scientific reports

Check for updates

OPEN

Research on shale TOC prediction method based on improved BP neural network

Chaorong Wu[1], Kaixing Huang[1]✉, Zhengtao Sun[2], Yizhen Li[3], Yong Li[1], Yuexiang Hao[3], Zhengxing Sun[2] & Ziqi Wang[1]

With the increasing attention to shale oil and gas in the field of oil and gas exploration and development, accurate prediction of TOC content has become the key to evaluating shale gas sweet spots. This paper studies a method for predicting shale TOC content using a BP neural network optimized by an improved cuckoo search algorithm. First, for the Longmaxi Formation shale, through logging sensitivity analysis, seven logging parameters sensitive to TOC content were determined: DEN, AC, RT, U, K, GR, and CNL. Using these parameters, a CSBP model was established and compared with the traditional BP neural network, multiple linear fitting method, and extended ΔlgR method. The results show that the CSBP model has higher prediction accuracy and generalization ability, with the mean absolute error and mean absolute percentage error being 0.38 and 15.00% respectively, which are significantly better than other methods. Further, the CSBP model was applied to predict the TOC content of Well W16 in the study area and verified by comparing with the measured TOC values. The correlation between the predicted and measured values is 0.89, and the change trends are consistent, confirming the applicability of the CSBP model. Finally, combined with the seismic waveform-guided simulation inversion technology, the planar and spatial distribution of TOC in the study area was predicted. The correlations between the predicted and measured values of four wells in the study area are all greater than 0.89. This method has high accuracy in the three-dimensional TOC content prediction of shale reservoirs and provides technical support for the evaluation of shale gas sweet spots in the work area.

In recent years, the escalating demands of resources and energy in the modern economy and society have underscored the significance of shale oil and gas as pivotal energy sources and potential substitutes for conventional oil and gas reserves. Consequently, shale oil and gas exploration has emerged as a focal point in the energy sector[1]. The total organic carbon (TOC) content serves as a critical geological parameter for assessing the viability of shale oil and gas reservoirs[2,3]. Hence, investigating the TOC content distribution holds paramount importance in evaluating shale gas reservoirs and devising strategies for horizontal well placement[4,5].

Currently, the TOC logging prediction methods mainly include geochemical experimental methods, conventional logging simulation methods, mathematical statistical methods, and empirical formula methods[6]. Among them: ① Although the geochemical experimental method can accurately obtain TOC values, the obtained data are discrete, making it difficult to characterize the spatial distribution of TOC, and the cost is high[7,8]; ② The conventional logging simulation method mainly uses the linear relationship fitted by TOC and logging curves such as acoustic time difference (AC), resistivity (RT), natural gamma (GR), and density (DEN) to predict TOC[9,10]; ③ The mathematical statistical method mainly uses mathematical statistical methods to analyze and process logging data, so as to establish a statistical model and then realize the prediction of TOC[11,12]; ④ The empirical formula method mainly realizes the prediction of TOC based on empirical formulas. For example, an empirical formula for predicting TOC is established based on the relationship between density and TOC[5]; the ΔlgR method[13] and the improved ΔlgR method[14] established based on acoustic time difference and resistivity are used to predict TOC.

In general, geochemical experimental methods, despite their high measurement accuracy, have difficulty in evaluating TOC across the entire work area. Conventional well-logging model methods are only applicable to TOC prediction at well locations. Mathematical statistical methods are suitable for strata with simple structures

[1]College of Geophysics, Chengdu University of Technology, Chengdu 610059, China. [2]Sinopec Key Laboratory of Geophysics, Sinopec Petroleum Geophysical Technology Research Institute Co., Ltd, Nanjing 211101, China. [3]CNPC Chuanqing Drilling Engineering Company Limited, Chengdu 610051, China. ✉email: 3167822060@qq.com

and shallow burial depths, and it is challenging to predict TOC in strata with complex structures and deep burial depths. Empirical formula methods are restricted by specific regions and conditions.

Seismic methods for predicting TOC: Predict TOC by establishing linear relationships between TOC content and parameters such as density, P-wave impedance, and Poisson's ratio[5,15,16]; Predict TOC using the pre-stack angle-divided seismic inversion method based on facies constraints[17].

However, the above seismic methods mainly utilize pre-stack data, which are difficult to collect. Moreover, the accuracy of pre-stack data is subject to issues such as the incident angle and the requirement that the vertical variation of elastic parameters should tend to be stable[18]. In post-stack seismic data processing, seismic waveform-guided simulation is based on the principle of facies constraint, making full use of well logging and seismic data to improve the accuracy of inversion[19,20]. This method can not only invert conventional parameters such as the wave impedance of the stratum, but also simulate various non-impedance parameters such as natural gamma and porosity, and has been widely used in thin reservoir prediction[21].

To address the issues such as the lack of inter-well sample data and thin reservoir thickness in shale reservoirs, this paper proposes a TOC quantitative prediction method based on the BP neural network optimized by the Cuckoo Search (CS) algorithm. First, combined with the actual geological conditions of the study area, a correlation analysis is conducted between the well-logging data and the core TOC data. Then, the well-logging method for TOC prediction is used to predict the TOC of the Longmaxi Formation shale in the study area, so as to determine the method with the optimal prediction effect. Finally, an exploration is carried out to realize the prediction of the spatial distribution of the TOC of the Longmaxi Formation shale in the study area by combining the optimal well-logging method with the seismic waveform-guided simulation method, which can provide assistance for the exploration and development of shale oil and gas in the Longmaxi Formation of the study area.

Results

(1) To address issues such as the lack of inter - well sample data and relatively thin reservoir thickness in shale reservoirs, this paper proposes a TOC quantitative prediction method based on a BP neural network optimized by the cuckoo search (CS) algorithm.
(2) Through sensitivity analysis of various logging parameters of the Longmaxi Formation shale, seven parameters, namely DEN, AC, RT, U, K, GR, and CNL, were determined. Then, high-resolution inversion of sensitive parameters in the study area was carried out in combination with waveform indicator simulation. Finally, with the inverted parameter data as input, three-dimensional high-precision quantitative prediction of shale TOC was achieved.
(3) This method is applicable to areas where there is a good response between well - logging parameters and TOC. Further research is still required for regions with complex and chaotic underground lithological and electrical characteristics.

Discussion

(1) In this study, the prediction accuracy of the CSBP model is higher than that of the BP model, the multiple linear fitting method, and the extended ΔlgR method. Specifically, the mean absolute error (MAE) and mean absolute percentage error (MAPE) of the CSBP model are 0.38 and 15.00% respectively, which are significantly lower than those of other methods. This is because traditional TOC prediction methods, such as the multiple linear fitting method and the extended ΔlgR method, rely on simplified geological models and limited datasets, which restrict their prediction accuracy and generalization ability. In contrast, the CSBP model utilizes more geological parameters and combines advanced optimization algorithms, thereby improving the accuracy and reliability of prediction.
(2) The CSBP model was used to predict Well W16 in the work area. The predicted TOC values showed a good correlation with the measured values, with a correlation coefficient of 0.89. Moreover, the MAE and MAPE were 0.47 and 37.57% respectively, indicating relatively small prediction errors.
(3) Although the CSBP model performed excellently in this study, there is still room for further improvement and expansion. First, the prediction ability of the model is limited by the quality and integrity of the input data. For example, when predicting the TOC values across the entire work area, although there is a good correlation between the predicted TOC values and the measured TOC values for Well W16, the predicted MAE and MAPE reached 0.67 and 106.3% respectively. Future research can improve the model's prediction ability by integrating more geological parameters and using more complex network structures. Second, the applicability of the CSBP model in different types of shale formations needs further verification.

Methodology

Sensitivity analysis of Toc logging in the study

The first sub-member of the first member of the Longmaxi Formation (S_1L_1) in the study area is close to the bottom of the Longmaxi Formation. It is currently the main target production interval for shale gas and also a high-quality shale interval. The measured minimum TOC value of the core is 0.08%, the maximum is 5.64%, and the average is 1.65% (Table 1), indicating that the TOC varies greatly vertically and the heterogeneity is strong.

In this study, various well-logging data of the first member of the Longmaxi Formation (S_1L_1) were collected, including acoustic travel time (AC), neutron porosity (CNL), compensated density (DEN), natural gamma ray (GR), potassium (K), photoelectric absorption cross-section index (PE), resistivity (RT), uranium (U), etc. In the following, the TOC logging sensitivity analysis of the S_1L_1 section will be carried out by combining the response mechanism and correlation analysis of logging parameters.

Parameter name	Notation	Unit	Minimum value	Maximum values	Average value
Acoustic time difference	AC	us·ft^{-1}	74.20	97.46	83.05
Neutron porosity	CNL	%	9.46	22.13	16.01
Compensation density	DEN	g cm^{-3}	2.42	2.75	2.62
Natural gamma	GR	API	91.14	211.40	135.05
Potassium	K	%	1.05	5.48	3.26
Resistivity	RT	Ω·m	5.83	31.84	14.08
Uranium	U	%	1.32	18.29	5.09
Measured TOC		%	0.08	5.64	1.65

Table 1. Basic data statistics of logging variables and core measured Toc.

Analysis of the response mechanism between TOC content and logging parameters

The high organic matter content in shale leads to an increase in porosity, a decrease in density, and a reduction in velocity[22,23]. Therefore, AC is directly proportional to the measured TOC content (Fig. 1A); DEN is inversely proportional to the measured TOC (Fig. 1C).

Neutron porosity tends to exhibit high values due to the hydrogen index. In high-TOC shales, organic matter may replace clay minerals with relatively high hydrogen content, resulting in a decrease in the overall hydrogen content of the formation. Consequently, the compensated neutron logging value decreases, showing a negative correlation. Additionally, high-TOC shales are often associated with pyrite (FeS_2). Its high density and neutron absorption characteristics may slightly suppress the neutron logging response. Therefore, the linear relationship between CNL and measured TOC is inverse (Fig. 1B).

When TOC is enriched, it can adsorb a large number of uranium ions. Source rocks rich in organic matter often have relatively high uranium contents and higher natural gamma values[24]. Therefore, GR and U are positively correlated with the measured TOC (Fig. 1D,G).

In a quiet and anoxic sedimentary environment of water bodies, it is conducive to the enrichment of organic carbon, but not conducive to the precipitation and preservation of potassium-rich minerals. Therefore, K is inversely proportional to the measured TOC (Fig. 1E).

The conductivity of the kerogen matrix is extremely low, far lower than that of formation water or the bound water in clay minerals. In high-TOC formations, organic matter replaces some conductive minerals (such as clay) or occupies pore spaces, which reduces the overall conductivity of the formation and leads to an increase in resistivity. Therefore, RT is directly proportional to the measured TOC (Fig. 1F).

Correlation analysis between TOC content and logging parameters

The correlation coefficient (r) is used to measure the sensitivity of TOC and logging parameters, which is calculated as Eq. (1). The coefficient of determination, R^2, is used to measure the degree of fit of the two types of data. The positive and negative values of r indicate the correlation between the parameters, with the size of the absolute value of r showing the degree of correlation. A higher absolute value of r signifies a stronger correlation between the parameters, while a larger value of R^2 indicates a better fitting effect.

As can be seen from Table 2, among the selected logging parameters, seven logging parameters have a good correlation with the measured TOC. The absolute values of the correlation coefficients in descending order are DEN, AC, RT, U, K, GR, and CNL, all of which are above 0.45. Among them, DEN has the largest absolute value of the correlation coefficient, which is 0.83, while K and GR have the smallest, which is 0.46. According to R^2, it can also be seen that DEN has the best fitting effect, while K has a relatively poor fitting effect.

$$\mathrm{r}(inx_q, d) = \frac{\sum_{i=1}^{n} (inx_{q,i} - \overline{inx_q})(d_i - \overline{d})}{\sqrt{\sum_{i=1}^{n} (inx_{q,i} - \overline{inx_q})^2 \sum_{i=1}^{n} (d_i - \overline{d})^2}} \quad (1)$$

where inx_q input logging parameters; d measured TOC; $inx_{q,i}$ ith value of input logging parameters; $\overline{inx_q}$ average value of the input logging parameter; d_i i th value of the measured TOC; $\overline{d}$ average value of the measured TOC.

Based on the above analysis, it is finally confirmed that seven logging parameters, namely DEN, AC, RT, U, K, GR, and CNL, are sensitive to TOC. These seven logging parameters are selected for subsequent research.

Analysis of prediction methods for shale TOC content

Conventional logging simulation method

After pre-processing the 195 sets of measured TOC data of cores collected in the work area and their corresponding DEN, AC, RT, U, K, GR, and CNL data, these 195 sets of data were divided into two categories: one was test data, and the other was training data. Among them, the test data accounted for 10% (20 sets of data), and the training data accounted for 90% (175 sets of data). The following formula of the multiple linear fitting model was obtained:

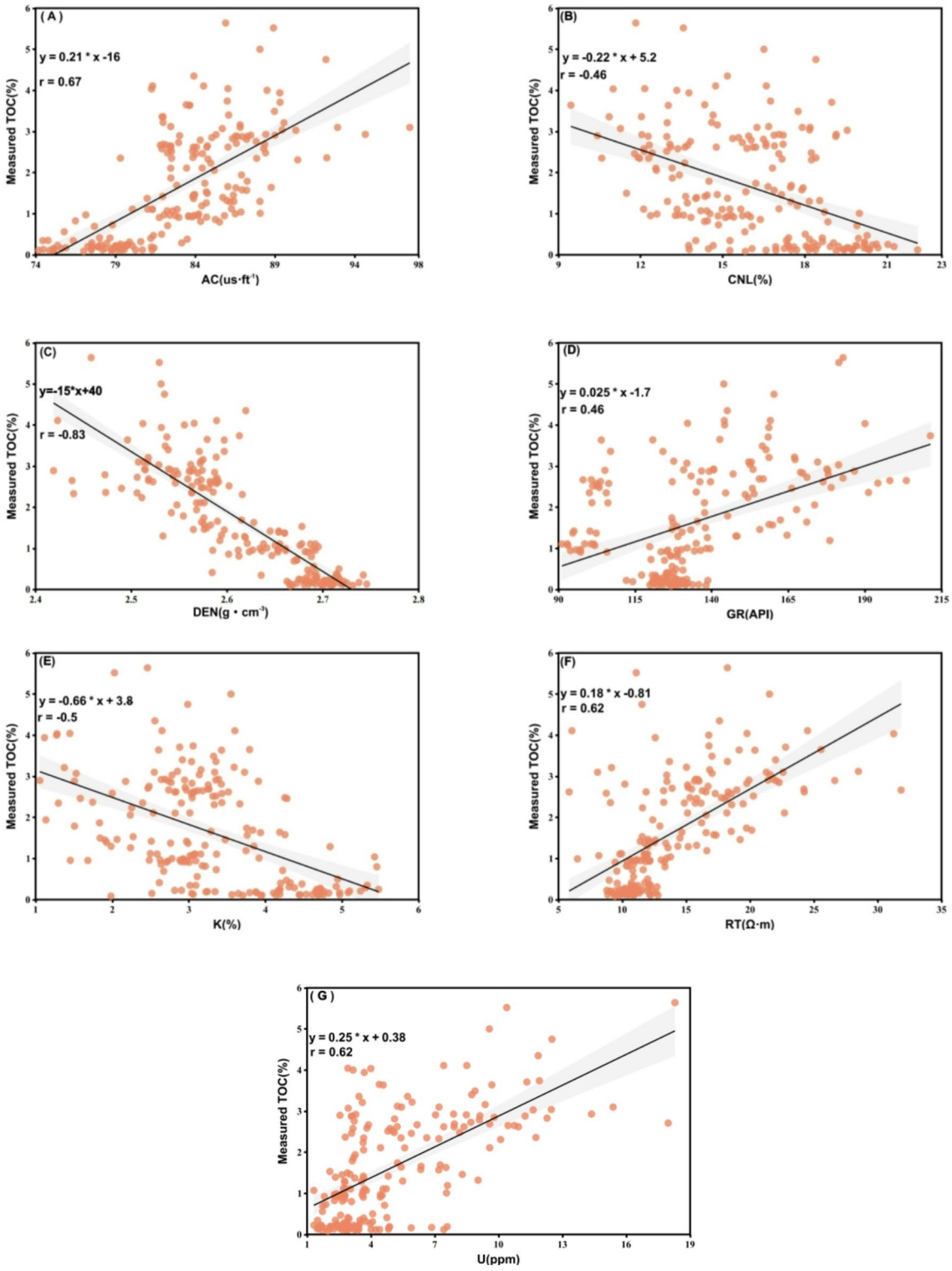

Fig. 1. Rendezvous of measured TOC with logging variables. (**A**) Rendezvous diagram of AC with TOC, (**B**) Rendezvous diagram of CNL with TOC, (**C**) Intersection plot of DEN with TOC, (**D**) Intersection plot of GR with TOC, (**E**) Rendezvous diagram of K with TOC, (**F**) Rendezvous diagram of RT with TOC, (**G**) Rendezvous diagram of U with TOC.

Typology	*r*	R^2
Measured TOC/AC	0.67	0.46
Measured TOC/CNL	− 0.46	0.21
Measured TOC/DEN	− 0.83	0.69
Measured TOC/GR	0.46	0.22
Measured TOC/K	− 0.50	0.25
Measured TOC/RT	0.62	0.39
Measured TOC/U	0.62	0.39

Table 2. Table of correlation coefficients between logging variables and measured Toc.

$$\begin{aligned} \mathrm{TOCDY} = {} & 0.0585 \times \mathrm{AC} - 0.153 \times \mathrm{CNL} - 6.2209 \times \mathrm{DEN} + 0.0032 \times \mathrm{GR} \\ & + 0.083 \times \mathrm{K} + 0.0528 \times \mathrm{RT} + 0.0978 \times \mathrm{U} + 13.5948 \end{aligned} \tag{2}$$

In the formula: TOCDY represents the TOC predicted by the multiple linear fitting method.

Empirical formula method

The traditional ΔlgR method uses the overlapping acoustic travel time (AC) and resistivity (RT) curves as the baseline, and the difference between the AC and RT curves is the ΔlgR value[25].

The formula for predicting TOC using the traditional ΔlgR method is as follows:

$$\Delta\mathrm{lgR} = \lg\left(\frac{\mathrm{RT}}{\mathrm{RT_b}}\right) + 00.02 \times (\mathrm{AC} - AC_{\mathrm{b}}) \tag{3}$$

$$TOC = (\Delta\mathrm{lgR}) \times 10^{2.297 - 0.1688 \times LOM} \tag{4}$$

where TOC is the TOC predicted by the conventional ΔlgR method; LOM is the cheese root maturity; AC_b is the baseline AC curve value; RT_b is the baseline AC curve value.

Since the traditional ΔlgR method requires maturity data and is based on medium to shallow, normally compacted strata, it is not suitable for the Longmaxi Formation shale in this work area, which lacks maturity parameters and is deeply buried.

According to previous studies, GR is less affected by compaction and is more sensitive to deep source rocks[26]. Therefore, GR can be considered to replace the maturity parameter. Based on this, the calculation formula of the extended ΔlgR method can be established:

$$\Delta\mathrm{lgRTOC} = (\mathrm{a} \times \mathrm{GR} + \mathrm{b})\ \mathrm{lgR} + \mathrm{c} \tag{5}$$

where ΔlgRTOC is the TOC predicted by expanding the ΔlgR method, and a, b, c are constants.

Expanded ΔlgR method model building again followed the same data division as in the multivariate linear fitting model building. The selected DEN, U, GR, AC, RT, and K logging parameters were used to build the expanded ΔlgR method model, and the following model equations were obtained:

$$\Delta\mathrm{lgRTOC} = (0.0149 \times \mathrm{GR} + 3.4239)\,\mathrm{lgR} - 4.4746 \tag{6}$$

BP neural network improved by CS algorithm

Currently, the emerging artificial intelligence methods are data-driven, and they search for the intrinsic mapping relationships among complex data through machine learning, which is highly suitable for handling such multi-dimensional mapping problems. Among numerous artificial intelligence methods, the BP neural network features a simple structure, strong anti-noise ability, and good generalization ability, making it well-suited for fitting various complex mappings. However, the BP neural network has the problem of easily converging to local minima instead of obtaining the global optimum[27]. To address this issue, the cuckoo search (CS) algorithm is employed to optimize the BP neural network.

BP neural networks: In 1986, a research team led by Rumelhart, McClelland et al. proposed a multi-layer feed-forward network model called the BP neural network, which learns through the back-propagation algorithm. The learning process of this network mainly consists of two stages: forward propagation and back-propagation. Structurally, the BP neural network is composed of an input layer, a hidden layer, and an output layer. In the forward propagation stage, the input data starts to be transmitted from the input layer. If there is a deviation between the actual output of the output layer and the expected output, the back-propagation process is initiated, and the connection weights of each layer are adjusted according to the error[28]. Thanks to this back-propagation ability, it is very easy to find the mapping relationship between the input and the output. The method to reduce the network output error is to adjust the weights and thresholds using the gradient descent method[29,30].

Figure 2 shows the structure of a BP neural network. Suppose the training data set $V = (X_1, X_2, ..., X_m)^T$ contains m parameters. That is, the training set contains m logging parameters, with each parameter containing q elements. In other words, $X_m = (x_1, x_2, ... , x_q)$.The number of nodes in the input layer is m, and the number of nodes in the hidden layer is h. The hidden layer input vector is $Z = (z_1, z_2, ... , z_h)^T$; the number of nodes in the output layer is p, then the output layer vector is $Y = (y_1, y_2, ... , y_q)^T$, i.e., the predicted value of BP neural network

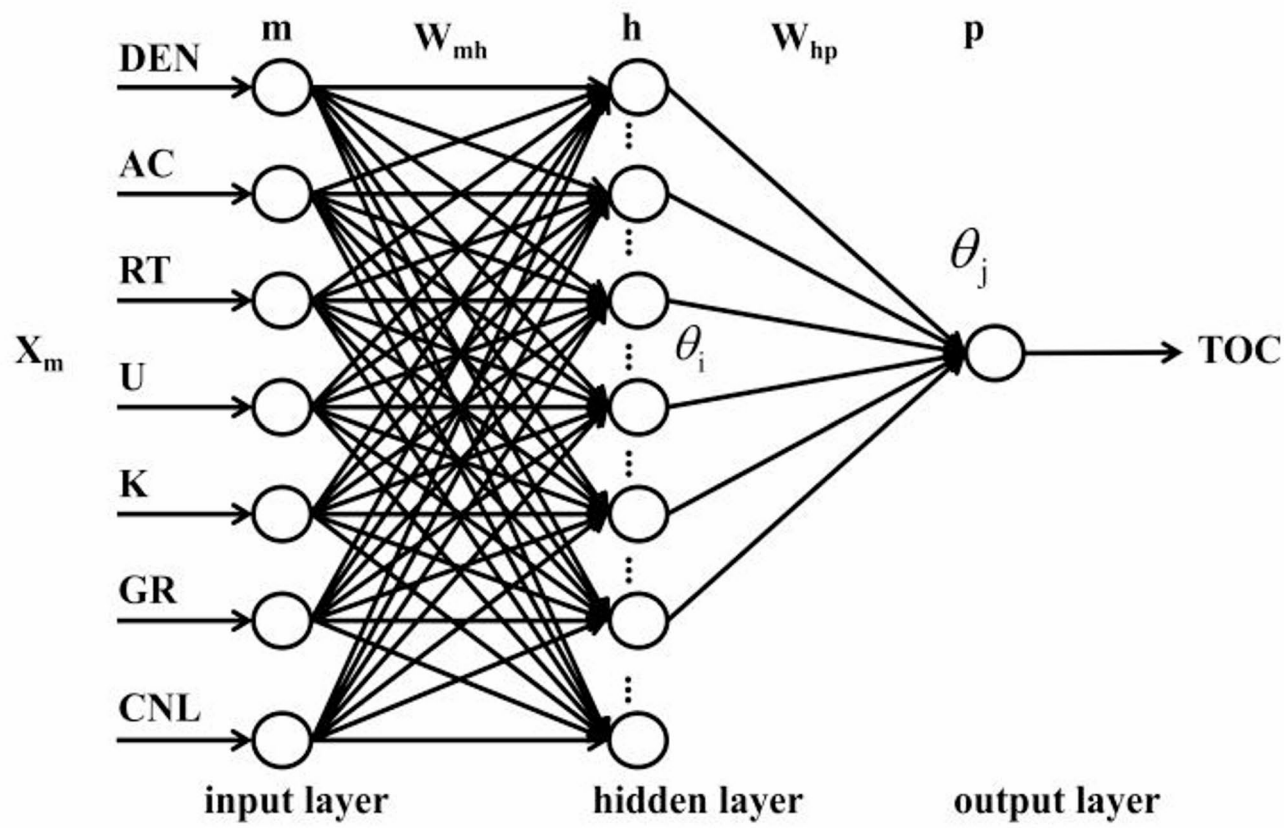

Fig. 2. Structural diagram of BP neural network.

about TOC TOC=$(y_1, y_2, \ldots, y_q)^T$; the desired output vector is D=$(d_1, d_2, \ldots, d_q)^T$, i.e., the measured TOC. The weight matrix between the input layer and the hidden layer is $W_{mh}=(w_{1h}, w_{2h}, \ldots, w_{mh})^T$, and the weights between the hidden layer to the output layer $W_{hp}=(w_{1p}, w_{2p}, \ldots, w_{hp})^T$; the threshold of the ith node of the hidden layer is θ_i, the threshold of the jth node of the output layer is θ_j, the activation function of the hidden layer is ξ(net), and the activation function of the output layer is φ(net), $\xi(\text{net}) = \phi(\text{net}) = \frac{1}{1+exp^{-net}}$.

The input layer signals input to the ith node of the hidden layer as:

$$net_i = \sum_{k=1}^{m} w_{ki} x_{ki} + \theta_i, i = 1, 2, \ldots, h \tag{7}$$

$$z_i = \xi(net_i) \tag{8}$$

where x_{ki} is the input layer composition matrix element.

The jth node of the hidden layer sends input to the output layer as follows:

$$net_j = \sum_{i=1}^{h} w_{ij} z_i + \theta_j, j = 1, 2, \ldots, p \tag{9}$$

$$y_q = \phi(net_j), j = 1, 2, \ldots, p \tag{10}$$

Equation (10) is the formula for calculating TOC using a BP neural network.

The error formula for sample q is as follows:

$$E = \frac{1}{2} \sum_{k=1}^{q} (d_q - y_q)^2 \tag{11}$$

The weights from the input layer to the hidden layer and from the hidden layer to the output layer are adjusted by the gradient descent method to reduce the total error. The formulas for adjusting the weights and thresholds between the input layer to the hidden layer are respectively (in the formula $\eta \in (0, 1)$ is the learning rate):

$$\Delta w_{mh} = -\eta \frac{\partial E}{\partial y_q} \frac{\partial y_q}{\partial z_i} \frac{\partial z_i}{\partial net_i} \frac{\partial net_i}{\partial w_{mh}} \tag{12}$$

$$\Delta \theta_i = -\eta \frac{\partial E}{\partial y_q} \frac{\partial y_q}{\partial z_i} \frac{\partial z_i}{\partial net_i} \frac{\partial net_i}{\partial \theta_i} \tag{13}$$

The weights and threshold adjustment formulas between the hidden layer and the output layer are as follows:

$$\Delta w_{hp} = -\eta \frac{\partial E}{\partial y_q} \frac{\partial y_q}{\partial net_j} \frac{\partial net_j}{\partial w_{hp}} \tag{14}$$

$$\Delta \theta_j = -\eta \frac{\partial E}{\partial y_q} \frac{\partial y_q}{\partial net_j} \frac{\partial net_j}{\partial \theta_j} \tag{15}$$

The final adjustment formulas for the weights and thresholds of the layers are obtained as:

$$\Delta w_{mh} = -\eta \sum_{k=1}^{m} \sum_{i=1}^{h} (d_q - y_q) \phi'(net_j) w_{hp} \xi'(net_i) x_{kq} \tag{16}$$

$$\Delta \theta_i = -\eta \sum_{k=1}^{m} \sum_{i=1}^{h} (d_q - y_q) \phi'(net_j) w_{hp} \xi'(net_i) \tag{17}$$

$$\Delta w_{hp} = -\eta \sum_{i=1}^{h} (d_q - y_q) \phi'(net_j) z_i \tag{18}$$

$$\Delta \theta_j = -\eta \sum_{i=1}^{h} (d_q - y_q) \phi'(net_j) \tag{19}$$

Cuckoo (CS) Algorithm: The cuckoo search algorithm is a novel heuristic optimization algorithm proposed based on the breeding characteristics of cuckoos and the flight mode of fruit flies. This method has significant advantages over the particle swarm optimization and genetic algorithms, mainly manifested in aspects such as stronger global search ability, faster convergence speed, fewer parameter requirements, and better generality and robustness[28]. Its concept originates from the Lévy flight behavior of birds and the parasitic behavior of cuckoos. The well-known scholar Yang from the University of Cambridge in the UK proposed the cuckoo search algorithm based on the following three assumptions: (1) Each cuckoo lays only one egg at a time and randomly selects a nest to lay the egg; (2) High-quality eggs in the best nests will be retained and hatch the next generation; (3) The probability that the host discovers a foreign egg is Pa. Once discovered, the host will directly abandon the egg or the nest. Under the above assumptions, the position update formula of the cuckoo search algorithm is proposed:

$$X_{ij}^{t+1} = X_{ij}^{t} + \alpha \times L(\lambda) \tag{20}$$

where X_{i+1j}^{t} and X_{ij}^{t} denote the i (i = 1,2 ,… ,m) nest at generation t and t + 1 in j (j = 1,2,… ,h)-dimensional positions; $L(\lambda)$ is the jump position of the Lévy (λ) flight random search jump path. α is the step size control quantity, this time set $\alpha = 0.01$.

Yang et al. simplified the $L(\lambda)$ distribution function was simplified and obtained by Fourier transform:

$$L(\lambda) \sim u = t^{-\lambda}, 1 < \lambda < 3 \tag{21}$$

To facilitate programming needs, in 1992 Yang used Mantegna's proposed simulation of the Lévy (λ) flight jump path formula:

$$s = \frac{\mu}{|\nu|^{\frac{1}{\lambda}}} \tag{22}$$

where $\lambda = 1.5$, μ and ν obeys the normal distribution:

$$\begin{cases} \mu \sim N(0, \sigma_\mu^2) \\ \nu \sim N(0, \sigma_\nu^2) \end{cases} \tag{23}$$

$$\begin{cases} \sigma_\mu = \left[\frac{\Gamma(1+\lambda) \sin(\frac{\pi\lambda}{2})}{\Gamma(\frac{1+\lambda}{2}) \lambda^{\frac{\lambda-1}{2}}} \right]^{\frac{1}{\lambda}} \\ \sigma_\nu = 1 \end{cases} \tag{24}$$

Thus Eq. (20) reduces to:

$$X_{ij}^{t+1} = X_{ij}^{t} + \alpha \times s \tag{25}$$

The CS algorithm enhances the performance of the BP neural network. The primary objective of integrating the CS algorithm with the BP neural network is to optimize the initial parameter settings of the gradient descent method, specifically the adjustment of the initial weights and bias values of the BP neural network. This integration is referred to as the CSBP neural network algorithm. The CSBP neural network algorithm is optimized at the end of the cuckoo after undergoing multiple iterations. The results obtained from the best search are then passed on to the BP neural network, resulting in a more accurate outcome with faster convergence and fewer errors.At the same time, it overcomes the dilemma of the traditional BP neural network that is trapped in local minima.

The flow of the whole algorithm is shown in Fig. 3. The basic idea is:

Firstly, the necessary pre-processing of the data (outliers removal, data normalisation) is carried out, and then the pre-processed data is used as input data to the CSBP neural network algorithm.

Randomly generate m bird nests with a given spatial extent. Each bird's nest represents a set of weights and thresholds of the neural network to be optimised, the parameters of the algorithm are set, and then the

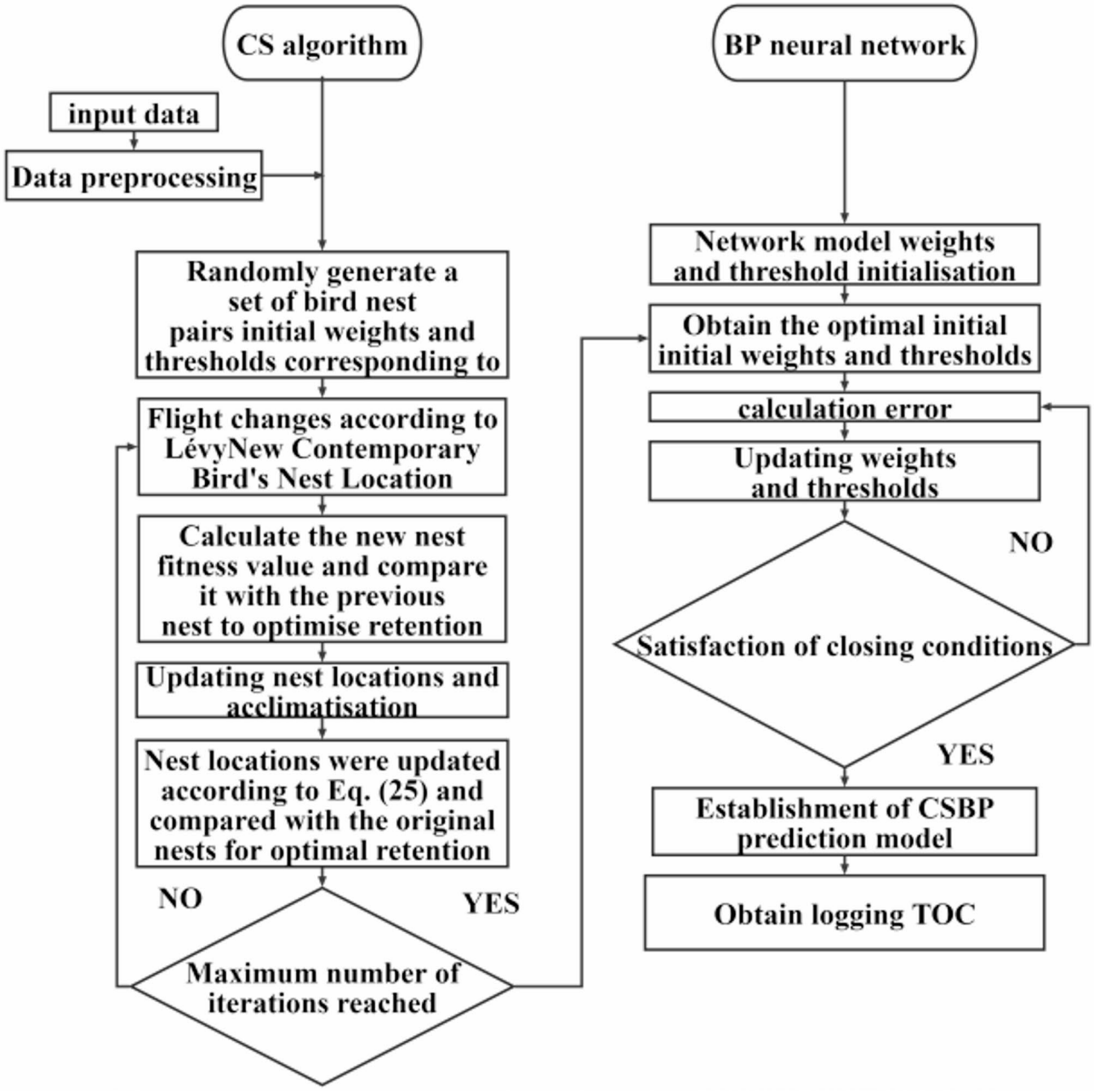

Fig. 3. Flowchart of CSBP neural network algorithm.

optimisation training is performed and calculated according to the fitness function to find the current optimal bird's nest location.

Retain the optimal bird nest position of the previous generation and update the position of n bird nests according to the Lévy flight pattern. Calculate the fitness value of the updated contemporary nests and compare it with the fitness value of the previous generation. If the updated position is better, then it is retained; otherwise, the position of the previous generation nests is kept.

To update the positions of all bird nests, generate random numbers k from the range (0,1) and set Pa = 0.25. If k is greater than Pa, the position of the original bird's nest is retained; if k is less than Pa, update the position of the bird's nest according to Eq. (25). At the same time, it is compared with the position of the original bird's nest; if it is better, it is retained. Otherwise, the original bird's nest is still used, and the updated m bird's nest positions are obtained.

Finally, the optimal bird's nest from many iterations is used as the optimal weights and thresholds for the BP neural network. The preprocessed data is inputted as the algorithm's input data, and the prediction results are obtained in the output layer.

Normalization of data and development of CSBP prediction model: Logging data scales are usually different. Therefore, it is necessary to normalize the input parameters before algorithm training. Normalize the logging parameters DEN, AC, RT, U, K, GR, and CNL selected in this study. The normalization formula is:

$$x_i^* = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{26}$$

where x_i^* ith normalised data, x_i ith original data, x_{max} maximum value in the data, x_{min} minimum value in the data.

Establishment of the CSBP prediction model: The relevant program for the CSBP was written using MATLAB software. Then, the selected logging parameters DEN, AC, RT, U, K, GR, and CNL were normalized and input into the program to train the CSBP neural network algorithm. During the algorithm training process, certain evaluation indicators are required to assess the reliability of the model and the accuracy of the prediction results. Among them, the mean absolute error (MAE) can avoid the problem of error cancellation, thus accurately reflecting the actual prediction error. The mean absolute percentage error (MAPE) reflects the relative size of the error, and a smaller value indicates higher prediction accuracy[31]. Therefore, in this paper, MAE and MAPE were used as the evaluation indicators for the CSBP prediction model, and their calculation formulas are as follows:

$$MAE = \frac{\sum_{i=1}^{N} |y_i - d_i|}{N} \tag{27}$$

$$MAPE = \frac{\sum_{i=1}^{N} \frac{|y_i - d_i|}{d_i}}{N} \times 100\% \tag{28}$$

where N is the number of samples.

Establishment of the CSBP prediction model: Similarly, using the data division method adopted in the establishment of the multiple linear fitting model, the selected logging parameters of DEN, AC, RT, U, K, GR, and CNL were used to train the CSBP prediction model.

The prediction results of different methods are shown in Fig. 4: The results of the BP model, CSBP model, and multiple linear fitting method are basically consistent with the measured TOC, and their changing trends are quite similar.

The MAE and MAPE of the CSBP prediction results are 0.38 and 15.00% respectively, with the smallest prediction error. Its correlation with the measured TOC is also the highest, reaching 0.85. The MAE and MAPE of the extended ΔlgR method prediction results are 0.62 and 24.83% respectively, with the largest prediction error. Its correlation with the measured TOC is also the poorest, at 0.48 (Table 3).

By comparing the results of TOC prediction using different methods, it can be known that the prediction effect follows the order: CSBP model > BP model > multiple linear fitting method > extended ΔlgR method. Therefore, it can be confirmed that the CSBP neural network is the optimal method, and the CSBP prediction model established by it will be used in the subsequent research.

Application

Practical application of the CSBP prediction model

Based on the previous research, the accuracy of the CSBP prediction model has been verified. To further validate the applicability of this method, it is used to predict the TOC of other wells. The TOC content of the reservoir section (S_1L_1) of Well W16 in the study area was predicted and compared with the TOC measured from the core. After pre - processing the seven logging variables (DEN, AC, RT, U, K, GR, and CNL) of Well W16, such as removing outliers and normalizing, they were input into the CSBP prediction model. Figure 5 shows the prediction results of the TOC content of Well W16, where TOCCSBP represents the predicted value by CSBP.

The prediction results of Well W16 show that: the predicted CSBP values of Well W16 have a good correlation with the measured TOC values, with a correlation coefficient of 0.89 (Fig. 5). The MAE and MAPE are 0.47 and 37.57% respectively. The variation trend of the predicted CSBP values is the same as that of the measured TOC

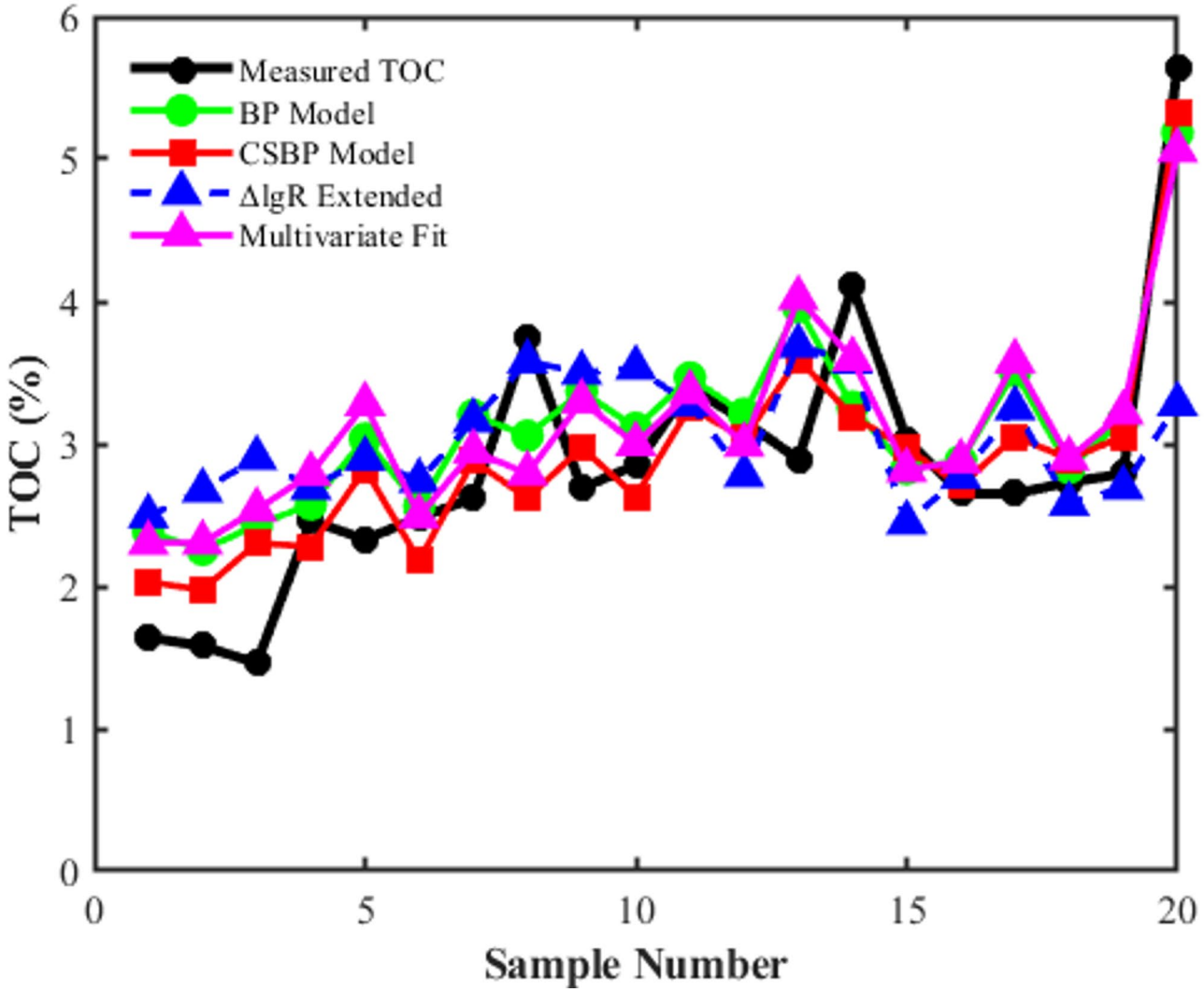

Fig. 4. Comparison of prediction results of different methods.

Forecasting methodology	*r*	MAE	MAPE
Multivariate Fit	0.79	0.51	21.07%
ΔlgR Extended	0.48	0.62	24.83%
BP Model	0.83	0.49	20.10%
CSBP Model	0.85	0.38	15.00%

Table 3. Statistical table of prediction errors of different methods.

values (Fig. 6), indicating that the overall prediction effect of the predicted CSBP values is good, and the CSBP prediction model has good applicability.

The CSBP prediction results show that the thickness of the S_1L_1 layer in Well W16 is 34.75 m, with an average TOC of 1.77%. The prediction results for each sub-layer are as follows: The thickness of the $S_1L_1^{2}$ layer is 19.5 m, with an average TOC of 0.19%; the thickness of the $S_1L_1^{14c}$ layer is 6 m, with an average TOC of 1.03%; the thickness of the $S_1L_1^{14b}$ layer is 17 m, with an average TOC of 2.62%; the thickness of the $S_1L_1^{14a}$ layer is 4.2 m, with an average TOC of 3.28%; the thickness of the $S_1L_1^{13}$ layer is 5.4 m, with an average TOC of 3.07%; the thickness of the $S_1L_1^{12}$ layer is 4 m, with an average TOC of 3.14%; the thickness of the $S_1L_1^{11}$ layer is 2.4 m, with an average TOC of 2.93%. According to the reservoir classification standard based on TOC content in the work area: shales with TOC content ≥ 3% are Class I high-quality shales; shales with 2% ≤ TOC content < 3% are Class II high-quality shales; shales with TOC content < 2% are Class III ordinary shales. It can be seen that the $S_1L_1^{2}$ and $S_1L_1^{14c}$ layers in Well W16 are Class III ordinary shales; the $S_1L_1^{14b}$ and $S_1L_1^{11}$ layers are Class II high-quality shales; the $S_1L_1^{14a}$, $S_1L_1^{13}$ and $S_1L_1^{12}$ layers are Class I high - quality shales.

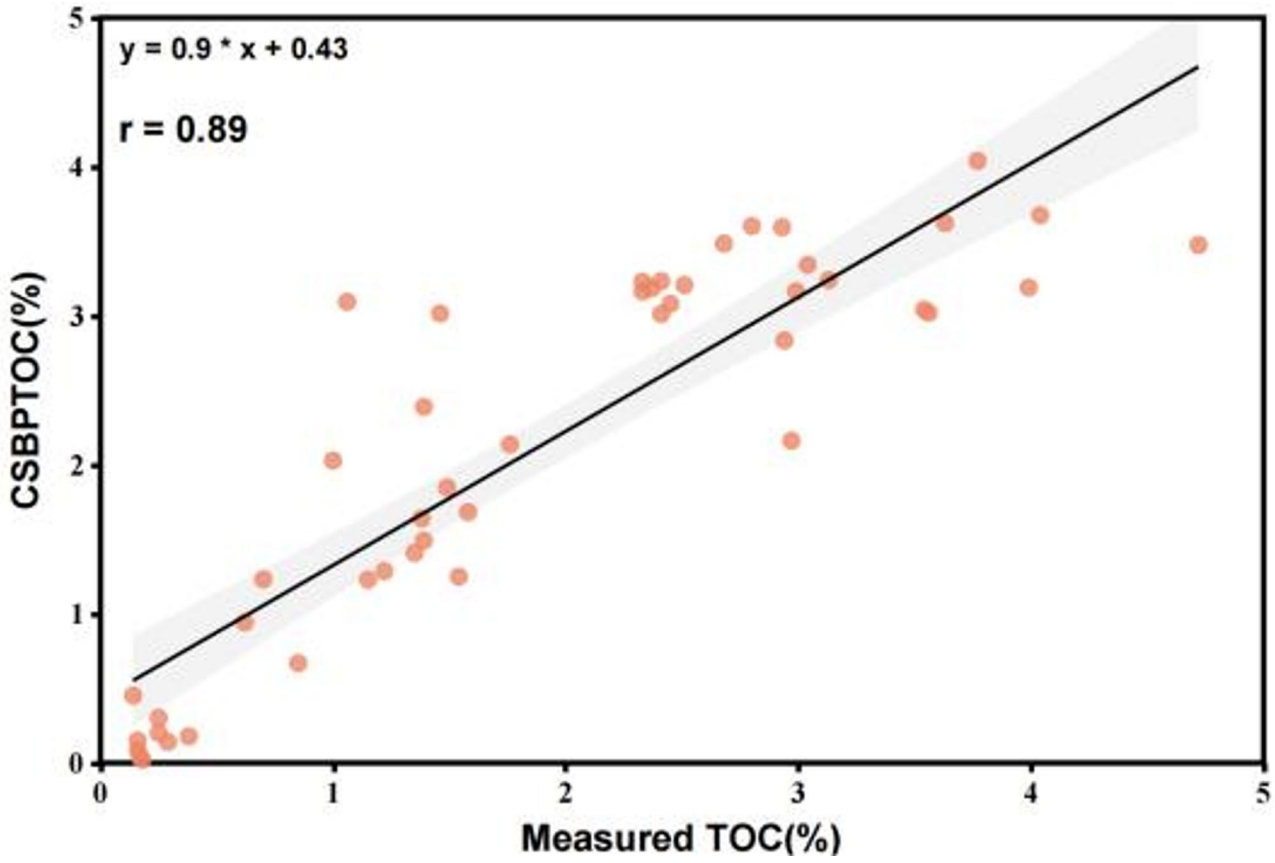

Fig. 5. Crossplot of predicted TOC values and measured TOC values in Well W16.

Application of CSBP prediction model and seismic waveform indicator simulation inversion

In the previous work, the CSBP prediction model was used to predict the TOC of a single well, yielding high-precision single-well prediction results. However, due to the limited drilling and logging data in the work area, it is impossible to obtain the TOC distribution characteristics of the entire work area. Therefore, in this study, an attempt was made to use the CSBP prediction model to predict the planar and spatial distribution of TOC in the study area based on the TOC-sensitive logging parameter volume obtained from high-resolution waveform indicator simulation inversion[19].

Basic principle of waveform-indicated simulation inversion: In the same sedimentary environment, seismic waveforms have similar characteristics. Similar-characteristic wells can be selected as effective samples for high-resolution well-seismic joint inversion[32]. Based on the above idea, a well-connected seismic section in the work area was selected for research (Fig. 7). There are two wells, W3 and W9, in the section. Then, the seismic waveforms of the well-side traces of the same layer in Wells W3 and W9 were extracted, and the seismic waveforms were compared by superposition (Fig. 8). The results show that the waveforms of the two are very similar, with a coincidence rate of over 95%. Therefore, the results of this study area show that the stratigraphic distribution areas with similar lithological combinations often have similar seismic waveform response characteristics, indicating that the seismic waveform-indicated simulation inversion is also applicable in this study area.

The seven logging parameters, namely DEN, AC, RT, U, K, GR, and CNL, confirmed in the previous section, were inverted using the seismic waveform-guided simulation method. As shown in Fig. 9, the inversion results of DEN, AC, RT, U, K, GR, and CNL along Well W9 match the logging curves. Furthermore, the inversion coincidence rates of each logging parameter were statistically analyzed. Wells W3 and W9 were used for the seismic waveform-guided simulation inversion of sensitive logging parameters, and Well W16 was used as a posterior well to verify the inversion results.

The results show that the inversion coincidence rates of all logging parameters are above 75.3% (Table 4). The average inversion coincidence rates of the seven logging parameters range from 84.99 to 88.96%. Among them, Well W3 has the highest average inversion coincidence rate, reaching 88.96%. For the four logging parameters, the average inversion coincidence rates range from 82.95 to 93.23%. Among them, the inversion coincidence rate of DEN is the highest, with an average value of 93.23%, indicating high inversion accuracy. Meanwhile, the average inversion coincidence rate of the logging parameters in the verification well W16 is 87.94%. The inversion coincidence rate of DEN is the highest at 95.60%, and the lowest is 78.90% for K. Therefore, in this study, the inversion of sensitive logging parameters has high accuracy, and the results of the verification well also prove the high reliability of the inversion.

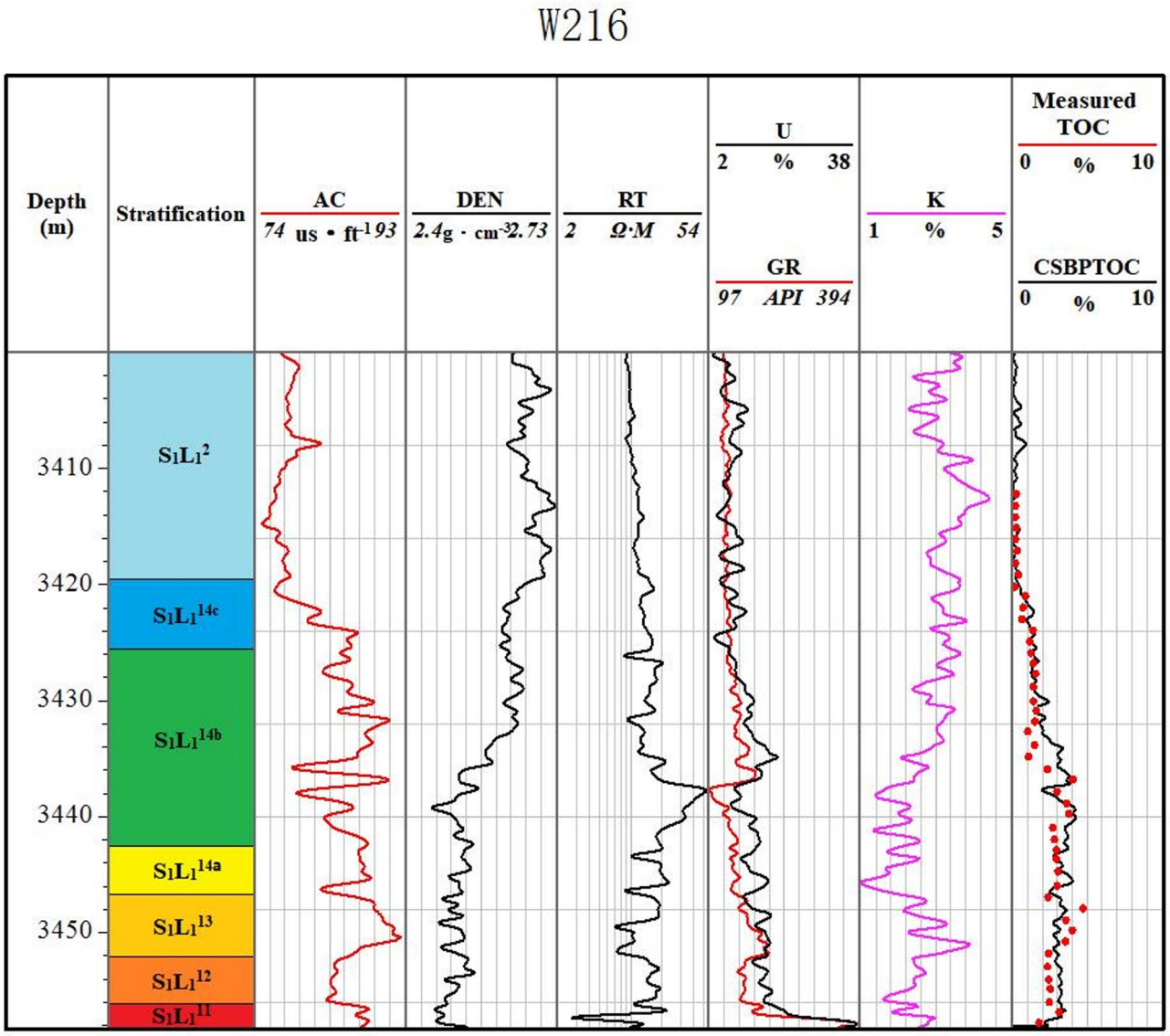

Fig. 6. Comparison between the predicted TOC values and the measured TOC values in Well W16.

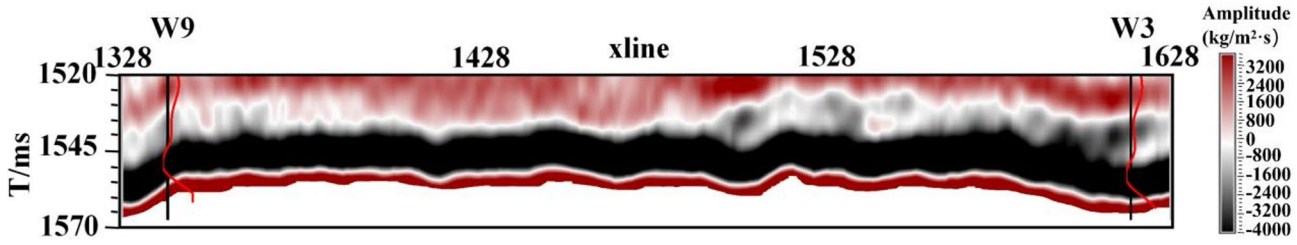

Fig. 7. Combined section of wells W3 and W9 in the study workings.

Therefore, the inverted data volumes of DEN, AC, RT, U, K, GR, and CNL were input into the CSBP prediction model to predict the planar and spatial distribution of TOC.

Figure 10 shows the slice map of TOC along the $S_1L_1^{12}$ layer. As can be seen from the figure, the overall TOC is greater than 2%. The areas where Wells W3, W9, and W16 are located have relatively high TOC values, while the area where Well W17 is located has a relatively low TOC value.

As can be seen from Table 5, there is a good correlation (above 0.89) between the predicted and measured TOC values of the four wells in the work area. Among them, Well W9 has the best correlation of 0.96, and its

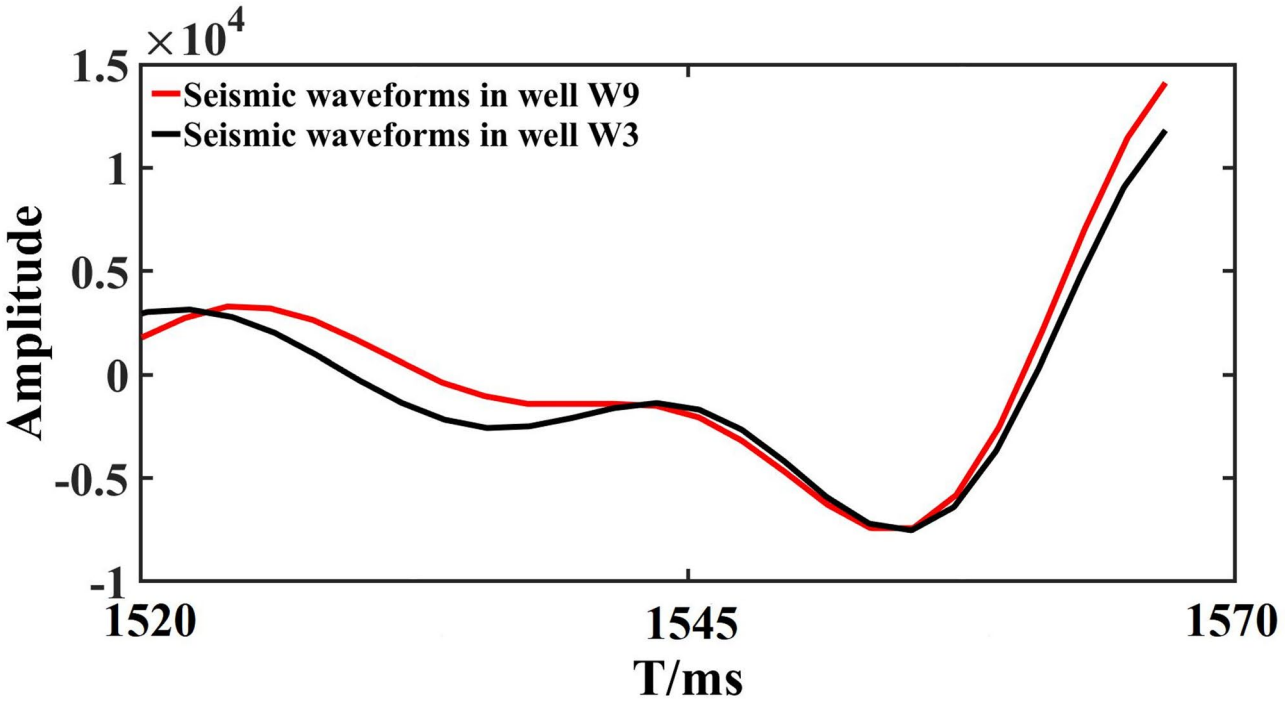

Fig. 8. Seismic waveform characteristics of wells W3 and W9 in the study work area.

prediction error is also smaller than those of the other three wells. Well W16 has the lowest correlation of 0.89 and the largest prediction error.

Figure 11 shows the comparison between the predicted and measured TOC values of Wells W9 and W16. As can be seen from the figure, the changing trends of the predicted and measured values for the two wells are consistent.

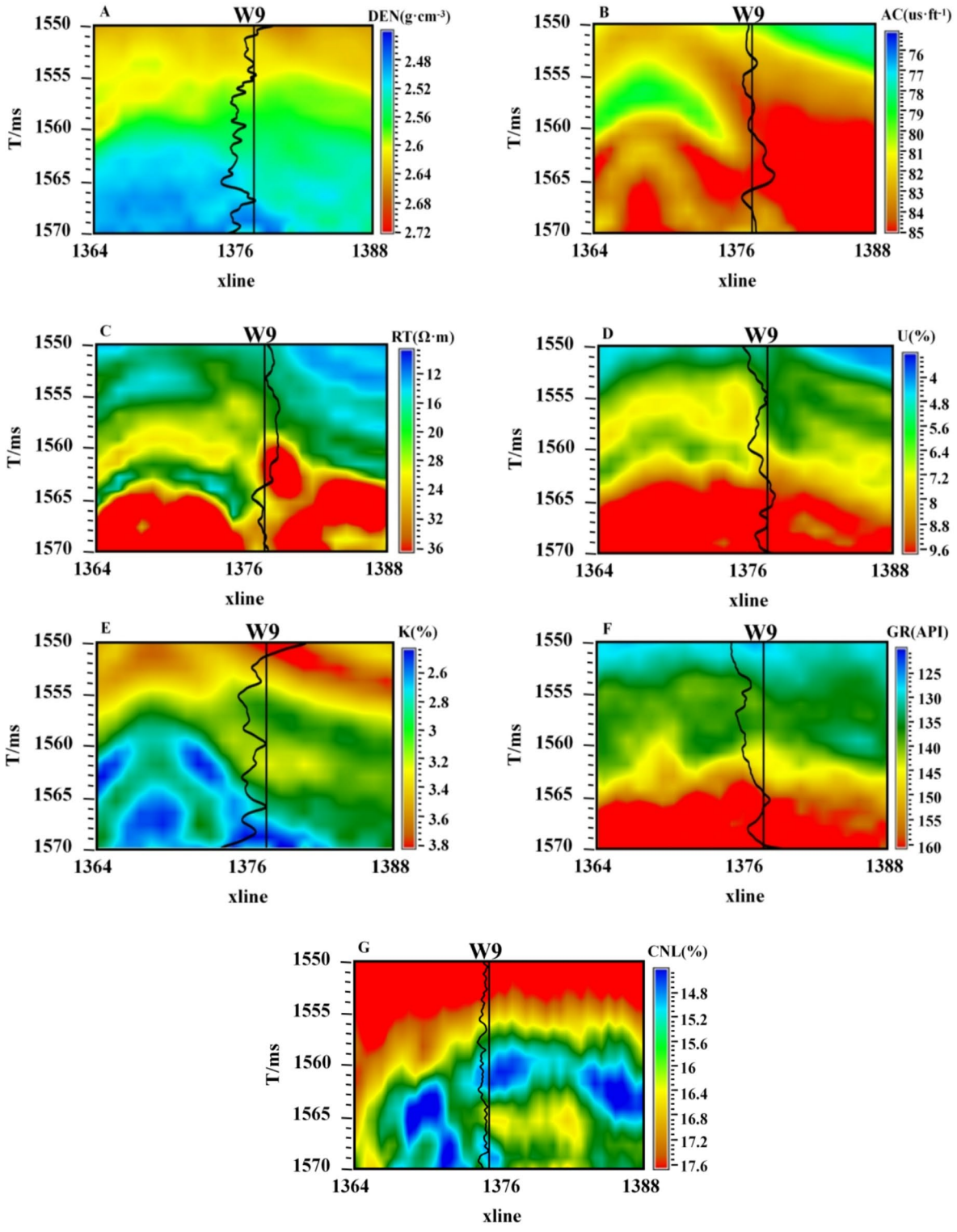

Fig. 9. Cross-sectional view of the well-logging sensitive parameter volume across Well W9 ((**A**) DEN section over W9 well, (**B**) AC section over W9 well, (**C**) RT profile over W9 well, (**D**) U profile over W9 well, (**E**) K profile over well W9, (**F**) GR profile over well W9,(**G**) CNL profile over well W9).

Logging	DEN	CNL	U	GR	AC	RT	K	Average of 7 logging parameters
W3	94.50%	84.90%	90.80%	79.60%	91.20%	88.50%	93.20%	88.96%
W9	94.70%	84.40%	85.40%	86.20%	89.40%	75.30%	88.40%	86.26%
W16	95.60%	89.80%	87.50%	89.10%	89.90%	84.80%	78.90%	87.94%
W17	88.10%	84.70%	82.90%	76.90%	85.70%	88.10%	88.50%	84.99%
4 Logging average	93.23%	85.95%	86.65%	82.95%	89.05%	84.18%	87.25%	87.04%

Table 4. Statistics of inversion compliance rate of logging parameters.

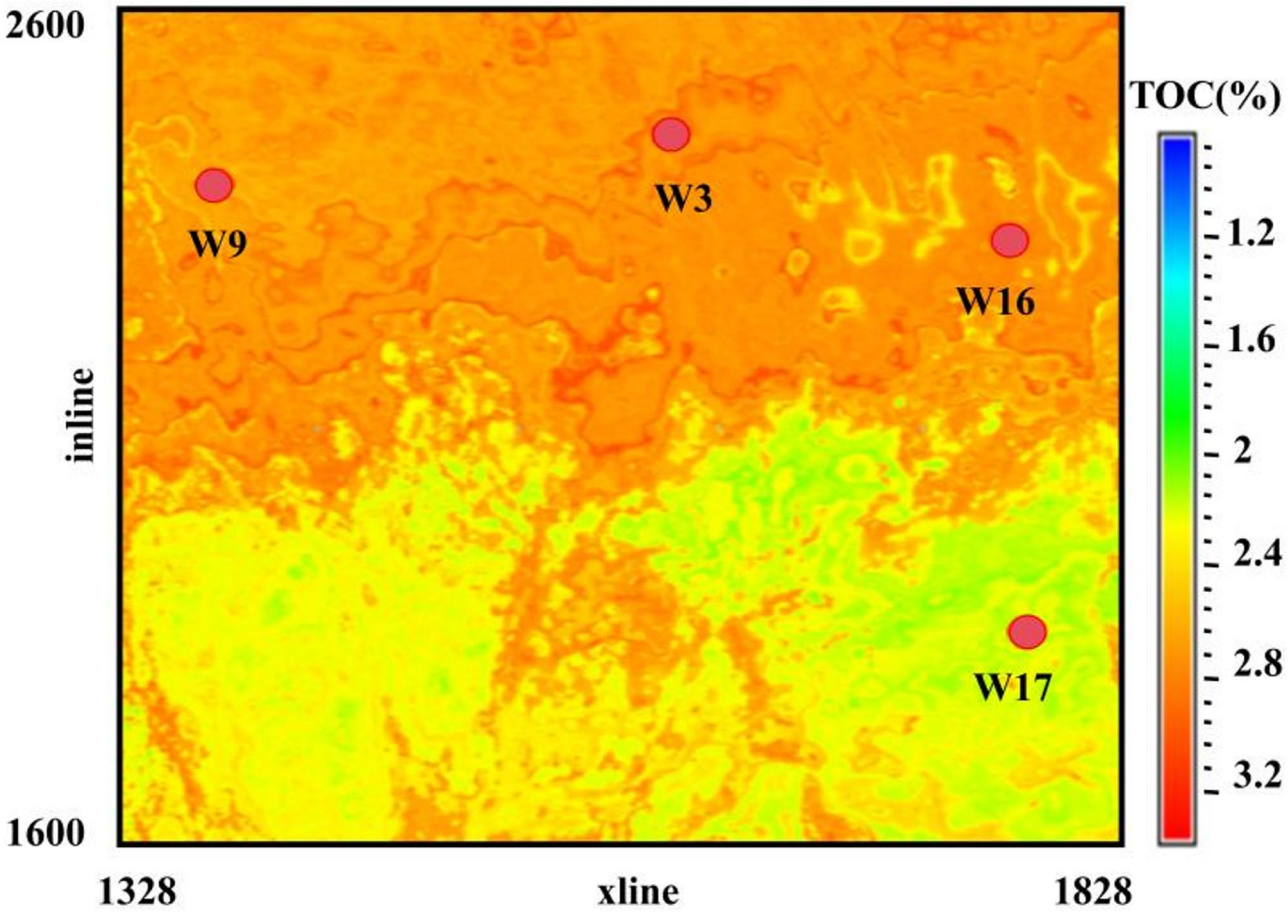

Fig. 10. TOC slice map along the S1L112 layer.

Logging	*r*	MAE	MAPE
W3	0.89	0.23	9.07%
W9	0.96	0.23	25.73%
W16	0.89	0.67	106.3%
W17	0.91	0.39	62.79%

Table 5. Statistical table of different logging prediction errors.

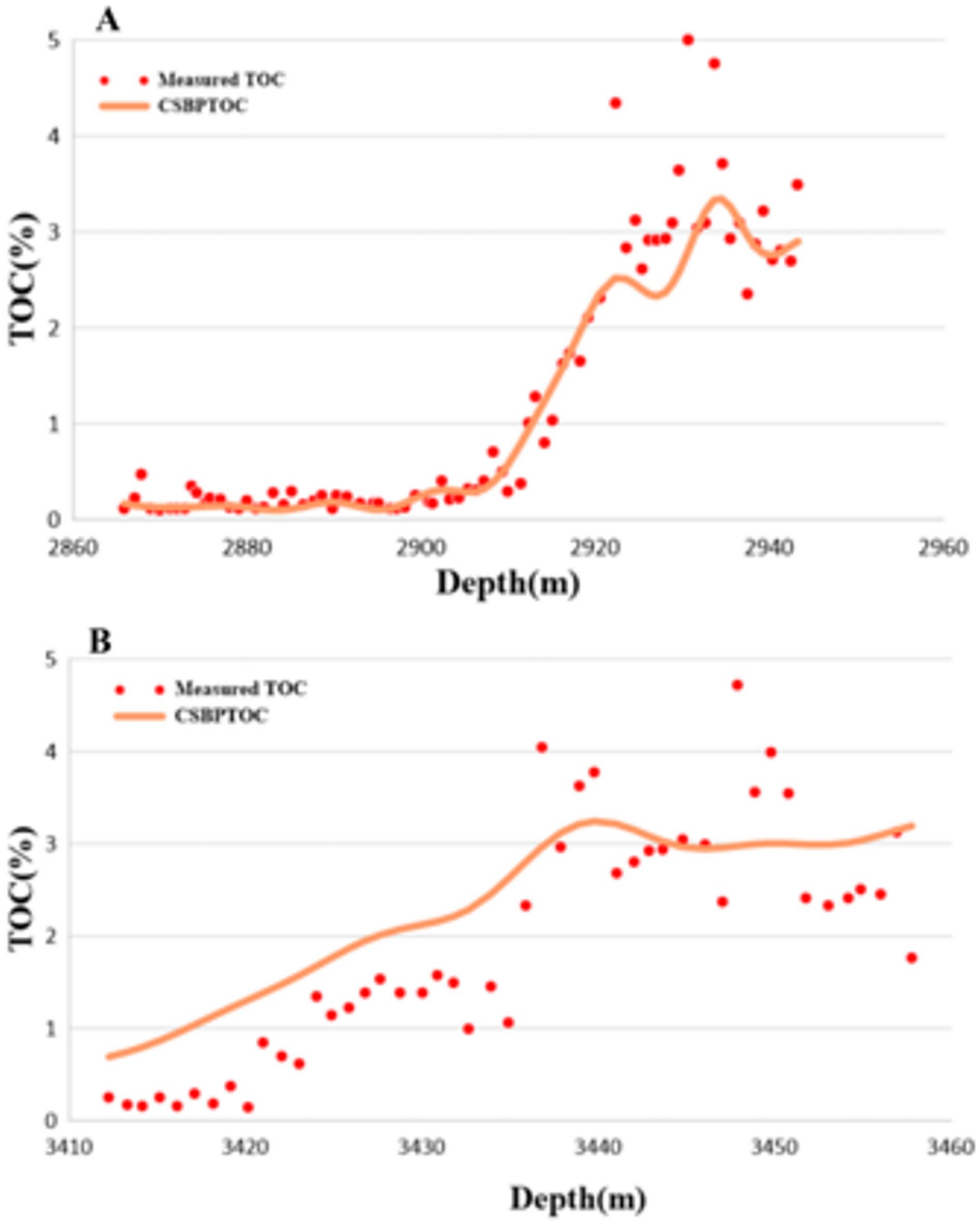

Fig. 11. Comparison between predicted and measured TOC values (**A** represents Well W9, **B** represents Well W16).

Data availability

The dataset used and analyzed during the current study period can be obtained from the corresponding author upon reasonable request.

Received: 19 December 2024; Accepted: 20 May 2025
Published online: 05 June 2025

References

1. Song, Y. J. et al. A method for evaluating the total organic carbon content of shale oil reservoirs based on ensemble learning using the stacking algorithm. *Logging Technol.* **02**, 163–178. https://doi.org/10.16489/j.issn.1004-1338.2024.02.004 (2024).
2. Chen, X. J. et al. Discussion on shale gas resource evaluation methods and key parameters. *Pet. Explor. Dev.* **5**, 566–571 (2012).
3. Du, H. et al. Research on logging interpretation and evaluation methods for dual sweet spots of shale oil in Dongying Sag. *Mineral. Explor.* **3**, 480–490. https://doi.org/10.20008/j.kckc.202303013 (2023).
4. Xu, J. et al. Geophysical prediction of organic carbon content in gas-bearing shale. *Petroleum Geophys. Explor.* **S1**, 64–68. https://doi.org/10.13810/j.cnki.issn.1000-7210.2013.s1.012 (2013).
5. Chen, Z. Q. Quantitative prediction technology of TOC seismic in marine shale and its application: A case study of Jiaoshiba area in Sichuan basin. *Nat. Gas. Ind.* **6**, 24–29 (2014).
6. Zhou, C. R. et al. The logging evaluation of organic carbon content based on ΔLogR-GR method: case study of the first member of Maokou formation in the southeastern Sichuan basin. *Nat. Gas Geosci.* **35** (3), 542–552 (2024).
7. Liu, H. et al. Source-to-sink System and Hydrocarbon Source Rock Prediction in low-exploration Areas of Faulted Lacustrine Basins: A Case Study of the Northern Subsag Area of the Zhu I Depression in the Pearl River Mouth Basin. *Oil Gas Geol.* **44**(3), 565–583 (2023).
8. Sun, J. T. *Study on TOC Prediction Method of Shale Based on Machine Learning* (Xi'an Shiyou University, 2023).
9. Chen, H. et al. Optimization and application of TOC logging prediction methods for source rocks in the Juyanhai depression of the Yin'e basin. *Progr. Geophys.* **34** (3), 1017–1024 (2019).
10. Jiang, D. X. et al. Discussion on the logging prediction model of total organic carbon content in source rocks: A case study of the Wenchang formation in the Lufeng Sag. *Lithol. Reserv.* **31** (6), 109–117 (2019).
11. Zhu L. et al. An improved method for evaluating the TOC content of a shale for mation using the dual-difference ΔlogR method. *Mar. Pet. Geol.*, **102**: 800–816. (2019).
12. He, Y. Prediction of the TOC content of shale based on support vector regression. *Petroleum Geophys.* **16** (3), 18–21 (2018).
13. Payyab, M. et al. TOC determination insource rocks using GR spectrometry and neuro-fuzzytechniques in aZagros basin oilfield. *Petroleum Sci. Technol.* **31**(12), 1268–1274 (2013).
14. Wang, X. et al. Prediction of total organic carbon content using a generalized ΔlogR method considering density factor: A case study of deep continental source rocks in the Southwestern Bozhong Sag. *Progr. Geophys.* **35** (4), 1471–1480 (2020).
15. Chen, S. et al. Identification of shale gas reservoir sweet spots using comprehensive geophysical prediction methods: A case study of the lower silurian longmaxi formation in the Changning block, Sichuan basin. *Nat. Gas. Ind.* **37** (5), 20–30 (2017).
16. Li, S. G. et al. Seismic prediction technology for double sweet spot parameters of deep shale gas. *Geol. Explor.* **39** (1), 113–116 (2019).
17. Yang, G. et al. Research on seismic inversion method for total organic carbon content of shale based on phase constraint. *Progr. Geophys.* **39** (3), 1038–1047 (2024).
18. Bi, C. C. et al. A nonlinear direct inversion method for brittleness index based on BI_Zoeppritz equation. In *China Geoscience Union Annual Meeting*, 72–73 (2018).
19. Zeng, H. L. Seismogenic sedimentology in China: review and outlook. *Acta Sedimentol. Sin.* **3**, 417–426. https://doi.org/10.14027/j.cnki.cjxb.2011.03.013 (2011).
20. Sheng, S. C. et al. Research on the seismic waveform-indicated stochastic modeling inversion (SMI) method. *Inner Mongolia Petrochem. Ind.* **41** (21), 147–151 (2015).
21. Du, J. et al. Tight sandstone reservoir prediction based on waveform indication simulation. *Bull. Geol. Sci. Technol.* **41** (05), 94–100 (2022).
22. Liu, C. et al. Research on the prediction method of organic carbon in the shale reservoir of the early paleozoic Qiongzhusi formation in the Weiyuan area. *Comput. Tech. Geophys. Geochem. Explor.* **43** (6), 705–714 (2021).
23. Zhao, L. Q. Logging evaluation of total organic carbon content in source rocks in the Southern Jiyang depression and its application in tight oil reservoir formation. *Daqing Petroleum Geol. Dev.* 1–9. https://doi.org/10.19597/J.ISSN.1000-3754.202308046 (2023).
24. Wei, M. Q. et al. Prediction model of total organic carbon content in shale gas based on machine learning. *Sci. Technol. Eng.* **23** (30), 12917–12925 (2023).
25. Passer, Q. R. et al. A practicalmodel for organic richness from porosity and resistivitylogs. *AAPG Bull.* **74**(12), 1777–1794 (1990).
26. Hu, H. T. et al. Method and application of generalized ΔLgR technique for predicting organic carbon content of continental deep source rocks. *Nat. Gas Geosci.* **27** (1), 149–155 (2016).
27. Zhou, Z. A review of the current development of BP neural networks. *Shanxi Electron. Technol.* **1** (2), 90–92 (2008).
28. Li, F. et al. Research on controlling environmental parameters based on improved cuckoo optimization BP neural network. *Comput. Digit. Eng.* **49** (8), 1505–1524 (2021).
29. Sun, C. et al. Stock price prediction based on BP neural network model optimized by cuckoo search algorithm. *Comput. Appl. Softw.* **33** (2), 276–279 (2016).
30. Chen, E. K. et al. Fault diagnosis of mining transformers based on cuckoo algorithm and BP neural network. *Coal Technol.* **6**, 223–224. https://doi.org/10.13301/j.cnki.ct.2018.06.084 (2018).
31. Ye, Y. et al. A new method to predict brittleness index for shale gas reservoirs: insights from well logging data. *J. Petrol. Sci. Eng.* **2** (08), 1–14 (2022).
32. Chen, Y. H. et al. Seismic waveform-guided inversion method and its application. *Pet. Explor. Dev.* **47** (6), 1149–1158 (2020).

Author contributions

H.K.X. completed the manuscript writing, W.C.R. provided the methods and data, and S.Z.T., L.Y.Z., H.Y.X., S.Z.X, and W.Z.Q. all reviewed and approved the final manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to K.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.