ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

Research on the Application of Integrated Learning Models in Oilfield Production Forecasting

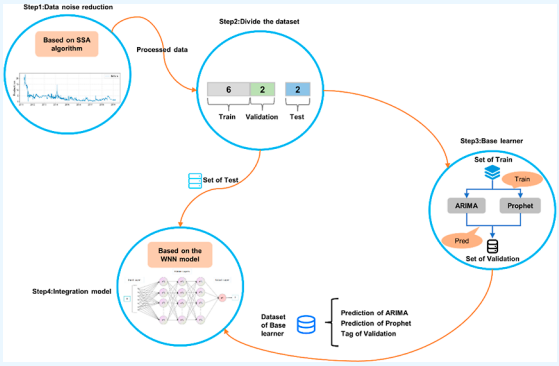
MingCheng Ni, XianKang Xin,* GaoMing Yu,* Yu Liu, and YuGang Gong

Cite This: *ACS Omega* 2023, 8, 39583−39595

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Forecasting oil production is crucially important in oilfield management. Currently, multifeature-based modeling methods are widely used, but such modeling methods are not universally applicable due to the different actual conditions of oilfields in different places. In this paper, a time series forecasting method based on an integrated learning model is proposed, which combines the advantages of linearity and nonlinearity and is only concerned with the internal characteristics of the production curve itself, without considering other factors. The method includes processing the production history data using singular spectrum analysis, training the autoregressive integrated moving average model and Prophet, training the wavelet neural network, and forecasting oil production. The method is validated using historical production data from the J oilfield in China from 2011 to 2021, and compared with single models, Arps model, and mainstream time series forecasting models. The results show that in the early prediction, the difference in prediction error between the integrated learning model and other models is not obvious, but in the late prediction, the integrated model still predicts stably and the other models compared with it will show more obvious fluctuations. Therefore, the model in this article can make stable and accurate predictions.



1. INTRODUCTION

Oil is a crucial strategic resource that supports the sustained development of humanity and the prosperity of nations. It plays a vital role in advancing social progress and economic growth, as well as enhancing national defense security.[1] Effective exploitation of this precious resource is of the utmost significance.

Improving economic efficiency through efficient energy use is a key aspect of oilfield production development. Maintaining a stable and efficient rate of oil production is essential for achieving this goal. Forecasting oilfield production serves as a foundation for the scientific management of oilfields and the creation of production plans.[2] These forecasts can be utilized to regulate oilfield production operations and adjust production levels to meet the changing economic market demands. The ultimate aim is to attain maximum economic benefits while ensuring the effective utilization of oil and energy resources.

In the realm of oil production forecasting, the existing literature primarily employs conventional reservoir engineering methods and statistical techniques such as Arps[3−6] and autoregressive integrated moving average (ARIMA).[7,8] However, these methods are limited by their reliance on linear assumptions and are unable to identify the underlying nonlinear properties of oil production data. As a result, relying solely on these traditional, statistical, and econometric methods may result in inadequate predictive performance. Thus, it is clear that these traditional approaches are inadequate for oil production forecasting.

With the advancements in forecasting algorithms,[9] nonlinear and artificial intelligence methods have gained increasing popularity in oil production forecasting. Techniques such as artificial neural networks (ANN),[10,11] genetic algorithms,[12] and long and short-term memory (LSTM)[13,14] have emerged as popular approaches. Negash and Yaw[11] proposed an ANN-based prediction model for water flooding reservoirs. AlKhammash[12] developed an optimized gradient model for crude oil production prediction that incorporates genetic algorithms. Sagheer and Kotb[13] applied a deep-length short-time memory structure for time series oilfield production prediction. Ning et al.[15] treated production data as time series data and investigated and compared three different algorithms to address the limitations of traditional production forecasting methods: ARIMA, LSTM, and Prophet. Ibrahim et al.[16] employed various artificial intelligence techniques, including

Received: July 26, 2023
Accepted: September 25, 2023
Published: October 10, 2023

https://doi.org/10.1021/acsomega.3c05422
ACS Omega 2023, 8, 39583−39595

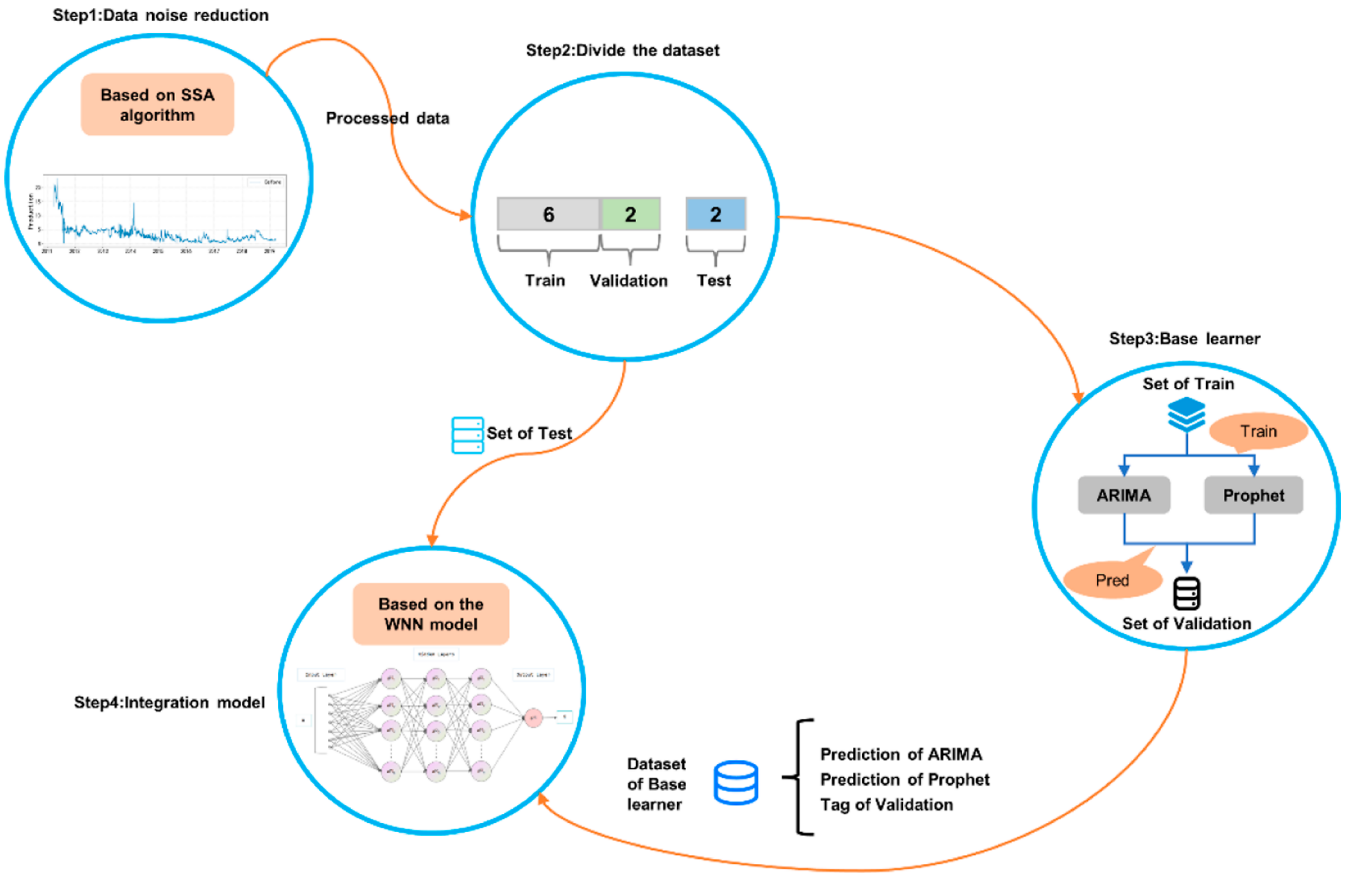


Figure 1. Flowchart of the model in this paper.

support vector machines (SVM) and random forests (RF), to predict production from wells with a high gas−oil ratio and water-cut. Khan et al.[17] aimed to provide a straightforward and widely applicable solution by conducting a comparative study between artificial neuro fuzzy inference systems (ANFIS) and SVM algorithms to accurately predict oil production from artificially lifted wells. Al Dhaif et al.[18] utilized ANFIS and function neural (FN) techniques to predict surface production from volatile oil and gas condensate reservoirs, enabling real-time production forecasting without additional operational costs.

However, these methods also have limitations. Nonlinear prediction models based on artificial intelligence are often computationally complex and can be prone to falling into "local minima" or "overfitting". To overcome these issues, hybrid models have been proposed that combine the strengths of different component models and improve the forecasting performance. Liu et al.[19] introduced a LSTM learning method for set and empirical pattern decomposition. Guo et al.[20] proposed a hybrid prediction model that combines an improved grouped data processing method (GMDH) with a back-propagation algorithm. A hybrid approach based on linear, statistical, and machine learning models has been successfully applied in many other fields[21,22] but has not received sufficient attention in well-production forecasting. This model combines the advantages of both linear and nonlinear models to improve the forecasting performance. The ARIMA model is currently the best-known and most effective linear statistical model, which filters out linear trends in time series well. At the same time, machine learning models are suitable for the accurate estimation of complex nonlinear relationships. Therefore, inspired by the success of this hybrid prediction model, an integrated model is proposed, which is based on singular spectrum analysis (SSA),[23] ARIMA,[24−26] Prophet,[15] and wavelet neural network (WNN),[27,28] for predicting daily oil production. The proposed method consists of the following main steps. First, considering the randomness and uncertainty of the raw oil production data, the SSA algorithm, a nonparametric time series preprocessing technique with data-driven and adaptive features, is used in the oil production data processing stage to extract and reduce noise from the main features of the original production time series. Second, the ARIMA model and the Prophet model are used as base learners by using blending integrated learning. The ARIMA model can well obtain periodic and trend information in the time series, while the Prophet model is good at handling the outliers and missing values in the time series. Finally, the WNN model is used as an integrated model to integrate learners' results. The model can handle nonlinear problems well and avoids falling into nonlinear optimization problems like local optimality.

The main contributions of this paper are as follows: (1) theoretically, the effectiveness of the proposed model as a convenient and superior method for production time series forecasting is demonstrated. (2) Practically, a useful tool for oilfield workers to handle complex oilfield production time series is provided. The SSA algorithm effectively reduces noise in time series data; the ARIMA method deals well with stable decreasing curves; the Prophet method handles outliers and missing values; and the WNN method avoids local optimization problems. By integrating these models, this novel proposed

https://doi.org/10.1021/acsomega.3c05422
ACS Omega 2023, 8, 39583−39595

method offers better performance when considering both linear and nonlinear aspects.

2. METHODOLOGY

2.1. Framework for Integrated Forecasting Methods.

As shown in Figure 1, the integrated learning prediction method proposed in this paper involves the following main steps: data noise reduction, segmentation of data, base learner prediction, and integrated prediction.

(1) The SSA algorithm is applied to extract the major features of the original oil production data and apply denoising to the data.

(2) The processed oil production data is partitioned into train, validation, and test data sets; base learners are built for the training set; and then the prediction results of the base learners on the validation set are combined with the real data in the validation set to form the learner data set.

(3) The ARIMA model and the Prophet model are used as base learners. The ARIMA model is good for obtaining periodic trend information in the oil production data, and the Prophet model is good for dealing with outliers and missing values in the oil production data.

(4) The WNN model was chosen as an integrated model to integrate the learners' results, which can handle nonlinear problems well.

2.2. Data Noise Reduction. The first step in the proposed integrated prediction method is data noise reduction. Here, the SSA algorithm is applied to extract the main features of the raw oil production data and reduce the noise in the data. By refining and reconstructing the original time series signal, the SSA effectively identifies its periodic and oscillatory components and creates a new time series that retains the essence of the original signal. This helps in mitigating the complexity of the prediction model, leading to improved accuracy and validity of the oil production prediction.

In general, the standard SSA algorithm is executed as follows: (1) it serves as a segment of length N time series $Y = \{y_1, y_2, \cdots, y_N\}$ for the embedding window width; (2) it performs singular value decomposition; (3) it helps in grouping; and (4) it helps in diagonal averaging. In particular, (1) and (2) are used to decompose the original oil production sequence and (3) and (4) are used to reconstruct the decomposed signal.

The pseudocode of the SSA algorithm is as follows.

Code for singular spectrum analysis

/* Decomposition phase */

$K = N - L_w + 1$

$$X = \begin{bmatrix} y_1 & \cdots & y_{N-L_w+1} \\ \vdots & \ddots & \vdots \\ y_{L_w} & \cdots & y_N \end{bmatrix}_{L_w \times K}$$

/*SVD*/

$[U, S, V] = svd(H)$

/* Reconfiguration phase */

Dividing two disjoint subsets $I = \{i_1, i_2, \cdots, i_d\}, I_1 = \{1,2,\cdots, r\}$ *and* $I_2 = \{r+1, r+2, \cdots, m\}$

$$X_1 = X_{I_1} + X_{I_2} = \sum_{i=1}^{r} \sqrt{\lambda_i} U_i V_i^T + \sum_{i=r+1}^{m} \sqrt{\lambda_i} U_i V_i^T$$

/* Diagonal average */

$\tilde{Y}(1) = \{\tilde{y_1}(1), \tilde{y_2}(1), \cdots, \tilde{y_N}(1)\}$ *and* $\tilde{Y}(2) = \{\tilde{y_1}(2), \tilde{y_2}(2), \cdots, \tilde{y_N}(2)\}$

Input:

$Y = \{y_1, y_2, \cdots, y_N\}$—Original time series

Output:

$\tilde{Y}(1) = \{\tilde{y_1}(1), \tilde{y_2}(1), \cdots, \tilde{y_N}(1)\}$—Time series after decomposition and noise reduction

Parameters:

N—Length of time series

L_w—Window width, [1,N]

r—Number of PCs

Y—Original time series

X—Track matrix

2.3. Segmentation of Data. The processed yield data is divided into three data sets: train, validation, and test. The ARIMA and Prophet models are then constructed using the train data set. The prediction results of these two models are computed on the validation data set, and the label values in the validation data set are combined to form a learner result data set. This learner result data set is once again divided into train and validation data sets, and the WNN model is constructed using the train data set. The end of training for the WNN model is determined based on the prediction results from the validation data set. Finally, the prediction accuracy of the integrated model is evaluated on the test data set.

2.4. Base Learner Prediction. In this study, ARIMA and Prophet models are employed as base learners in the integrated prediction method. The ARIMA method is shown to be effective in capturing periodic and trend-related information present in the time series data, while the Prophet model is demonstrated to be well-suited for handling outliers and missing values in such data.

2.4.1. ARIMA. The ARIMA model is a popular time series forecasting technique that combines an autoregressive (AR) term, a moving average (MA) term, and a differencing operation to model and forecast the future values of a time series. The ARIMA (p, d, and q) model comprises an AR of order p, an MA term of order q, and a difference order of d, applied to the original time series to make it smoother.

2.4.1.1. AR Model. This model describes the relationship between oil production data at the current moment and historical oil production data. It uses historical data to make predictions about the data at the current moment. A p-order AR model with a perturbation term can be written as

$$X_t = \propto_1 X_{t-1} + \propto_2 X_{t-2} + \cdots + \propto_p X_{t-p} + u_t = \sum_{i=1}^{p} \propto_i X_{t-i} + u_t \quad (1)$$

where: X_t is the water quality data at time t; p denotes the number of moments used for prediction; u_t denotes the random perturbation term; and $\propto_i$ is the model parameter.

2.4.1.2. MA Model. This model considers the oil production series data to be smooth; the random disturbance term u_t in the AR model is regarded as a MA term of order q. The MA model expression is as follows

https://doi.org/10.1021/acsomega.3c05422
ACS Omega 2023, 8, 39583−39595

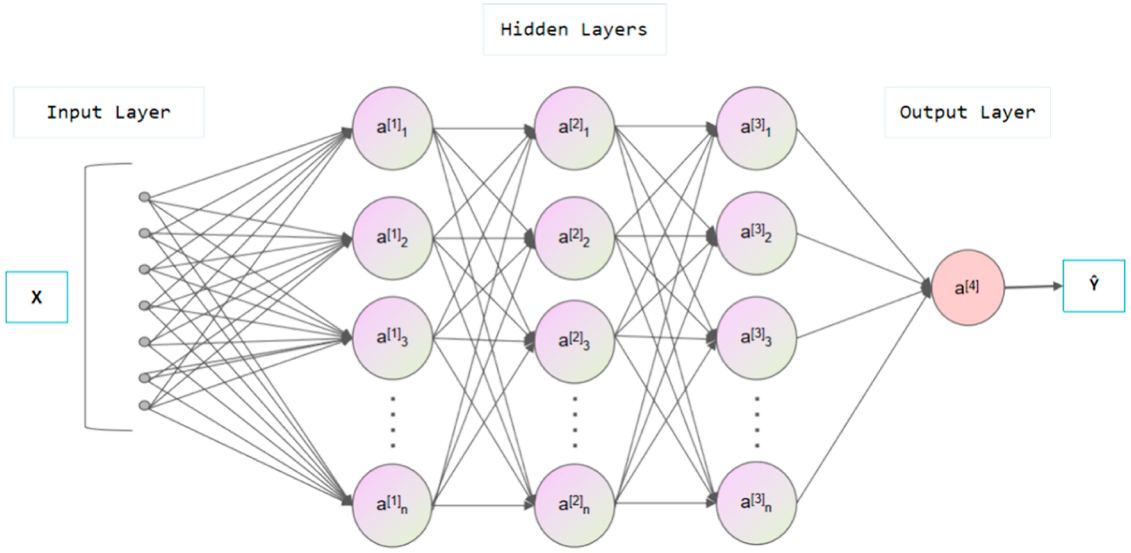


Figure 2. Structure of WNNs.

$$u_t = \varepsilon_t + \beta_1\varepsilon_{t-1} + \beta_2\varepsilon_{t-2} + \cdots + \beta_q\varepsilon_{t-q} = \varepsilon_t + \sum_{j=1}^{q}\beta_j\varepsilon_{t-j} \tag{2}$$

where: ε_t is the white noise series; q denotes the number of moments affected by the white noise series; and β_j is the model parameter.

2.4.1.3. ARIMA Model. For nonstationary oil production data using differencing to obtain a stationary series, where the number of differences is d. Combining AR(p) and MA(q) gives the expression for the ARIMA (p, d, and q) model

$$X_t = \varepsilon_t + \sum_{i=1}^{p}\propto_i X_{t-i} + \sum_{j=1}^{q}\beta_j\varepsilon_{t-j} \tag{3}$$

2.4.2. Prophet. The Prophet model is a highly flexible and efficient forecasting tool for time series data. It effectively handles missing values and outliers, making it ideal for forecasting oil production data. This open-source model has been designed specifically to handle time series data, providing robust results for a variety of forecasting tasks.

The Prophet model is expressed in the following form

$$x(t) = g(t) + s(t) + h(t) + \varepsilon_t \tag{4}$$

where: $x(t)$ denotes the daily oil production data at moment t; $g(t)$ denotes the trend term, which is the part of the time series that shows a nonperiodic trend; $s(t)$ denotes the period term, which is the part of the time series that shows a periodic variation; $h(t)$ denotes the holiday term, which is the part of the series that is affected by holidays, as the effect of holidays is small in production forecasting and this term is not considered in this paper; ε_t is the residual term.

2.4.2.1. Trend Terms. In the Prophet algorithm, the trend term can be calculated using two distinct functions: one based on a logistic regression function and the other on a segmented linear function. Taking into account the characteristics of the oil production data itself, this paper uses a segmented linear function to simulate the trend terms.

The model based on the segmented linear function is as follows

$$g(t) = (k + a(t)\delta)\cdot t + (m + a(t)^T\gamma) \tag{5}$$

$$\gamma = (\gamma_1, \gamma_2, \cdots, \gamma_s)^T, \gamma_i = -s_j\delta_j \tag{6}$$

where: k denotes the growth rate; $a(t) \in \{0,1\}^S$ is the indicator function; $\delta \in R^S$ denotes the amount of change in the growth rate; δ_j denotes the amount of change in the growth rate at the time stamp s_j; m denotes the compensation parameter; and S denotes the number of variation points.

2.4.2.2. Periodic Terms. Using the Fourier series to simulate the periodicity of a time series, the periodicity can be written in the following form

$$s(t) = \sum_{n=1}^{N}\left(\alpha_n \cos\left(\frac{2\pi nt}{p}\right) + b_n \sin\left(\frac{2\pi nt}{p}\right)\right) = w(t)\beta \tag{7}$$

$$\beta = (\alpha_1, b_1, \cdots \alpha_N, b_N)^T \tag{8}$$

where: p denotes the period of the time series and β is initialized as $\beta \sim$ normal $(0, \sigma^2)$, with larger values of σ indicating more pronounced seasonal effects and smaller values indicating less pronounced seasonal effects.

2.5. Integrated Forecasting. The WNN model was selected as the integration model in this study due to its ability to effectively handle nonlinear problems and avoid issues associated with nonlinear optimization, such as local optimality. Blending integration learning was used to integrate the results of the base learners.

2.5.1. WNN. The WNN is a combination of wavelet transformation (WT) and ANNs. This network is formed by replacing the discrete wavelet transform coefficients with the weights of an ANN. This combination not only preserves the localized properties of wavelet transform in the time and frequency domains but also incorporates the autonomous learning capability of ANNs. Unlike the traditional back propagation neural network (BPNN), the activation function in WNN is the Morlet function as opposed to the Sigmoid function used in BPNN. Figure 2 depicts the topology of the WNN. The main principles and steps for its application to time series prediction are as follows

Step 1: Network initialization. First, set the input of the network as $X = \{x_1, x_2, \cdots, x_k\}$, and the output of the network as Y

https://doi.org/10.1021/acsomega.3c05422
ACS Omega 2023, 8, 39583−39595

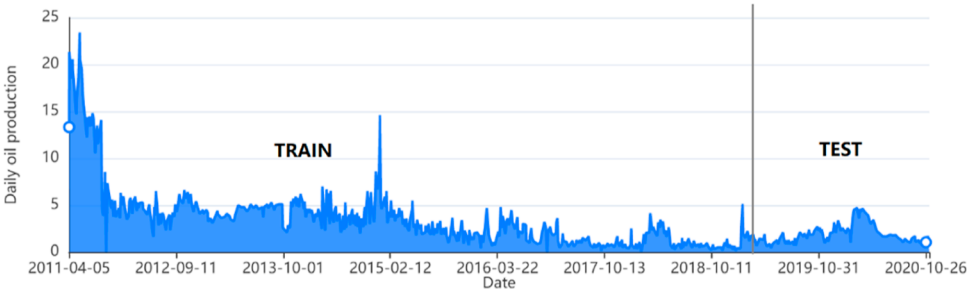


Figure 3. Data on daily oil production from the J oilfield in China.

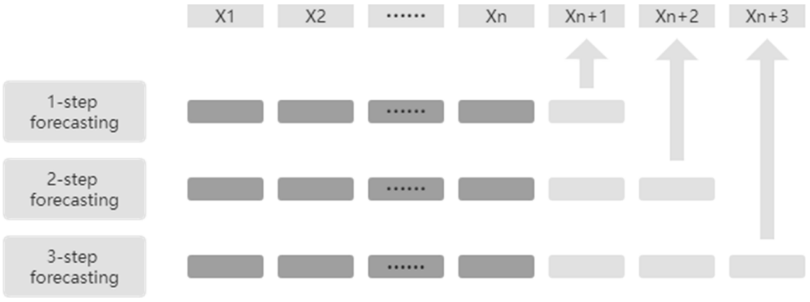


Figure 4. Data structures for multistep prediction.

$= \{y_1, y_2, \cdots, y_m\}$, and set the number of nodes of the input, hidden, and output layers as k, l, and m. In addition, the connection weights w_{ij} and W_{jk} between the input layer−implicit layer and the implicit layer−output layer are initialized, as are the scaling parameter a_j of the wavelet function and the translation parameter b_j of the wavelet function.

Step 2: Calculation of the output of the implied layer. The output h_j of the implied layer is calculated as follows

$$h_j = F_j\left\{\frac{\sum_{i=1}^{k} w_{ij}y_i - b_j}{a_j}\right\}, \qquad j = 1, 2, \cdots, l \tag{9}$$

where: h_j is the output of node j in the hidden layer; b_j is the translation factor of the wavelet basis function h_j; a_j is the scaling factor of the wavelet basis function h_j; and F_j is the wavelet basis function.

Step 3: Calculation of the output layer. The output layer is calculated as follows

$$y'_k = \sum_{j=1}^{l} W_{jk}h_j, \qquad k = 1, 2, \cdots, m \tag{10}$$

Step 4: Network error calculation. The error between the output layer and the original time series is calculated with the following equation

$$e_k = y_k - y'_k, \qquad k = 1, 2, \cdots, m \tag{11}$$

Step 5: Weights' update. The connection weights w_{ij} and W_{jk} between the input layer−implicit layer and the implicit layer−output layer are updated based on the errors calculated in step 4.

3. RESULTS AND DISCUSSION

In this study, daily oil production data from the J oilfield in China was employed as sample data to assess the performance of the integrated model. To provide a comprehensive evaluation, three single prediction models based on the SSA algorithm that comprised the integrated model were selected for comparison. These models are the SSA−ARIMA, the SSA−Prophet, and the SSA−WNN models, respectively. To further demonstrate the advantages of the integrated model, we also conducted comparisons with the Arps model and current mainstream time series forecasting models.

In addition, it is important to note that our model analyzes historical production data and predicts the future production trends without considering the variations in the reservoir conditions.

3.1. Experimental Design. In this section, we describe the details of the data set utilized in the study, including its characteristics and preprocessing steps. We also outline the evaluation criteria and experimental parameters that were used to assess the performance of the integrated prediction model.

3.1.1. Data Descriptions. We utilize daily oil production data from the J oilfield in China as the test data, as depicted in Figure 3. The daily oil production data of the well spans from 2011 to 2021, totaling 3000 observations.

In addressing the issue of data discontinuity caused by missing values in the data set, we employ the Prophet model for imputation. The Prophet model utilizes a segmented linear model to capture the trends in time series data, fitting linear models within each segment to approximate the data trends. When missing values are present in the time series, Prophet utilizes its fitted segmented linear model and seasonal components to impute these missing values.

Next, the processed sample data is partitioned into three subsets: the training set, validation set, and testing set, with a split ratio of 6:2:2. Furthermore, to evaluate the robustness of the integrated model, we perform multistep prediction on the model, the structure of which is presented in Figure 4.

3.1.2. Evaluation Criteria. In the evaluation of the proposed integrated model, four metrics have been selected to assess its performance and efficiency. These metrics, namely, the mean absolute error (MAE), mean square error (MSE), mean absolute percentage error (MAPE), and the improvement ratio of index (I_{index}), are used to compare the results of the

https://doi.org/10.1021/acsomega.3c05422
ACS Omega 2023, 8, 39583−39595

predictive model. The expressions for each index are listed in Table 1. A lower value of MSE, MAE, and MAPE indicates a higher accuracy in prediction.

Table 1. Error Evaluation

evaluation	description	expressions
MAE	average absolute error of N predicted values	$\frac{1}{N}\sum_{i=1}^{N}\lvert x_i^{\text{real}} - x_i^{\text{forecast}}\rvert$
MSE	mean squared error of N predicted values	$\frac{1}{N}\sum_{i=1}^{N}(x_i^{\text{real}} - x_i^{\text{forecast}})^2$
MAPE	mean absolute percentage error of N predicted values	$\frac{1}{N}\sum_{i=1}^{N}\left\lvert\frac{x_i^{\text{real}} - x_i^{\text{forecast}}}{x_i^{\text{real}}}\right\rvert \times 100\%$
I_{index}	predicted rate of improvement of outcome evaluation indicators	$\frac{\text{INDEX}_{\text{comp}} - \text{INDEX}_{\text{prop}}}{\text{INDEX}_{\text{comp}}} \times 100\%$

3.1.3. Setting of Model Parameters. In this article, all methods were implemented in Python and run on a computer equipped with a 2.7 GHz CPU, 16GB RAM, and Windows 11. To mitigate the impact of random factors on the prediction results, each model was executed independently for 10 runs, and the results were averaged over 10 trials. The initial parameters for each algorithm in the proposed integrated model were set as follows

(1) For the SSA algorithm, the singular value decomposition (SVD) process can be computed by calling linalg.svd in NumPy. The window width L_w is generally chosen to be $<N/2$, at most half as long as the time series data of production.
(2) For the WNN model, the architecture was set with 2 nodes in the input layer, 4 nodes in the hidden layer, and 1 node in the output layer. The learning efficiency was set to 0.1, the maximum number of iterations to 100, and the training precision to 0.0001.
(3) For the ARIMA model, according to the criteria of the Akaike information criterion (AIC) and Bayesian information criterion (BIC), a combination of parameters is used such that p, q that minimize the sum of AIC and BIC are used as the parameters of the ARIMA model for this indicator. In this paper, p- and q-values are chosen in the range 0−6 and d-values in the range 0−2.
(4) For the Prophet model, it requires the adjustment of two hyper-parameters. The first one is "changepoint_prior_scale", which determines the elasticity of the trend, and the default value is 0.05. The second hyper-parameter, "seasonality_prior_scale" controls the degree of flexibility of seasonality, and the default value is 10. Cross-validation is employed to optimize both the hyper-parameters.

For the coefficient complexity of the integrated models, as we employ the blending integrated learning approach, which leverages the integration of predictions from multiple base models to enhance the overall performance, the coefficients associated with the involved models are independent. For instance, the order of ARIMA and the relevant parameters of Prophet are examples of such coefficients. The primary coefficient involved in the blending integrated learning approach is the weight coefficient used to combine the predictions from different base models, determining the contribution of each base model to the final prediction.

The allocation of the weight coefficient is typically determined through cross-validation. It involves weighting the predictions of each base model by candidate weight coefficients, calculating errors, and ultimately determining the optimal weight combination based on the actual data set and the predictive direction emphasized by the model. In this article, the weights for ARIMA and Prophet are both set to 0.5.

3.2. Results of Experiment and Discussion. In this section, we present the calculation process of the proposed integrated learning model, followed by prediction experiments performed on real daily oil production data. The results are then compared to various benchmark models, and the comparisons are presented.

3.2.1. Results of SSA and Division of Data. In Section 2, the proposed integrated learning model is described as consisting of four steps: (1) noise reduction of the data using the SSA algorithm; (2) division of the data into training, validation, and testing sets; (3) prediction of the base learner; and (4) prediction of the integration. Figure 5 presents the results of

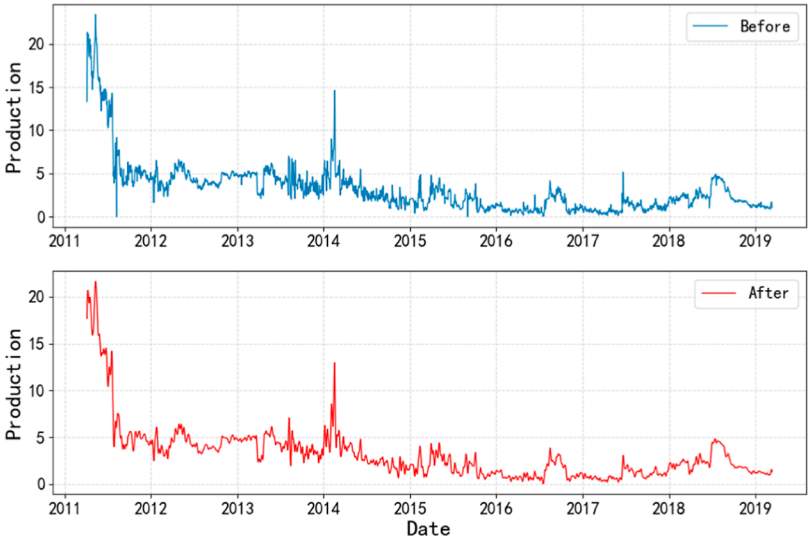


Figure 5. Comparison of data before and after noise reduction by SSA.

https://doi.org/10.1021/acsomega.3c05422
ACS Omega 2023, 8, 39583−39595

the noise reduction step, where it is demonstrated that the denoised data are smoother than the raw data while retaining their overall variability.

In the second step of the proposed integrated learning model, the noise-reduced data are divided into three parts: training, validation, and test data sets. The proportion of the training and testing sets to be split is a crucial factor that affects the performance of the model, but there is currently no theoretical standard. If the sample size is too small, it can result in poor performance of the model, while if it is too large, it can lead to overfitting problems during model training. In light of this, a 6:2:2 split ratio was chosen with the first 1740 daily oil production data serving as the training set and the next 1160 daily oil production data serving as the validation and test sets. The statistical results are presented in Table 2.

Table 2. Specific Division of Sample Data

	sets of train	sets of validation	sets of test
ARIMA, Prophet	(1740) 2011−2016	(580) 2016−2019	(580) 2019−2021
	sets of train	sets of validation	sets of test
integrated learning model	(464) 2016−2018	(116) 2018−2019	(580) 2019−2021

3.2.2. Results of ARIMA. Before building the ARIMA model, it is crucial to test the smoothness of the training data to avoid the occurrence of the pseudoregression phenomenon, which can render the model useless in practical applications. In this study, we assess the smoothness of the series by examining the autocorrelation function (ACF), partial autocorrelation function (PACF) plots, and augmented Dickey−Fuller test (ADF) unit roots. As presented in Figure 6, the autocorrelation plot shows slow decay to 0 without tail dragging or truncation, indicating that the series is initially nonstationary. The ADF test results in Table 3 reveal a p-value significance level greater than 0.05, which does not pass the significance test, further confirming that the series is not smooth.

After conducting a difference operation on the training data of this sequence, the first-order difference plot is displayed in Figure 7. The ACF and PACF plots of the resulting series are illustrated in Figure 8. The autocorrelation and partial autocorrelation plots demonstrate a rapid decay to zero without any noticeable trends, suggesting that the series is initially deemed smooth. The ADF test results in Table 4 reveal a p value of 0.00 and a significant level value much lower than 0.05, indicating that the series has passed the significance test and is not a white noise series. Moreover, the values of the three serial intervals of ADF are greater than the ADF value of −7.388, further confirming that the series is smooth. Thus, the first-order difference series can be effectively analyzed and modeled.

Table 3. Undifferentiated ADF Results

		t-statistic	P
augmented Dickey−Fuller test statistic		−2.82	0.055*
test critical values	1% level	−3.447	
	5% level	−2.869	
	10% level	−2.571	

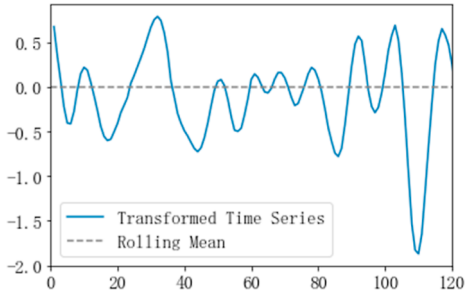


Figure 7. First-order difference.

As previously mentioned, the ARIMA model involves three hyperparameters: p, q, and d. In this study, the order of difference is 1, indicating that the hyperparameter d is equal to 1. To determine the optimal values of p and q, the PACF and ACF plots are initially examined. As shown in Figure 3, both the autocorrelation and partial autocorrelation coefficients exhibit trailing behavior, suggesting an order of 4. However, the

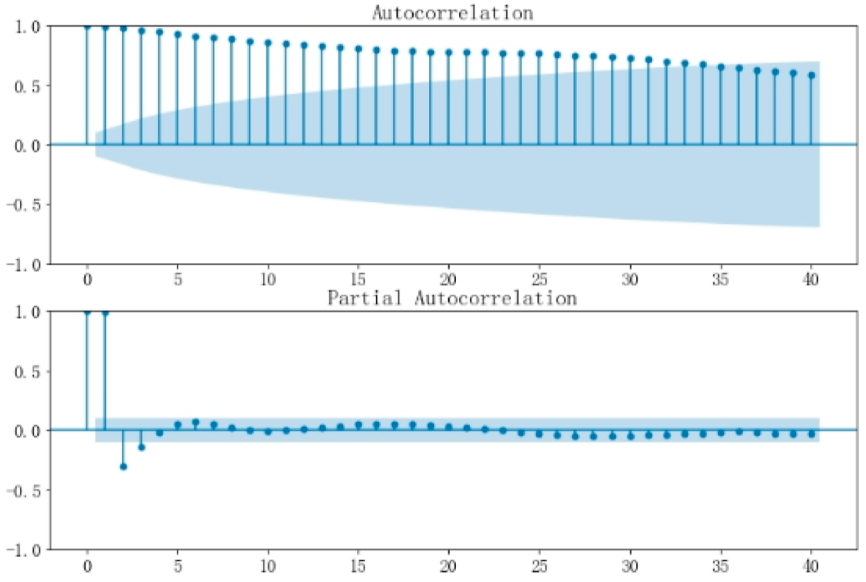


Figure 6. ACF and PACF plots of the original data.

https://doi.org/10.1021/acsomega.3c05422
ACS Omega 2023, 8, 39583−39595

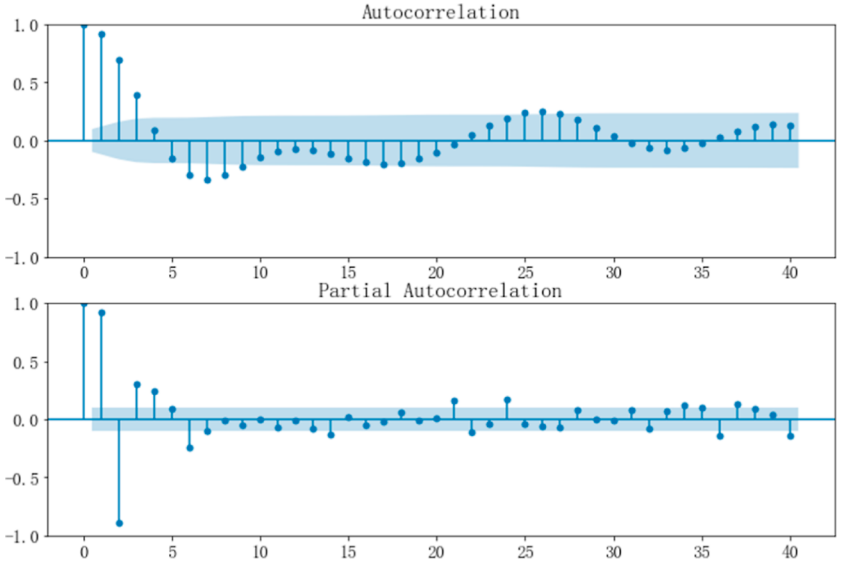


Figure 8. ACF and PACF plots of first-order difference.

Table 4. First-Order Differential ADF Results

		t-statistic	*P*
augmented Dickey–Fuller test statistic		−7.388	0.000***
test critical values	1% level	−3.447	
	5% level	−2.869	
	10% level	−2.571	

autocorrelation coefficients reach a critical point at order 7, while the partial autocorrelation coefficients reach a critical point at order 6, making it challenging to determine the order of the model using the traditional Box–Jenkins method. Therefore, the AIC and BIC statistical methods were employed to identify the optimal values of (p, q) as $(2, 5)$. Additionally, the Ljung–Box test was conducted on the residuals of the ARIMA model to verify if they are random. The Ljung–Box statistic yielded a *p*-value greater than 0.05, indicating that the residuals are white noise.

3.2.3. Results of Prophet. The fitting results of Prophet are presented in Figure 9, where the blue solid line represents the fitted value, the black point depicts the real value, and the purple area represents the 95% confidence interval of the sequence. The sequence's abnormal value falls outside the interval, while the red dotted line represents some mutations of the sequence point. As shown in the figure, Prophet exhibits excellent outlier and missing value handling capabilities.

3.2.4. Performance Comparison of Single Forecasting Methods. As shown in Table 5, the integrated learning prediction model proposed in this paper demonstrates superior performance over the three single prediction models based on the SSA algorithm in terms of the prediction metrics MAE, MSE, and MAPE for 1-step forecasting. The minimum values of these metrics corresponding to the integrated prediction model are 0.02, 0.0005, and 2.88%, respectively, indicating a high level of prediction accuracy. Through Figures 10–12, it can be more intuitively concluded that the integrated learning model has better prediction performance than the single model.

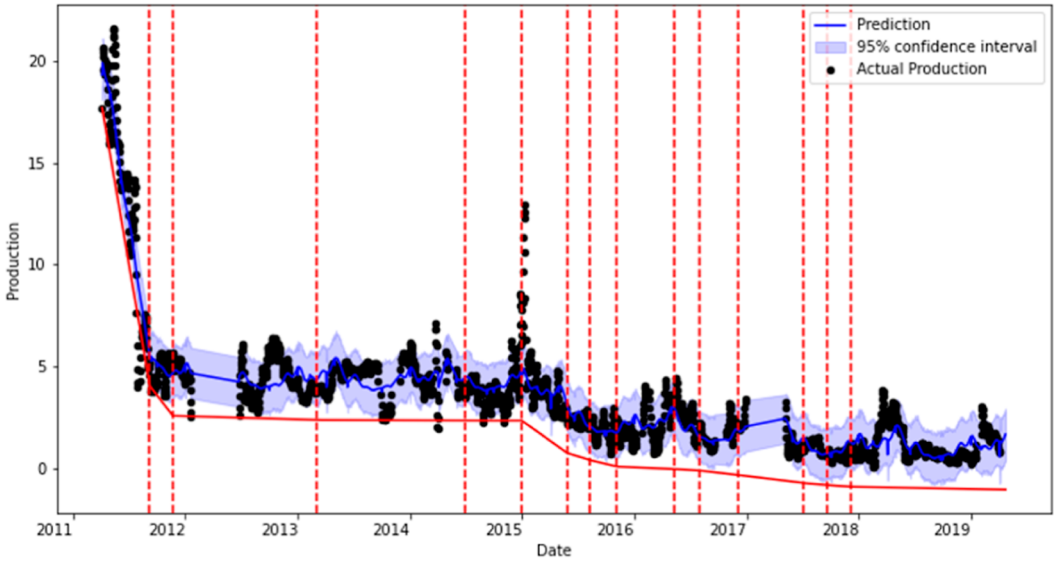


Figure 9. Results of Prophet fitting, identification of sequence mutation points, and complement of missing values.

https://doi.org/10.1021/acsomega.3c05422
ACS Omega 2023, 8, 39583–39595

Table 5. Performance Comparisons of Different Steps

evaluation	model	one-step-ahead	two-step-ahead	three-step-ahead	six-step-ahead
MAE	SSA−WNN	0.030935	0.059888	0.093878	0.184509
	SSA−ARIMA	0.063004	0.102662	0.151203	0.305079
	SSA−Prophet	0.063307	0.088036	0.131225	0.314093
	integration	**0.021793**	**0.034168**	**0.033133**	**0.097522**
MSE	SSA−WNN	0.000957	0.004425	0.011683	0.044284
	SSA−ARIMA	0.003970	0.012112	0.028623	0.122781
	SSA−Prophet	0.004007	0.008361	0.021359	0.139577
	integration	**0.000475**	**0.001321**	**0.001202**	**0.015586**
MAPE	SSA−WNN	0.041450	0.079152	0.121183	0.215962
	SSA−ARIMA	0.088210	0.144784	0.214387	0.435853
	SSA−Prophet	0.075315	0.114228	0.175127	0.467449
	integration	**0.028847**	**0.043419**	**0.040208**	**0.100929**

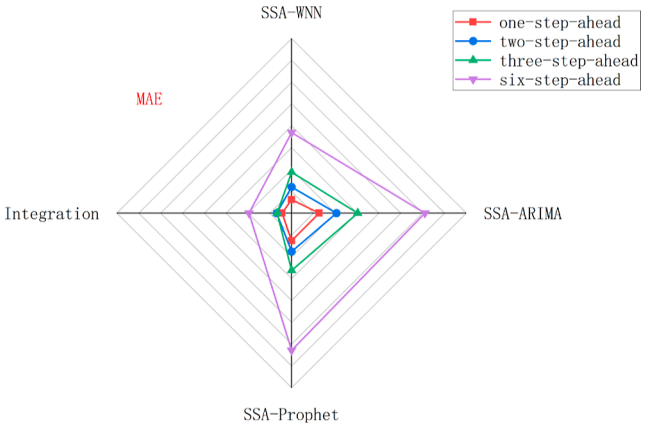


Figure 10. Comparison of MAE results in different time steps.

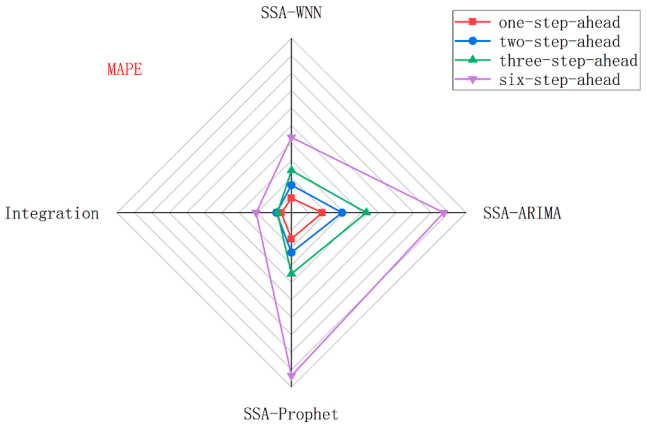


Figure 11. Comparison of MAPE results in different time steps.

By comparing the integrated learning prediction model with three other single prediction models based on the SSA algorithm, it can be concluded that the integrated learning prediction model proposed in this article has better prediction performance. In addition, in the case of multistep prediction of time series, the prediction performance of the model deteriorates with increasing prediction step length. Figure 13 shows the multistep prediction results for each model.

Table 6 records the degree of improvement in the prediction performance of the integrated learning prediction model compared to that of the remaining three single prediction models based on the SSA algorithm. For one-step forecasting, the integrated model reduced the MAE metric by 29.55, 65.41,

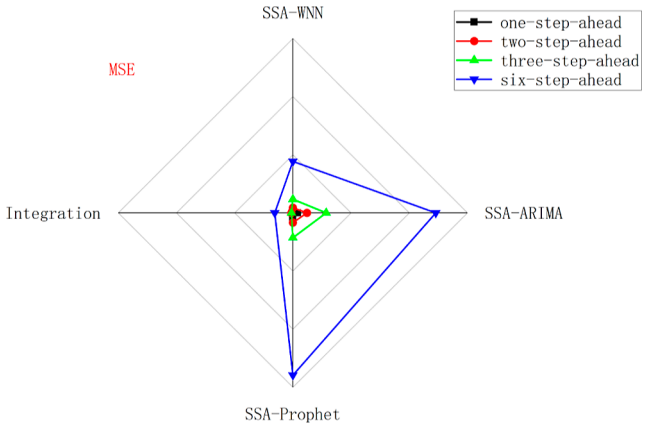


Figure 12. Comparison of MSE results in different time steps.

and 65.58% compared to SSA−WNN, SSA−ARIMA, and SSA−Prophet, respectively, and reduced the MSE metric by 50.37, 88.04, and 88.15% compared to the three single models, respectively, and reduced the MAPE metric by 30.4, 67.3, and 61.7%, respectively. However, as the prediction step length increases, the improvement of the integrated model in comparison to the single models begins to decrease. This is shown in Figure 14, where the degree of improvement in the MAE metric starts to decline significantly for 6-step prediction.

By comparing the prediction performance improvement of the integrated prediction model with the other three single prediction models based on the SSA algorithm, it can be concluded that the performance improvement of the integrated learning prediction model improves in the short term; however, the improvement decreases as the prediction step size increases. The advantages and disadvantages of each model are shown below: (1) the ARIMA model is simple to implement, and the decreasing production curve can be transformed into a smooth series using the first-order difference. The results in Figure 13 show that the ARIMA model works better in short-term forecasting and has smooth forecasting results. (2) The Prophet model is convenient due to its fixed structure and can detect abrupt change points and implied seasonality effects in the series. However, it can be seen from the results in Figure 13 that the prediction results of the Prophet model are more volatile, which may exaggerate the seasonal effects or the absence of seasonal effects in the series, leading to variation of the prediction results. (3) The WNN model is able to simulate the nonlinear downward trend, but it is not simple to choose the appropriate network architecture. According to the results in

https://doi.org/10.1021/acsomega.3c05422
ACS Omega 2023, 8, 39583−39595

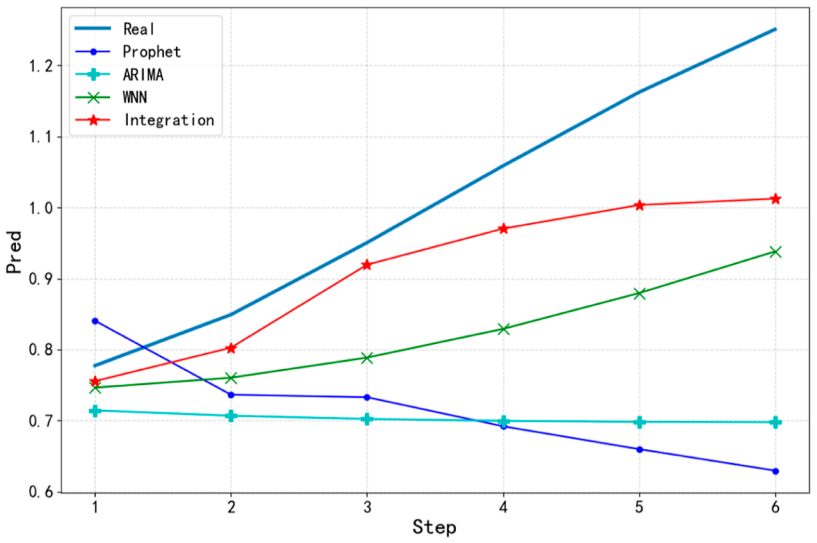


Figure 13. Prediction results of the single model and integrated model.

Table 6. Rate of Improvement of Integrated Model Evaluation Indicators

comparison	evaluation	one-step-ahead	two-step-ahead	three-step-ahead	six-step-ahead
integration vs WNN	I_{MAE} (%)	29.55%	42.95%	64.71%	47.15%
	I_{MSE} (%)	50.37%	70.16%	89.71%	64.81%
	I_{MAPE} (%)	30.40%	45.14%	66.82%	53.27%
integration vs ARIMA	I_{MAE} (%)	65.41%	66.72%	78.09%	68.03%
	I_{MSE} (%)	88.04%	89.10%	95.80%	87.31%
	I_{MAPE} (%)	67.30%	70.01%	81.24%	76.84%
integration vs Prophet	I_{MAE} (%)	65.58%	61.19%	74.75%	68.95%
	I_{MSE} (%)	88.15%	84.21%	94.37%	88.83%
	I_{MAPE} (%)	61.70%	61.99%	77.04%	78.41%

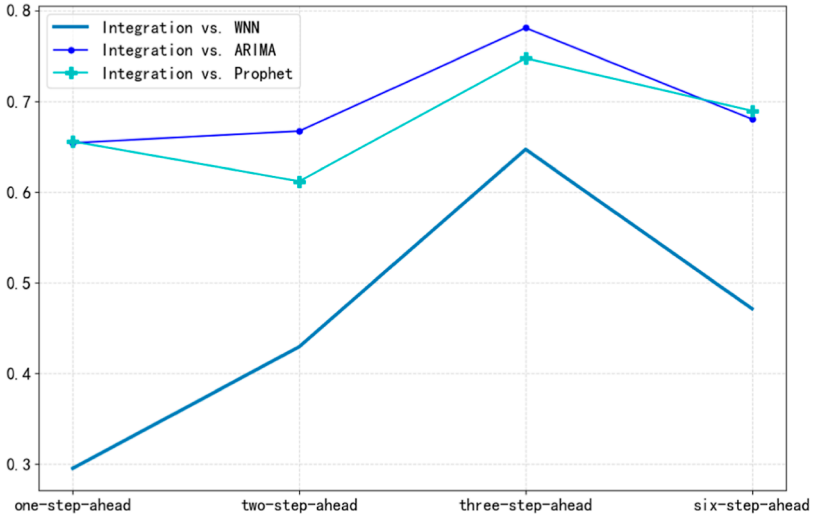


Figure 14. Degree of improvement of the integrated model compared to that of the single model.

Figure 13, WNN can capture the fluctuation of oil production, and the prediction results are similar to the trend of real oil production. (4) The integrated model combines the advantages of the above three models, and the predicted results are smoother and closer to the true oil production values.

3.2.5. DM Test of Single Models and Integrated Models. The performance of the integrated prediction model for daily oil production time series, proposed in this paper, is further evaluated and analyzed through the use of the Diebold–Mariano (DM) test.[29] This method is primarily employed to determine if there is a significant difference in prediction accuracy between the integrated model and other single models. The fundamental principles of the DM test are described below.

3.2.5.1. DM Test. In the DM statistic, $\{x_i; i = 1, 2, \cdots, t + p\}$ denotes a set of observation series, and $\{x_i^{(1)}; i = 1, 2, \cdots, t + p\}$ and $\{x_i^{(2)}; i = 1, 2, \cdots, t + p\}$ denote two sets of prediction value series obtained by two different prediction models, respectively. The prediction errors obtained from the two prediction models

https://doi.org/10.1021/acsomega.3c05422
ACS Omega 2023, 8, 39583−39595

Table 7. DM Test Results

comparison	one-step-ahead	two-step-ahead	three-step-ahead	six-step-ahead
integration vs Prophet	−2.8327 (0.003)	−2.5154 (0.004)	−2.5387 (0.004)	−3.1595 (0.002)
integration vs WNN	−1.7847 (0.037)	−1.7196 (0.038)	−1.6786 (0.042)	−0.9989 (0.047)
integration vs ARIMA	−1.9264 (0.035)	−1.7775 (0.036)	−1.9946 (0.03)	−2.1339 (0.009)

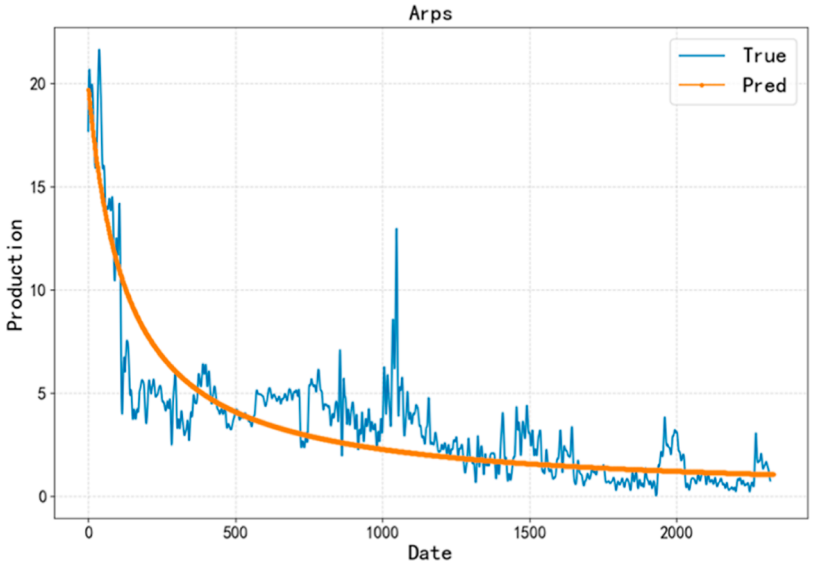


Figure 15. Prediction results of the Arps model.

are $e_{t+m}^{(1)} = x_{t+m} - x_{t+m}{}^{(1)}$, $m = 1, 2, \cdots, p$ and $e_{t+m}{}^{(2)} = x_{t+m} - x_{t+m}{}^{(2)}$, $m = 1, 2, \cdots, p$, respectively. The accuracy of each prediction model is estimated using the loss function $L(e_{t+m}{}^{(1)})$, $i = 1, 2$. The DM statistic enables the prediction model to be evaluated based on any loss function $L(g)$, and the expression for the DM statistic is shown below

$$\mathrm{DM} = \frac{s^2 \sum_{m=1}^{p} (L(e_{t+m}^{(1)}) - L(e_{t+m}^{(2)}))/p}{\sqrt{s^2/p}} \tag{12}$$

where: s^2 is an estimate of the variance of $d_m = L(e_{t+m}{}^{(1)}) - L(e_{t+m}{}^{(2)})$.

Here, the asymptotic distribution of the DM statistic obeys $N(0,1)$. If the calculated value of DM does not lie within the interval $-z_{\alpha/2}$ to $z_{\alpha/2}$, the two prediction methods are indistinguishable and the original hypothesis will be rejected at this point.

3.2.5.2. DM Test Results. The results of the DM test, presented in Table 7, demonstrate that the integrated forecasting model proposed in this paper outperforms the three single models in terms of forecasting accuracy at the 1-step, 2-step, 3-step, and 6-step prediction horizons. The results indicate that there is a significant difference in the forecasting accuracy of the integrated model compared with the single models at a 10% significance level. This suggests that the integrated forecasting model has a superior forecasting ability compared with some of the single forecasting models.

The results of the DM test indicate that the integrated prediction model proposed in this study outperforms the three single forecast models based on the SSA algorithm in predicting daily oil production time series. This suggests that the integrated model could be a more effective tool for actual oilfield production forecasting, providing highly accurate results for oil production time series.

3.2.6. Comprehensive Forecasting Studies with the Integrated Predictive Model. In order to evaluate the integrated SSA-based forecasting model proposed in this paper more comprehensively, the Arps model and the current popular time series forecasting algorithms are further selected and compared in this section.

3.2.6.1. Comparison of the Arps Model. The Arps prediction model[30] is a common empirical model used for forecasting oil and gas well production. In the application process of this article, the parameter estimation of the Arps decline model is conducted using the method of linear fitting between cumulative production and production. The specific steps for parameter calculation and production forecasting are as follows: According to the relationship between cumulative production "N_p" and production "q" given by $N_p = \frac{q_i}{d}\frac{1}{1-n} - \frac{q_i^n}{d}\frac{1}{1-n}q^{1-n}$. Let $x = q^{1-n}$, $y = N_p$, $b_0 = \frac{q_i}{d}\frac{1}{1-n}$, $b_1 = \frac{q_i^n}{d}\frac{1}{1-n}$. Thus, the linear regression equation is given by $y = b_0 + b_1 x$. The production data are then fitted using this equation to obtain the regression coefficients b_0 and b_1. From b_0 and b_1, the values of n (decline exponent), q_i (initial production), and d (decline rate) are calculated.

The prediction results of the Arps model are shown in Figure 15, with an MSE of 0.03, MAE of 0.15, and MAPE of 0.14, from which it can be known that the model in this paper is more advantageous than the Arps model. The specific fitting prediction process of the Arps model and the parameters can be found at https://github.com/SallBryant/ACS.git.

3.2.6.2. Comparison of Current Popular Time Series Forecasting Algorithms. In this section, three time series prediction algorithms, LSTM,[13] gated recurrent unit (GRU),[31] and temporal convolutional network (TCN),[32] are selected for comparison, and the prediction results of these three types of algorithms are demonstrated in Figure 16, from which it can be

https://doi.org/10.1021/acsomega.3c05422
ACS Omega 2023, 8, 39583−39595

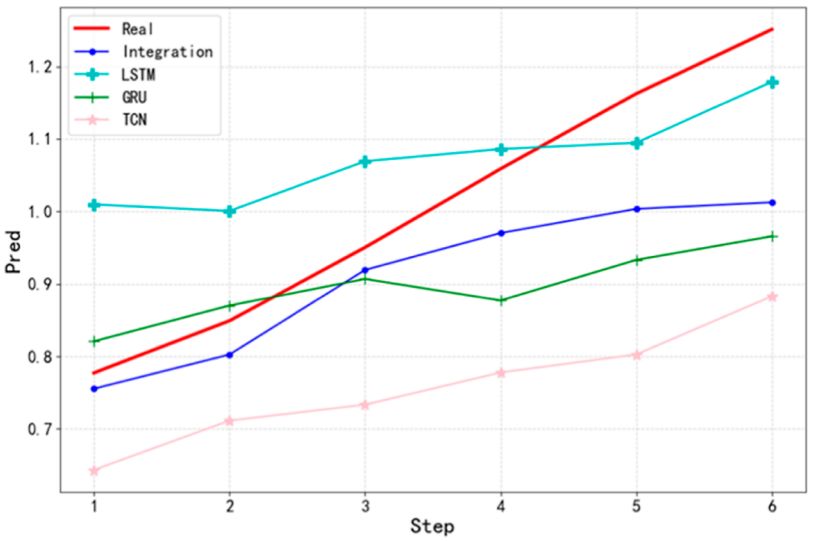


Figure 16. Prediction results of mainstream models and integrated models.

seen that the results of GRU and TCN are on the low side in the late stage of prediction, while the overall prediction results of LSTM are on the high side. Also, from Table 8, it can be more directly concluded that the model in this paper has reliable results compared to mainstream prediction models. The implementation-specific details of the model prediction and the setting of parameters can be found at https://github.com/SallBryant/ACS.git.

Table 8. Performance Comparisons of Mainstream Models and Integrated Models

evaluation	model	six-step-ahead
MAE	SSA−LSTM	0.111644
	SSA−GRU	0.133981
	SSA−TCN	0.249435
	integration	**0.033135**
MSE	SSA−LSTM	0.016931
	SSA−GRU	0.028495
	SSA−TCN	0.071241
	integration	**0.001202**
MAPE	SSA−LSTM	0.124012
	SSA−GRU	0.120519
	SSA−TCN	0.238513
	integration	**0.028847**

4. CONCLUSIONS

The integrated learning time series prediction method proposed in this paper combines the advantages of linear and nonlinear models and is universal without considering other factors in the oil field. The modeling process is more convenient compared to reservoir simulation methods. For illustration and validation, production data from the J oilfield in China is used for calculation and testing.

From the results of the tests, it is evident that the integrated prediction model exhibits improved performance in terms of multistep forecasting compared to the three single prediction models based on the SSA algorithm. The model also demonstrates less fluctuation in its MAPE values. To further verify the efficacy of the integrated prediction model, a DM test was conducted, which confirmed that it has a relatively good prediction performance and could serve as an effective production time series model in oilfield operations. In addition, comparisons were made with mainstream time series forecasting models and the traditional Arps model, robustly demonstrating the accuracy and novelty of the model proposed in this paper.

The methods employed in this study have certain limitations. First, the accuracy of predictions is highly dependent on data quality, with factors such as data missingness, outliers, or instability having an impact on the results. Second, models like ARIMA and Prophet require appropriate parameter configuration, and the choice of parameters may affect the accuracy of the forecasts.

Future directions for development are suggested as follows: first, one must consider the use of automated model selection and parameter tuning methods, such as utilizing automated machine learning (AutoML) tools, to reduce the need for manual adjustments. Second, one must actively explore additional feature engineering techniques, considering the incorporation of useful information related to production to further enhance the accuracy of production predictions.

AUTHOR INFORMATION

Corresponding Authors

XianKang Xin − *School of Petroleum Engineering, Yangtze University, Wuhan, Hubei 430100, China; Hubei Provincial Key Laboratory of Oil and Gas Drilling and Production Engineering (Yangtze University), Wuhan, Hubei 430100, China; School of Petroleum Engineering, Yangtze University: National Engineering Research Center for Oil and Gas Drilling and Completion Technology, Wuhan, Hubei 430100, China;* orcid.org/0000-0002-8133-3602; Email: xiankang.xin@hotmail.com

GaoMing Yu − *School of Petroleum Engineering, Yangtze University, Wuhan, Hubei 430100, China; Hubei Provincial Key Laboratory of Oil and Gas Drilling and Production Engineering (Yangtze University), Wuhan, Hubei 430100, China; School of Petroleum Engineering, Yangtze University: National Engineering Research Center for Oil and Gas Drilling and Completion Technology, Wuhan, Hubei 430100, China;* Email: ygm1210@vip.sina.com

https://doi.org/10.1021/acsomega.3c05422
ACS Omega 2023, 8, 39583−39595

Authors

MingCheng Ni − *School of Petroleum Engineering, Yangtze University, Wuhan, Hubei 430100, China;* orcid.org/0000-0003-1710-1206

Yu Liu − *School of Petroleum Engineering, Yangtze University, Wuhan, Hubei 430100, China*

YuGang Gong − *School of Petroleum Engineering, Yangtze University, Wuhan, Hubei 430100, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.3c05422

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This research was funded by the National Natural Science Foundation of China, grant number 52104020.

■ REFERENCES

(1) Jiménez-Rodríguez, R. Oil shocks and global economy. *Energy Econ.* **2022**, *115*, 106373.

(2) Werneck, R. d. O.; Prates, R.; Moura, R.; Gonçalves, M. M.; Castro, M.; Soriano-Vargas, A.; Ribeiro Mendes Júnior, P.; Hossain, M. M.; Zampieri, M. F.; Ferreira, A.; Davólio, A.; Schiozer, D.; Rocha, A. Data-driven deep-learning forecasting for oil production and pressure. *J. Pet. Sci. Eng.* **2022**, *210*, 109937.

(3) Ling, K.; He, J. Theoretical Bases of Arps Empirical Decline Curves, in: All Days. *Presented at the Abu Dhabi International Petroleum Conference and Exhibition, Abu Dhabi, UAE*; Society of Petroleum Engineers, 2012; p 161767.

(4) Sureshjani, M. H.; Gerami, S. A New Model for Modern Production-Decline Analysis of Gas/Condensate Reservoirs. *J. Can. Pet. Technol.* **2011**, *50*, 14−23.

(5) Wachtmeister, H.; Lund, L.; Aleklett, K.; Höök, M. Production Decline Curves of Tight Oil Wells in Eagle Ford Shale. *Nat. Resour. Res.* **2017**, *26*, 365−377.

(6) Zhang, H.; Rietz, D.; Cagle, A.; Cocco, M.; Lee, J. Extended exponential decline curve analysis. *J. Nat. Gas Sci. Eng.* **2016**, *36*, 402−413.

(7) Ayeni, B. J.; Pilat, R. Crude oil reserve estimation: An application of the autoregressive integrated moving average (ARIMA) model. *J. Pet. Sci. Eng.* **1992**, *8*, 13−28.

(8) Yusof, N. M.; Rashid, R. S. A.; Mohamed, Z. Malaysia crude oil production estimation: an application of ARIMA model. *Presented at the 2010 International Conference on Science and Social Research (CSSR)*; Institute of Electrical and Electronics Engineers, 2010; pp 1255−1259.

(9) Chahar, J.; Verma, J.; Vyas, D.; Goyal, M. Data-driven approach for hydrocarbon production forecasting using machine learning techniques. *J. Pet. Sci. Eng.* **2022**, *217*, 110757.

(10) Li, W.; Wang, L.; Dong, Z.; Wang, R.; Qu, B. Reservoir production prediction with optimized artificial neural network and time series approaches. *J. Pet. Sci. Eng.* **2022**, *215*, 110586.

(11) Negash, B. M.; Yaw, A. D. Artificial neural network based production forecastingfor a hydrocarbon reservoir under water injection. *Pet. Explor. Dev.* **2020**, *47*, 383−392.

(12) Alkhammash, E. H. An Optimized Gradient Boosting Model by Genetic Algorithm for Forecasting Crude Oil Production. *Energies* **2022**, *15*, 6416.

(13) Sagheer, A.; Kotb, M. Time series forecasting of petroleum production using deep LSTM recurrent networks. *Neurocomputing* **2019**, *323*, 203−213.

(14) Yang, R.; Liu, X.; Yu, R.; Hu, Z.; Duan, X. Long short-term memory suggests a model for predicting shale gas production. *Appl. Energy* **2022**, *322*, 119415.

(15) Ning, Y.; Kazemi, H.; Tahmasebi, P. A comparative machine learning study for time series oil production forecasting: ARIMA, LSTM, and Prophet. *Comput. Geosci.* **2022**, *164*, 105126.

(16) Ibrahim, A. F.; Al-Dhaif, R.; Elkatatny, S.; Shehri, D. A. Applications of Artificial Intelligence to Predict Oil Rate for High Gas−Oil Ratio and Water-Cut Wells. *ACS Omega* **2021**, *6*, 19484−19493.

(17) Khan, M. R.; Alnuaim, S.; Tariq, Z.; Abdulraheem, A. Machine Learning Application for Oil Rate Prediction in Artificial Gas Lift Wells, in: Day 3 Wed, March 20, 2019. *Presented at the SPE Middle East Oil and Gas Show and Conference*; Society of Petroleum Engineers: Manama, Bahrain, 2019; p D032S085R002.

(18) Al Dhaif, R.; Ibrahim, A. F.; Elkatatny, S. Prediction of Surface Oil Rates for Volatile Oil and Gas Condensate Reservoirs Using Artificial Intelligence Techniques. *ASME: J. Energy Resour. Technol.* **2022**, *144* (3), 033001.

(19) Liu, W.; Liu, W. D.; Gu, J. Forecasting oil production using ensemble empiricalmodel decomposition based Long Short-Term Memory neural network. *J. Pet. Sci. Eng.* **2020**, *189*, 107013.

(20) Guo, J.; Wang, H.; Guo, F.; Huang, W.; Yang, H.; Yang, K.; Xie, H. The backpropagation based on the modified group method of data-handling network for oilfieldproduction forecasting. *J. Pet. Explor. Prod. Technol.* **2019**, *9*, 1285−1293.

(21) Luo, Z.; Guo, W.; Liu, Q.; Zhang, Z. A hybrid model for financial time-series forecasting based on mixed methodologies. *Expert Syst.* **2021**, *38*, No. e12633.

(22) Xu, W.; Peng, H.; Zeng, X.; Zhou, F.; Tian, X.; Peng, X. A Hybrid Modeling Method Based on Linear AR and Nonlinear DBN-AR Model for Time Series Forecasting. *Neural Process. Lett.* **2022**, *54*, 1−20.

(23) Kuang, W.; Ling, B. W.-K.; Yang, Z. Reconstructing signal from quantized signal based on singular spectral analysis. *Digit. Signal Process.* **2018**, *82*, 11−30.

(24) de Araújo Morais, L. R.; da Silva Gomes, G. S. Forecasting daily Covid-19 cases in the world with a hybrid ARIMA and neural network model. *Appl. Soft Comput.* **2022**, *126*, 109315.

(25) de O Santos Júnior, D. S.; de Oliveira, J. F. L.; de Mattos Neto, P. S. G. An intelligent hybridization of ARIMA with machine learning models for time series forecasting. *Knowl.-Based Syst.* **2019**, *175*, 72−86.

(26) Panigrahi, S.; Pattanayak, R. M.; Sethy, P. K.; Behera, S. K. Forecasting of Sunspot Time Series Using a Hybridization of ARIMA, ETS and SVM Methods. *Sol. Phys.* **2021**, *296*, 6.

(27) Chen, Q.; Song, Y.; Zhao, J. Short-term traffic flow prediction based on improved wavelet neural network. *Neural. Comput. Appl.* **2021**, *33*, 8181−8190.

(28) Herrera, O.; Priego, B. Wavelets as activation functions in Neural Networks. *J. Intell. Fuzzy Syst.* **2022**, *42*, 4345−4355.

(29) Diebold, F. X.; Mariano, R. S. Comparing Predictive Accuracy. *J. Bus. Econ. Stat.* **2002**, *20*, 134−144.

(30) Arps, J. J. Analysis of Decline Curves. *Trans. AIME* **1945**, *160*, 228−247.

(31) Ma, X.; Hou, M.; Zhan, J.; Zhong, R. Enhancing Production Prediction in Shale Gas Reservoirs Using a Hybrid Gated Recurrent Unit and Multilayer Perceptron (GRU-MLP) Model. *Appl. Sci.* **2023**, *13*, 9827.

(32) Zhang, L.; Dou, H.; Wang, T.; Wang, H.; Peng, Y.; Zhang, J.; Liu, Z.; Mi, L.; Jiang, L. A production prediction method of single well in water flooding oilfield based on integrated temporal convolutional network model. *Pet. Explor. Dev.* **2022**, *49*, 1150−1160.