Energy 220 (2021) 119708

Contents lists available at ScienceDirect

Energy

journal homepage: www.elsevier.com/locate/energy

Well production forecasting based on ARIMA-LSTM model considering manual operations

Dongyan Fan [a, b, c], Hai Sun [a, c, *], Jun Yao [a], Kai Zhang [a], Xia Yan [a], Zhixue Sun [a]

[a] School of Petroleum Engineering, China University of Petroleum (East China), Qingdao, 266580, PR China
[b] Key Laboratory of Unconventional Oil & Gas Development (China University of Petroleum (East China)), Ministry of Education, Qingdao, 266580, PR China
[c] Department of Petroleum Engineering, University of Houston, Houston, TX, 77204, USA

ARTICLE INFO

Article history:
Received 26 September 2020
Received in revised form
12 December 2020
Accepted 22 December 2020
Available online 24 December 2020

Keywords:
Production forecasting
Hybrid model
ARIMA
LSTM
Daily production time series

ABSTRACT

Accurate and efficient prediction of well production is essential for extending a well's life cycle and improving reservoir recovery. Traditional models require expensive computational time and various types of formation and fluid data. Besides, frequent manual operations are always ignored because of their cumbersome processing. In this paper, a novel hybrid model is established that considers the advantages of linearity and nonlinearity, as well as the impact of manual operations. This integrates the autoregressive integrated moving average (ARIMA) model and the long short term memory (LSTM) model. The ARIMA model filters linear trends in the production time series data and passes on the residual value to the LSTM model. Given that the manual open-shut operations lead to nonlinear fluctuations, the residual and daily production time series are composed of the LSTM input data. To compare the performance of the hybrid models ARIMA-LSTM and ARIMA-LSTM-DP (Daily Production time series) with the ARIMA, LSTM, and LSTM-DP models, production time series of three actual wells are analyzed. Four indexes, namely, root mean squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and similarity (Sim) values are evaluated to calculate the prediction accuracy. The results of the experiments indicate that the single ARIMA model has a good performance in the steady production decline curves. Conversely, the LSTM model has obvious advantages over the ARIMA model to the fluctuating nonlinear data. And coupling models (ARIMA-LSTM, ARIMA-LSTM-DP) exhibit better results than the individual ARIMA, LSTM, or LSTM-DP models, wherein the ARIMA-LSTM-DP model performs even better when the well production series are affected by frequent manual operations.

1. Introduction

Accurate prediction of well production is essential and challenging for the efficient development and management of oil and gas resources. As the most significant energy, oil and gas provide up to 58% contribution to the total global energy consumption in the year 2019 [1]. Production forecast is always playing a significant role in the entire life cycle of oil and gas wells including early resource assessment, middle technology adjustment, and later enhancing recovery. Meanwhile, oil and gas resources development is affected by various factors [2,3] such as reservoir heterogeneity, flow mechanism complexity, development method diversity and manual interference, which make accurate production forecasting becomes more complicated and challenging.

Typically, three methods are used for building well production forecasting models. Numerical simulation is the most common method for predicting oil and gas production. The process is based on numerical models [4–6] and provides good results and fully describes the geological heterogeneity of the reservoir. However, the models are tedious and time-consuming [7] and involve establishing the geological model, numerical model and history matching. They also require various types of formation and fluid data, such as logging, permeability, porosity, and saturation. Additionally, analytical methods are used to calculate different types of wellbore flow rate changes. To obtain the analytical solution, some assumptions based on formation heterogeneity, complex well structure, and boundary conditions are required [8–10]. Although these assumptions can simplify complex reservoir models, the analytical results may not match the actual production changes, especially frequent manual operations and underground multiphase flow. Additionally, accurate analytical solutions are based on

* Corresponding author. School of Petroleum Engineering, China University of Petroleum (East China), Qingdao, 266580, PR China.
E-mail address: sunhaiupc@sina.com (H. Sun).

https://doi.org/10.1016/j.energy.2020.119708
0360-5442/

correct formation and fluid data, which often require long-lasting and expensive physical experiments. The traditional decline curve method [11,12] can predict production performance from the analysis of long-term oil and gas production data. The basis of decline curve analysis involves matching past actual production rate/time data with a "model", such as exponential, harmonic, and hyperbolic models. These models are all ideal curves and cannot consider the actual formation factors. Hence, it is difficult to guarantee the correct performance using this method. Therefore, a more effective and convenient method should be established, which can consider the internal decline by various formation and fluid factors, as well as the external influence by manual operations.

The production data is a typical time series structure, which is affected by numerous internal and external factors. Hence, many researchers [13–15] use time series exploration methods to extract hidden information from past time series to predict the future behavior of well production. ARIMA is one of the most well-known and successful linear statistical models for time series prediction. It was proposed by Box and Jenkins [16] in the early 1970s. Based on the perspective of forecasting characteristics, ARIMA model can successfully address linear sequences and has been applied in many fields, such as electricity [17], agriculture [18,19], weather [20], as well as petroleum industry [21,22]. Due to the shortcomings in nonlinear feature extraction, this method has not been used widely in production forecasting.

In recent years , advancement of artificial intelligence and big data collection in the oilfield, researchers are increasingly focusing on machine learning (ML) methods to solve production forecasting problems [23–26], such as artificial neural network (ANN), recurrent neural network (RNN), long short term memory (LSTM), convolutional neural network (CNN), gated recurrent units (GRU) and so on. Among the above models, LSTM, a modified RNN architecture is introduced by Hochreiter and Schmidhuber in 1997 [27]. The LSTM algorithm has attracted much attention for its sufficiency to capture nonlinear trends and dependencies. The results [25,26,28] indicate that LSTM performs better than traditional decline curve analysis methods considering the influence of multiple factors simultaneously. However, some shortcomings emerge in single AI models such as low convergence, outliers influence, loss of time, local minima and so on. In order to overcome those problems, considerable hybrid models are proposed to take advantage of each component model and improve forecasting performance.

To improve convergence speed and global searching abilities of the single neural network, many hybrid models [26,29,30] have been proposed to enhance production time series performance prediction, in which genetic algorithm (GA), particle swarm optimization (PSO) or imperialist competitive algorithm (ICA) methods are applied to optimize the coefficients and configuration of neural network. Then to overcome the loss of time problems, in Refs. [31–33], the CNN layer and principal component analysis (PCA) are employed to extract features of short-term time series and neural network is to perform long term prediction of well production. The reduced-order models [34,35] such as extended dynamic mode decomposition (EDMD), proper orthogonal decomposition and discrete empirical interpolation method (POD-DEIM) are applied to construct robust deep learning architectures, which combine the physical model with deep learning methods for accurate approximation of oil and gas production. Based on the principle of "decomposition and ensemble", the machine learning method LSTM has been performed by ensemble empirical mode decomposition (EEMD) for the sake of improving prediction accuracy [36].

For the other time series problems, hybrid models have also been widely applied to overcome the shortcomings of the single neural network model. For example, Seckin [37,38] proposed support vector regression (SVR) with a wrapper-based feature selection approach to predict volatility in crude oil price time series. A hybrid neural network named deep CNN and gated recurrent unit network (DCGNet) is designed for sintering temperature forecasting [39]. Mi [40] built a wind speed multi-step prediction model based on the singular spectrum analysis (SSA), empirical mode decomposition and convolutional support vector machine (CNNSVM). Aytac [41] developed a hybrid forecasting model based on LSTM and empirical wavelet transform (EWT) decomposition for digital currency time series.

However, another hybrid approach based on the linear statistical model with machine learning model has been applied to many other fields successfully [42–46], but has not received enough attention in well production forecasting. This model takes advantage of linear and nonlinear models to improve forecasting performance. As the most well-known and effective linear statistical model, ARIMA model is good at filtering linear tendency in the time series. At the same time, the RNN model suit to exact the complicated non-linear relations due to the special memory and determine structure. Hence, we are inspired by the success of this hybrid prediction model. Besides, the well production data are affected by the internal flow mechanisms of underground oil and gas wells and external manual operations. So these complex factors can be used to construct well production time series, which contain linear and nonlinear behaviors. Therefore, we propose the hybrid model that combines ARIMA with LSTM machine learning method to build well production forecasting models.

Additionally, the frequent manual operation is an important reason for the nonlinear trend of well production data, as shown in Fig. 1. In the actual production process, oil wells are often faced frequent opening and closing operations. Manual operations cannot be predicted via modeling. Thus, daily production time series are added as a component of nonlinear parts. Subsequently, we obtained another coupling model, which can be named ARIMA-LSTM considering daily production time (ARIMA-LSTM-DP) model.

The main goal of this study is to build reliable and accurate forecasting models for oil and gas well production time series. The hybrid models of ARIMA-LSTM and ARIMA-LSTM-DP are studied through two components for modeling linear and non-linear parts, and verified by three actual well production data. The main contribution of this study includes (1) in the theoretical aspect, it is demonstrated that the hybrid model ARIMA-LSTM is a powerful and convenient method for production forecasting. Manual operations are the important sources of nonlinear fluctuation, which can be considered in the LSTM network effectively; (2) in the practical aspect, it provides a guide to engineers on how to choose the suitable method to deal with complicated oil and gas rate time sequence. The ARIMA method is a good tool for stable decline rate curves, and LSTM is good at solving the nonlinear fluctuating data. The hybrid model ARIMA-LSTM has a better performance while considering linear and nonlinear models. The ARIMA-LSTM-DP is the best choice when the production time series have obvious nonlinear fluctuation.

The rest of this paper is organized as follows: In Section 2, we describe our proposed models in detail. In Section 3, we present the analysis process of well production, namely linear prediction of ARIMA model, nonlinear prediction of the LSTM model, and coupling prediction results. In section 4, comparisons between hybrid models and individual models are made. In Section 5, we present the conclusions of the study.

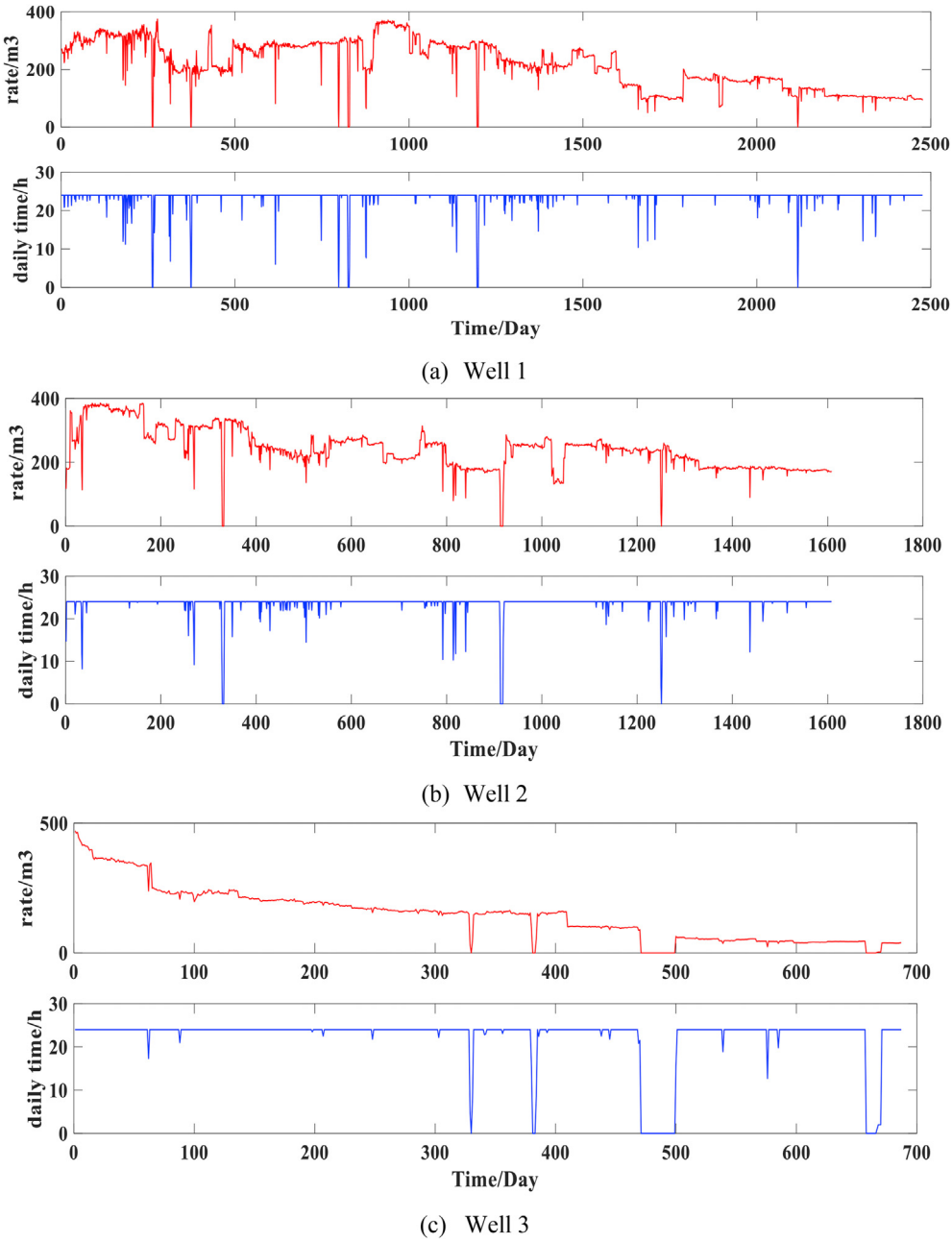


Fig. 1. Production rate and daily production time curves.

2. Methodology

The well production data are typical times series data and can be assumed to contain linear and nonlinear components. As per previous studies, ARIMA is a traditional and effective linear statistical method for time series forecasting. Conversely, LSTM can capture nonlinear features in a dataset. Given that the manual open-shut well operations from the surface lead to nonlinear fluctuations in production data, we propose a coupling of ARIMA-LSTM-DP models to incorporate linear and nonlinear parts, which in turn indicates the effect of the open-shut manual operations.

2.1. Well production data with open-shut operations

The actual well production time series are often accompanied by frequent open-shut manual operations wherein the shut time can vary from a few hours to days. The relationship curves (Fig. 1) indicate that the trend of changes in well production is evidently affected by daily production time. There are some studies [29,47] that delete the shut-in day points and move the production time series forward. However, the length of the shut-in time determines the degree of formation pressure recovery, which significantly impacts the well production in the future. Therefore, the simple deletion process leads to the omission of information, which in turn ignores the pressure recovery process. In addition to shutting a well all day, some hours of shut-in operations are also common, as shown in Fig. 1. Hence, the daily production time series should be considered in the model training process.

Given that the open-shut manual operations mainly affect the nonlinear fluctuations of well production curves, the daily

production time series are recorded for LSTM deep learning to capture nonlinear production features, which in turn can lead to accurate production forecasts.

2.2. Autoregressive integrated moving average model, ARIMA

The ARIMA model is one of the most popular linear regression models for forecasting stationary time series. The model is expressed as ARIMA (p, D, q), where the parameters p, D and q denote the structure of the forecasting model, which is a combination of auto-regression AR(p), moving average MA(q) and differencing degree D. The mathematical formula of the ARIMA (p, D, q) can be described as follows:

$$\left(1-\sum_{i=1}^{p}\varphi_i L^i\right)(1-L)^{\mathrm{D}}x_t=\left(1+\sum_{i=1}^{q}\theta_i L^i\right)\varepsilon_t \tag{1}$$

where L denotes the lag operator, $\varphi_i$are the parameters of the autoregressive part of the model, $\theta_i$are the parameters of the MA part, and $\varepsilon_i$are error terms.

Box and Jenkins [16] proposed a general process to build an ARIMA model, which involves three iterative steps. The first step involves the identification and selection of the type of model. To judge the best fitting model, stationary time series are essential, in which the basic statistical properties, such as mean, variance, and covariance or autocorrelation, are constant over time. To construct the stationary time series, an appropriate degree (D) of differencing is used. Then, the autocorrelation function (ACF) and partial autocorrelation function (PACF) are examined to select the model type. The second step involves parameter estimation. To select order q and p of the ARIMA model, many methods have been developed based on Akaikes information criterion (AIC) [48], minimum description length (MDL) [49], AIC, and Bayesian information criterion (BIC) [50] or fuzzy systems [51]. In this study, we use the AIC and BIC metric to estimate the parameters. The last step involves diagnostic checking of residual analysis. The errors are examined via some diagnostic statistics and plots of the residuals. Thus, we use the residual time series from the ARIMA model as the input for the subsequent LSTM model.

2.3. The long short term method, LSTM

Long short term memory (LSTM) network is an extension of the recurrent neural network (RNN). Due to its versatility in addressing parameters with large dimensions and the use of nonlinear activation functions in each layer, the LSTM model can capture nonlinear trends in data and remember previous information over a long time. Hence, LSTM has been successfully applied to numerous time series problems. The advantage of the LSTM structure is that it contains three types of gates, including the input, forget, and output. As shown in Fig. 2, the LSTM solves the vanishing gradient problem of RNN and allows the storage of information over a long-term period.

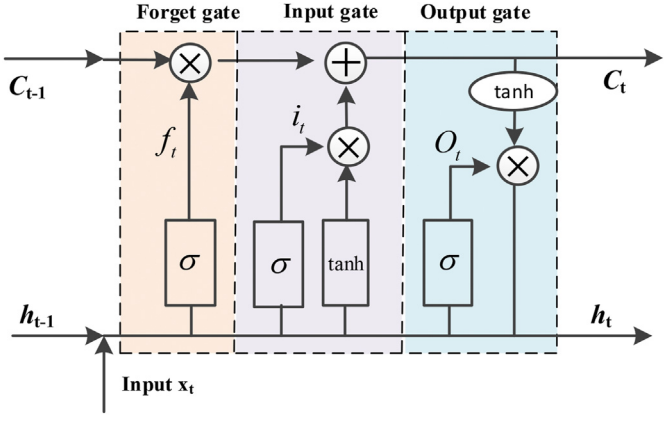


Fig. 2. LSTM cell structure.

The main information flow of the LSTM cell (Fig. 2) can be described mathematically. The symbols ⊕ and ⊗ denote addition and multiplication in the model, and the arrow denotes the flow direction of information. The first layer of memory gate determines removing unnecessary information to the cell state and can be expressed as follows:

$$f_t=\sigma\left(W_f\times x_t+U_f\times h_{t-1}+b_f\right) \tag{2}$$

where f_t denotes forgetting threshold at time t, σ denotes the sigmoid activation function, $W_f$and U_f denote the weights, $x_t$denotes the input value, h_{t-1}denotes the output value at time t-1, and $b_f$denotes the bias term.

The second input gate decides which information should be stored in the cell state from the current input vector. This includes decisioni_t, which updates the value and *tanh* layer for generating a new state value C_t. the specific expressions are as follows:

$$i_t=\sigma(W_i\times x_t+U_i\times h_{t-1}+b_i) \tag{3}$$

$$\overline{C}_t=\sigma(W_c\times x_t+U_c\times h_{t-1}+b_c) \tag{4}$$

where $i_t$denotes the input threshold at time t, W_i,U_i,W_c , and$U_c$are the weights, $b_c$and $b_i$are bias terms. To update the state of the cell at time t, the expressions is as follows:

$$C_t=f_t\times C_{t-1}+i_t\times\overline{C}_t \tag{5}$$

The third layer is produced as output information in the current time step and can be expressed as follows:

$$O_t=\sigma(W_o\times x_t+U_o\times h_{t-1}+b_o) \tag{6}$$

where$O_t$denotes the output threshold at time t, $W_o$and $U_o$are the weights , and $b_o$is the bias term. Then, the output value of the cell can be expressed as follows:

$$h_t=O_t\times\tanh(C_t) \tag{7}$$

where h_t denotes the output value of the cell at time t, *tanh* denotes the activation function, and $C_t$denotes the state of the cell at time t. After the data are passed through the three gates, the effective information is the output, and the invalid information is forgotten.

2.4. The proposed hybrid models

The well production data are in the form of time series data and can be assumed to consist of a linear portion and nonlinear portion, which can be expressed as follows:

$$x_t=L_t+N_t+\varepsilon_t \tag{8}$$

where L_t denotes the linearity of the data at time t, N_t signifies nonlinearity, and ε_t denotes the error term. The ARIMA method can successfully model nonlinear relationships in the time series data, and LSTM can successfully model nonlinear components. To reach the best forecasting results, we construct two hybrid models, as shown in Fig. 3, which combine the advantages of ARIMA and LSTM

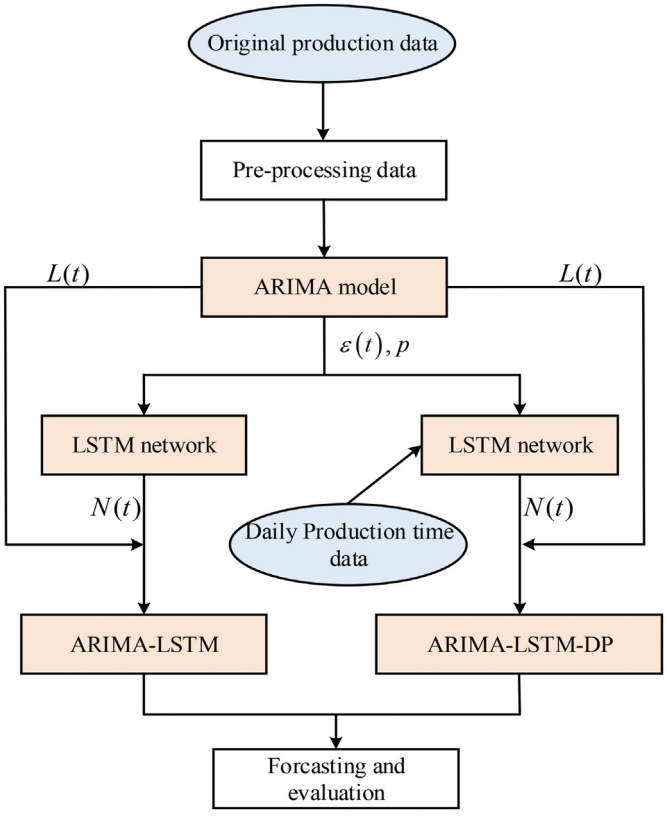


Fig. 3. The working flow of the hybrid models.

methods.

Based on the working flow of the proposed method in Fig. 3, the hybrid models can be divided into four steps: (1) recording of the raw data. The data used in this paper are the original production data from actual oilfields. Hence, their results can show actual well production changes. (2) Linear prediction of the ARIMA model. The statistical model of ARIMA is applied to extract the linear portion L_t of production time series and return the residual items ε_L and auto-regression order p, which are the input terms for the next step. (3) Nonlinear prediction of LSTM modeling. We build two hybrid models with or without considering manual open−shut operation as shown in Fig. 3. In model 1 (ARIMA-LSTM), the residuals from the ARIMA model are the only inputs for the LSTM machine learning model. Hence, we forecast the nonlinear data as$N_\varepsilon = f(\varepsilon(t-1), \varepsilon(t-2), \cdots, \varepsilon(t-p))$. In model 2 (ARIMA-LSTM-DP), the input times series data include the residual terms and daily production time series. Hence, nonlinear data forecasting can be expressed as$N_\varepsilon = f(\varepsilon(t-1), \varepsilon(t-2), \cdots, \varepsilon(t-p); h(t-1), h(t-2), \cdots, h(t-p))$. (4) Coupling and evaluating the final result of ARIMA-LSTM model. The final result of fitting the production time series can be obtained by adding the ARIMA model's forecast results to the LSTM network's forecast results. Subsequently, the performance assessment is conducted.

2.5. Evaluation indicators

To assess the prediction performance under different experimental scenarios, scientific evaluation indicators are selected for the time series prediction. We choose root mean squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and similarity (Sim) as evaluation metrics, which are used for evaluating the performance of different models in forecasting results and can be expressed as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i(t) - y_i(t))} \tag{9}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i(t) - y_i(t)| \tag{10}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_i(t) - y_i(t)}{x_i(t)}\right| \tag{11}$$

$$Sim(x,y) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{1+\frac{|x_i(t)-y_i(t)|}{\max(x)-\min(x)}} \tag{12}$$

where $x_i(t)$ denotes the original well production data value, $y_i(t)$ denotes the production value predicted by different models, n denotes the number of time series. Generally, the lower values of RMSE, MAE, and MAPE lead to better performance of the forecasting task. Furthermore, the Sim value is in the interval [0, 1], and higher Sim values indicate better fitting results.

3. Analysis of well production

In this section, production data from three actual wells are used to illustrate the prediction modeling steps. Based on the working flow process of hybrid models in Fig. 3, we first describe the production datasets, then use ARIMA and LSTM methods to forecast the linear and nonlinear parts, respectively, and finally, the coupling results are obtained and evaluated via different criteria.

3.1. Raw data description

To investigate the prediction performance of the proposed hybrid models, the oil production and daily production time series of three wells are gathered from an oilfield in China, as shown in Fig. 1. With respect to well 1, 2476 observations are recorded from July 20, 2013 to September 1, 2019, and the first 1980 samples are used for training. These samples used for training constitute 80% of the entire time series length. The remaining 496 samples are used for testing. The production time series for well 2 and well 3 include 1608 and 687 samples, respectively. They are also divided into training set and testing set with a ratio of 4:1. The statistical information of the three wells is calculated and shown in Table 1.

Unlike other time series data, well production data are not only affected by geological factors, such as permeability, porosity, and saturation, but also by manual operations, especially the shut-in operations. As shown in Fig. 1, the open-shut operations are frequently one of the main causes that lead to the fluctuations in the well rate curves. Hence, the hybrid model (ARIMA-LSTM-DP), including the daily production time series, is also considered in the paper.

3.2. Linear prediction of ARIMA model

As shown in 2.2, the ARIMA method is applied to the well rate series to obtain the linear predicted values $L(t)$ and their residuals. This consists of three steps: identification, estimation, and prediction. Firstly, the differencing process is applied to ensure that the well rate time series is stationary. This in turn aids in obtaining the degree of integration D, where is unusual to perform the differencing operation for more than two times. Hence, the common

Table 1
The Statistical information of three well production data.

	Time series data	Count	Mean	Min	Max	Standard derivation
Well 1	Production rate	2476	216.617	0	374.97	81.1337
	Daily production time	2476	23.559	0	24	2.476
Well 2	Production rate	1608	241.6439	0	385.78	62.9745
	Daily production time	1608	23.5878	0	24	2.4375
Well 3	Production rate	687	146.2542	0	470.81	100.6431
	Daily production time	687	22.231	0	24	6.0661

value of D corresponds to 0, 1, or 2. To satisfy the stationarity conditions, the D values of these three wells are 1, 1 and 2, respectively. Then, the ACF and PACF curves are calculated to aid the selection of model type, and p and q maximum values are determined. By using the minimum value of the AIC error metric, the orders of p and q are decided for the ARIMA model of each well as shown in Table 2. As observed, the ARIMA (4, 1, 3), ARIMA (1, 1, 1), ARIMA (1, 2, 3) models are obtained to forecast the linear production part of three wells.

After determining the parameters of the ARIMA(p, D, q) model, the production of all three wells is forecasted, as shown in Fig. 4, and the evaluation indicators are calculated and listed in Table 3. In Fig. 4, the linear modeling results of each well, based on the ARIMA method, are shown in two parts. The upper part of each well curve shows the results of the well production training and testing, and the first 80% training data are fitted via the ARIMA model. The remaining 20% testing data are forecasted (the red line denotes the original data, the black line denotes forecasting outcomes, and the black dots indicate 95% probability area). The results indicate that the declining trend of well production is successfully predicted, especially with respect to the steady decrease in well 3 wherein the Sim value is as high as 0.9096. The results of other wells also indicated good performance with a similarity of more than 80%. And all actual production values are within 95% probability range. The RMSE values of three wells are 27.7097, 26.96, and 14.5281, respectively. It shows that the ARIMA method is an effective tool for the relatively stable linear trend, but nonlinear fluctuation is not preferred.

Furthermore, the residual time series are obtained and are shown in the lower part of each well curves. The blue line indicates the fluctuating error wherein the ARIMA model is used to fit the input training data for the subsequent LSTM model, and the black line denotes the residual ARIMA model forecasting of the test set, which is considered as the test data of the LSTM model.

3.3. Nonlinear prediction of LSTM modeling

The LSTM models are developed to discover the nonlinear relations of well production. In the hybrid model (ARIMA-LSTM), we use the residual values that are derived from the ARIMA model as the sole inputs for the LSTM model. In the ARIMA-LSTM-DP model, the daily production time series are added into the inputs for the LSTM model. In this paper, the multi-step ahead prediction models based on one-step ahead prediction are used. This implies that to forecast the value of y(t), p previous value y(t-1), y(t-2), y(t-3),…,y(t-p) are used. Subsequently, to forecast the value of y(t+1), y(t) is used as one of the input values along with the previous terms as mentioned above.

Table 2
The p, D, q values of three well production data.

Well name	p value	D value	q value
Well 1	4	1	3
Well 2	1	1	1
Well 3	1	2	3

To ensure a better fit and prevent the training from diverging, we standardize the training data such that its mean and unit variance corresponds to zero. The normalized parameters are calculated as follows:

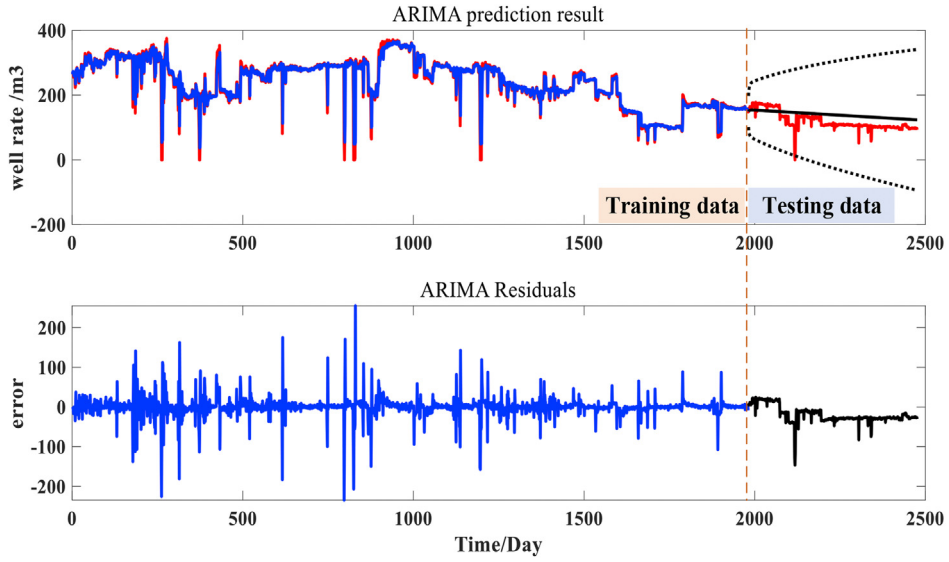
$$X_{t,N} = \frac{X_t - \overline{X_t}}{Std(X_t)} \quad Y_{t,N} = \frac{Y_t - \overline{Y_t}}{Std(Y_t)} \tag{13}$$

where $X_{t,N}$ and $Y_{t,N}$ denote normalized input time series and output production series, respectively, X_t and Y_t denote the original input time series and output production data, $\overline{X_t}$ and $\overline{Y_t}$ denote the average values of the inputs and outputs. Furthermore, $Std(X_t)$ and $Std(Y_t)$ denote the standard deviations.

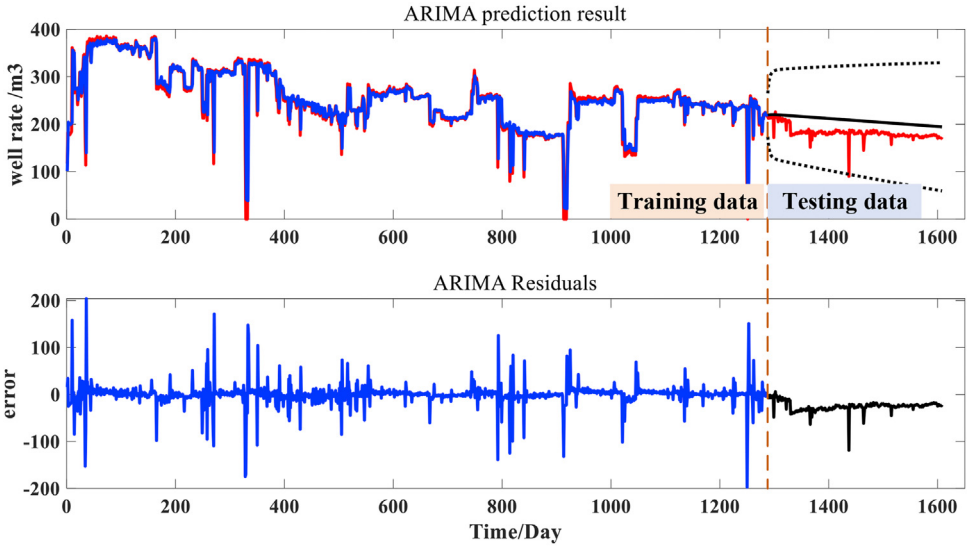
Overfitting is a common problem in a training process wherein the predictive performance on the training data set is good. However, the predictive performance is poor on the newly forecasted data. To prevent this type of problem, a dropout layer and ridge regularization (L2) layer are added in the LSTM model. Our designed neural network architecture consists of input, LSTM, dropout, fully connected layer, L2 regularization, and regression Layer. Due to the complex architecture of the LSTM model, the parameters play an important role in forecasting accuracy. Through trial and error, we set the number of iterations to [100, 200, 300, 400, 500, 600, 700, 800, 900, 1000]. Each iteration is calculated 10 times. The optimal number of iterations is obtained with minimum average RMSE values. In our model, it turns out that it is not the larger of iteration number, the better performs. The final outcome for three wells is set as 100, 200, and 500. The other hyperparameters are applied in the training process with a mini-batch size of 20, an Adam optimization function, a dropout rate of 0.2, an L2-norm regularization parameter of 0.0001, and an initial learning rate of 0.005. The results of the LSTM prediction with and without considering daily production time series are shown in Fig. 5, and the error values are listed in Table 4.

Fig. 5 compares the LSTM predicted results for ARIMA fitting residuals of the three wells. Each well graph indicates that the ARIMA residuals of well 1 and well 2 fluctuate frequently, and the change in well 3 is relatively small. It is evident that the ARIMA-LSTM-DP yields the best results wherein the forecast values are approximately the same as the ARIMA residuals. The non-linear fluctuations, caused by manual operations, can be predicted effectively. For example, in well 3 during time period of 658−667 Days, the ARIMA-LSTM-DP model has obvious grooves that indicate the shut-in operations. The results of the ARIMA-LSTM model are not as good when compared to that of the previous model. However, the ARIMA-LSTM model still captures the overall trend of the ARIMA residual very well. It is difficult to capture severe changes via the ARIMA-LSTM model because the sudden fluctuations are mainly due to manual operations.

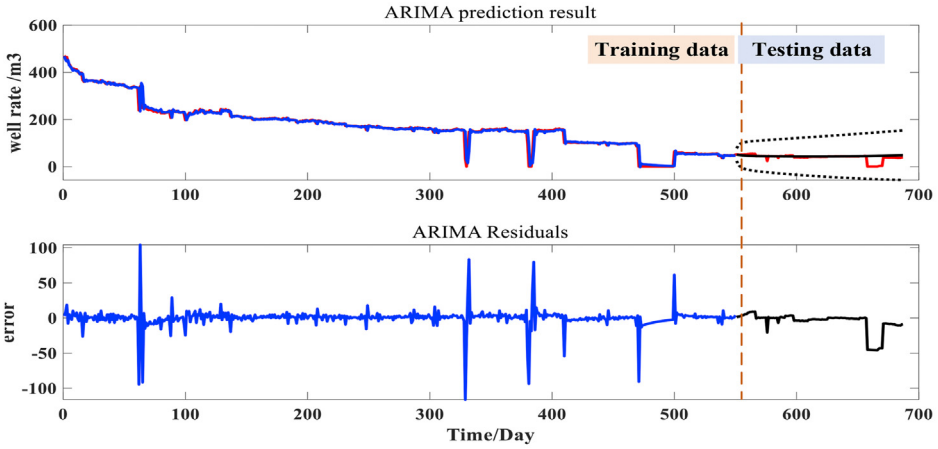
Table 4 lists the errors that are obtained using the LSTM and LSTM-DP algorithms on the ARIMA residuals time series of the three wells. The RMSE and MAE values decrease after using the



(a) ARIMA modeling result of well 1



(b) ARIMA modeling result of well 2



(c) ARIMA modeling result of well 3

Fig. 4. The linear modeling results of ARIMA method.

Table 3
The errors of three well production data using ARIMA.

Well name	RMSE	MAPE	MAE	Sim
Well 1	27.7097	0.1791	24.5913	0.8815
Well 2	26.96	0.1202	24.9592	0.8479
Well 3	14.5281	0.1574	7.1728	0.9096

LSTM-DP hybrid model. It indicates that the LSTM model, which considers oil rate and daily production time series, outperforms the methods of the single oil rate time series. However, the MAPE values of well 2 and 3 increase from 0.6931 to 0.7426, and from 1.7198 to 4.8546, respectively. Similarly, the Sim value of well 3 slightly increases from 0.9093 to 0.9108. On the one hand, the curve becomes more volatile after the manual operations are considered in the model. On the other hand, the values of ARIMA residuals are relatively small around zero, so the denominator values of x_i and max-min difference in the MAPE and SIM expressions are also tiny. Therefore, a more volatile outcome leads to increase of the MAPE and SIM values.

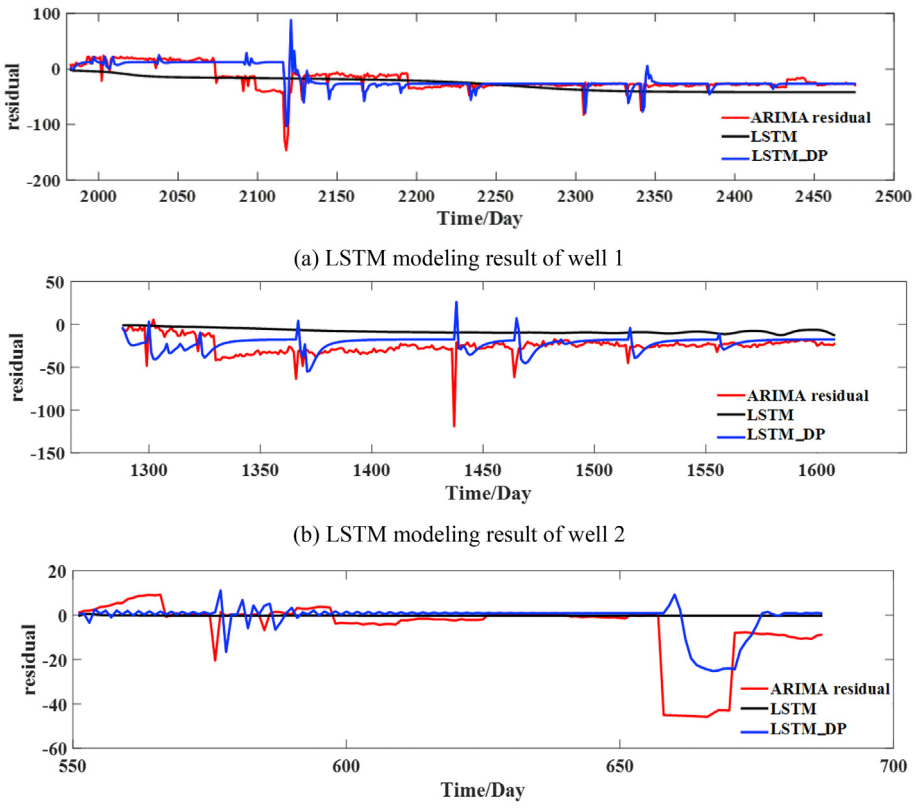
3.4. Coupling prediction results

In this section, the production rates of the three wells, as predicted via different methods, are shown in Fig. 6. The results of the hybrid models involve the coupling of linear part of ARIMA method and nonlinear part of LSTM models. Furthermore, shut-open operations are also considered. In each well graph, three predicted curves are compared with the raw testing data (red line), including individual ARIMA (green line), hybrid ARIMA-LSTM (black line), and hybrid ARIMA-LSTM-DP methods (blue line). As shown in Fig. 6, the ARIMA model can predict the declining trend of the three wells. However, it is impossible to grasp the nonlinear fluctuations. The LSTM model can be used to learn the oscillation information, but the forecasting results exhibit often cyclical changes, as shown in the results of well 2 (Fig. 6b). This does not match the variation in the production of the actual well. Given that the changes in the actual production are mainly due to manual operations, especially frequent open-shut well operation, compound models that consider daily production times are more effective in forecasting well production rates.

Table 4
The errors of three wells residuals using LSTM modeling.

Well name	model	RMSE	MAPE	MAE	Sim
Well 1	LSTM	18.7197	0.7043	14.3244	0.9258
	LSTM-DP	17.7403	0.5703	10.33	0.9479
Well 2	LSTM	19.8856	0.6931	17.3267	0.8815
	LSTM-DP	14.156	0.7416	10.1636	0.9287
Well 3	LSTM	14.4337	1.7198	7.1118	0.9108
	LSTM-DP	11.5840	4.8546	6.5376	0.9093

The predicted cumulative productions of the three wells are shown in Fig. 7. Furthermore, the cumulative production of the compound ARIMA-LSTM-DP model is closer to actual cumulative production when compared to those of the other methods. The output obtained via the ARIMA method is often higher and overestimates the well production. The results of the ARIMA-LSTM model are better than those of the ARIMA model but cannot reflect sudden changes during the production process. Therefore, it is necessary to consider the effect of manual operations in the production forecasting process.

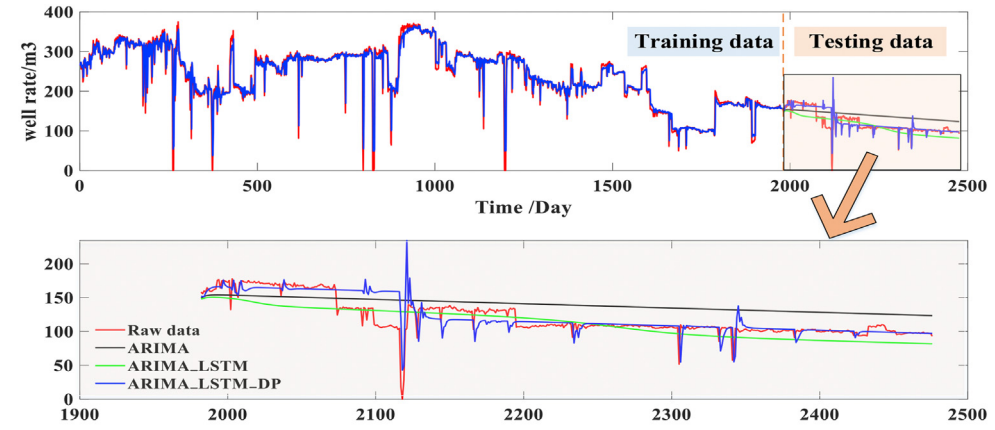


(a) LSTM modeling result of well 1

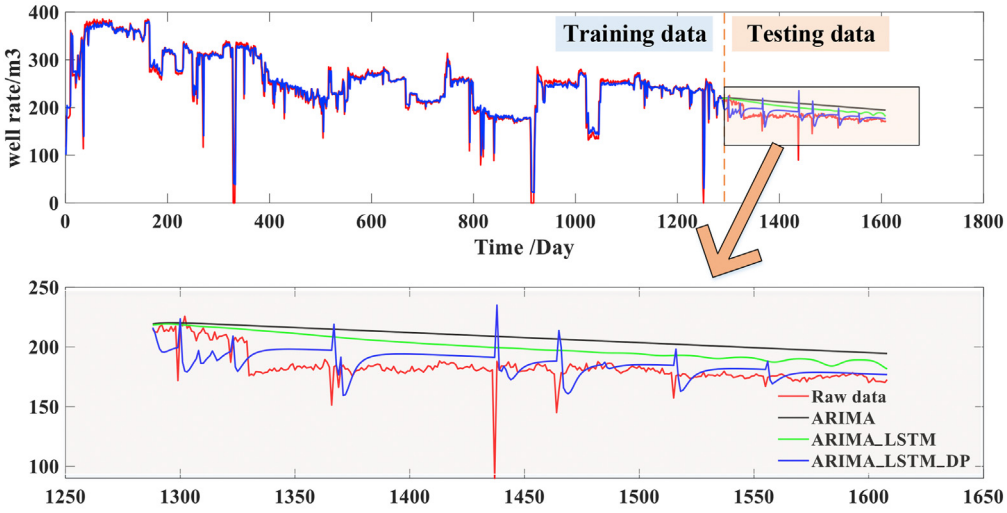
(b) LSTM modeling result of well 2

(c) LSTM modeling result of well 3

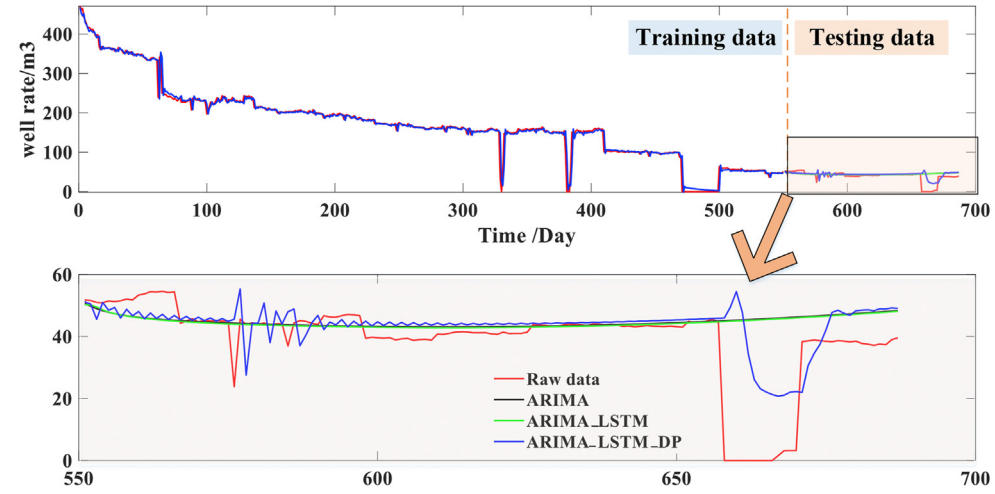
Fig. 5. Comparison of LSTM modeling results for ARIMA residuals.



（a） Comparison of well 1 production forecasting results



（b）Comparison of well 2 production forecasting results



（c） Comparison of well 3 production forecasting results

Fig. 6. Hybrid modeling results for three wells.

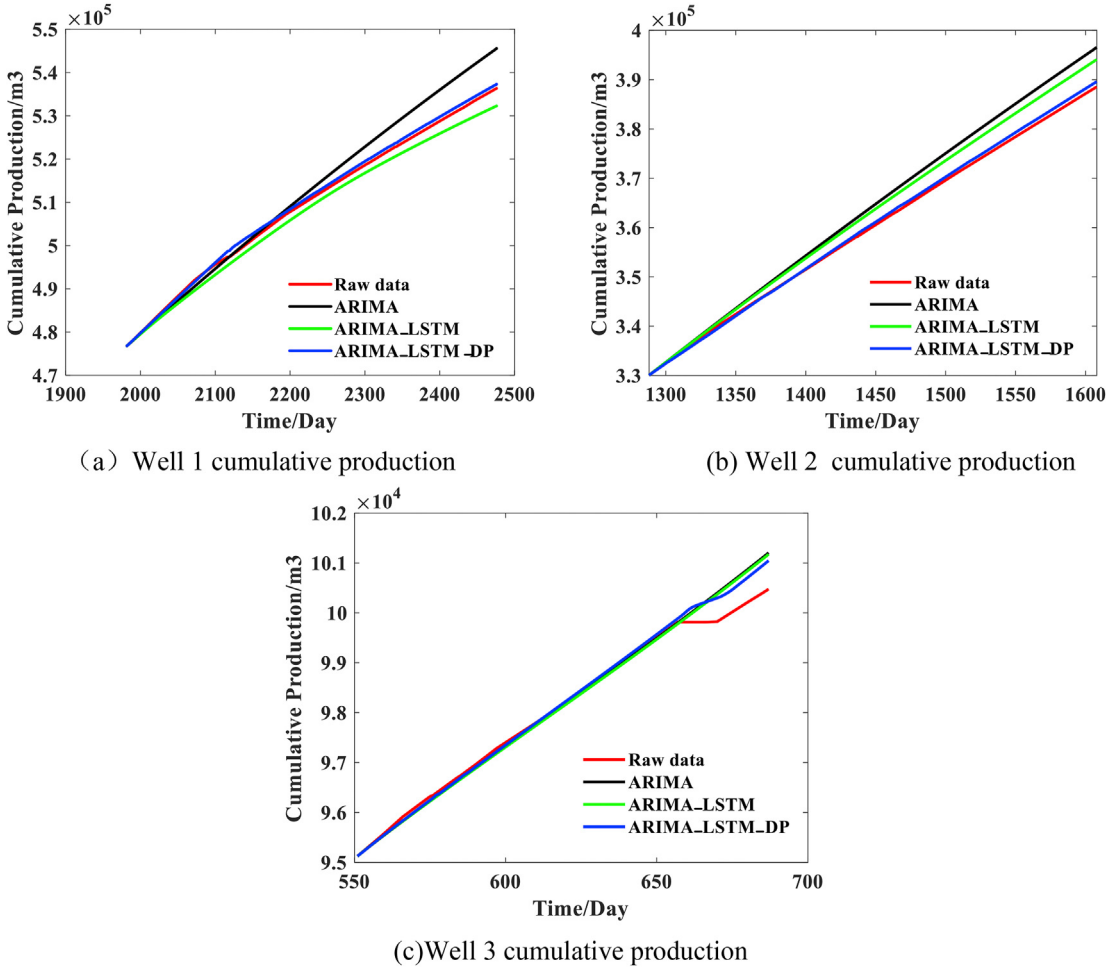


Fig. 7. Comparison of predicted cumulative production for three wells.

4. Comparisons and discussion

To take advantage of linear and nonlinear models, the coupling models of ARIMA-LSTM and ARIMA-LSTM-DP are introduced into oil and gas well production forecasting. Daily production time series are investigated effectively by multivariate inputs in LSTM model, which provide a powerful tool for reservoir engineer to solve the frequent manual operations. In addition to the above three models (ARIMA, ARIMA-LSTM, and ARIMA-LSTM-DP), the individual LSTM method and LSTM model that consider manual operations (LSTM-DP) are used to evaluate the forecasting performances for the three wells. The RMSE, MAPE, MAE, and Sim values of the prediction are calculated. Table 5 lists all the metric values of the average forecast results with respect to different methods, and Fig. 8 shows the metric values across different models.

As we know, the lower values of RMSE, MAE, MAPE, and the higher Sim value mean better performance. Different evaluation indices for three wells across five different models are shown in Fig. 8 and Table 5. The compound model of ARIMA-LSTM-DP exhibits the lowest RMSE, MAPE, and MAE and the highest Sim (with the exception of well 3) values, wherein the Sim value of well 3 is 0.9087, slightly smaller than the optimal value 0.9102 of the ARIMA-LSTM model. Therefore, we can make a conclusion that the proposed model ARIMA-LSTM-DP can be adapted well to predict the oil well production time series, which provides a reliable and effective methodology for engineers to make decisions for improving economic efficiency.

Table 5
The errors of three wells prediction using different methods.

Well	Model	RMSE	MAPE	MAE	Sim
Well1	ARIMA	27.7097	0.1791	24.5913	0.8815
	LSTM	24.7169	0.1675	21.2195	0.8503
	LSTM-DP	29.3475	0.1185	15.8972	0.8971
	ARIMA-LSTM	18.7197	0.1021	14.3244	0.9285
	ARIMA-LSTM-DP	17.7403	0.0722	10.33	0.9497
Well2	ARIMA	26.96	0.1202	24.9592	0.8479
	LSTM	19.9372	0.1137	18.2868	0.878
	LSTM-DP	18.139	0.1065	17.2754	0.8884
	ARIMA-LSTM	19.8856	0.0831	17.3267	0.89
	ARIMA-LSTM-DP	14.1560	0.0481	10.1636	0.9339
Well3	ARIMA	14.5281	0.1574	7.1728	0.9096
	LSTM	19.8972	0.5214	17.3648	0.7725
	LSTM-DP	20.8056	0.3315	15.785	0.7977
	ARIMA-LSTM	14.4337	0.1560	7.1118	0.9102
	ARIMA-LSTM-DP	11.5840	0.1443	6.5376	0.9087

Compared with well 1 and 2, well 3 has less frequent manual operations (as shown in Fig. 1). And the corresponding production time series show a smoother declining trend. Under the circumstances with less non-linear manual operations, the traditional linear statistical model ARIMA performs well with the values of RMSE, MAPE, MAE and SIM being 14.5281, 0.1574, 7.1728 and 0.9096, respectively. It is obviously better than the single non-

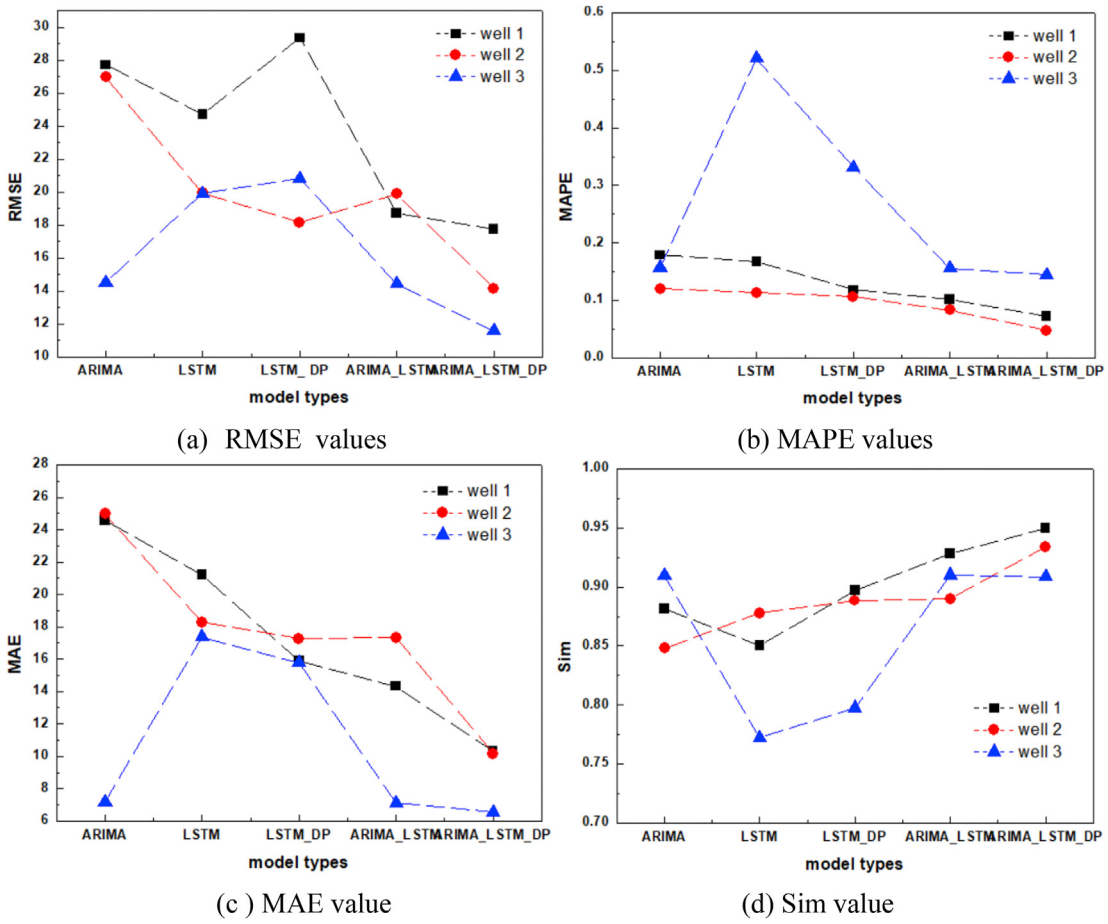


Fig. 8. The metric evaluation indicators of five models performance.

linear LSTM model (the RMSE, MAPE, MAE and SIM are 19.8972, 0.5214, 17.3648 and 0.7725, respectively) and LSTM-DP model (the RMSE, MAPE, MAE and SIM are 20.8056, 0.3315, 15.785 and 0.7977, respectively). The forecasting performance improves significantly after using the hybrid model of ARIMA and LSTM, including the ARIMA-LSTM model (the RMSE, MAPE, MAE and SIM are 14.4337, 0.1560, 7.1118 and 0.9102, respectively) and ARIMA-LSTM-DP model (the RMSE, MAPE, MAE and SIM are 11.5840, 0.1443, 6.5376 and 0.9087, respectively). The reason is that the well 3 production time series is a stable linear decline without frequent non-linear interference, the individual ARIMA method is more effective to grasp linear features than LSTM method. The hybrid models (ARIMA-LSTM and ARIMA-LSTM-DP) take advantages of linear and nonlinear models, which exhibit an increase in performance to the individual ARIMA and LSTM models. Meanwhile, after considering manual operations in well 3, the coupled model of ARIMA-LSTM-DP has no obvious improvement than the ARIMA-LSTM model. This also demonstrates that fewer non-linear operations have a minor impact on the production of well 3.

With respect to well 1 and 2, there are more frequent manual operations (as shown in Fig. 1), so the corresponding production time series show more dramatic nonlinear fluctuations. At this time, the LSTM model performs better than the ARIMA model. For example, in well 1, the RMSE, MAPE, MAE and SIM values of the LSTM model are 24.7169, 0.1675, 21.2195 and 0.8503, respectively, while the corresponding values of ARIMA model are 27.7097, 0.1791, 24.5913 and 0.8815, respectively. Simultaneously, the performances of composite models (ARIMA-LSTM and ARIMA-LSTM-DP) are also more stable and better than single ARIMA and LSTM models. This is because the LSTM deep learning is good at capturing nonlinear production features by manual operations, while the ARIMA model cannot. So the hybrid models perform better by taking advantage of linear and nonlinear models to improve the forecasting performance. Meanwhile, after considering manual operations in well 1and well 2, the coupled model performance of ARIMA-LSTM-DP has obvious promotion than the ARIMA-LSTM model. This demonstrates that frequent non-linear operations have a bigger impact on productions of well 1 and 2.

In summary, compared with the single ARIMA and LSTM models, the hybrid models exhibit higher prediction accuracy because they separate the time series into a linear part for the ARIMA model and a nonlinear part for the LSTM model. When the well production time series are smoothly decreasing and the well has less interference by manual operations during the production period, the ARIMA model exhibit better than the LSTM model to capture this linear decline, and the influence of manual operation is not obvious. On the contrary, the LSTM model perform better for nonlinear fluctuation, and higher impact caused by manual operations. Overall, the coupled model ARIMA-LSTM-DP is more adaptable and possesses higher efficiency, which can be used for guiding engineers to choose the best method for production forecasting.

5. Conclusion

The focus of the study involves providing a reliable and accurate

model for actual well production time series forecasting, which is very important albeit a challenging task in petroleum engineering. We adopt the ARIMA-LSTM hybrid model wherein ARIMA models the linear part and the LSTM recurrent neural network predicts the nonlinear part. Considering the nonlinear fluctuations of production curves due to manual operations, the compound ARIMA-LSTM-DP model is developed to obtain a better prediction accuracy where daily production time series are added to the inputs for the LSTM modeling.

The testing results of three actual wells' production time series indicate that the performance of hybrid models (ARIMA-LSTM, ARIMA-LSTM-DP) are better and more reliable than the single traditional methods, i.e., ARIMA, LSTM, and LSTM-DP. In addition, when the production data is affected by frequent manual operations as in well 1 and well 2, the ARIMA-LSTM-DP model has a better performance than the hybrid model of ARIMA-LSTM. And single model LSTM performs better than the ARIMA model to capture nonlinear fluctuation in well 1 and well 2. Conversely, in well 3, the production data is smoothly decreasing with fewer manual operations. The ARIMA-LSTM-DP model has little improvement compared to the ARIMA-LSTM model. And the individual ARIMA model has better performance than the LSTM model. Overall, the ARIMA-LSTM-DP hybrid model is more outstanding and reliable than the hybrid ARIMA-LSTM model as well as the single traditional methods, i.e., ARIMA, LSTM, and LSTM-DP.

Author statement

Dongyan Fan: Conceptualization, Methodology, Software, Writing- Reviewing and Editing. Hai Sun: Supervision, Methodology, Writing – original draft. Jun Yao: Supervision, Investigation, Funding acquisition. Kai Zhang: Formal analysis, Data curation. Xia Yan: Software, Validation. Zhixue Sun: Visualization, Resources, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors are thankful for financial support from the National Natural Science Foundation of China (Grant No.51774308, No.51774317), the Natural Science Foundation of Shandong Province, China (Grant No. ZR2020ME087), and the Science and Technology Support Plan for Youth Innovation of University in Shandong Province, China (Grant No. 2019KJH002). The authors would also like to appreciate the reviewers and editors whose critical suggestions were very helpful to this article.

References

[1] Waqas HH. Estimates of total oil &gas reserves in the world, future of oil and gas companies and SMART investments by E&P companies in renewable energy sources for future energy needs.the international petroleum conference 2020; Dhahran, Sudio Arabia.
[2] Shuai Y, Chao H, Wu ZH, Li QM. Developmental characteristics, influencing factors and prediction of fractures for a tight gas sandstone in a gentle structural area of the Ordos Basin, China. J Nat Gas Sci Eng 2019;72:103032.
[3] Muhanmmad RA, Malcolm G, Rigby Sean P. Numerical simulation of the impact of geological heterogeneity on performance and safety of Thai heavy oil production process. J Petrol Sci Eng 2019;173:1130–48.
[4] Kazemi H, Merrill LS, Porterfield KL, et al. Numerical simulation of water-oil flow in naturally fractured reservoirs. SPE J 1976;16(6):317–26.
[5] Mohammad M, Vatani A, Majid S, et al. Parallel processing of numerical simulation of two-phase flow in fractured reservoirs considering the effect of natural flow barriers using the streamline simulation method. Int J Heat Mass Tran 2019;131:574–83.
[6] Zhang J, Liu X, Chen Z, et al. Numerical simulation of the improved gas production from low permeability hydrate reservoirs by using an enlarged highly permeable well wall. J Petrol Sci Eng 2019;183:106404.
[7] Nwaobi U, Anandarajah G. Parameter determination for a numerical approach to undeveloped shale gas production estimation: the UK Bowland shale region application. J Nat Gas Sci Eng 2018;58:80–91.
[8] Chen KP. Production from a fractured well with finite fracture conductivity in a closed reservoir: an exact analytical solution for pseudosteady-state flow. SPE J 2016;21(2):550–6.
[9] Ji J, Yao Y, Huang S, et al. Analytical model for production performance analysis of multi-fractured horizontal well in tight oil reservoirs. J Petrol Sci Eng 2017;158:380–97.
[10] Mohammad B, Morteza D, Sohrab Z. Semi-analytical solution for productivity evaluation of a multi-fractured horizontal well in a bounded dual-porosity reservoir. J Hydrol 2020;581:124288.
[11] Henrik W, Linnea L, Kjell A, Mikael H. Production decline curves of tight oil wells in Eagle ford shale. Nat Resour Res 2017;26(3):365–78.
[12] Zhang H, Dean R, Adam C, et al. Extended exponential decline curve analysis. J Nat Gas Sci Eng 2016;36:402–13.
[13] Gupta S, Fuehrer F, Benin C. Production forecasting in unconventional resources using data mining and time series analysis. In: The SPE/CSUR unconventional resources conference; 2014. Alberta, Canada.
[14] Amirmasoud KD, Shahab M, Soodabeh E. Coupling numerical simulation and machine learning to model shale gas production at different time resolutions. J Nat Gas Sci Eng 2015;25:380–92.
[15] Geng ZQ, Li YN, Han YM, Zhu QX. A novel self-organizing cosine similarity learning network: an application to production prediction of petrochemical systems. Energy 2018;142:400–10.
[16] Box GEP, Jenkins GM, Reinsel GC. Time series analysis forcasting and control. 1994. p. 238–42. vol. 37(2) Oakland, California.
[17] Wang YJ, Wang GZ, Dong Y. Application of residual modification approach in seasonal ARIMA for electricity demand forecasting: a case study of China. Energy Pol 2012;48:284–94.
[18] Mishra P, Sarkar C, Vishwajith KP, et al. Instability and forecasting using ARIMA model in area, production and productivity of onion in India. Journal of Crop and Weed 2013;9(2):96–101.
[19] Hossain MM, Faruq A. Forecasting the sugarcane production in Bangladesh by ARIMA model. J Stat Appl 2015;4(2):297–303.
[20] Shamsnia SA, Shahidi N, Ali L, et al. Modeling of weather parameters using stochastic methods (ARIMA model)(case study: Abadeh region, Iran). In: International conference on environment and industrial innovation IPCBEE; 2011. Singapore.
[21] Ayeni B, Pilat R. Crude oil reserve estimation: an application of the autoregressive integrated moving average (ARIMA) model. J Petrol Sci Eng 1992;8:13–28.
[22] Yusof NM, Ruzaidah SA, Zamzuiani M, et al. Malaysia crude oil production estimation: an application of ARIMA model. In: IEEE. Malaysia: Kuala Lumpur; 2010.
[23] Sun J., Ma X., Kazi M. Comparison of Decline curve analysis DCA with recursive neural networks RNN for production forecast of multiple wells. The SPE western regional meeting 2018; California, USA.
[24] Hamzeh A, Hamid R, Nancy C. Multivariate time series modelling approach for production forecasting in unconventional resources. Denver Colorado, USA: The SPE Annual Technical Conference & Exhibition; 2020.
[25] Xuanyi S, Yuetian L, Liang X, et al. Time-series well performance prediction based on Long Short-Term Memory (LSTM) neural network model. J Petrol Sci Eng 2020;186:106682.
[26] Sagheer A, Mostafa K. Time series forecasting of petroleum production using deep LSTM recurrent networks. Neurocomputing 2019;323:203–13.
[27] Hochreiter S, Jrgen S. Long short term memory. Neural Comput 1997;9(8):1735–80.
[28] Kyungbook L, Jungtek L, Daeung Y. Prediction of shale-gas production at duvernay formation using deep-learning algorithm. SPE J 2019;24(6):2423–37.
[29] Song X, Liu YT, Xue L, et al. Time-series well performance prediction based on Long Short-Term Memory (LSTM) neural network model. J Petrol Sci Eng 2020;186:106682.
[30] Shahram MB, Mehdi S. An imperialist competitive algorithm artificial neural network method to predict oil flow rate of the wells. Int J Comput Appl 2011;26(10):47–50.
[31] Fargana A, Yadigar I. Development of oil production forecasting method based on deep learning. Stat., Optim. Inf. Comput. 2019;7:826–39.
[32] Li YJ, Sun RX, Roland H. Deep learning for well data history analysis. In: SPE annual technical conference and exhibition. Calgary, Alberta: Canada; 2019.
[33] Shaban H, Tavoularis S. Measurement of gas and liquid flow rates in two-phase pipe flows by the application of machine learning techniques to differential pressure signals. Int J Multiphas Flow 2014;67:106–17.
[34] Zhang JY, Cheung SW, Efendiev Y, et al. Deep model reduction-model learning for reservoir simulation. In: SPE reservoir simulation conference. Texas, USA: Galveston; 2019.
[35] Klie H, Florez H. Data Driven prediction of unconventional shale reservoir dynamics. SPE J 2020;25(5):2564–81.
[36] Liu W, Liu D, Gu J. Forecasting oil production using ensemble empirical model

decomposition based Long Short-Term Memory neural network. J Petrol Sci Eng 2020;189:107013.
[37] Karasu S, Altan A, Bekiros S, Ahmad W. A new forecasting model with wrapper-based feature selection approach using multi-objective optimization technique for chaotic crude oil time series. Energy 2020;212:118750.
[38] Altan A, Karasu S. The effect of kernel values in support vector machine to forecasting performance of financial time series and cognitive decision making. The Journal of Cognitive Systems 2019;4(1):17–21.
[39] Zhang XG, Lei YY, Chen H, et al. Multivariate time series modeling for forecasting sintering temperature in rotary kilns using DCGNet. 2020. https://doi.org/10.1109/TII.2020.3022019.
[40] Mi XW, Liu H, Li YF. Wind speed prediction model using singular spectrum analysis, empirical mode decomposition and convolutional support vector machine. Energy Convers Manag 2019;180:196–205.
[41] Altan A, Karasu S, Bekiros S. Digital currency forecasting with chaotic meta-heuristic bio-inspired signal processing techniques. Chaos, Solit Fractals 2019;126:325–36.
[42] Phan TTH, Nguyen XH. Combining statistical machine learning models with ARIMA for water level forecasting: the case of the Red river. Adv Water Resour 2020;142:103656.
[43] Temür AS, Akgün M, Temür G. Prediction housing sales in Turkey using ARIMA, LSTM and Hybrid models. J Bus Econ Manag 2019;20(5):920–38.
[44] Ma R, Li ZL, Breaz E, et al. Data-fusion prognostics of proton exchange membrane fuel cell degradation. IEEE Trans Ind Appl 2019;55(4):4321–31.
[45] Sun Y, Zhao Z, Ma X, Du Z. Hybrid model for efficient anomaly detection in short-timescale GWAC light curves and similar datasets. Program Comput Software 2019;45(8):600–10.
[46] Ji L, Zou YC, He KJ, Zhu BZ. Carbon futures price forecasting based with ARIMA-CNN-LSTM model. Procedia Computer Science 2019;162:33–8.
[47] Bagheri M, Zhao H, Sun M, Li H. Data conditioning and forecasting methodology using machine learning on production data for a well pad. In: The offshore technology conference; 2020. Houston, TX, USA.
[48] Shibata R. Selection of the order of an autoregressive model by akaike's information criterion. Biometrika 1976;63(1):117–26.
[49] Hurvich CM, Tsai CL. Regression and time series model selection in small samples. Biometrika 1989;76(2):297–307.
[50] Aho K, Derryberry D, Peterson T. Model selection for ecologists: the worldviews of AIC and BIC. Ecology 2014;95(3):631–6.
[51] Haseyama M, Kitajiman H. An arma order selection method with fuzzy reasoning. Signal Process 2001;81:1331–5.