



The back propagation based on the modified group method of data-handling network for oilfield production forecasting

Jia Guo¹ · Hongmei Wang¹ · Fajun Guo¹ · Wei Huang¹ · Huipeng Yang¹ · Kai Yang¹ · Hong Xie²

Received: 7 August 2018 / Accepted: 2 November 2018 / Published online: 19 November 2018
© The Author(s) 2018

Abstract

In this paper, a novel hybrid forecasting model combining modified group method of data handling (GMDH) and back propagation (BP) is introduced for time series oilfield production forecasting. The proposed model takes advantages of both the modified GMDH networks in effective parameter selection and the BP network in excellent nonlinear mapping and provides a robust simulation ability for oilfield production with higher precision. Various production parameters of an actual oilfield were utilized to analyze and test the annual output predicted by proposed model (modified GMDH-BP). The performance of the proposed model was compared with the multiple linear regression (MLR), GMDH, modified GMDH, BP, and the hybrid model combining group method of data handling and back propagation (GMDH-BP) using time series annual production data. The relative error, correlation coefficient (R), root mean square error, mean absolute percentage of error, and scatter index were utilized to investigate the performance of the presented models. The evaluation results indicate that the hybrid model provides more accurate production forecasts compared to other models and exhibits a robust simulation ability for capturing the nonlinear relation of complex production time series prediction of oilfield.

Keywords Oilfield production · Modified GMDH · BP · Variable selection · Forecasting

Introduction

The production of oil field is a fundamental indicator that providing decision-making basis for oil production investment and adjustment disposition. The productivity of oilfield is a complex nonlinear system with comprehensive characterization gathering geological cognition, development policies, and production performance, which is becoming an important topic and generated abundant research achievements (Lia et al. 1997; Bardi 2005; Wang et al. 2015). The forecast methods of oilfield production can be classified into two categories: the traditional approach mainly based on knowledge driven and the artificial intelligence algorithm presented by the data-driven models (Samsudin et al. 2011). The traditional methods include experience formula (Tang et al. 2010), physical simulation (Liu and McVay DA 2009), and curve fitting (Arps

1945; Hubbert 1980; Wang and Feng 2016) and have been well developed in the past decades. This method is widely used for the advantage of simple principle and handy calculation, but has a larger prediction error when dealing with complex nonlinear system.

Recent years have witnessed the vigorous development of the artificial neural network (ANN) and it is widely applied in engineering, computer science, and information science (Haykin 1998). Although lacking of physical interpretation and insight of production rules, but it can provide sufficiently accurate and reliable results. In the field of oil exploitation, the ANN models have recently accepted as an effective tool to predict reservoir properties such as permeability (Ahmadi and Shadzadeh 2012; Ahmadi et al. 2013; Ahmadi 2012; Ahmadi and Goudarzi 2013), minimum miscible pressure (Ahmadi 2012), asphaltene precipitation (Ahmadi and Golshadi 2012), condensate-to-gas ratio (Zendehboudi et al. 2012), and forecast oil flow (Jr et al. 2009; Mollaiy Berneti and Shahbazian 2013). Especially, for the back propagation (BP) neural network, which is one of the most popular algorithm in ANN, has been proved with excellent advantages in the aspects of reservoir dynamic performance from single pattern recognition (Balch et al. 1999; Tapias et al. 2001) to

✉ Jia Guo
yjy_gj@petrochina.com.cn

¹ Exploration and Development Institute, PetroChina Huabei Oilfield Company, Renqiu, China

² CNPC Bohai Drilling Engineering Company Second Logging Company, Renqiu, China

multi-factor forecasting (Yi-Bao et al. 2005; Yu et al. 2008). For complex nonlinear systems, the major advantage of BP is strong adaptive, power fault tolerance, and high fitting accuracy. However, this method is sensitive to the topological construction and different types and quantities of input factors may lead to different results (Yu et al. 2008).

It is very difficult to describe the nonlinear system, since a priori knowledge is needed for system identification (Kashyap 1973). However, the group method of data-handling (GMDH) algorithm has a automatically selection ability and higher forecast ability without any prior knowledge (Mori and Tsuzuki 1990). The GMDH algorithm was first developed by Ivakhnenko as a tool for identifying the relationship between the input layer and output layer in nonlinear system. In the past decades, the GMDH model has been successfully applied in a widely fields such as engineering (Najafzadeh and Barani 2011), education (Abdel-Aal and El-Alfy 2009), medicine (Abdel-Aal 2005), and economy (Parks et al. 1975). In petroleum engineering, the GMDH model has been paid little attention and only a few applications to predict oil price (Mohsen et al. 2010) and determine physical properties (Al-Ajmi et al. 2012; Ghorbani et al. 2014) have been carried out. However, conventional GMDH has shortcomings of difficulty in determining the best partition of data sets and elimination of effective parameters untimely, which induce the model precision exists subtle differences.

In this paper, the GMDH algorithm is improved by randomly drawing method and original variable preservation method to improve the selection performance. Then, the

optimized GMDH model is combined with the Back Propagation (BP) neural network, and the effective parameters selected by the modified GMDH algorithm are used as the input neurons of the BP network. Excellent mapping modeling ability of complex systems of BP network makes it available to accurately predict oil production.

Definition of parameters for reservoir production forecasting

The production of oilfield is one of the most direct indicators of the oilfield development which is a complicated system engineering with various development indexes. This paper makes a qualitative analysis of the factors influencing oil production and initially determined the representative factors based on reservoir engineering principle. In theory, the production of oil fields will pass through rising, stable, and declining periods successively during development process, as shown in Fig. 1a. For a particular oilfield, the actual oil production in recent years generally followed this rule, in which stable production period and decline period contributed most of the oil production (Fig. 1b). It is well known that the factors influencing oil reservoir production can be divided into geological factors and development factors. For a reservoir with certain reserves, the dynamic oil production is closely related to the development of technical policies. Specifically, the development technology policy involves exploitation method, well type, well density, injection

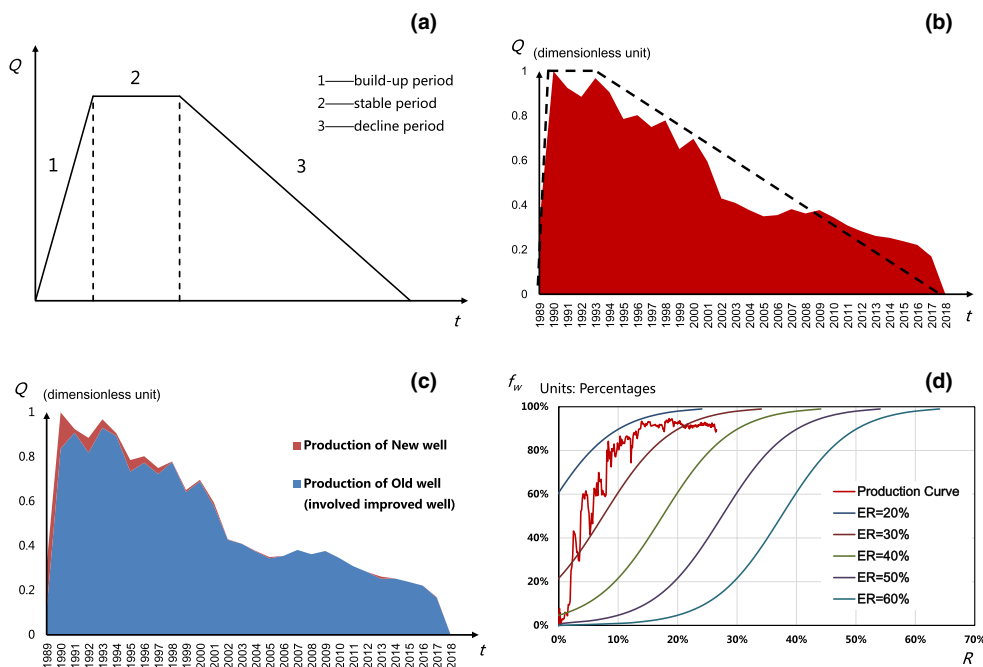


Fig. 1 Conventional development modes of oilfield

Figure a: Concept graphs of oil field development stage

Figure b: Past production for a certain oilfield

Figure c: Production composing for both new wells and old wells

Figure d: The relation curve of water cut and recovery percent

strength, deployment scale for new wells, and other factors, which has a different impact on the oil production.

The dynamic change of oil production is mainly affected by changes in liquid production, water injection rate, and reservoir water cut, which will have a lasting influence upon the whole life of petroleum recovery. Especially, in the later period of development, the amount of liquid production is basically maintained in a stable state and the speed of oil production decline keeps rising in the same time. Thus, the change of oil production is only affected by the change of water cut in this period. The production composition will change dramatically in the medium-later stage of oilfield development, which is mainly manifests in two aspects: one is that the proportion of new wells in oil production will decreased significantly with the speed of oil and water wells growing slower, as shown in Fig. 1c; the other is that is the principal strategies of oilfield development shifting from controlling water cut rising too fast to seek optimizing recovery percent of geological reserves and oil recovery rate, as shown in Fig. 1d. Based on the analysis above, we have initially take these influence factor into consideration: the total amount of oil and water wells (x_1), the number of active wells (x_2), the number of new wells (x_3), the injection rate of last year (x_4), water cut (x_5), the oil production rate (x_6), recovery percent of reserves (x_7), and oil production of last year (x_8).

Methodology

Modified group method of data handling

The algorithm of group method of data handling (GMDH) is a feed forward neural network for modeling and identification of complex systems, which was proposed by Ivakhnenko in the 1960s (Ivakhnenko 1971). General form of the network can be expressed by a complicated polynomial series in the form of the Volterra series, known as the Kolmogorov–Gabor polynomial:

$$\bar{y} = a_0 + \sum_{i=1}^n a_i x_i + \sum_{i=1}^n \sum_{j=1}^n a_{ij} x_i x_j + \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n a_{ijk} x_i x_j x_k + \dots \quad (1)$$

where x represents the input of the system, n is the number of the inputs, and a is coefficient.

In general, however, the polynomial proposed above is used in the form of multiple binary quadratic equations in each layer just like

$$\bar{y} = a_1 x_1^2 + a_2 x_2^2 + a_3 x_1 + a_4 x_2 + a_5 x_1 x_2 + a_6. \quad (2)$$

All pairs of the neurons in each layer are calculated in the form (2), and then, the difference between the actual output y and the fitted value \bar{y} can be obtained. By introducing mean square error (MSE) as the principle of screening for

each layer, the difference mentioned above can be minimized continuously until the downtrend stopped:

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2. \quad (3)$$

Compared to the traditional neural network, the GMDH network has two significant benefits: (1) determine automatically the number for both network layers and neurons in each layer, which diminish the artificiality successfully in the simulation process and (2) build the connection between the selected parameters and the output in the form of polynomials, which differs from other neural networks with black box model. However, different division modes for data sets lead to multifarious results, which cannot get global optimum solution. Meanwhile, fluctuant threshold based on selection rules will produce the possibility that different parameters in each layer are eliminated untimely.

In this section, the algorithm of GMDH network was improved specifically in two ways. First, to avoid the case that different division modes of the data sets lead to various constructed models with distinct differences, we use randomly drawing method to realize the division of the data sets into training sets and testing sets. The constructed model can get global optimum ultimately with higher precision than traditional partition manner. Second, the intermediate polynomials in each layer only related to the upper layer by screening with external criterion, so the selection of effective variables is independent among the network layers, which lead to the circumstance that partial variables are estimated with less influence and eliminated untimely in some layers. Point to this situation, we introduced original variable preservation method to optimize the selection of variables establishing the GMDH network. More details of the optimum method were presented in the lecture (Guo et al. 2017). In this investigation, the function of the modified GMDH algorithm was set to provide effective variables as the input of the back propagation.

Hybrid modified GMDH-BP algorithm

The modified GMDH network is good at estimating the relationship between the effective variables and the output in higher precision, and the back propagation network has great advantages in regression problems. In this section, the combination of the modified group method of data handling (GMDH) and back propagation (BP) as a hybrid model is proposed to improve the precision of oil production, which overcomes the shortcomings of back propagation for variable selection problems. First, the input parameters are selected by the function of the modified GMDH network, which has been enhanced the ability of screening variables

by the improved algorithm. Then, based on the selected variables, the BP network is used to forecast the output of oilfield. The whole procedure of the proposed hybrid model can be described in the following manner:

Step 1: The original data sets are normalized first and then separated into the training sets and testing sets with the randomly drawing sample method.

Step 2: With the input variables $\{x_1, x_2, \dots, x_m\}$, each two of them are generated and the number of the combination are $C_m^2 = \frac{m(m-1)}{2}$. Compare the value \bar{y} calculated by the formula (2) with the true value y and determine new input variables $\{x_{11}, x_{12}, \dots, x_{1j}\}$ for next layer.

Step 3: Merge the new input variables $\{x_{11}, x_{12}, \dots, x_{1j}\}$ and the original variables $\{x_1, x_2, \dots, x_m\}$ into a new pair of input variables and repeat step 2. Sort the MSE of all pairs of the neurons in each layer and select the variables with fewer error into next layer.

Step 4: The iteration stops when the smallest MSE of each layer cannot keep decreasing and record the MSE of the last layer. Return to the step 1, repeat steps 1–4 until the number of iterations reach the upper limit or the required precision of MSE is obtained.

Step 5: The process of the modified GMDH network ends and export the selected effective variables $\{x_1, x_2, \dots, x_n\}$ to the BP network. The whole structure of the network is designed as single-hidden layer and the number of hidden nodes is computed based on the actual situation.

Step 6: The input variables are propagated forward through the designed network in each layer; when it reached the output layer, the difference between the output and the

true value are calculated by a loss function. Then, the error is propagated from the output layer back through the same network and completes the weight update. The configuration of the proposed hybrid model is shown in Fig. 2.

Application of hybrid algorithm based on modified GMDH and BP networks

Data collection

The data sets provided by Guo (2009) were utilized in this study, which present the dynamic process of a low permeability reservoir development with a time span of 1980–2006. Based on the research conducted by Wang et al. for using multiple linear regression (MLR) to predict the output of oilfield, Guo built an improved MLR model for predicting annual output of oilfield by analyzing the statistics and the important information from the regression parameters. The detailed data sets are shown in Table 1; meanwhile, the performance of the MLR model is shown in Table 2.

Effective parameter selection based on the GMDH-type algorithm

After the normalization procedure, the data sets were used in the proposed hybrid model discussed above. The modified GMDH algorithm was used to perform effective parameter selection, and the corresponding polynomials of oilfield dynamic output for selective parameters are as follows:

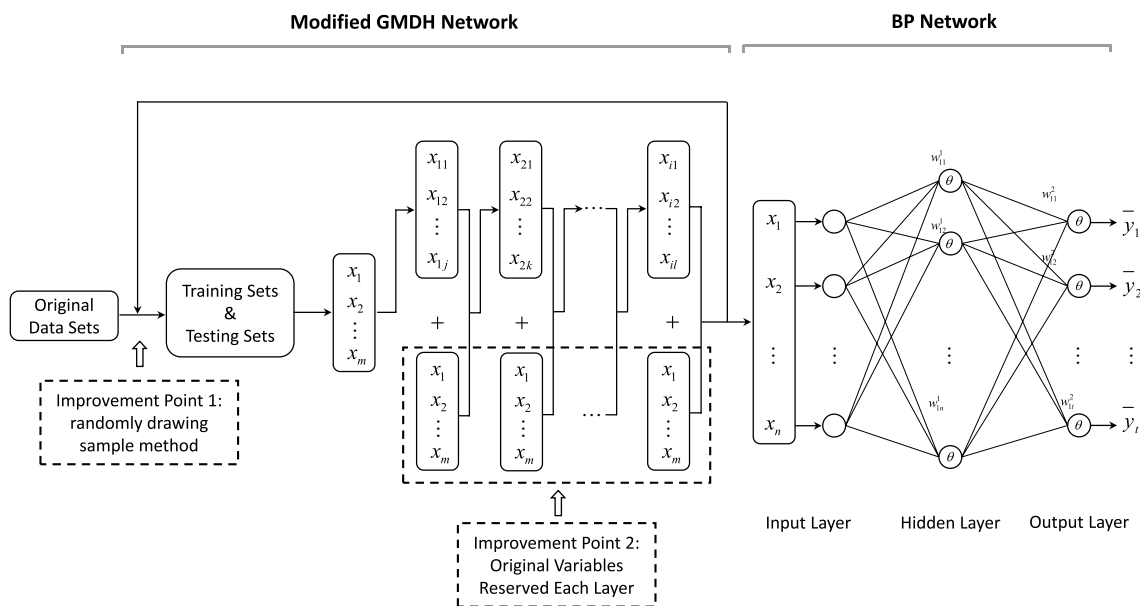


Fig. 2 Proposed structure of the hybrid model for predicting oil production

Table 1 Original data sets of parameters for production forecasting of oilfield

Year	N_{total} X_1	N_{act} X_2	N_{new} X_3	q_{inj} X_4	f_w X_5	V X_6	R X_7	Q_{last} X_8	Q Y
1980	379	309	136	119.18	35.6	1.87	7.05	126.39	136.75
1981	455	403	157	161.83	38.4	1.68	8.49	136.75	139.63
1982	569	495	207	184.56	39.4	1.77	9.61	139.63	142.19
1983	689	612	311	237.59	41.8	1.45	9.07	142.19	144.28
1984	855	720	351	230.5	42.33	1.53	9.54	144.28	141.72
1985	1028	874	426	276.59	42.93	1.6	9.49	141.72	146.61
1986	1268	1087	472	330.64	46.21	1.55	10.25	146.61	145.45
1987	1446	1197	652	398.14	45.8	1.49	9.35	145.45	148.94
1988	1705	1417	486	455.1	47.8	1.43	9.08	148.94	155.92
1989	1892	1524	458	526.91	49.3	1.31	9.31	155.92	165.23
1990	2113	1761	473	602.04	52.15	1.37	10.13	165.23	202.46
1991	2372	1903	506	740.62	55.46	1.26	10.88	202.46	217.59
1992	2641	2123	705	867.65	59.83	1.18	11.54	217.59	260.64
1993	3090	2574	689	987.98	60.87	1.11	12.07	260.64	302.53
1994	3603	2826	964	1110.87	63.39	1.11	12.96	302.53	349.31
1995	3987	2878	1073	1183.27	63.12	1.2	13.57	349.31	372.58
1996	4530	3002	1003	1309.18	64.79	1.2	14.76	372.58	403.76
1997	4872	3172	1044	1406.31	67.45	1.07	14.59	403.76	420.05
1998	5110	3260	854	1576.06	68.89	1.01	14.88	420.05	439.82
1999	5400	3375	686	1676.03	70.12	0.95	15.4	439.82	464.97
2000	5524	3497	758	1651.9	71.88	0.88	15.82	464.97	471.25
2001	5653	3704	891	1808.34	71.88	0.91	16.46	471.25	520.5
2002	6958	5523	1043	1926.73	72.95	0.83	17.22	520.5	611.55
2003	8680	7805	1181	1958.05	72.83	0.83	17.74	611.55	715.87
2004	9864	8263	1319	2536.5	72.28	0.89	17.71	715.87	810.95
2005	11,805	9522	1946	3003.2	72.01	0.84	16.98	810.85	905.1
2006	12,314	11,092	2347	3298.7	72.31	0.85	17.2	905.1	962.3

Table 2 Summary of the evaluation results for different models

Methods	Relative error	R	RMSE	MAPE	SI
MLR (Guo 2009)	0.0935	0.9923	42.0044	4.5892	0.1124
GMDH	0.0724	0.9883	41.6208	8.2109	0.1051
BP	0.0249	0.9969	21.1922	3.2006	0.0575
Modified GMDH (Guo et al. 2017)	0.0629	0.9984	13.9440	3.0268	0.0378
GMDH-BP	0.0197	0.9949	24.8495	5.4415	0.0675
Modified GMDH-BP	0.0099	0.9985	13.8481	3.0293	0.0378

$$Q_4^2 = 0.31998(Q_4^1)^2 + 0.12064(Q_{12}^1)^2 + 1.1412Q_4^1 \cdot Q_{12}^1 + 0.51057Q_4^1 + 0.031821Q_{12}^1 + 0.0079427 \tag{12}$$

$$Q_{17}^2 = R \tag{13}$$

$$Q_3^3 = -0.28403(Q_4^2)^2 - 0.070727(Q_{17}^2)^2 + 0.31146Q_4^2 \cdot Q_{17}^2 + 0.93718Q_4^2 + 0.022181Q_{17}^2 - 0.0010487 \tag{14}$$

$$Q_{18}^3 = Q_{last} \tag{15}$$

$$Q_1^4 = -3.7867(Q_3^3)^2 - 3.4284(Q_{18}^3)^2 + 7.2835Q_3^3 \cdot Q_{18}^3 + 1.1992Q_3^3 - 0.22921Q_{18}^3 + 0.002322 \tag{16}$$

$$Q_{16}^4 = v \tag{17}$$

$$Q_2^5 = -0.0067055(Q_1^4)^2 + 0.0037029(Q_{16}^4)^2 - 0.010687Q_1^4 \cdot Q_{16}^4 + 0.99635Q_1^4 - 0.0086437Q_{16}^4 + 0.004273 \tag{18}$$

$$Q_4^1 = -0.24729N_{new}^2 - 0.75502Q_{last}^2 + 0.1695N_{new} \cdot Q_{last} + 0.062865N_{new} + 1.2909Q_{last} - 0.014572 \tag{10}$$

$$Q_{12}^1 = N_{act} \tag{11}$$

$$Q_{11}^5 = N_{\text{total}} \quad (19)$$

$$Q_1^6 = -0.63468(Q_2^5)^2 - 0.7048(Q_{11}^5)^2 + 1.3469Q_2^5 \cdot Q_{11}^5 + 0.94474Q_2^5 + 0.052598Q_{11}^5 - 0.00012446. \quad (20)$$

As a contrast, the traditional GMDH algorithm was considered and taken to predict the relationship between the output and the parameters, and the results are presented as follows:

$$Q_4^1 = -0.24729N_{\text{new}}^2 - 0.75502Q_{\text{last}}^2 + 0.1695N_{\text{new}} \cdot Q_{\text{last}} + 0.062865N_{\text{new}} + 1.2909Q_{\text{last}} - 0.014572 \quad (21)$$

$$Q_7^1 = 2.1287N_{\text{new}}^2 + 0.86016v^2 + 0.69915N_{\text{new}} \cdot v - 0.87815N_{\text{new}} - 1.4141v + 0.57713 \quad (22)$$

$$Q_{10}^1 = 0.92075N_{\text{new}}^2 + 0.4325Q_{\text{inj}}^2 + 0.16661N_{\text{new}} \cdot Q_{\text{inj}} - 0.38459N_{\text{new}} + 0.63473Q_{\text{inj}} + 0.0014875 \quad (23)$$

$$Q_1^2 = 0.65114(Q_4^1)^2 + 0.02469(Q_7^1)^2 + 0.19724Q_4^1 \cdot Q_7^1 + 0.69265Q_4^1 - 0.012697Q_7^1 + 0.01097 \quad (24)$$

$$Q_{10}^2 = -0.15054(Q_4^1)^2 - 0.13066(Q_{10}^1)^2 + 0.74814Q_4^1 \cdot Q_{10}^1 + 0.55802Q_4^1 + 0.03793Q_{10}^1 + 0.0075964 \quad (25)$$

$$Q_1^3 = 6.569(Q_1^2)^2 + 5.9129(Q_{10}^2)^2 - 12.4899Q_1^2 \cdot Q_{10}^2 + 0.77332Q_1^2 + 0.22963Q_{10}^2 - 8.1951 \times 10^{-5}. \quad (26)$$

As can be seen from the analytical equations, the modified GMDH algorithm runs within six rounds of iteration, with the MSE of each layer reduces from 0.1999 to 0.0231. Meanwhile, the effective parameters selected by the modified GMDH algorithm are $\{N_{\text{total}}, N_{\text{act}}, N_{\text{new}}, v, R, Q_{\text{last}}\}$. By contrast, the traditional GMDH algorithm runs within three rounds of iteration, with the MSE reduces from 0.1999 to 0.0631, and the output parameters are $\{N_{\text{new}}, Q_{\text{inj}}, v, Q_{\text{last}}\}$.

The output prediction based on the BP algorithm

Different types and numbers of the input nodes lead to differentiated outcomes. As mentioned above, the effective parameters selected by the traditional/modified GMDH network fed into the BP network for analysis and prediction.

The structure of the BP network adopted single-hidden layer network. The number of the input nodes equals to the sum of parameters selected by GMDH type model,

respectively. Determining the number of nodes in hidden layer is a crucial point for BP network prediction. Either too many or too few nodes in hidden layer will increase the simulating error. Major methods for solving this problem conclude the following rules:

Liu (2008):

$$l = 2m + 1 \quad (27)$$

or

$$l = \sqrt{mn} \quad (28)$$

Fu and Zhao (2010):

$$l = \sqrt{n + m} + a. \quad (29)$$

In which l is the number of nodes in hidden layer, n and m are the number of nodes in the input and output layers, respectively, and a is constant ranging from 0 to 10. In this paper, the actual number of nodes in hidden layer is computed to be $\sqrt{6 + 1} + 10 = 13$.

For the hybrid model of modified GMDH and BP network, a total of 22 records used for oilfield output prediction were gathered from 27 records with the production history from 1980 to 2006. The learning rate was 0.1, the stop criterion of error function was set to 0.001 and the maximum number of iteration was 1000. The initial weights and threshold were randomly generated by the computer. In the process of model operation, 22 data samples (81%) were randomly selected for training the BP network and the remaining 5 data samples (19%) were used as testing data sets for model evaluation, the average results of residual error obtained from the six prediction models are presented in Table 2.

Results and discussion

The outputs predicted by the hybrid model are presented in Fig. 3, compared with the traditional equations (MLR) and the artificial neural network models (GMDH, modified

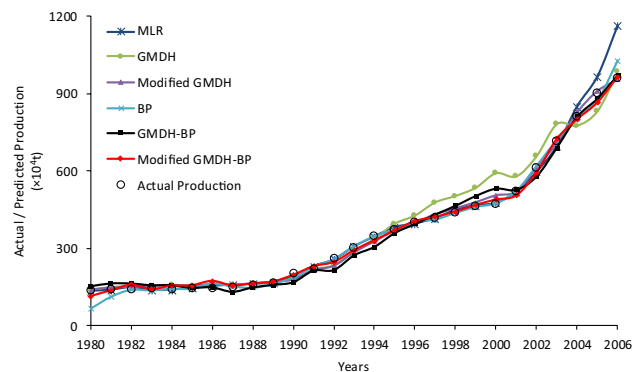


Fig. 3 Comparison chart for actual/predicted production by different models

GMDH, BP, GMDH-BP). The comparison chart indicates that the hybrid model combining the modified GMDH network and BP algorithm is more approximate to the actual production for the oilfield than other models listed. To investigate the precision in-depth analysis, error (mean relative error), R (correlation coefficient), RMSE (root mean square error), MAPE (mean absolute percentage of error), and SI (scatter index) are utilized to investigate the performance of the presented models:

$$\text{error} = \frac{1}{M} \cdot \frac{|Y_{i(\text{Actual})} - Y_{i(\text{model})}|}{Y_{i(\text{Actual})}} \quad (30)$$

$$R = \frac{\sum_{i=1}^M (Y_{i(\text{Actual})} - \bar{Y}_{(\text{Actual})})(Y_{i(\text{Model})} - \bar{Y}_{(\text{Model})})}{\sqrt{\sum_{i=1}^M (Y_{i(\text{Actual})} - \bar{Y}_{(\text{Actual})})^2 \cdot \sum_{i=1}^M (Y_{i(\text{Model})} - \bar{Y}_{(\text{Model})})^2}} \quad (31)$$

$$\text{RMSE} = \left[\frac{\sum_{i=1}^M (Y_{i(\text{Model})} - Y_{i(\text{Actual})})^2}{M} \right]^{1/2} \quad (32)$$

$$\text{MAPE} = \frac{1}{M} \left[\frac{\sum_{i=1}^M |Y_{i(\text{Model})} - Y_{i(\text{Actual})}|}{\sum_{i=1}^M Y_{i(\text{Actual})}} \times 100 \right] \quad (33)$$

$$\text{SI} = \frac{\sqrt{(1/M) \sum_{i=1}^M ((Y_{i(\text{Model})} - \bar{Y}_{(\text{Model})}) - (Y_{i(\text{Actual})} - \bar{Y}_{(\text{Actual})}))^2}}{(1/M) \sum_{i=1}^M Y_{i(\text{Actual})}} \quad (34)$$

where $Y_{i(\text{Model})}$ and $Y_{i(\text{Actual})}$ are the forecasted and observed values, respectively, $\bar{Y}_{(\text{Model})}$ and $\bar{Y}_{(\text{Actual})}$ are the average of the forecasted and observed values, and M is the total of events.

The statistical results of the proposed traditional equations and artificial intelligence approaches with training and testing data are presented in Table 2. It is clear at a glance that the hybrid model combining modified GMDH network and BP algorithm (modified GMDH-BP) is more accurate than other models with higher correlation ($R=0.9986$) and lower error (error = 0.0099, RMSE = 13.7979, MAPE = 2.9889, SI = 0.0376). In general, the neural network tools perform better with a relatively higher correlation and lower error in prediction precision. Comparing with the hybrid model combining traditional GMDH network and BP algorithm (GMDH-BP), the proposed model (modified GMDH-BP) improves overall precision in forecasting, with higher correlation from 0.9949 to 0.9986 and lower error from 0.0197 to 0.0099 (error), from 24.8495 to 13.7979 (RMSE), 5.4415 to 2.9889 (MAPE), and 0.0675 to 0.0376 (SI), respectively.

Figure 4 shows the comparison of the six models mentioned above with time series and scatter plots for predicting

the production of oilfield. The results derived from the whole six models are in agreement with the actual production, indicating that these prediction algorithms are applicable for modeling oilfield production series data. However, the dashed line generated from modified GMDH-BP is the closest than other models to the solid line which indicating the actual output of oilfield. By means of using correlation coefficient to evaluate the degree of fitting, the hybrid model combining the modified GMDH network and BP algorithm (modified GMDH-BP) is slightly superior to other models taken into consideration. The successful performance obtained in this paper indicates that the hybrid model (modified GMDH-BP) is a powerful tool to simulate the oilfield production time series and has the ability to provide a better prediction performance. In conclusion, the evaluation results suggest that the best performance can be obtained by the hybrid model (modified GMDH-BP), followed by modified GMDH, BP, GMDH-BP, MLR and GMDH models in turn.

Conclusion

Yearly production estimation of oilfield is vital in oilfield development programming and plenty of models predicting the dynamic output have been proposed in recent years. In this paper, we have demonstrated systematically how the yearly production of oilfield could be represented by a hybrid model combining the modified GMDH and BP models. To illustrate the capability of the hybrid model (modified GMDH-BP), an actual oilfield with various production parameters was chosen to be analyzed and used to test the annual output predicted by six models. The forecast of the oilfield production is a complex issue, including various parameters in which share different impact on each other. Therefore, the first step of model construction is the effective parameter selection on the basis of using modified GMDH network. The modified GMDH network proposed in this paper performs better in selecting input variables, owing to the improvement of introducing randomly drawing method for original datasets and original variable preservation method to optimize the selection of variables in each layer. In addition, the excellent precision of the BP algorithm is another favorable factor for obtaining the best degree of fitting for the modified GMDH-BP model. By comparing the performances of the six models, the proposed hybrid model (Modified GMDH-BP) provides the best forecast precision with highest correlation ($R=0.9986$) and lowest error (error = 0.0099, RMSE = 13.7979, MAPE = 2.9889, SI = 0.0378). It should be mentioned that the hybrid model (modified GMDH-BP) provides a robust simulation ability

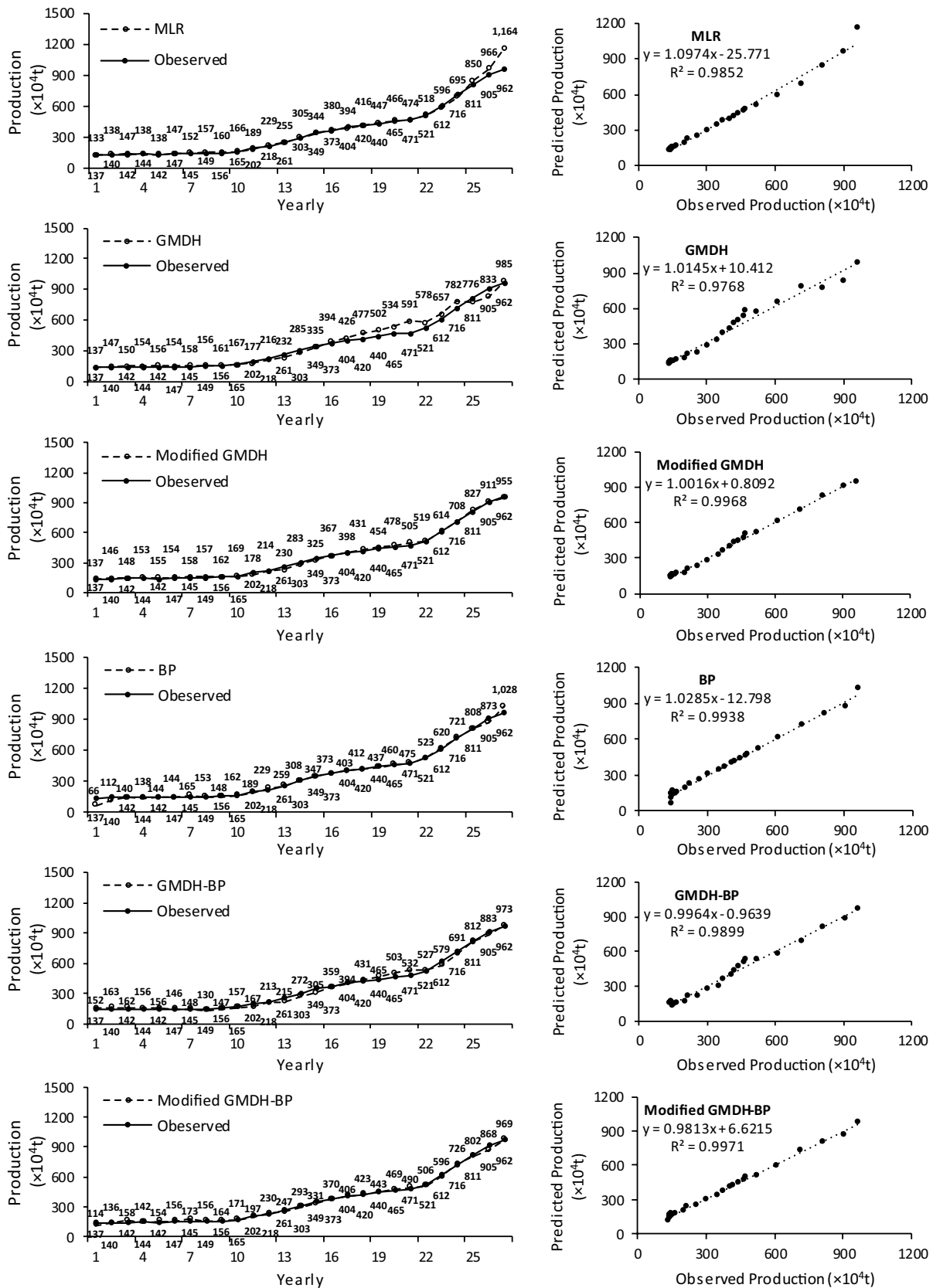


Fig. 4 Comparison of the performances of MLR, GMDH, modified GMDH, BP, GMDH-BP, and modified GMDH-BP models for oilfield forecasting

of capturing the nonlinear relation of complex production time series prediction of oilfield and thus producing more accurate forecasts.

Acknowledgements The authors gratefully acknowledge the support received from the Major Science and Technology Projects of Petro-China Co Ltd under Grant numbers 2017E-1505.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Abdel-Aal RE (2005) GMDH-based feature ranking and selection for improved classification of medical data. *J Biomed Inform* 38(6):456–468
- Abdel-Aal RE, El-Alfy ESM (2009) Constructing optimal educational tests using GMDH-based item ranking and selection. *Neurocomputing* 72(4):1184–1197
- Ahmadi MA (2012) Withdrawn: prediction of minimum miscible pressure by using neural network based hybrid genetic algorithm and particle swarm optimization. *J Pet Sci Eng*. <https://doi.org/10.1016/j.petrol.2012.08.015>
- Ahmadi MA, Golshadi M (2012) Neural network based swarm concept for prediction asphaltene precipitation due to natural depletion. *J Pet Sci Eng* 98–99(6):40–49
- Ahmadi MA, Goudarzi A (2013) Retracted article: combining artificial neural network and unified particle swarm optimization for oil flow rate prediction: case study. *Neural Comput Appl* 23(2):565–565
- Ahmadi MA, Shadzadeh SR (2012) Permeability prediction of carbonate reservoir by combining neural network and shuffled frog-leaping. *J Am Sci* 8(2):529–534
- Ahmadi MA, Zendejboudi S, Lohi A, Elkamel A, Chatzis I (2013) Reservoir permeability prediction by neural networks combined with hybrid genetic algorithm and particle swarm optimization. *Geophys Prospect* 61(3):582–598
- Al-Ajmi RM, Abou-Ziyan HZ, Mahmoud MA (2012) Performance evaluation of 14 neural network architectures used for predicting heat transfer characteristics of engine oils. *Int J Comput Methods Eng Sci Mech* 13(1):60–75
- Arps JJ (1945) Analysis of decline curves. *Pet Trans* 160:228–247
- Balch RS, Stubbs BS, Weiss WW, Wo S (1999) Using artificial intelligence to correlate multiple seismic attributes to reservoir properties. *Soc Petrol Eng*
- Bardi U (2005) The mineral economy: a model for the shape of oil production curves. *Energy Policy* 33(1):53–61
- Fu H, Zhao H (2010) Application design of MATLAB neural network. Machinery Industry Press, Beijing
- Ghorbani B, Ziabasharhagh M, Amidpour M (2014) A hybrid artificial neural network and genetic algorithm for predicting viscosity of Iranian crude oils. *J Nat Gas Sci Eng* 18(18):312–323
- Guo L (2009) Application of improved multivariate linear regression to output prediction of an oilfield. *J Xidian Univ* 17(7):1251–1257
- Guo J, Huang W, Mao Q, Wang X, Wang X, Song T (2017) Modified GMDH networks for oilfield production prediction. *Geosyst Eng* (7), 1–9
- Haykin S (1998) *Neural networks: a comprehensive foundation*, 3rd edn. Macmillan, London
- Hubbert MK (1980) Techniques of prediction as applied to production of oil and gas. In: *Proceedings, US Department of Commerce symposium*, Washington, DC, vol 16(6), pp 8–20
- Ivakhnenko AG (1971) Polynomial theory of complex systems. *IEEE Trans Syst Man Cybern SMC-1*(4):364–378
- Jr OL, Nunes U, Rui A, Schnitman L, Lepikson HA (2009) Applications of information theory, genetic algorithms, and neural models to predict oil flow. *Commun Nonlinear Sci Numer Simul* 14(7):2870–2885
- Kashyap R (1973) System identification. *Autom Control IEEE Trans* 18(1):85–86
- Lia O, Omre H, Tjelmeland H, Holden L, Egeland T (1997) Uncertainties in reservoir production forecasts. *AAPG Bull* 81(5):775–802
- Liu X (2008) *Artificial neural network and particle swarm optimization*. Beijing University of Posts and Telecommunications Press, Beijing
- Liu C, McVay DA (2009) Continuous reservoir simulation model updating and forecasting using a markov chain monte carlo method. *Soc Petrol Eng*
- Mohsen M, Nafiseh B, Mehdi A, Mohsen M (2010) Forecasting volatility of crude oil price using the GMDH neural network. *Quart Energy Econ Rev* 7(25):89–112
- Mollaiy Berneti S, Shahbazian M (2013) An imperialist competitive algorithm artificial neural network method to predict oil flow rate of the wells. *Int J Comput Appl* 26(10):47–50
- Mori H, Tsuzuki S (1990) Comparison between backpropagation and revised GMDH techniques for predicting voltage harmonics. In: *IEEE international symposium on circuits and systems*, vol 2, pp 1102–1105. IEEE, New York
- Najafzadeh M, Barani GA (2011) Comparison of group method of data handling based genetic programming and back propagation systems to predict scour depth around bridge piers. *Sci Iran* 18(6):1207–1213
- Parks PC, Ivakhnenko AG, Boichuk LM, Svetalsky BK (1975) A self-organizing model of the British economy for control with optimal prediction using the balance-of-variables criterion. *Int J Comput Inf Sci* 4(4):349–379
- Samsudin R, Saad P, Shabri A (2011) River flow time series using least squares support vector machines. *Hydrol Earth Syst Sci* 15:1835–1852
- Tang X, Zhang B, Höök M, Feng L (2010) Forecast of oil reserves and production in Daqing oilfield of China. *Energy* 35(7):3097–3102
- Tapias O, Soto CP, Sandoval J, Perez HH, Bejarano A (2001) Reservoir engineer and artificial intelligence techniques for data analysis. *Soc Petrol Eng*
- Wang J, Feng L (2016) Curve-fitting models for fossil fuel production forecasting: key influence factors. *J Nat Gas Sci Eng* 32:138–149
- Wang J, Feng L, Steve M, Xu T, Gail TE, Mikael H (2015) China's unconventional oil: a review of its resources and outlook for long-term production. *Energy* 82:31–42
- Yi-Bao LI, Zhang XY, Jian-Guo MA, Wang LJ (2005) Study of improving algorithms based on the bp neural network. *J Hefei Univ Technol* 28(6):668–671
- Yu S, Zhu K, Diao F (2008) A dynamic all parameters adaptive bp neural networks model and its application on oil reservoir prediction. *Appl Math Comput* 195(1):66–75
- Zendejboudi S, Ahmadi MA, James L, Chatzis I (2012) Prediction of condensate-to-gas ratio for retrograde gas condensate reservoirs using artificial neural network with particle swarm optimization. *Energy Fuels* 26(6):3432–3447

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