



Intelligent Prediction Technology for Production of Tight Oil Based on Data Analysis

Ning Li^{1,2}(✉), Xiang-hong Wu^{1,2}, Xin Li^{1,2}, Zhi-ping Wang^{1,2}, Yue-zhong Wang^{1,2}, Li-ao Zhao^{1,2}, Liang Ren^{1,2}, Hong-liang Wang^{1,2}, Hong-yu Tian¹, Shu-hang Ren^{1,2}, and Si-rui Jiang³

¹ Research Center of Artificial Intelligence, Research Institute of Petroleum Exploration and Development, PetroChina, Beijing, China

lining_riped@petrochina.com.cn

² Artificial Intelligence Technology R&D Center for Exploration and Development, CNPC, Beijing 100083, China

³ Department of Asia-Pacific E&P, Research Institute of Petroleum Exploration and Development, PetroChina, Beijing, China

Abstract. China is rich in tight oil resources, with a wide distribution range and a large amount of resources, making it one of the key areas for strategic replacement of future oil reserves and production. In response to issues such as strong heterogeneity of terrestrial tight oil reservoirs, difficulty in drilling high-quality oil layers, large production differences, and unclear main control factors for production capacity, a detailed analysis of dynamic and static data of production wells was conducted to analyze production performance and decline patterns. Production wells were classified according to production characteristics, and development indicators at different stages were statistically analyzed based on actual production days. Using a combination of principal component analysis and Pearson correlation coefficient, based on multiple dynamic and static data such as geological factors, fracturing factors, and development factors, and analyzing the correlation between different single and combined factors and cumulative oil production at different stages, the main control factors for different production stages of tight oil were obtained. A production capacity prediction model for tight oil fracturing horizontal wells was established based on machine learning intelligent algorithms,

Copyright 2023, IFEDC Organizing Committee.

This paper was prepared for presentation at the 2023 International Field Exploration and Development Conference in Wuhan, China, 20-22 September 2023.

This paper was selected for presentation by the IFEDC Committee following review of information contained in an abstract submitted by the author(s). Contents of the paper, as presented, have not been reviewed by the IFEDC Technical Team and are subject to correction by the author(s). The material does not necessarily reflect any position of the IFEDC Technical Committee its members. Papers pre-sented at the Conference are subject to publication review by Professional Team of IFEDC Technical Committee. Electronic reproduction, distribution, or storage of any part of this paper for commercial purposes without the written consent of IFEDC Organizing Committee is prohibited. Permission to reproduce in print is restricted to an abstract of not more than 300 words; illustrations may not be copied. The abstract must contain conspicuous acknowledgment of IFEDC. Contact email: paper@ifedc.org.

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 78–98, 2024.

https://doi.org/10.1007/978-981-97-0272-5_7

A production capacity evaluation and prediction technology for tight oil fracturing horizontal wells has been developed. By comparing with actual production data, the accuracy of the predicted results can meet production needs, providing a strong technical foundation for precise prediction and guidance of tight oil production in China.

Keywords: Tight oil · Production forecast · Data analysis · Analysis of main control factors · Intelligent algorithms

1 Introduction

Rich tight oil resources have been discovered in terrestrial sedimentary reservoirs of multiple basins in China, with a total resource volume exceeding 11 billion tons, making tight oil a major development replacement field and a new strategic growth point for crude oil production in China. Compared with North American marine tight oil, Chinese terrestrial tight oil has the characteristics of “multiple types, low porosity, low fluidity, and relatively poor oil properties”. The geological conditions of terrestrial tight oil in China are complex, with multiple types and complex resource composition. The distribution of sand bodies is scattered, the vertical and horizontal continuity of reservoirs is poor, the reservoirs are dense and heterogeneous, and there are significant differences in single well drilling rates. The source reservoir relationship is mainly dominated by the intra source type, accounting for approximately 77.7%, the sub source type accounting for 18.2%, and the above source type accounting for 4.1%. The lithology is mainly composed of sandstone, accounting for about 69%, carbonate rock accounting for about 29.8%, and sedimentary volcanic rock accounting for about 1.3%. The pressure coefficient is mainly high pressure, 64.8% of which is >1.2 , 29.3% of which is $0.8\text{--}1.2$, and 22% of which is <0.8 . The physical properties of crude oil are mainly low viscosity crude oil, with 41.2% having a viscosity of <2 mPa.s, 31.7% having a viscosity of $2\text{--}10$ mPa.s, and 27.1% having a viscosity of >10 mPa.s.

Through the analysis of development effectiveness, domestic tight oil development currently faces two challenges in terms of production and efficiency: firstly, the large difference in single well production capacity, rapid decline, and low EUR of tight oil, which poses challenges to the effective utilization of tight oil. The second is the high cost and poor efficiency of using horizontal wells and volume fracturing for development. In the current context of low oil prices, how to reduce costs and improve development efficiency faces serious challenges. Through research, it has been found that the strong heterogeneity of the physical properties and oil-bearing properties of tight oil reservoirs is the fundamental reason for the significant productivity differences in horizontal wells. The significant difference in the effectiveness of tight oil fracturing is an important factor affecting production capacity. The production of tight oil in a single well depends on the production of each fracturing section, which is mainly controlled by the oil-bearing, physical properties, fluid properties, and fracturing effect of the reservoir; The organic matching of high-quality reservoir drilling rate and effective fracturing interval number is the main controlling factor for single well productivity. The strong heterogeneity of the reservoir is an important factor affecting the drilling rate. The low drilling rate and

low saturation of movable fluids in Class I high-quality reservoirs are the fundamental reasons for the failure to achieve the expected production of horizontal wells.

In order to effectively predict the decline law of tight oil production, analytical and numerical calculation methods are currently mainly used. Among them, analytical calculation methods mainly include Arps decline curve method, typical decline curve chart method, relative permeability curve method, etc. Numerical calculation methods mainly refer to reservoir numerical simulation methods. However, each of these two methods has its advantages and disadvantages: the analytical calculation method has a fast calculation speed and can quickly provide a rough curve trend pattern. However, the decline pattern of tight oil is complex, and a single decline pattern formula is difficult to describe the overall decline process, and the calculation accuracy is not very accurate. However, the reservoir numerical simulation method can accurately calculate numerical solutions, but generally takes a long time and has high calculation costs.

With the gradual rise of artificial intelligence technology and the significant improvement of computer computing power, artificial intelligence prediction technology has emerged. From the perspective of big data analysis, this technology considers more influencing factors and is more comprehensive compared to traditional analytical methods. At the same time, compared to reservoir numerical simulation methods, it does not require global direct numerical simulation of the flow field values at each time step, greatly improving the calculation speed. Hamid Rahmanifard [1] made a detailed comparative analysis of the performance of ML algorithms and statistical methods, and then used two statistical methods (exponential smoothing and seasonal autoregressive comprehensive moving average) to make a comparative study of six kinds of modern ML networks, including multilayer perceptron (MLP), long short-term memory (LSTM), bidirectional LSTM (BiLSTM), convolutional neural network (CNN), long-term recursive convolutional network (LRCN) and gated recursive unit (GRU). In order to determine the relationship between static and dynamic data of some development units in the oilfield and the decline rate of oil production, Zhang Yan [2] used data-driven methods to identify the correlation between post fracturing production and production influencing factors by analyzing the geological properties and fracturing construction parameters of tight sandstone in Changqing Oilfield. Elastic networks, decision tree regression, support vector regression have been used to establish prediction models from reservoir properties and fracturing construction parameters to production. Liang Tao [3] established an initial cumulative oil production mixing model for Multi Fractured Horizontal Wells (MFHWs) that considers both geological and volumetric fracturing factors. Based on big data, a multi-level evaluation system has been established using Analytic Hierarchy Process. Calculate the weighting factor to reveal the key factors affecting the productivity of MFHWs. Using fuzzy logic method to calculate Euclidean distance and quantitatively predict the production of any horizontal well. Zainab Al Ali Hussain Al Ali [4] used two deep learning models, namely, Long short-term memory (LSTM) and N-BEATS, to predict the oil recovery data of two wells in Norway's Norne Oilfield. The use of pre-trained N-BEATS models overcomes the shortcomings of LSTM models that previously required feature selection and rich training history, and the performance of N-BEATS meta learning methods is superior to LSTM models. The LSTM neural network model

has been used multiple times to predict the trend of monthly oil production and water content in high water cut old oilfield blocks [5–11].

The eXtreme Gradient Boosting (XGBoost) algorithm is a scalable distributed gradient boosting decision tree (GBDT) machine learning library. XGBoost provides parallel tree enhancement function and is an advanced machine learning library for regression, classification, and ranking problems. XGBoost was initially initiated as a research project by Tianqi Chen as part of the Distributed (Deep) Machine Learning Community (DMLC) group. It is an optimized distributed gradient enhancement library designed for efficiency, flexibility, and portability. XGBoost is a tool for large-scale parallel boosting trees, which is more than 10 times faster than common toolkits. In terms of large-scale data in the industry, the distributed version of XGBoost has extensive portability, supporting running on various distributed environments such as Kubernetes, Hadoop, SGE, MPI, Dask, etc., making it a good solution to the problem of large-scale data in the industry.

This paper adopts the XGBoost algorithm to establish a corresponding single well production decline prediction model based on the characteristics of tight oil reservoirs in China. Through practical application in a tight oil field in China, the superiority and correctness of this method in predicting single well production capacity have been confirmed, meeting the urgent needs of oilfield dynamic analysis, development planning, and decision-making.

2 Analysis of the Declining Law of Tight Oil Production

Although the overall changes in production characteristics of horizontal wells in each block are consistent, there are certain differences in the changes in daily liquid production, daily oil production, water content, production casing pressure, and other characteristics of each horizontal well based on the analysis of single well development performance data. Through literature research, it was found that most tight oil reservoirs are analyzed for production characteristics based on the variation of daily oil production with mining time. Therefore, this article will classify and analyze the production and mining characteristics of horizontal wells in the study area based on the variation of daily oil production. According to the curve characteristics of the daily oil production of a single well changing with mining time, the production and mining characteristics of horizontal wells can be divided into four categories:

2.1 Type 1: Rapid Increase in Initial Production and Short Stable Production Period

The overall performance is that the daily oil production capacity of horizontal wells continues to increase in the initial stage of production, and reaches the highest daily oil production level (10t–15t) within about 10 months. However, the stable production period is relatively short, and after 1 year of production, the daily oil production begins to decrease. After 2 and a half years, the daily oil production of a single well decreases to about 5t; The change in daily liquid production is similar to that of daily oil production; The water cut changes in the opposite direction and fluctuates within the range of 80% to 100%. The fracture network formed by horizontal well fracturing is the reason for high

production in the initial stage of production, and the high production period is generally maintained between the second month and the sixth month, after which it enters the decreasing stage. The typical production curve is shown in Fig. 1(a).

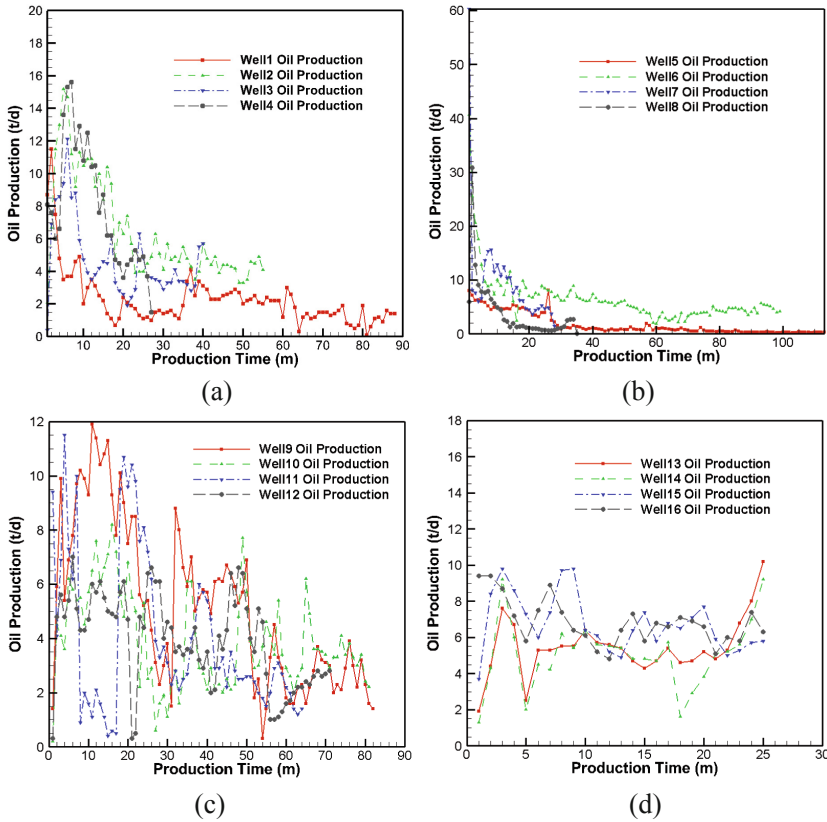


Fig. 1. Curve of Daily Oil Production of a Single Well Changing with Production Time.

2.2 Type 2: High Initial Production and Rapid Decline in Later Stages

The overall performance is that the horizontal well has a high daily oil production capacity in the early stage of production, but the stable production period is extremely short. Generally, the daily oil production level starts to decrease within one month, and the daily oil production in the first three months drops to about 50% of the initial production, with a very fast decline rate. Generally, the daily oil production of the well drops to below 5t within 2–3 years of production. The changes in daily liquid production and production casing pressure of horizontal wells are similar to daily oil production. The typical production curve is shown in Fig. 1(b).

2.3 Type 3: The Fluctuation of Production is Large and Showing Multiple “Peaks”

The performance is that the daily oil production capacity of horizontal wells gradually increases in the initial stage of production, but the stable production period is short. During the production time, the daily oil production continuously fluctuates up and down. Overall, the daily oil production level is the strongest in the initial stage, and the daily production in the later stage shows a downward trend fluctuation. Generally, the daily production of wells decreases to below 5t after 4–5 years of production. The typical production curve is shown in Fig. 1(c).

2.4 Type 4: No Significant Fluctuations in Production and Maintaining Stable Production

The performance is that the daily oil production capacity of horizontal wells gradually increases in the initial stage of production, reaching its maximum in about 3 months, and the daily oil production is relatively stable throughout the entire production period, maintaining between 5–10t/d; The changes in water content and daily liquid production are similar to the daily oil production. The maximum daily oil production of this type of horizontal well is within the range of 5–10t/d, which is at a moderate level. At the same time, the production time is relatively short, mostly within two years. The daily production is still in a stable period, so there is no significant fluctuation and stable production has been maintained. The typical production curve is shown in Fig. 1(d).

3 Introduction of XGBoost Algorithm

XGBoost, as one of the Boosting algorithms, is a lifting tree model that integrates many tree models. By adding a regular term to the loss function, the complexity of the model is controlled to prevent overfitting. It can achieve parallel processing, which has greatly improved the speed compared to GBDT. XGBoost is essentially k decision trees (k is a positive integer), and the output of the regression tree is a real number (continuous variable). Boosting method is to combine multiple weak learners to give the final learning results, and take the output results of each weak learner as continuous values. The purpose of this is to accumulate the results of each weak learner, and better use the loss function to optimize the model.

Let $f^t(x_i)$ is the output result of the t -round weak learner, $\hat{y}_i^{(t)}$ it is the output result of the model, y_i it is the actual output result, and the expression is as follows:

$$\hat{y}_i^{(t)} = \sum_{k=1}^t f^k(x_i) = \hat{y}_i^{(t-1)} + f^t(x_i) \quad (1)$$

The objective function, that is, the loss function, builds the optimal model by minimizing the loss function. The loss function should add a regular term representing the complexity of the model, and the model corresponding to XGBoost contains multiple CART trees. Therefore, the objective function of the model is:

$$obj(\theta) = \sum_i^n L(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^t \Omega[f^k(x_i)] \quad (2)$$

The above formula is the regularization loss function. The first part on the right side of the equation is the training error of the model, and the second part is the regularization term. The regularization term here is the sum of the regularization terms of k trees. The specific form is:

$$\Omega[f^k(x_i)] = \gamma T + \frac{1}{2} \lambda \|w\|^2 \quad (3)$$

where: T is the number of leaf nodes, $\|w\|$ is the modulus of the leaf node vector, γ it indicates the difficulty of node segmentation, indicates L2 regularization coefficient.

According to the expansion rule of the second derivative of the Taylor formula, the training error is further deduced and expanded to obtain:

$$obj(\theta)^{(t)} = \sum_{j=1}^T \left[G_j w_j + \frac{1}{2} (H_j + \lambda) w_j^2 \right] + \gamma T \quad (4)$$

where: G_j represents the sum of the first derivative of all input samples mapped as leaf node j , H_j represents the sum of second derivative of all input samples mapped to leaf node j .

In summary, we have introduced the main algorithms of XGBoost, which lays a theoretical foundation for subsequent prediction applications.

4 Workflow

For the prediction of well production in tight oil fields, first of all, data collection and pre-processing should be carried out, including the static and dynamic data of the reservoir, and the corresponding sample database should be established. Then, closely combining with the field data of the oilfield, and making full use of geological, engineering and development data, based on the production performance analysis and production decline law analysis in the study area, Identify the relevant influencing factors that affect the production capacity of horizontal wells for volume fracturing in tight oil reservoirs, calculate the partial correlation coefficient between the two factors, screen out independent influencing factors, and conduct single factor and multiple combination factor analysis from three aspects: geological parameters, engineering parameters, and development factors. Through Principal Component Analysis (PCA) & Pearson Correlation Coefficient Analysis method (PCCA) methods, comprehensively analyze multiple/single factors to screen out the main controlling factors for production capacity; Establish a prediction model based on XGBoost, which requires training and tuning the model to ultimately form the optimal XGBoost tight oil field well production prediction model. The specific process is shown in Fig. 2.

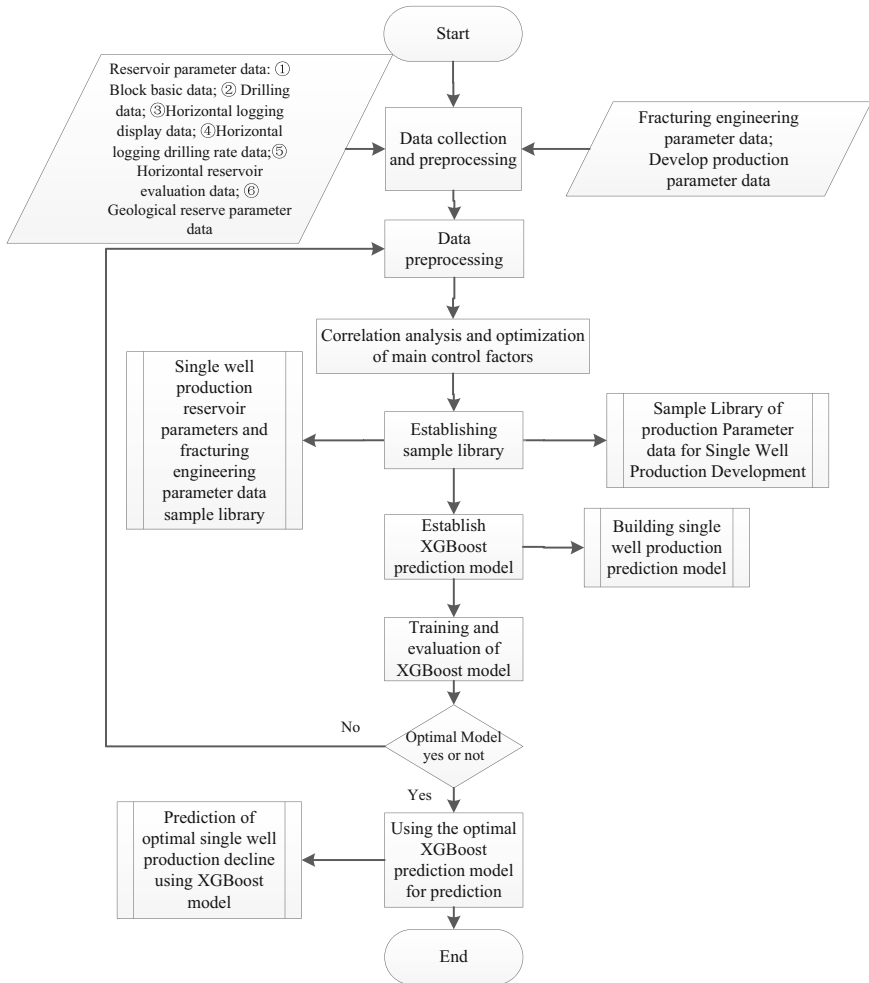


Fig. 2. Technical workflow.

4.1 Data Collection and Preprocessing

Collection and Organization of Data. The production of a single well in a tight oil field is influenced by various factors, mainly including reservoir parameter data, fracturing engineering parameter data, and development and production parameter data. In terms of reservoir parameter data, it also includes block basic data, drilling data, horizontal section logging display data, horizontal section drilling rate data, horizontal section reservoir evaluation data, geological reserve parameter data, etc. The specific relevant parameters are shown in Table 1.

Table 1. Collected dynamic and static parameters.

Data classification	Related parameters	
Reservoir parameter data	Block basic data	a) Block, well number, horizon, sublayer, reference well, interpretation layer, effective thickness of each sublayer, designed production capacity, designed well depth, and designed horizontal section length
	Drilling data	b) First drilling time, completion time, year of production, completion method, cycle, well depth, artificial bottom hole, oblique depth of point A during actual drilling, vertical depth of point A during actual drilling, length of horizontal section, length of horizontal section used
	Horizontal logging display data	c) The length of sandstone encountered in horizontal section logging, the length of oil layer encountered in horizontal section logging, the drilling rate of sandstone encountered in horizontal section logging, the drilling rate of oil layer encountered in horizontal section logging, the oil immersion length of horizontal section logging, the length of oil spot in horizontal section logging, the length of oil stains in horizontal section logging, the fluorescence length of horizontal section logging, and the total length of horizontal section logging

(continued)

Table 1. (continued)

Data classification	Related parameters
Horizontal logging drilling rate data	d) The length of sandstone encountered during horizontal logging, the length of oil layer encountered during horizontal logging, the drilling rate of sand-stone encountered during horizontal logging, and the drilling rate of oil layer encountered during horizontal logging
Horizontal reservoir evaluation data	e) Horizontal Section I Reservoir Length, Horizontal Section II Reservoir Length, Horizontal Section III Reservoir Length, Horizontal Section IV Reservoir Length, Horizontal Well Classification Evaluation Category, Horizontal Well Classification Evaluation Index
Geological reserve parameter data	f) Production thickness, fracture length, porosity, average saturation, density, volume coefficient, controlled reserves, production reserves
Fracturing engineering parameter data	g) Fracturing completion structure, number of fracturing segments, number of fracturing clusters, average interval spacing, average cluster spacing, total fracturing fluid volume, total fracturing sand volume, fluid intensity, sand intensity, single stage fluid volume, single stage sand volume, fracturing completion time, soaking time after fracturing, single cluster fluid volume, single cluster sand volume

(continued)

Table 1. (continued)

Data classification	Related parameters
Develop production parameter data	h) Block, well number, well pattern type, production time, oil production method, pump diameter, oil pressure, casing pressure, dynamic liquid level, production days, cumulative production days, monthly oil production, monthly water production, verification of monthly oil production, verification of monthly water production, verification of cumulative oil production, verification of cumulative liquid production, number of wells opened, daily liquid production capacity, daily oil production capacity, water content, verification of daily liquid production capacity, verification of daily oil production capacity, depth of middle oil layer, flow pressure, storage and production coefficient Return rate, deficit, recovery degree, upward pumping time, and number of months of self-production

In addition to the single factor mentioned above, in order to highlight the impact of different factors and have a greater correlation with production capacity, the following multiple factors have been added according to the needs of the research problem, including:

Among them, the effective length of the horizontal well L_{eh} is

$$L_{eh} = a_1L_{oi} + a_2L_{osp} + a_3L_{ost} + a_4L_f \tag{5}$$

And the oil-bearing S_{ob} is

$$S_{ob} = \frac{L_{eh}}{L_{oi} + L_{osp} + L_{ost} + L_f} \tag{6}$$

where: L_{oi} stands for the length of oil immersion, m; L_{osp} stands for the length of oil spot, m; L_{ost} stands for the length of oil stains, m; L_f stands for the length of fluorescence, m; $a_i, i = 1, 2, 3, 4$ stands for the weight.

Table 2. Added multiple factor parameters.

Data classification	Related parameters
Reservoir parameter multiple data	a) Oil-bearing b) Effective length of horizontal well c) Permeability \times Thickness used d) Permeability \times Thickness used \times Effective length of horizontal well e) Utilized reserves \times Permeability f) Utilized reserves \times Permeability/Viscosity g) Oil-bearing \times Production reserves \times Permeability/Viscosity h) Reservoir quality \times Oil-bearing \times Produced reserves \times Permeability/Viscosity
Multiple data of fracturing engineering	a) Liquid strength \times Sand strength b) Liquid strength \times Sand strength \times Number of segments c) Liquid strength \times Sand strength \times Number of segments \times Effective length of horizontal well d) The amount of liquid added in single stage \times Effective length of horizontal well e) The amount of sand added in single stage \times Effective length of horizontal well

Note: Liquid strength equals to the amount of total liquid/Utilized reserves. Sand strength equals to the amount of total sand/Utilized reserves

Data Preprocessing. For different types of data in tight oil well areas, data cleaning is carried out based on their data volume, data type, data quality, etc., eliminating duplicate well information, completing missing data, data integration, data transformation, and other processes, and corresponding preprocessing is carried out for each data item.

- (1) Correction of flowback period data: After fracturing construction, the production during the flowback period is very low, which is not a normal industrial oil flow. Therefore, it is necessary to remove the time period of the flowback period and the oil production below a certain amount. The specific quantitative values vary from different oilfields;
- (2) Reorganize production data based on differences: Due to the cleaning of flowback period data and time periods, it is necessary to recalculate the cumulative oil production, cumulative liquid production, and water content for different production time periods;
- (3) Removal of abnormal well data: Based on expert experience and data analysis, identify wells with abnormally high or low production by drawing charts, and eliminate them according to specific circumstances;
- (4) Pre processing of specific data tables:
 - (a) Reservoir static data: porosity, formation pressure, and other data, with fixed values for each block. Based on the collected data, these types of data are supplemented in the data table.

- (b) Developing dynamic data: Dynamic data such as extraction degree and dynamic liquid level vary, varies with production time, and need to be recalculated and organized based on expert experience and specific formulas.
- (c) The combination of dynamic and static data: Through difference calculation, the production dynamic data has been reorganized and calculated. Merge the newly generated development dynamic data into a static data table according to the well name.

4.2 Data Correlation Analysis

After sorting out the influencing factors of production capacity and preprocessing the data, it is necessary to conduct correlation analysis between the influencing factors and production capacity, and screen out the main controlling factors of production capacity. This article adopts a combination of principal component analysis (PCA) and Pearson correlation coefficient, the method of combining PCA and Pearson is adopted.

Principal Component Analysis (PCA). The principal component analysis method is to transform multiple existing indicators into a few well representative comprehensive indicators, which can reflect most of the information of the original indicators and maintain independence between each indicator to avoid overlapping information. Principal component analysis mainly plays a role in reducing dimensionality and simplifying data structures.

- (a) Standardize indicator data, collect p -dimensional random vectors X , n samples,

$$X_i = \{X_{i1}, X_{i2}, \dots, X_{ip}\}^T, (i = 1, 2, \dots, n) \quad (7)$$

Construct a sample matrix and perform standardized transformation on the sample matrix;

$$Z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, (i = 1, 2, \dots, n; j = 1, 2, \dots, p) \quad (8)$$

- (b) Calculate correlation coefficient matrix based on standardized matrix;

$$R = [r_{ij}]_{p \times p} = \frac{Z^T Z}{n - 1} \quad (9)$$

- (c) Solve the characteristic equation of the sample correlation matrix R , obtain p characteristic roots, and determine the principal components;

$$U_{ij} = z_i^T b_j^0, (j = 1, 2, \dots, m) \quad (10)$$

- (d) Convert the standardized indicator variables into main components;
- (e) Perform a comprehensive evaluation of m principal components, sum them with weights, and obtain the final evaluation value. The weight is the variance contribution rate of each principal component.

Pearson Correlation Coefficient Analysis Method (PCCA). Pearson correlation coefficient analysis is a method used to measure the degree of correlation between two variables X and Y , with values between -1 and $+1$. Defined as the quotient of covariance and standard deviation between two variables.

$$\rho_{X,Y} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (11)$$

By using the above method, the correlation coefficients between the factors in Tables 1 and 2 and the cumulative oil production at different stages were obtained, as shown in Table 3.

Table 3. Correlation analysis results of various factors and oil production.

Influence factor	Correlation coefficient (Weight)
Effective length of horizontal section	0.304
Oil-bearing	0.212
Thickness used	0.152
Controlled reserves	0.152
Permeability \times Thickness used	0.05
Controlled reserves \times Permeability/Viscosity	0.05
Permeability \times Thickness used /Viscosity	0.015
Number of fracturing segments	0.222
The amount of sand added in single cluster	0.14
Liquid strength	0.098
The amount of liquid added in single stage	0.091
Storage and production correlation coefficient	0.08
The amount of sand added in single cluster	0.079
Post-pressure soaking time	0.07
Total amount of sand added	0.05
Production pressure difference	0.048
Return rate before oil breakthrough	0.035
Sanding strength	0.023
Liquid strength \times Sand strength	0.017
Number of fracturing clusters	0.017

(continued)

Table 3. (continued)

Influence factor	Correlation coefficient (Weight)
Bottom hole pressure	0.012
The amount of sand added in single stage	0.01
Total amount of liquid added	0.005
Water displacement before oil exposure	0.002

4.3 Selecting Main Controlling Factors for Tight Oil Production

Through the above data analysis and combined with expert experience, the following parameters were ultimately selected as the main control factors (Table 4):

Table 4. Results of main control factors for tight oil production.

Classification	Main control factors
Main control factors for geology	a) Effective length of horizontal section
	b) Oil-bearing
	c) Thickness used
	d) Controlled reserves \times Permeability/Viscosity
Main control factors for fracturing and development	e) Number of fracturing segments
	f) The amount of sand added in single cluster
	g) Liquid strength
	h) The amount of liquid added in single stage
	i) The amount of liquid added in single cluster
	j) Total amount of sand added
	k) Production pressure difference

4.4 Constructing a Typical Well Production Sample Library

Combining professional knowledge and expert experience, based on correlation analysis results and combined with cumulative production data from different stages, a sample library reflecting the changes in single well production was established. Through the sample library, expert experience was reflected.

4.5 Establishing a Multi Parameter Prediction Model for Well Production

Model Construction. Due to the fact that this article only involves three blocks of an oil field with a small sample size, it belongs to the small sample problem. Therefore, in the design of the prediction model, the concept of cyclic input is considered, which is to establish production prediction models according to different stages. When predicting the current stage of production, the cumulative output value of the previous production stage is input, as follows (Table 5):

Table 5. Prediction model input and output values.

No.	Input values	Output value
1	Main control factors	Q_1
2	Main control factors, Q_1	Q_3
3	Main control factors, Q_1, Q_3	Q_6
4	Main control factors, Q_1, Q_3, Q_6	Q_9
5	Main control factors, Q_3, Q_6, Q_9	Q_{12}
6	Main control factors, Q_6, Q_9, Q_{12}	Q_{18}
7	Main control factors, Q_9, Q_{12}, Q_{18}	Q_{24}
8	Main control factors, Q_{12}, Q_{18}, Q_{24}	Q_{36}
9	Main control factors, Q_{18}, Q_{24}, Q_{36}	Q_{48}

Note: Q_i stands for the accumulated oil production until the i^{th} month.

As shown in the above table, this article adopts the concept of “equal dimensional replenishment”, which refers to the dimension of input data. Except for the initial three stages as the initiation stage, all other stages use fixed four dimensional data input, always using the latest stage production data as the input of the model, and establishing a mapping relationship with the accumulated oil production in the next stage.

Model Training. Configure algorithm parameters and conduct model training.

- (1) Max_depth: The maximum depth of each tree. When establishing each tree, achieving the expected accuracy or maximum depth will proceed to the next tree model construction. The default value is 6.
- (2) Learning rate: learning rate is one of the most important hyperparameter. After each new tree model is established, the prediction results of the new model are given based on the previous prediction results and the interaction between the leaf output and the learning rate calculated this time. For different problems, the ideal learning rate will fluctuate between 0.05 and 0.3.
- (3) Booster model: There are two models to choose: gbtrees and gblines. Gbtrees uses a tree based model for lifting calculations, while gblines uses a linear model for lifting calculations. The default is gbtrees.
- (4) Gamma: The minimum “loss reduction” required for further splitting at leaf nodes, with a default of 0.

- (5) `Min_child_weight`: It can be understood as the minimum number of samples for leaf nodes, with a default of 1.
- (6) `Subsample`: The sampling ratio of the training set. Before fitting a tree, this sampling step will be performed, with a value range of (0, 1]. The default is 1.
- (7) `Colsample_bytree`: Before fitting a tree each time, determine how many features to use, with a value range of [0, 1], and the default value is 1.
- (8) `Reg_alpha`: Tuning of regularization parameters. The alpha parameter can reduce the complexity of the model, thereby improving its performance.
- (9) `Reg_lambda`: Tuning regularization parameters. Lambda parameters can reduce the complexity of the model and improve its performance.
- (10) `Random_State`: Random seed, 0 by default.

Model Evaluation. Based on parameters such as the error and root mean square error between the predicted and actual data of the model, model optimization is carried out to provide the optimal model for predicting single well oil production. The calculation method for model accuracy is:

- (1) Calculate the data of individual well oil production over time for each well sample in the test set;
- (2) Calculate the average absolute percentage error between all predicted data points and actual data points, which is the model prediction accuracy.

During the calculation process, the following error calculations were used [12–18]. Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE):

$$MAE = \frac{1}{n} \sum_{i=1}^n |y'_i - y_i| \quad (10)$$

$$MAPE = \frac{1}{n} \sum_{i=1}^n \frac{|y'_i - y_i|}{y_i} \quad (11)$$

Coefficient Determination (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - y'_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2} \quad (12)$$

4.6 Using Optimal Intelligent Models for Indicator Prediction

In response to the problem of predicting single well oil production, the optimized and trained XGBoost prediction model for single well oil production was used to carry out prediction work, obtaining future trends of single well oil production that can be used to guide actual production and conform to production laws. This plays a positive guiding role in production operation scheduling and adjustment of work systems.

5 Calculation Results and Analysis

5.1 Overview of FY Oilfield Work Area

The FY oil layer is the earliest discovered, most abundant, and widely distributed oil layer in the southern part of the SL Basin. The FY oil layer is distributed in the CL depression, HG terrace, and western region of the FX uplift zone in the central depression area. FY oilfield includes three blocks: Q block, R1 block, and R2 block.

5.2 Establishing a Multi Parameter Intelligent Prediction Model for Single Well Indicators

Model Construction. The XGBoost model for predicting single well oil production in tight oil fields was constructed using the XGBoost model introduced in the previous section. Establish models for different production stages.

Model Training. According to the basic content of the model training parameters mentioned in the previous section, parameter tuning tests were conducted with the accuracy of the test set as the evaluation label. There are a total of 84 wells in the sample set, with a ratio of 8:2 for training + validation sets, and testing set. This means that there are a total of 67 wells in the training + validation set, and 17 wells in the testing set. Compared through testing, `max_depth` is 15, `learning_rate` is 0.1, `boost_model` is `gbtree`, `gamma` is 0, `min_child_weight` is 1, `subsample` is 1, `colsample_bytree` is 1, `reg_alpha` is 0, `reg_lambda` is 1, `random_state` is 0.

Model Evaluation. The prediction model is constructed based on different sample types, and the final 12 to 48 months prediction model R^2 has an average accuracy of 86%, an average MAPE value of 13%, and an average MAE value of 351t.

5.3 Model Prediction Results and Analysis Discussion

Based on the optimal oil production prediction model in this article, relevant prediction work was carried out for 17 wells in three blocks of FY Oilfield. The comparison between the predicted results and actual production data of four wells is listed below, as shown in Figs. 3 and 4.

Figure 3 shows the predicted results of cumulative oil production at different production stages of wells WQ1 and WQ2. It can be seen that at the beginning of production, the predicted results are in good agreement with the actual production curve. During the production period of 20 to 48 months, the predicted values were slightly higher than the actual production data. The predicted value of WQ2 well in the mid-term production stage is slightly lower than the actual production data, and then the predicted value and production value continue to increase by the same magnitude.

Figure 4 shows the results of cumulative oil production predictions for WQ3 and WQ4 wells at different production stages. It can be seen that the predicted value of WQ3 well is generally lower than the actual production value, but the difference is relatively small. However, in the early stage of production, the predicted value of WQ4 well

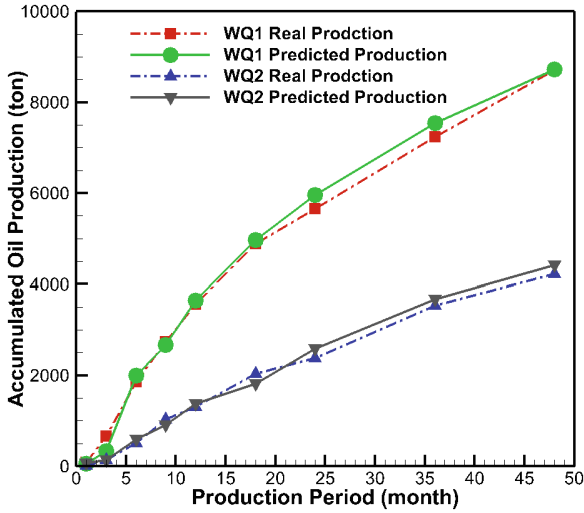


Fig. 3. Comparison between the predicted and real cumulative oil production with wells WQ1 and WQ2.

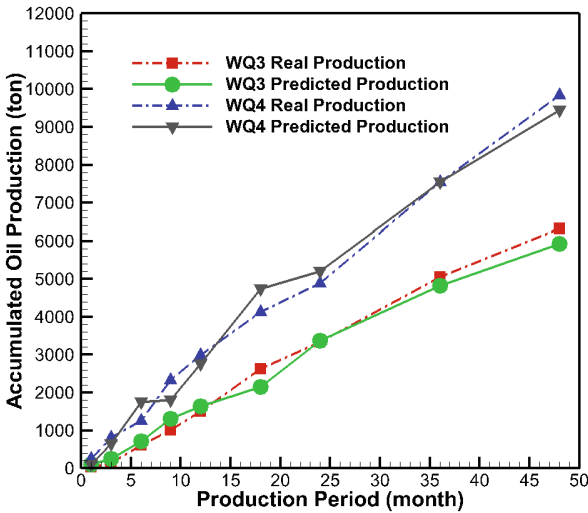


Fig. 4. Comparison between the predicted and real cumulative oil production with wells WQ3 and WQ4.

increases alternately with the actual production data, and remains basically consistent after 25 months of production.

From the comparison between the predicted results in Figs. 3 and 4 and the actual production curve trend, as well as the model error evaluation results, it can be seen that the prediction accuracy of the model in this paper is relatively high in predicting the cumulative oil production over 48 months. This indicates that the prediction model

established through a series of methods and techniques introduced in this article is more effective in predicting the cumulative production of a single well, thus achieving multi-dimensional tight oil single well production prediction, This provides a strong technical foundation for precise prediction and reasonable optimization of tight oil production in China.

6 Conclusion

Based on the XGBoost model, a typical tight oil well production sample library was constructed through data collection, organization, and preprocessing. Correlation analysis of influencing factors was conducted, and a multi-parameter intelligent prediction model for single well oil production indicators was established. The development indicators were predicted, and the conclusion is as follows:

- (1) Established a complete and effective method for predicting development indicators of tight oil fields based on XGBoost model;
- (2) The XGBoost cumulative oil production prediction model established is suitable for predicting the trend of cumulative oil production in tight oil fields, and the established model has a high accuracy in predicting single well production;
- (3) The methods and techniques introduced in this article are not only limited to tight oil fields, but can also be applied to the production prediction of unconventional oil and gas fields.

In summary, the artificial intelligence model established in this article has achieved multi-dimensional prediction of single well tight oil production, improved the dynamic management level of oil well production, improved the accuracy of single well measure decision-making, and improved the ultimate oil recovery rate and production efficiency of the oilfield. This provides a strong technical foundation for precise prediction of tight oil production and reasonable optimization of production allocation in China.

Acknowledgments. The work is supported by the following projects: ① Jilin Oilfield Branch project “Research on Capacity Evaluation and Parameter Optimization of Typical Tight Oil Blocks” (Number: JS2022-W-13-JZ-37-47); ② CNPC RIPED project “Research on integrated mathematical model of fracture network and software module development” (Number: YGJ2019-07-04). ③ CNPC major scientific and technological project of upstream key core technology “Volume fracturing optimization design software” sub-project “EDFM-based Component reservoir simulation software” (Number: 2020B-4118). ④ CNPC RIPED project “Joining 5 industrial alliance organizations such as the University of Calgary for exchange and cooperation research” sub-project “Joining the MMRD Consortium at Colorado School of Mines” (Number: 2021DQ0105-03). ⑤ CNPC Project “Research on Intelligent Interpretation Technology and Knowledge Graph of Oil and Gas Exploration” (Number: 2021DJ7003). And in the process of completing this paper, the author has received technical support and help from researchers, including Li Chen, Zhong-cheng Li, Hong-yu Zhao, Chang-chun Dong, Qi Liu, Yi-meng Wang of Jilin Oilfield Exploration and Development Research Institute, and experts, including Bin Bai, Xiu-Lin Hou, Shu-jian Liu of Research Institute of Petroleum Exploration & Development PetroChina, Feng-peng Lai, Lin-lin Zhang of China University of Geosciences, and the author would like to express the sincere gratitude.

References

1. Rahmanifard, H., Gates, I., Asl, A.S.: Comparison of machine learning and statistical predictive models for production time series forecasting in tight oil reservoirs. In: SPE/AAPG/SEG Unconventional Resources Technology Conference, Houston, Texas, USA (2022). <https://doi.org/10.15530/urtec-2022-3703284>
2. Zhang, Y., Zheng, Y., Sun, S., et al.: Data driven production prediction of tight sandstone after compression in Changqing Oilfield. *Energy Environ. Protect.* **43**(10), 96–101127 (2021)
3. Tao, L., et al.: A new productivity prediction hybrid model for multi fractured horizontal wells in tight oil reservoirs. In: SPE/IATMI Asia Pacific Oil&Gas Conference and Exhibition, Virtual (2021). <https://doi.org/10.2118/205620-MS>
4. Al Ali Hussain Al Ali, Z., Horne, R.: Meta learning using deep N-BEATS model for production forecasting with limited history. In: Gas&Oil Technology Showcase and Conference held in Dubai, UAE (2023)
5. Understanding LSTM Networks. https://web.stanford.edu/class/cs379c/archive/2018/Class_Messages_Listing/content/Important_Neural_Network_Technology_Tutorials/Olah/LSTM_Neural_Network_Tutorial-15.pdf
6. Wang, Y., Wang, C., Zhang, H., et al.: Automatic ship detection based on RetinaNet using multi resolution Gaofen-3 image. *Remote Sens.* **11**(5), 531 (2019)
7. Chen, L., Wang, Z., Wang, G.: Application of LSTM network in short-term power load forecasting under deep learning framework. *Power Inf. Commun. Technol.* **15**(5), 8–11 (2017)
8. Li, N., Gong, R., Liu, Z., Mi, L., Liu, L.: Application of artificial intelligence technology in single well production and water cut prediction. In: Lin, J. (ed.) IFEDC 2021. Springer Series in Geomechanics and Geoengineering, pp. 512–528. Springer, Singapore (2021). https://doi.org/10.1007/978-981-19-2149-0_47
9. Ma, Q., Guo, J., Li, N.: Load forecasting methods for urban gas pipeline networks. *J. Anshan Univ. Sci. Technol.* **27**(2), 101–105 (2004)
10. Li, N.: Research on load forecasting of urban gas pipeline networks. Master's thesis, Liaoning University of Science and Technology (2004)
11. Ojedapo, B., Ikiensikama, S., Wachikwu, V.U.: Elechi petroleum production forecasting using machine learning algorithms. In: SPE Nigeria Annual International Conference and Exhibition held in Lagos, Nigeria (2022). <https://doi.org/10.2118/212018-MS>
12. Gong, R., Li, X., Li, N., et al.: Artificial Intelligence for Oil and Gas, pp. 9–10. Petroleum Industry Press (2021)
13. Li, N., Gong, R., Li, X., Li, W., Wu, B., Ren, S.: Factor analysis of affecting the accuracy for intelligent picking of seismic first arrivals with deep learning model. In: Lin, J. (ed.) IFEDC 2022. Springer Series in Geomechanics and Geoengineering, pp. 7042–7062. Springer, Singapore (2023). https://doi.org/10.1007/978-981-99-1964-2_598
14. Li, N., Li, L., Wu, S., Wu, Y.: Numerical simulation of the effect of nanoconfinement on hydrocarbon phase behavior in nanometer scale pores. In: Lin, J. (ed.) IFEDC 2019. Springer Series in Geomechanics and Geoengineering, pp. 162–174. Springer, Singapore (2020). https://doi.org/10.1007/978-981-15-2485-1_18
15. Li, N., Ran, Q.Q., Li, J.F., Yuan, J.R., Wang, C., Wu, Y.S.: A multiple-continuum model for simulation of gas production from shale gas reservoirs. SPE165991 (2013)
16. Li, N., Yan, L.: Direct numerical simulation of a mixed-media model for efficient developing shale gas reservoirs. In: Lin, J. (ed.) IFEDC 2020. Springer Series in Springer Series in Geomechanics and Geoengineering, pp. 1993–2009. Springer, Singapore (2021). https://doi.org/10.1007/978-981-16-0761-5_189
17. Li, N., Yan, L., Li, L., et al.: Numerical simulation of triple media percolation mechanism of shale gas reservoir. In: 10th National Symposium on Efficient Development Technology of Natural Gas Reservoir, pp. 342–349 (2019)