise and recoverable reserves in oil fields include four common water drive characteristic curves, which are type A, type B, type C and type D. These characteristic curves are derived from the oilfield practice of former Soviet Union and Chinese scholars.

Table 1 Water drive characteristic curve

Water drive characteristic curve	formula
Type A	$\log W_p = A + BN_p$
Type B	$\log L_p = A_2 + B_2 N_p$
Type C	$L_p / N_p = A_3 + B_3 L_p$
Type D	$L_p / N_p = A_4 + B_4 W_p$

3. DATA PREPROCESSING AND FEATURE OPTIMIZATION

Q oilfield is located in the middle of Bohai Sea, in the middle of Shimoluo bulge high bulge area. Shimoluo protrusions are located in the northwest of Bozhong Depression, and the distribution direction is east-west. There are major reservoir boundary faults on the north and south sides, adjacent to Qinnan Depression and Bozhong Depression. The lithomolar bulge is complicated by faults. The Q field began production in August 2002. In the initial stage of operation, the production condition of the oilfield is not good, the period of anhydrous oil production is short, the recovery degree of the oilfield is very low, the water cut rises rapidly, the production declines rapidly, and the characteristics of rapid coning of bottom water have been shown after many years of directional well mining, and now the oilfield has entered the development period of high water cut.

3.1 Data preprocessing

Dynamic Time Warping (DTW) method were used to calculate the correlation degree of each variable with daily oil production, and dynamic data of 132 Wells were collected. The annual average daily oil production of 132 Wells over time was shown in Figure 3. It can be seen from Figure 3 that the daily oil production per well of this oilfield is relatively high from the first year of production to the second year of production, and tends to be stable after the fifth year of production. The initial Wells of the reservoir are few and the production is high. In the middle and late period, the output gradually remained stable.

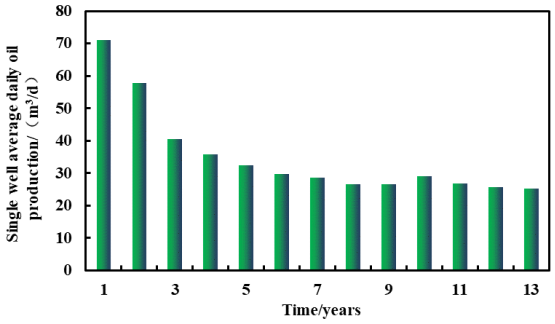


Fig. 3 Single well average daily oil production per year

Offshore oil fields are mostly produced by electric pumping, and the types of dynamic data are less than those of onshore oil fields. The pump head and pump displacement in the working condition data are analyzed and processed in this paper. Permeability and porosity are static parameters. Considering the change of perforating conditions and other working conditions, the thickness of the oil reservoir used for oil production will change, and the weighted average of the corresponding porosity permeability will also change. In this process, the weighted properties of the exploited oil reservoir can be regarded as dynamic parameters. The total reservoir

thickness and the thick-weighted average permeability, porosity, oil saturation and mud content of each well for oil production at different time periods were calculated according to the change of the wellbore of the oil well. Due to the limited sand body data of the oil well, a total of 30,000 pieces of data from 8 Wells were processed.

3.2 Optimization of dynamic parameters of oil well production prediction

The DTW distance between time series data is calculated using the dynamic time warping method, and the calculation results are shown in the following table.

Table 2 DTW distance between time series data and daily oil production

Features	DTW distance
Time	1160.24
Daily liquid production	199.59
Water cut	795.45
Reservoir thickness	300.24
Permeability	304.08
Porosity	461.88
Oil saturation	2626.50
Mud content	1172.32

According to the calculation results, the DTW distance between each feature and daily oil is as follows: daily fluid < reservoir thickness < permeability < porosity < water content < time < mud content < oil saturation.

In the two similarity measurement methods, the distance between daily fluid production, reservoir thickness, permeability, time and daily oil production is relatively small, and the similarity between daily oil production and daily oil production is high, which can be used as time series prediction. However, considering the strong correlation between daily fluid production and daily oil production and water cut, daily fluid production is not used as the characteristic data. Because the reservoir thickness and permeability are weighted values of the development zone calculated from the well history and only change after each job, they are not used as characteristic data.

3.3 Determining the number of characteristic variables of oil well dynamic production prediction

In the process of time series prediction, the feature data has one more time dimension than the conventional prediction, which further enhances the complexity of data features. Data with too many feature variables are easy to interfere with each other and affect the prediction effect when establishing the model, while data with too few feature variables are easy to miss some important features, resulting in poor prediction effect. Therefore, it is necessary to consider the influence of the number of characteristic variables on the model prediction effect.

The data of well C01 are sorted according to the importance of the preferred feature data in Chapter 2, and the least important data features are deleted successively to compare the prediction effect under different numbers of data features. The DTW distance in Section 2 of Chapter 2 from small to large is: daily fluid < reservoir thickness < permeability < porosity < water content < time < mud content < oil saturation. Among these eight factors, daily fluid production is a parameter directly related to daily oil production and water content. Since the model will use the daily oil production of the previous few days as the feature data, the daily fluid production with strong positive correlation is not used as the model training feature data. As for reservoir thickness, permeability, porosity, mud content and oil saturation, they are obtained by adjusting the thickness of the production zone during well operation and can only change after each well operation. The maximum number of well operations in each well in this oilfield is less than 10. The number of changes in these data is too small to reflect the dynamic characteristics of the well, so they are not important parameters. In addition, considering that time and water cut are very important dynamic parameters for oilfield development, they are removed at the end of the feature number optimization, in order to delete: oil saturation, shale content, porosity, permeability, oil layer thickness, time and water cut. The above dynamic characteristics plus daily oil production total 8 characteristics.

The well pattern where well C01 was located began flooding on April 14, 2008, so data from well C01 commissioned (May 30, 2002) through April 13, 2008 was used as evaluation data. The data from production to December 31, 2006 was used as the training set, and the data from December 31, 2006 to April 13, 2008 was used as the test set. The data training of each feature number can get its own model, and the number of features can be selected by evaluating the prediction effect of these models on the test set. The training model adopts the neural network of long and short term memory.

The forecast results are evaluated in the following table

Table 3 Evaluation results of feature number

Feature number	Training set R2	Training set MSE	Test set R2	Test set MSE
8	0.1021	473.99	-5.883	531.83
7	-0.2955	683.87	-7.500	656.73
6	0.2033	420.58	-4.056	390.69
5	0.8872	59.52	0.795	15.87
4	0.8909	56.29	0.850	12.03
3	0.8947	55.57	0.855	11.2
2	0.892	57.02	0.816	14.23
1	0.8858	60.3	0.833	12.93

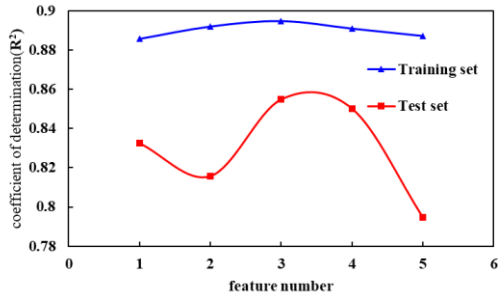


Fig.4 The change of the coefficient of determination with the feature number

It can be seen from Table 3 and Figure 4 that when the feature number is 3, the determination coefficient is the largest and mean square error and mean absolute error are the smallest. Therefore, time, water cut and daily oil production are selected as the final features.

4. MODEL DESIGN AND APPLICATION

4.1 The determination of structure of long and short term memory neural network

4.1.1 Number of hidden layers

Three variables of oil production, water cut and time were used for single step prediction. The input data has 3 characteristics. Firstly, the data with input step size 10 is used for prediction and neural network structure optimization.

Using more hidden layers in a neural network can fit more complex nonlinear functions. For a general data set, one or two layers are sufficient. The number of layers in the network and the complexity of the fitting function are shown in the table below.

Table 4 Relationship between the number of network layers and the fittable function

Number of hidden layers	What can be done
0	Can only represent linearly separable functions or decisions
1	Any function that "contains a continuous mapping from one finite space to another finite space" can be fitted
2	With the appropriate activation function, any decision boundary can be represented with any precision, and any smooth map can be fitted with any precision
>2	The extra hidden layers can learn complex descriptions (some kind of automatic feature engineering)

4.1.2 Optimization of the number of LSTM neurons in a single hidden layer

Taking the data from well C01 as the training and test data, the first 80% of the training set and the last 20% of the test set were taken as the input step size of 2. The prediction effect of the single hidden layer neural network was first explored, and the number of single hidden layer neurons was 1,2,4,8,16,32,64,128,256,512,1024 were selected as the range of parameter variation. Figure 5 shows the prediction effect evaluation of the three metrics as the number of neurons changes. It can be seen that the mean square error and mean absolute error of the training set are greater than that of the test set in this training. This is due to the high number of shut-ins and dramatic production changes in the training set data and the smooth and segmented processing of the final data used for training. When the number of neurons increases from 1 to 32, the coefficient of determination of training set and test set increases rapidly, and the mean absolute error and mean square error decrease rapidly. After that, the change rate of the three parameters is small and basically unchanged. When the number of neurons is 32, the neural network achieves the best prediction effect, and the coefficient of determination is 0.8911. Therefore, the number of neurons 32 is selected as the optimal number of single hidden layer neural networks. As can be seen from Figure 5, the prediction effect when the number of neurons is 32 is basically the same as that when the number of neurons is 64 and 128.

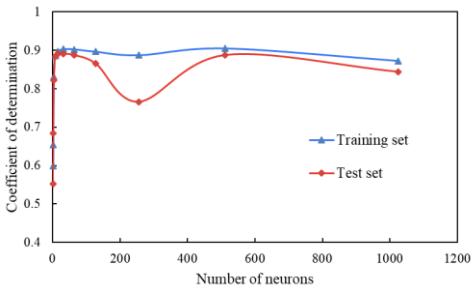


Fig.5 The coefficient of determination varies with the number of neurons

4.1.3 Optimization of the number of LSTM neurons in double hidden layers

Figure 6 shows the evaluation diagram of the number of neurons in two hidden layers of the double hidden layer neural network and the prediction results of the model. The evaluation index is the determination coefficient of the prediction results, which is similar to the change characteristics of the prediction of the single hidden layer neural network. When the number of neurons is between 1 and 32, the determination coefficient changes faster. When it is greater than 32, the rate of change is small and almost unchanged. When the number of neurons in the first hidden layer is 64 and the number of neurons in the second hidden layer is 32, the coefficient of determination is the largest, which is 0.8968. The optimal value of the determination coefficient of the double hidden layer is 0.8968, which is not much different from the optimal value of the single hidden layer 0.8911. Considering the difference in calculation amount, the single hidden layer neural network is preferred.

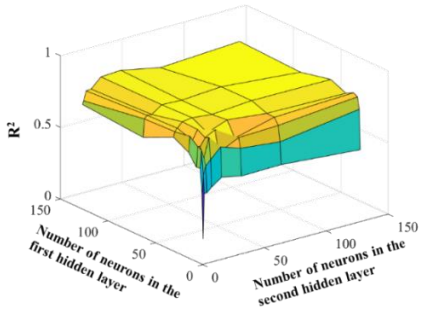


Fig. 6 Influence of the number of neurons in the double-hidden layer neural network on the prediction effect

4.2 Recursive multi-step forecast method for daily production of oil Wells

The data of multiple oil Wells in platform C were collected and the model parameters were optimized. The optimal neural network architecture is adopted. Due to the complexity of the field working conditions and the difficulty of quantification, this prediction method only predicts the working condition invariant segment between each two working conditions, and the production dynamics should be approximately similar to the constant liquid production, and the influence of human operation factors is not within the prediction range. The error of recursive multi-step prediction method will increase gradually with the progress of prediction.

Recursive multi-step prediction needs to output three parameters (daily oil production, water content and time), and the output result of the neural network is two (daily oil production and water content), and the time is updated by adding one day each time to achieve three-variable recursive multi-step prediction.

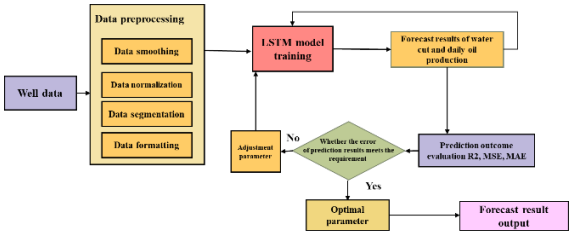


Fig. 7 LSTM recursive multi-step prediction diagram

Input step size from 2 to 16 is selected for input step size optimization, and the data test set still accounts for 20%. The prediction results of the next ten days after the training set are used for evaluation. As shown in Figure 8 ,when the input step is longer than 9, the fluctuation increases and then decreases. When the input step is 9, the error is the smallest, the average absolute error is 0.177, and the mean square error is 0.0374, which is the optimal input step.

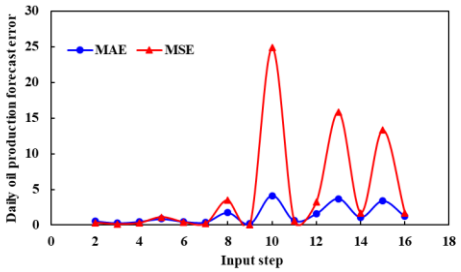


Fig. 8 Variation of long and short-term memory neural network error with step size

4.3 Integrated model of data driven and water drive feature

4.3.1 Integrated method

When the pure data-driven model predicts the change of water cut, it cannot reflect the rising process of water cut in the development process. Therefore, the production prediction model constrained by water drive characteristics is established, and the model prediction results obey the law of water cut rise. Figure 9 shows the integration of water drive features and data-driven method.

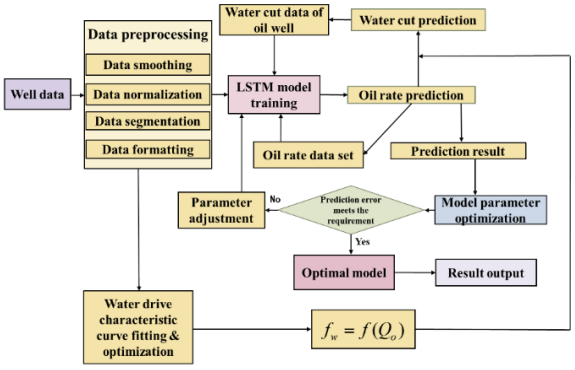


Fig. 9 Data-driven model and water drive characteristic curve integrated model prediction method

Figure 9 shows the schematic diagram of the method, which is based on LSTM recursive multi-step prediction model and uses water drive characteristic curve to calculate water content. LSTM recursive multi-step prediction is to add the prediction result to the data set in the prediction step and then make the next prediction. Integrated method prediction steps are as follows:

(1) Organize well data into time series prediction format and construct prediction data set;

(2) Complete model training and test evaluation with this data set;

(3) Based on the well data set and the LSTM model, the daily oil production of the next day is obtained.

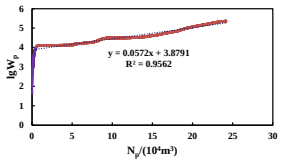
(4) Use the obtained daily oil production to calculate the water content of the next day through the water drive characteristic curve, and the update time is added one day forward;

(5) Add the updated daily oil production, water cut and time to the forecast data set, and repeat steps 3 and 4 to complete the daily oil production of the next day;

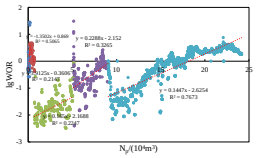
(6) Repeat steps 3, 4 and 5 to complete recursive multi-step prediction, which can predict any number of days backward.

4.3.2 Selection of water drive characteristic curve

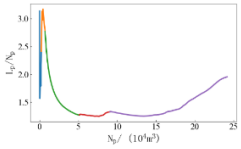
The accuracy of different water drive characteristic curves was evaluated by comparing the field oil well data. Figure 10 shows the fitting effect of different water drive characteristic curves, in which the fitting effect of A water drive characteristic curve and D water drive characteristic curve is better; B water drive characteristic curve can be fitted linearly after subsection; C water drive characteristic curve does not follow the linear law after subsection. Considering that the D-type water drive characteristic curve cannot calculate the water yield by oil production, the water-drive characteristic curve A is chosen as the water-drive characteristic curve integrated with data drive.



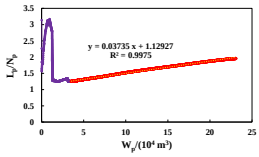
(a)A water drive characteristic curve



(b)Characteristic curve of water drive B



(c) C type water drive characteristic curve



(d)D-type water drive characteristic curve

Fig.10 Fitting effect of four kinds of water drive characteristic curves

4.3.3 Integrated model parameter optimization

Under different input steps, the prediction accuracy of the model is different. Evaluate the prediction effect under different input steps and select the optimal input step. Figure 11 shows the prediction error of the integrated model under different input steps. The results show that when the input step size is 19, the error is the smallest, the average absolute error is 0.1429, and the mean square error is 0.0262. Therefore, enter step 19 as the final input step.

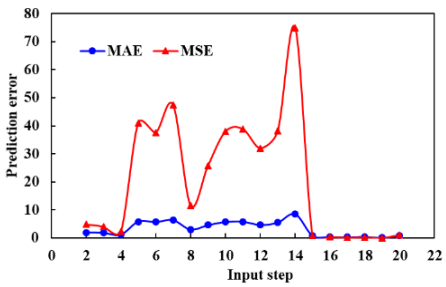


Fig.11 Integrated model error curve with input step size

4.4 Evaluation and comparison of model application effect

Taking Wells C02 and C09 as examples, the prediction effect of the model was evaluated. By comparing the prediction effect of the pure data-driven model, the numerical simulation model and the Integrated model, the prediction effect of different models on the field data was evaluated. Figure 12-15 shows the prediction effect and prediction error of the three models on the two Wells. Figure 12 and Figure 13 show that the predictive ability of the integrated model for the change trend of daily oil production is stronger than that of the numerical simulation model and the pure data-driven model. Figure 14 and Figure 15 show that the integrated model has the smallest prediction error, followed by the pure data-driven model, and the numerical simulation model has the largest prediction error. The overall prediction results show that the integrated model can predict oil production more accurately, and the prediction accuracy of the data-driven model can be improved by embedding water drive characteristics.

The integrated model predicts the production of a well over the next 30 days, and the forecast results can guide the working system of the well, predicting and adjusting the irrational production system in advance.

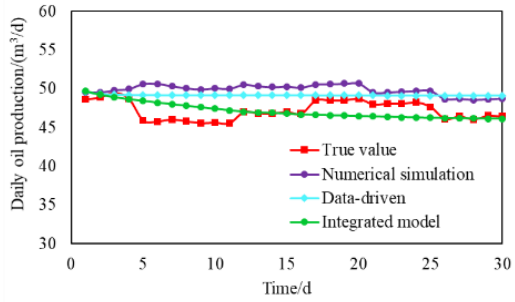


Fig.12 Prediction effect of multiple models in well C02

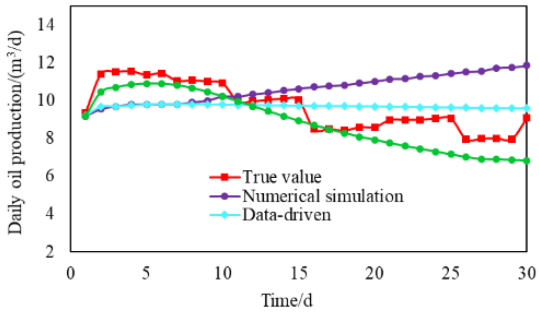


Fig.13 Prediction effect of multiple models in well C09

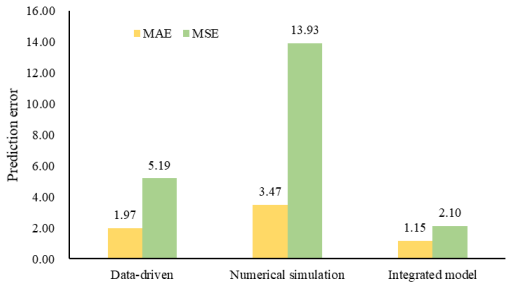


Fig.14 Prediction error of well C02

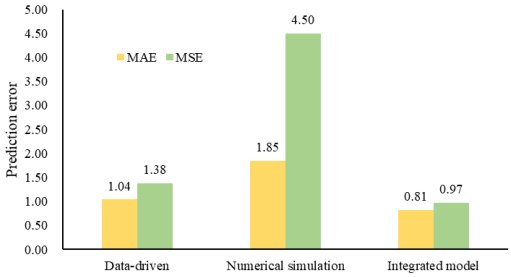


Fig.15 Prediction error of well C09

5. CONCLUSIONS

(1)LSTM feature number optimization results show that for QHD32-6 oilfield, the three features have the best prediction effect, so oil production, water cut and time are selected as the final features;

(2) When the input step of pure data-driven LSTM model is 9, the prediction error of daily oil production is the smallest; When the input step size of integrated model is 19, the prediction error is minimum.

(3)The integrated model improves the interpretability and prediction accuracy of the model. Compared with the numerical simulation model and the pure data-driven model, the integrated model has a stronger prediction ability for the change trend of production than the numerical simulation model and the pure data-driven model. Among the three models, the integrated model has the smallest prediction error, followed by the pure data-driven model, and the numerical simulation model has the largest prediction error.

REFERENCE

[1] Arps JJ. Analysis of decline curves. Transactions of the AIME, 1945.160(1):228-247.

[2] Fetkovich M J. Decline curve analysis using type curves. Journal of PetroleumTechnology,1980,32(6): 1065-1077

[3] Blasingame T A, McCray T L, Lee W J. Decline curve analysis for variable pressure drop/variable flowrate systems. Paper SPE 21513 presented at the SPE Gas Technology Symposium, Houston, Texas, USA, 1991.
[4] Agarwal R G, Gardner D C, Kleinsteiber S W, et al. Analyzing well production data using combined-type-curve and decline-curve analysis concepts. SPE Reservoir Evaluation& Engineering,1999,2(05):478-486
[5] Mattar L, Anderson D M. A systematic and comprehensive methodology for advanced analysis of production data. Paper 84472 presented at the SPE Annual Technical Conference and Exhibition, Denver, Colorado, USA, 2003.
[6] SCHILTHUIS R J. Active oil and reservoir energy. Transactions of the AIME,1936,118(1):33-52.
[7] LIU Dehua, LIU Zhisen. Foundation of reservoir engineering. Beijing: Petroleum Industry Press,2004.
[8] WEI Yi, RAN Qiquan, LI Ran, et al. Determination of dynamic reserves of fractured horizontal wells in tight oil reservoirs by multi region material balance method. Petroleum Exploration and Development,2016,43(3) :448-455.
[9] ZHANG Lixia, GUO Chunqiu, JIANG Hao, et al. Gas in place determination by material balance-quasi pressure approximation condition method. Acta Petrolei Sinica, 2019,40(3):337-349.
[10] LI Jianglong, ZHANG Hongfang. Application of the material balance method to the energy evaluation of fractured-vuggy carbonate reservoirs. Oil &. Gas Geology,2009 ,30(6) :773-778.
[11] ZHENG Songqing, CUI Shuyue, MU Lei. Material balance equation and driving energy analysis of fracture-cave oil reservoir Special Oil and Gas Reservoirs,2018 ,25(1):64-67.
[12] LI Chuanliang. Principle of reservoir engineering. 2nd ed. Beijing: Petroleum Industry Press, 20114
[13] JI Bingyu, Some understandings on the development trend in research of oil and gas reservoir engineering methods. Acta Petrolei Sinica,2020,41(12) :1774-1778.
[14] HAVLENA D,ODEH A S The material balance as an equation of a straight line. Journal of Petroleum Technology, 1963, 15(8) :896-900.
[15] NADER W. An investigation concerning the material balance equation part one: the linear form of the equation. Journal of Canadian Petroleum Technology,1964,3(1):28-32..
[16] TERRY R E，ROGERS J B. Applied petroleum reservoirs engineering. Zhu Daoyi trans. Beijing :Petroleum Industry Press ,2017.
[17] Tamhane D., et al. Soft Computing for Intelligent Reservoir Characterization. Springer Berlin Heidelberg, 2000.
[18] Nguyen, Ha H. , C. W. Chan , and M. Wilson . Prediction of oil well production: A multiple-neural-network approach. Intelligent Data Analysis 8.2 (2004):183-196.
[19] Bhattacharya, Shuvajit, Srikanta Mishra. Applications of machine learning for facies and fracture prediction using Bayesian Network Theory and Random Forest: Case studies from the Appalachian basin, USA. Journal of Petroleum Science and Engineering 170 (2018): 1005-1017.
[20] Noshi, Christine Ikram, Marco Risk Eissa, and Ramez Maher Abdalla. An intelligent data driven approach for production prediction. Offshore Technology Conference. OTC, 2019.
[21] Niu, Wente, Jialiang Lu, and Yu** Sun. Development of shale gas production prediction models based on machine learning using early data. Energy Reports 8 (2022): 1229-1237.
[22] Marvin, Mahlon Kida, et al. An Echo State Network Approach to Data-Driven Modelling and Optimal Control of Carbonate Reservoirs with Uncertainty Fields. Geoenergy Science and Engineering (2024): 212996.
[23] Chen, M, et al. Production Prediction Model of Tight Gas Well Based on Neural Network Driven by Decline Curve and Data. Processes 12.5 (2024): 932.
[24] Li Decai. Research on Correlation Analysis and prediction Method based on multivariate time series [D]. Dalian University of Technology,2012.
[25] Khan R. Image-to-text: Image captioning based on bidirectional LSTM and attention mechanism. University of Science and Technology of China,2022.
[26] PENG Hong. Research on Structural damage identification based on improved LSTM. Shantou University,2022.
[27] Wang Wei, Hu Caibo, Zhao He, et al. Application of LSTM neural network in clock difference prediction during fast change of satellite clock frequency. Geodesy and Geodynamics, 2019,43(04):369-373.