Received: 17 June 2019 | Revised: 13 December 2019 | Accepted: 22 February 2020

DOI: 10.1002/for.2677

RESEARCH ARTICLE

WILEY

On the forecasting of high-frequency financial time series based on ARIMA model improved by deep learning

Zhenwei Li[1] | **Jing Han**[2] | **Yuping Song**[1]

[1]School of Finance and Business, Shanghai Normal University, Shanghai, PR China

[2]School of Finance and Management, Shanghai University of International Business and Economics, Shanghai, PR China

Correspondence
Yuping Song, School of Finance and Business, Shanghai Normal University, Shanghai, 200234, PR China.
Email: songyuping@shnu.edu.cn

Funding information
Academic Innovation Team of Shanghai Normal University, Grant/Award Number: 310-AC7031-19-004228; Key Subject of Quantitative Economics of Shanghai Normal University, Grant/Award Number: 310-AC7031-19-004221; Academic Innovation Team, Grant/Award Number: 310-AC7031-19-004228; Key Subject of Quantitative Economics, Grant/Award Number: 310-AC7031-19-004221; General Research Fund of Shanghai Normal University, Grant/Award Number: SK201720; Youth Academic Backbone Cultivation Project of Shanghai Normal University, Grant/Award Number: 310-AC7031-19-003021; National Statistical Science Research Project, Grant/Award Number: 2018LZ05; Ministry of Education, Humanities and Social Sciences project, Grant/Award Number: 18YJCZH153; National Natural Science Foundation of China, Grant/Award Number: 11901397

Abstract

Through empirical research, it is found that the traditional autoregressive integrated moving average (ARIMA) model has a large deviation for the forecasting of high-frequency financial time series. With the improvement in storage capacity and computing power of high-frequency financial time series, this paper combines the traditional ARIMA model with the deep learning model to forecast high-frequency financial time series. It not only preserves the theoretical basis of the traditional model and characterizes the linear relationship, but also can characterize the nonlinear relationship of the error term according to the deep learning model. The empirical study of Monte Carlo numerical simulation and CSI 300 index in China show that, compared with ARIMA, support vector machine (SVM), long short-term memory (LSTM) and ARIMA-SVM models, the improved ARIMA model based on LSTM not only improves the forecasting accuracy of the single ARIMA model in both fitting and forecasting, but also reduces the computational complexity of only a single deep learning model. The improved ARIMA model based on deep learning not only enriches the models for the forecasting of time series, but also provides effective tools for high-frequency strategy design to reduce the investment risks of stock index.

KEYWORDS

ARIMA model, high-frequency financial time series, LSTM model, SVM model

1 | INTRODUCTION

According to the historical data, an appropriate forecasting model can be constructed to capture the fluctuating signals of the underlying time series and characterize their trend, which can provide a reliable basis for investors' decision making. For example, through accurate forecasting of the stock index, investors can roughly grasp the overall trend of the market to effectively capture trading opportunity and make reasonable asset allocations. As for the forecasting models of time series, based on classical models such as autoregressive

(AR; Yule, 1927) and moving average (MA; Walker, 1931), the autoregressive moving average model (ARMA) and autoregressive integrated moving average model (ARIMA) were proposed (Box, Jenkins, Reinsel, & Ljung, 2015). After making the original nonstationary time series to be stationary after *d*-order difference, the ARIMA model then can estimate and test the stationary sequence. It has become one of the more widely used methods in the study of forecasting models for time series. The ARIMA model has been used to predict sales of retail footwear products in one step and multiple steps and it was found that the ARIMA model had a good fitness for the forecasting of time series (Ramos, Santos, & Rebelo, 2015).

For financial time series, using the autocorrelation function, more scholars have verified that financial time series were time varying (Ding, Granger, & Engle, 1993), and that the financial data presented the characteristics of nonlinearity (Chevallier & Sévi, 2012; Giot, Laurent, & Petitjean, 2010; Slim & Dahmene, 2016). The time-varying and nonlinear properties and the large stochastic volatility of the sample data in the financial market have posed certain difficulties for quantitative forecasting based on only a single model. Many scholars have improved the ARIMA model to enhance the accuracy of forecasting. Based on the linear error correction model, the ARIMA model was modified by using the support vector machine (SVM) model to forecast financial time series and improve forecasting accuracy (Van Gestel et al., 2006). Particle swarm optimization (PSO) was adopted to modify the ARIMA-SVM combination model and to improve the accuracy of forecasting for time series (de Oliveira & Ludermir, 2014). Accuracy of predictive power demand was improved by the error correction model based on PSO optimal Fourier and seasonal ARIMA (Wang, Wang, Zhao, & Dong, 2012). By considering a hybrid correction method based on the ARIMA model, SVM, and cuckoo search algorithm (CSA), the ARIMA model was modified to predict the power load (Kavousi-Fard & Kavousi-Fard, 2013).

The above research was mainly focused on daily or weekly low-frequency data for modeling and forecasting. However, with the development of science and technology, the era of big data had arrived and the storage and computing power for high-frequency data were improved. In addition, by using intraday high-frequency data to estimate volatility, Andersen, Bollerslev, Diebold, and Labys (2003) found that high-frequency data contained more market information than low-frequency data, which could improve the accuracy of estimation. High-frequency data could provide more arbitrage space and more possibilities for strategy design (Hanson & Hall, 2012). High-frequency financial time series also changed the philosophy of investment strategy design and the investor's investment style. Up to 2012, the total transaction volume of high-frequency transactions accounted for 50–80% of the total transaction volume of US equity (Barrales, 2012; Laughlin, Aguirre, & Grundfest, 2014), and the proportion reached 45% in Europe, 40% in Japan, and about 12% in other Asian countries (Menkveld, 2014). In 2014, the development of high-frequency trading made even greater progress. Through the study of a large number of financial intermediaries, Biais and Foucault (2014) confirmed that in the process of running capital for financial intermediaries, although they were not named as high-frequency transactions, they were also consistent with the characteristics of high-frequency trading strategies.

ARIMA models and their modified models in the existing literature are mostly used in the nonfinancial field, and sample frequencies for forecasting have been relatively low. However, with the increase in frequency of financial data, high-frequency data are highly nonlinear (Jobson & Korkie, 1981) and nonnormal (Jacquier, Polson, & Rossi, 2002). Owing to these characteristics of high-frequency data that do not conform to the traditional low-frequency model hypothesis, the forecasting error for high frequency data based on low-frequency financial time series model has gradually become larger. How to modify the ARIMA model and migrate it to high-frequency forecasting has a significant and practical research value. Currently, few studies have used the deep learning method to correct the ARIMA model error or to improve the forecasting accuracy of high-frequency financial data. Therefore, this paper aims to modify the traditional ARIMA model by using the machine-derived deep learning long short-term memory (LSTM) model. Compared with the machine learning SVM model and other modified models, the deep learning corrected model not only can reduce the error of the forecasting model and improve forecasting accuracy, but it can also reduce model complexity and improve predictive performance. Methodologically, the ARIMA-LSTM model not only preserves the stability and interpretability of the traditional ARIMA model, but also absorbs the long memory of the deep learning model for time series. Practically, the ARIMA-LSTM model can reduce the complexity of the deep learning correction process, and guarantee timeliness for high-frequency financial time series.

Section 2 introduces relevant models. In Section 3 a Monte Carlo numerical simulation is constructed to discover the space that can be improved by the traditional ARIMA model, and then we combine machine learning such as SVM and deep learning such as LSTM to correct the residuals. Finally, empirical high-frequency data of

the CSI 300 index are analyzed to verify the better properties of the proposed ARIMA-LSTM method. The robustness of the improved method proposed in this article is shown in Section 4. Section 5 concludes and gives other extensions.

2 | COMPONENT MODELS

2.1 | ARIMA model

The ARIMA(p, d, q) model is an important method for analyzing time series. It is a combination of the AR model and the MA model, where d is the difference term and p, q are delay parameters.

2.1.1 | Autoregressive moving average model

The structure of the ARMA(p, q) model is defined as follows:

$$y_t = \varphi_0 + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \ldots + \varphi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \epsilon_{t-q}, \quad (1)$$

where

$$\begin{cases} \varphi_i \neq 0, \theta_q \neq 0 \\ \mathrm{E}(\varepsilon_t) = 0, \mathrm{var}(\varepsilon_t) = {\sigma_u}^2 \\ \mathrm{E}(\varepsilon_t \varepsilon_S) = 0, s \neq t \\ \mathrm{E}(y_s \varepsilon_t) = 0, \forall s < t, \end{cases} \quad (2)$$

y_t is a dependent variable representing the observed value of the time series at the time t; y_{t-1}, y_{t-2}... are independent variables representing a lag sequence of the time series y_t; ε_t represents a residual or white noise sequence; φ_i is the self-regressive parameter to be estimated; and θ_q represents the moving average parameter to be estimated.

2.1.2 | Autoregressive integrated moving average model

The above models (Equation 1) are used for modeling stationary sequences. However, Visser (2010) found that the financial time series was not stationary. For the non-stationary sequence, it can be converted into stationary time series through d-order difference, and then established as a differential autoregressive moving average model, namely the ARIMA(p, d, q) model. The converted stationary sequence can be expressed as follows:

$$Y_t^* = (1-B)^d Y_t. \quad (3)$$

Using ARIMA(p, d, q) to fit time series data means employing a combination of AR, MA, and ARMA with different orders to make the model express various information for time series, so as to achieve effective forecasting of time series.

2.2 | SVM model

SVM model is a kind of generalized linear classifier for binary classification of data according to supervised learning. The decision boundary is the largest margin hyperplane for the learning sample, which can be used for both classification and regression analysis. In terms of classification, SVM can perform nonlinear classification by using the kernel function, which transforms samples by a nonlinear mapping method, and converts linear indivisible samples of low-dimensional input space into high-dimensional feature space to make it linearly separable. As shown in Figure 1, the hyperplane is found in high-dimensional space for binary classification space partitioning.

The training sample is {(x_i, y_i),i = 1,2,...,l}; the regression model is then $f(x) = [\omega * \varphi(x)]+b$, and the corresponding optimization problems are as follows:

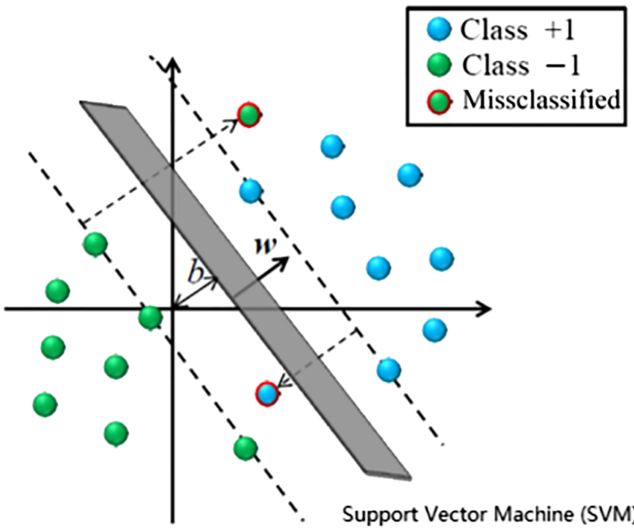

FIGURE 1 Schematic diagram of the SVM [Colour figure can be viewed at wileyonlinelibrary.com]

$$\min\left(\frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{l}\left(\xi_i + \xi_i^*\right)\right), \tag{4}$$

$$\text{s.t.}\begin{cases} y_i - [\omega^*\varphi(x_i)] - b \le \varepsilon + \xi_i \\ [\omega^*\varphi(x_i)] + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \qquad i = 1, 2, ..., l, \end{cases} \tag{5}$$

where ξ_i and ξ_i^* are slack variables; C is a penalty function; ε is the estimation precision, and the above quadratic programming problem can be converted into a dual problem:

$$\max\left(-\frac{1}{2}\sum_{i,j=1}^{l}\left(\alpha_i - \alpha_i^*\right)\left(\alpha_j - \alpha_j^*\right)\left[\varphi(x_i)^*\varphi(x_j)\right] + \sum_{i=1}^{l} y_i\left(\alpha_i - \alpha_i^*\right) - \varepsilon\sum_{i=1}^{l}\left(\alpha_i + \alpha_i^*\right)\right) \tag{6}$$

$$\text{s.t.}\begin{cases} \sum_{i=1}^{l}\left(\alpha_i + \alpha_i^*\right) = 0 \\ \alpha_i, \alpha_i^* \in [0, C] \;\; i = 1, 2, ..., l. \end{cases} \tag{7}$$

The kernel function that defines the inner product of the high-dimensional feature space is $K(x_i, x) = [\varphi(x_i), \varphi(x)]$, and the quadratic programming problem is solved to obtain the nonlinear mapping $f(x) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)[\varphi(x_i), \varphi(x)]$.

The SVM model avoids the complexity of high-dimensional space; it directly uses the kernel function of this space and adopts the solution method in the case of linear separability to solve the decision problem of corresponding high-dimensional space. The SVM model can be mapped to high-dimensional space by using kernel functions, which simplifies the difficulty of solving. At the same time, the kernel function can be used to settle the nonlinear classification problem. In addition, the SVM model adopts the classification idea of maximizing the interval between the sample and the decision surface. The principle is simple and easy to understand and the classification effect is good. However, the SVM model also possesses certain defects in the forecasting of high-frequency time series. It reduces the complexity of high-dimensional space while increasing the complexity of solving kernel functions. At the same time, SVM does not take into account the characteristics of long-term and short-term memory for high-frequency data, and cannot effectively describe the inherent correlation of time series. Furthermore, the SVM model presents difficulty in training large-scale data; in addition, the SVM model can only deal with multi-classification problems by using indirect methods.

2.3 | Long short-term memory model

LSTM is a long and short-term memory network, which is a method of deep learning. It is a time-cycle neural network suitable for processing and forecasting important events with relatively long intervals and delays in time series. Based on the traditional recurrent neural network, LSTM adds a "processor" to the algorithm to judge whether the information is useful or not, and then to settle the problem of gradient disappearance and explosion and the problem of long-term dependence of information.

For the time series $t = 1, 2, 3, ...$, the output of the previous LSTM unit is combined with the current point in time data as an input, and each time step has an output. At the same time, the memory unit generates a state vector of the current time step. In the structure of LSTM, the cell state carries important information and its mechanism is shown in Figure 2. There are three important gates in the model to realize the transformation of

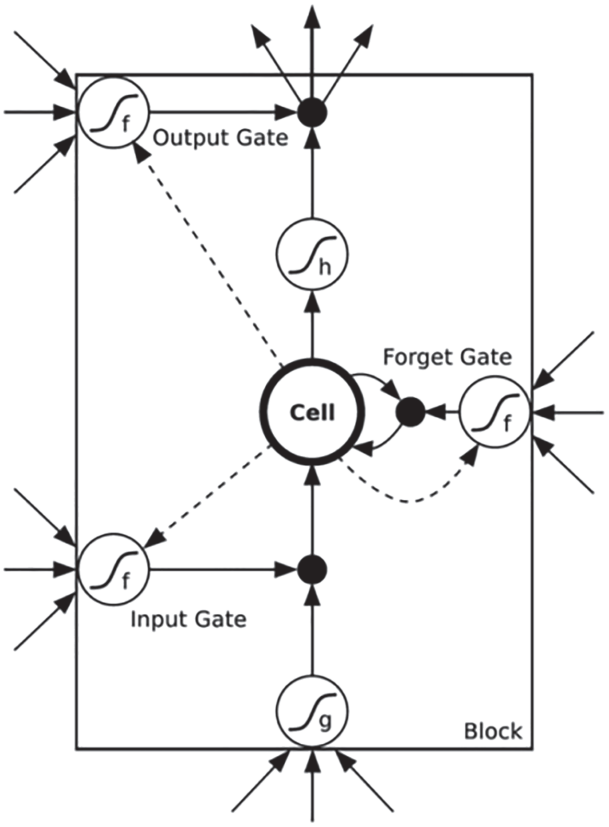

FIGURE 2 Schematic diagram of the LSTM

information, including forget gate i, input gate j, and output gate o. The cell state formula is as follows:

$$c_t = f_t c_{t-1} + i_t \tanh(W_{xc} x_t + W_{hc} h_{t-1} + b_c), \quad (8)$$

where c_t represents the cell state at t, f_t represents the current forget gate, c_{t-1} represents the cell state at $t-1$, i_t represents the current input layer, x_t represents the token of the current input sentence, h_{t-1}indicates the hidden layer output at $t-1$, W_{xy} represents the connection weight from neuron x to y, and b_c represents the offset amount.

The forget gate f_t implements the abandonment of information, which retains the input gate x_t and the hidden layer h_{t-1} to determine whether to completely retain or discard information about each element of the previous cell:

$$f_t = \sigma\left(W_{xf} x_t + W_{hf} h_{t-1} + W_{cf} c_{t-1} + b_f\right). \quad (9)$$

The input gate i_t implements the update of the information:

$$i_t = \sigma(W_{xi} x_t + W_{hi} h_{t-1} + W_{ci} c_{t-1} + b_i). \quad (10)$$

Finally, the output gate o_t implements the output of the information:

$$o_t = \sigma(W_{xo} x_t + W_{ho} h_{t-1} + W_{co} c_{t-1} + b_o). \quad (11)$$

The hidden layer output is

$$h_t = o_t \tanh(c_i). \quad (12)$$

Figure 3 shows the solution process for a block of LSTM cells. The large black circle (middle) represents the memory cell and stores the state information of the LSTM, the open circle represents the multiplication, and the small black circle represents the activation function.

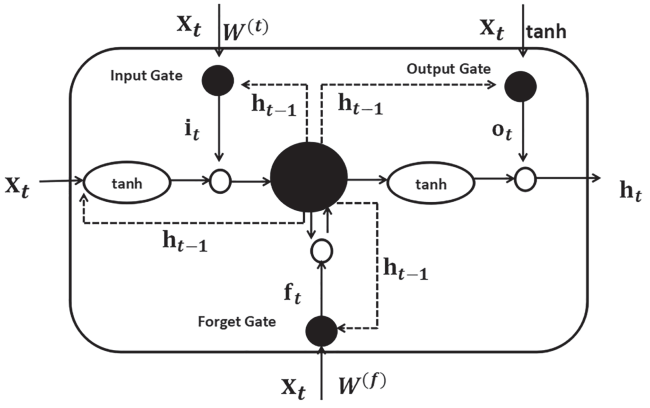

FIGURE 3 Schematic diagram for LSTM memory unit

2.4 | Algorithm procedure for ARIMA modified model

The forecasting of high-frequency financial time series using the ARIMA modified model can be divided into the following steps:

1. Observe the distribution state of high-frequency financial time series through data descriptive statistics, and preprocess the data according to the model requirements; specifically, for example, all values are divided by the initial value of the sequence.
2. Construct the ARIMA model for the processed data. First, determine the difference order d by the augmented Dickey–Fuller (ADF) test, and then determine the p and q values by the autocorrelation function (ACF) and partial autocorrelation function (PACF) tests.
3. After determining the parameters of ARIMA, search for various possible values to find the optimal value. Construct an ARIMA model with each combination of [0, p] and [0, q] until the Akaike information criterion (AIC) reaches a minimum, and select the corresponding p and q as parameters.
4. Calculate the residual of the ARIMA model of the optimal parameters, and model the residual through the SVM model and the LSTM model.
5. Combine the residual training SVM model and the LSTM model with the optimal ARIMA model to obtain the final modified model.

2.5 | Evaluation criteria

In order to evaluate each forecasting model, the following evaluation criteria are introduced, where f_i is the model forecasting value, y_i is the actual value, and n is the sample data volume:

MSE (mean squared error):

$$\mathrm{MSE} = \frac{1}{n}\sum (f_i - y_i)^2; \quad (13)$$

MAE (mean absolute error):

$$\mathrm{MAE} = \frac{1}{n}\sum |f_i - y_i|; \quad (14)$$

RMSE (root mean squared error):

$$\text{RMSE} = \sqrt{\text{MSE}}; \tag{15}$$

MAPE (mean absolute percentage error):

$$\text{MAPE} = \frac{100}{n}\sum\frac{|f_i - y_i|}{y_i}; \tag{16}$$

R^2 (goodness of fit):

$$R^2 = 1 - \frac{\sum(f_i - y_i)^2}{\sum(y_i - \bar{y})^2}; \tag{17}$$

RMSPE (root mean square percent error):

$$\text{RMSPE} = \sqrt{\frac{1}{n}\sum\left(1 - \frac{y_i}{f_i}\right)^2}. \tag{18}$$

3 | NUMERICAL ANALYSIS

3.1 | Monte Carlo simulation

3.1.1 | Data generation

In order to explore the application of the ARIMA modified model in financial time series, it is necessary to generate relevant data for testing and modeling. Based on the Black–Scholes model, the stochastic differential equation is as follows:

$$dS_t = rS_t dt + \sigma S_t dB_t, \tag{19}$$

where S_t represents the target asset price level at time t, σ is the fixed volatility of the target; r is the fixed risk-free short-term interest rate; B is a Brownian motion. High-risk objects (such as stock price or index level) follow the Brownian motion expressed by the stochastic differential equation SDE in the case of risk neutrality. Their specific forms are denoted as follows:

$$S_t = S_{t-\Delta t}\exp\left\{\left(r - \frac{1}{2}\sigma^2\right)\Delta t + \sigma\sqrt{\Delta t}B_t\right\}. \tag{20}$$

In order to study the statistical characteristics of high-frequency financial time series, the above model is parametrized (Hilpisch, 2014). Assume that the initial price of the object is $S_0 = 1$. The study duration is set to 1 day, the time frequency is 5 minutes data, then $T = 1$, $\Delta t = \frac{1}{48}$, so there is a total of 48 data per day. The fixed volatility $\sigma = 0.25$, the risk-free short-term interest rate $r = 5\%$, and they are brought into Equation 20 to obtain 2000 sets of sample data.

As can be seen from Figure 4, the financial time series paths of Monte Carlo simulation comprehensively contain the actual situation of the actual stock price trend. Monte Carlo simulation of financial data provides a reliable source of data for exploring the application of ARIMA modified models in financial time series.

Figure 5 shows the distribution of the value of the simulation data at the end of the period, which is also a side reflection of the rate of return. Unlike the normal distribution hypothesis of traditional research, the simulated data show a realistic "fat-tail distribution" situation. The final price is mostly close to the initial price, the

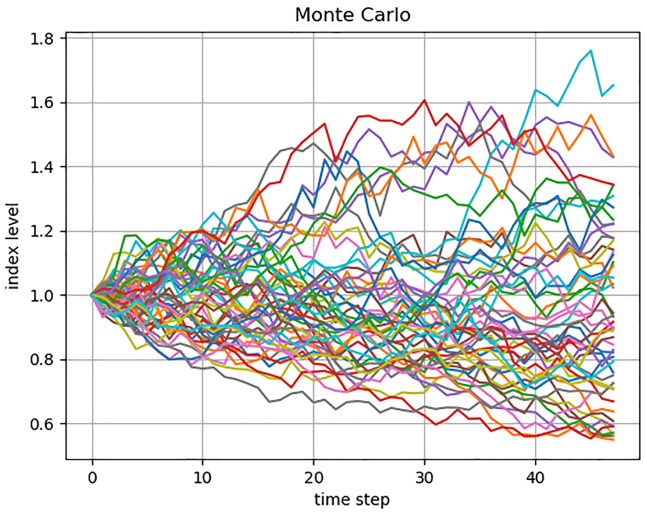

FIGURE 4 Fifty simulated index horizontal paths [Colour figure can be viewed at wileyonlinelibrary.com]

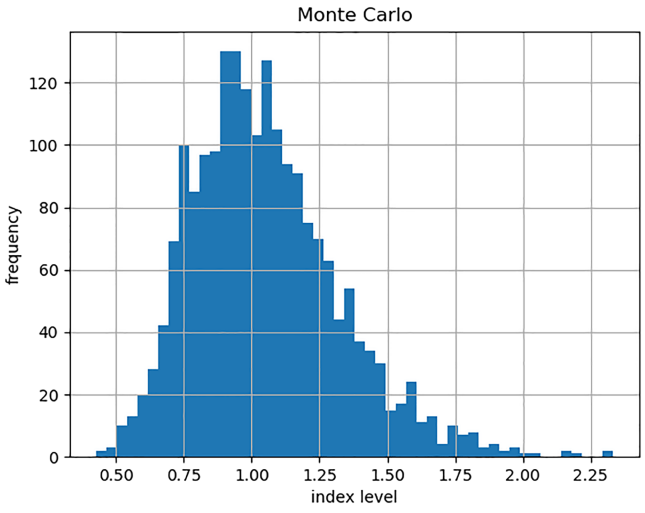

FIGURE 5 Histogram of end-values in all simulation paths [Colour figure can be viewed at wileyonlinelibrary.com]

high-loss and high-returns distribution are asymmetrical, and the probability of high returns is much lower than that of high losses. The real state of the securities market is similar to that of Figure 5. Investors who can make a profit in the securities market can sharply grasp high-returns stocks but, on the contrary, most investors tend to fall into no-change areas or loss areas.

3.1.2 | Parameter selection

Since the ARIMA correction model involves the traditional ARIMA model, the machine learning SVM model, and the deep learning LSTM model, the parameters need to be set in advance when constructing the model.

The parameter involved in the ARIMA model is the autoregressive term p, the difference order d and the moving average term q of the stationary sequence. According to the actual situation of the financial time series, the time series after one difference will pass the ADF stationarity test, so the parameter $d = 1$ is set in this paper. The setting of the autoregressive term p and the moving average term q are selected according to the actual situation of the simulated data. Through the ACF and PACF, the initial p and q are selected, and the final p and q are found by cyclic optimization based on the AIC. Finally, all the parameters of ARIMA are obtained.

When using the machine learning SVM model for regression forecasting, the important parameter is the choice of kernel function. Alternative kernel functions are linear kernel functions, polynomial kernel functions, Gaussian (radial basis function, RBF) kernel functions, and sigmoid kernel functions, respectively. The Gaussian kernel function is highly local and can map a sample into a higher dimensional space. Moreover, it is applicable to both large samples and small samples, and it has fewer parameters than polynomial kernel functions. The number of models in this paper is small and the number of samples is normal, so RBF is chosen as the model kernel function. The penalty parameter of the SVM model is set to the default value $C = 1$, and the gamma is set to gamma = " *auto*". The dynamic mean and volatility of the previous data and the current data are introduced into the model as independent variables.

The LSTM model of deep learning has the ability to memorize the long-term or short-term time series, and the parameters of the neural network also need to be set. The activation function of the LSTM module uses "tahn," and the activation function of the fully connected artificial neural network that receives the output also uses "linear." To prevent overfitting, the discard rate for each layer of network nodes uses 0.2. The number of neurons is set to 32 and a hidden layer is added. The mean square error is used as the calculation method for determining the error. The iterative update method for determining the weight parameter adopts the RMSprop algorithm, and the number of iterations is epoch = 100.

3.1.3 | Model comparison

The traditional ARIMA model was constructed for simulation data and the mean value of the results was counted. It can be seen from Table 1 that the errors (MSE, MAE, RMSE, RMSPE) and loss rate (MAPE) are generally too large and the goodness of fit (R^2) is significantly smaller, indicating that the traditional ARIMA model leaves much room for improvement in the forecasting of high-frequency financial time series. The SVM model and the LSTM model are combined with the ARIMA model to correct the error term of the ARIMA model, which can also improve the accuracy of the model. This method of combining a traditional linear model with a nonlinear model can explain both the linear part of the data and the nonlinear part of the data.

It can be seen from Figure 6 that the ARIMA model has large deviations and only predicts the overall trend of the original data. Although the ARIMA-SVM model has improved performance on the ARIMA model, the large fluctuations cannot be well identified. ARIMA-LSTM can correct the error term through deep learning after forecasting the overall trend. The curve predicted by the model based on deep learning is more consistent with the curve of the original data.

TABLE 1 Model evaluation criteria

MODEL	MSE (rank)	MAE (rank)	RMSE (rank)	MAPE (rank)	R^2 (rank)	RMSPE (rank)	Average rank
ARIMA	0.1996 (3)	0.1238 (3)	0.1533 (3)	11.91% (3)	14.09% (3)	0.1460 (3)	3
ARIMA-SVM	0.0068 (2)	0.0510 (2)	0.0629 (2)	5.13% (2)	38.98% (2)	0.0633 (2)	2
ARIMA-LSTM	0.0041 (1)	0.0291 (1)	0.0381 (1)	2.84% (1)	68.76% (1)	0.0369 (1)	1

Note. Rank is the order of model performance in evaluation criteria and Average rank is the sample mean for ranks.

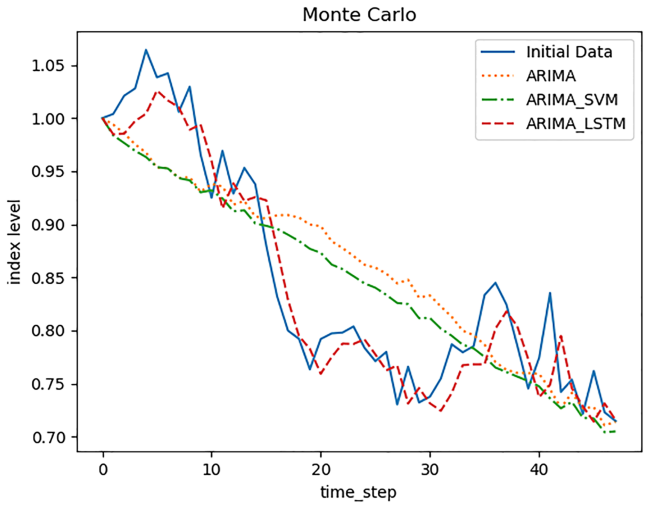

FIGURE 6 Model forecasting comparison [Colour figure can be viewed at wileyonlinelibrary.com]

3.2 | Empirical analysis

This section will employ real data from the securities market to verify the better performance of the proposed model. It is divided into three parts: fitting, forecasting, and strategy design. In the subsection of fitting, the ARIMA, SVM, LSTM, ARIMA-SVM, and ARIMA-LSTM models are constructed respectively. Furthermore, the five models are analyzed and compared from the perspective of price time series. In the subsection of forecasting, the prediction effects of these models are demonstrated, and the ARIMA, ARIMA-SVM, and ARIMA-LSTM models are constructed for comparative analysis. In the subsection of strategy design, the ARIMA-LSTM model is designed as a strategy for the practical application.

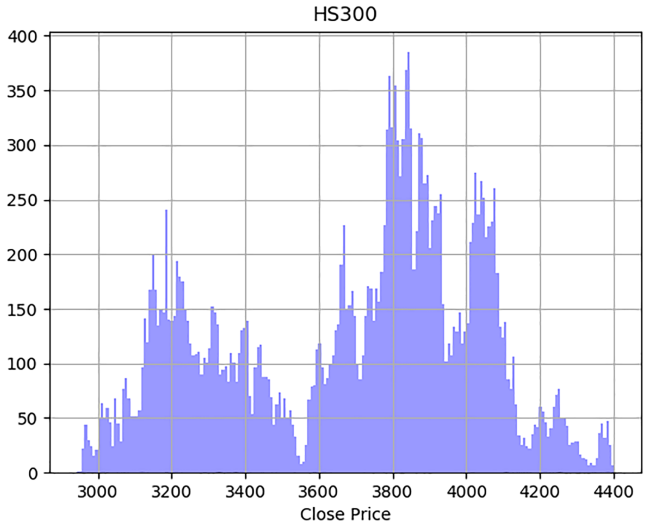

FIGURE 7 Distribution histogram for closing prices of CSI 300 index [Colour figure can be viewed at wileyonlinelibrary.com]

3.2.1 | Data description

In order to demonstrate the better performance of the ARIMA model improved by deep learning, this paper selects 5-minute high-frequency data from the CSI 300 index as the data set. The data of total 2 years is selected from November 1, 2017, to November 1, 2019. The price index is the closing price of 23,472 data samples, which is obtained from the Wind financial database (see Wind, 2019). One can refer to the source for the data used here in Supporting Information as Data S1. The software used for analysis is Python 3.6.

Figure 7 shows the distribution of the closing prices of the CSI 300 index from November 1, 2017, to November 1, 2019. It can be seen from the figure that the closing prices of the CSI 300 index are not normally distributed and present a double-peak trend. Furthermore, this double-peak trend is not symmetrical and is significantly different.

Table 2 is the descriptive statistics of the CSI 300 index. From the perspective of price time series, the overall price level fluctuates around 3,700 points, with a fluctuation range of about 300 points, and the price difference is about 1,400 points. From the perspective of returns time series, market returns have less fluctuation than individual stock returns. Although the return fluctuates around 0 and the amplitude is small, it shows a loss trend. The loss and profit are significantly asymmetrical; that is, the absolute value of the minimum and maximum differs by about 1%. This indicates that there exists an investment risk in the index, and the pursuit of high returns must bear greater risks. Through the discovery of the market index pattern, constructing a reasonable model and formulating effective strategies can reduce investment risks to a certain extent and ensure profit safety.

TABLE 2 Descriptive statistics of the CSI 300 index

	Mean	Median	Min	Max	SD	Skew	Kurt
Price	3,688.96928	3,784.845	2,941.05	4,401.74	339.522	−0.344	−0.939
Return	−0.000001	−0.000026	−0.033724	0.024812	0.002	−1.550	47.416

3.2.2 | Fitting for high-frequency data

In order to verify the fitting effect of the improved ARIMA model on the actual high-frequency data, this subsection selects the data from May 1, 2018, to May 1, 2019, as the fitting data set. Due to the large sample size of the data set, fitting with the total data will miss the characteristics of the sample in some specific time intervals. Therefore, the 48 high-frequency data in a day are used as a standard interval for modeling by rolling fitting.

The residuals are calculated based on the fitting result of the ARIMA model, whose distribution is shown in Figure 8. It can be seen in the figure that the residual sequence shows a trend of approximately normal distribution, which ranges from −100 to 100. By observing its positive and negative symmetry, it is found that the range of the positive value of the residual is wider than that of negative value. Such a prediction result will give investors the illusion of rising saturation, causing investors to make wrong investments and thus magnifying investment risks.

Figure 9 shows the trend chart between the price and the residuals of the ARIMA model for the CSI 300 index. It can be seen from the figure that there is no obvious linear relationship between residuals and prices. Although the price in the next period changed based on the price in the previous period and contained the information of the previous period, the index price is the overall performance of the market. The index price is also affected by economic factors, policy influence, and investor sentiment. The mechanism of these influencing factors is usually nonlinear, which indicates the importance of models improved to correct residuals based on deep learning

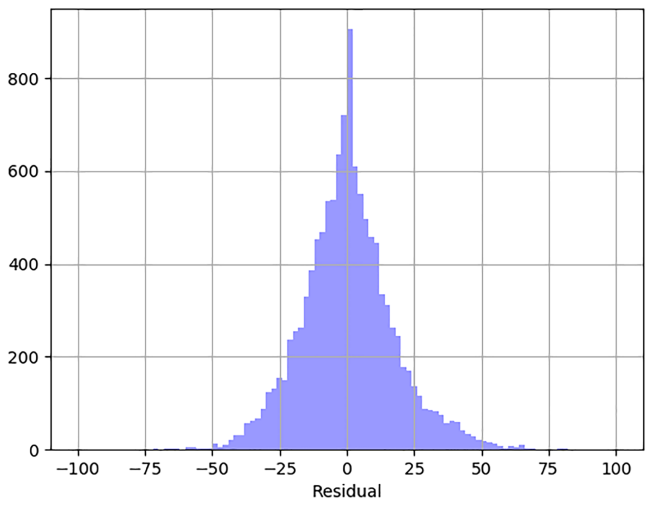

FIGURE 8 Distribution histogram for fitting residuals of price based on ARIMA [Colour figure can be viewed at wileyonlinelibrary.com]

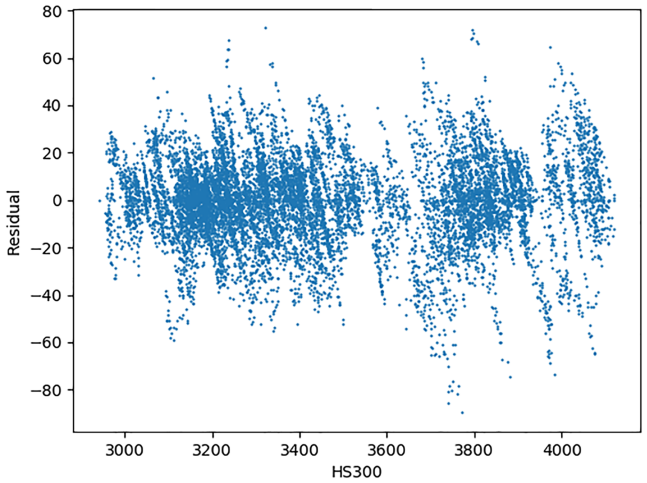

FIGURE 9 Price versus residual of CSI 300 index [Colour figure can be viewed at wileyonlinelibrary.com]

models. In addition, the recharacterization of nonlinearity based on deep learning tools can enrich the method for nonlinear data fitting, and has better practical application value for improving model performance.

Figure 10 depicts the fitted line of the price time series, in which the left is for all the data and the right is a partial enlarged view for the last 100 data. From the fitting trend of the left-hand figure, one can see that the traditional ARIMA model can roughly fit the overall trend of the time series, but the residuals with the original data are large and the direction is unstable. These problems can be verified from the partial enlarged image. This indicates that the fitting curve based on the ARIMA model leaves much room for improvement. Although machine learning methods such as the SVM model can significantly improve the fitting effect of the trend, the accuracy of the fitting results still needs to be improved. The deep learning approach, such as the LSTM model, is not only improved in the trend, but also relatively good in accuracy. Furthermore, the single machine learning model or deep learning model has not reached the best fitting effect. By modifying the residuals of traditional models, combining ARIMA models with deep learning models will employ the advantages of their respective advantage to improve model performance. For example, the ARIMA-SVM model and ARIMA-LSTM model in the figure both demonstrated good fitting performance.

The results in Table 3 mean that the performance of the ARIMA, SVM, LSTM, ARIMA-SVM, and ARIMA-LSTM models is improved in turn. Compared with the ARIMA model, the various evaluation criteria for the ARIMA-LSTM model are improved by 10.76, 2.43, 2.48, 0.29%, 2.48, and 2.42; compared with the LSTM model, the various evaluation indicators for the ARIMA-LSTM model are improved by 2.04, 0.74, 0.87, 0.06%, 0.85, and

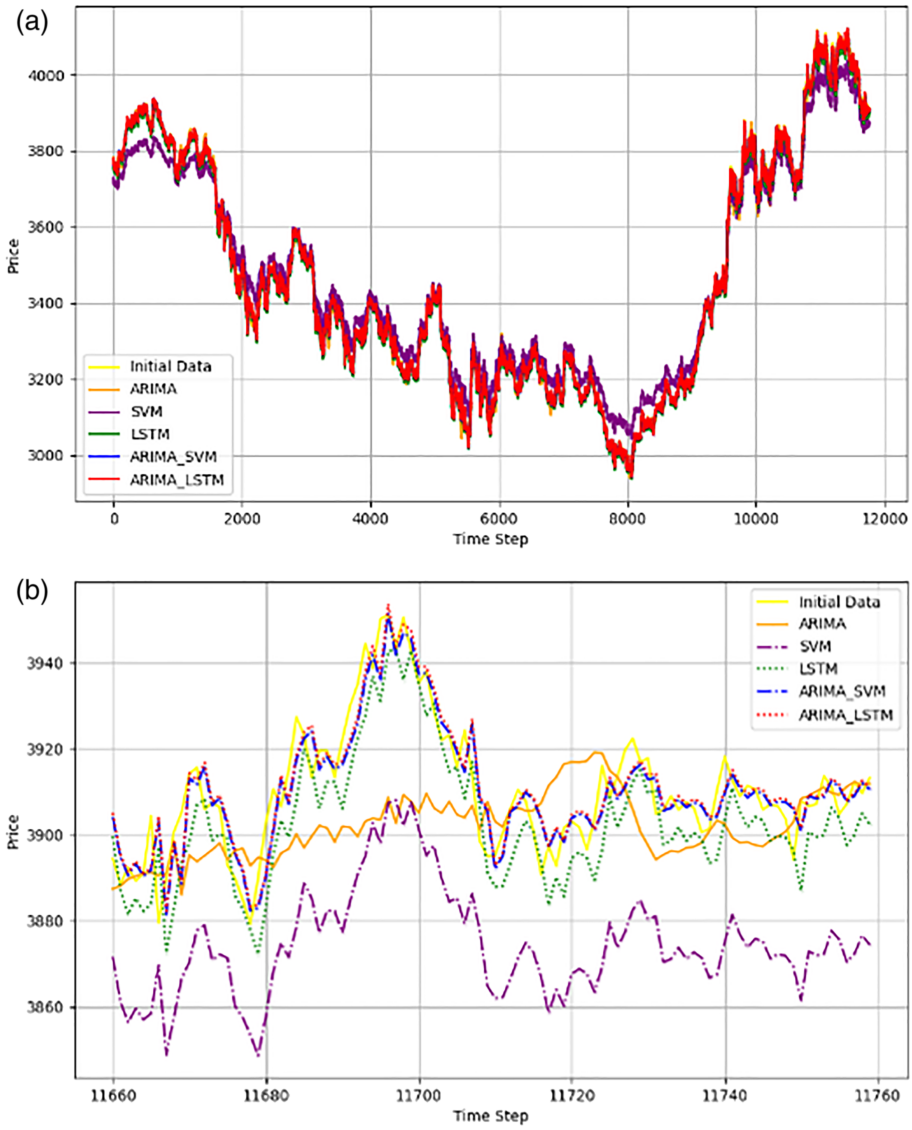

FIGURE 10 Fitting for closing price series of CSI 300 index [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 3 Model evaluation criteria for fitting of price series

	MSE (rank)	RMSE (rank)	MAE (rank)	R^2 (rank)	MAPE (rank)	RMSPE (rank)	Average rank
ARIMA	301.43095 (5)	17.36177 (5)	12.91725 (5)	0.99679 (5)	0.00372 (5)	0.00496 (5)	5
SVM	90.16773 (4)	9.49567 (4)	8.55822 (4)	0.99903 (4)	0.00249 (4)	0.00276 (4)	4
LSTM	77.81687 (3)	8.82139 (3)	6.92266 (3)	0.99917 (3)	0.00198 (3)	0.00251 (3)	3
ARIMA_SVM	25.76548 (2)	5.07597 (2)	3.71517 (2)	0.99973 (2)	0.00107 (2)	0.00145 (2)	2
ARIMA_LSTM	25.63969 (1)	5.06356 (1)	3.70907 (1)	0.99973 (1)	0.00107 (1)	0.00145 (1)	1

Note. Rank is the order of model performance in evaluation criteria and Average rank is the sample mean for ranks.

0.73. The ARIMA-LSTM model shows a greater improvement in performance for the single traditional ARIMA model or deep learning model, which can effectively characterize the trend of price time series.

The significance of model modification lies in the difference between the traditional model and the modified model. In order to verify the differences between the five models, a two-sample Kolmogorov–Smirnov (KS) test is used here to perform a pairwise test on the model residual series. The null hypothesis of the KS test is that the two sequences have the same distribution, and the calculated p-value is compared with the significance level $\alpha = 0.05$. If the p-value is less than 0.05, the null hypothesis is rejected and the series are considered to have different distribution characteristics, indicating that there are differences between the models.

Table 4 shows the *p*-statistics of the KS test for differentiation of the five models. From the statistical data, one can see that the *p*-values of the test for model differentiation between the residual series of each model and itself are 1 and those between different models are approximately 0. The *p*-value of the test between different models is less than 0.05, which rejects the null hypothesis and means that the two groups of data do not follow the same distribution; that is, the difference between the two models is large. The ARIMA, SVM, and LSTM models possess model differentiation, which indicates that there is a significant difference in performance among the traditional model, machine learning model and deep learning model. Moreover, the ARIMA-SVM and ARIMA-LSTM models are also different, which means that although the machine learning model and the deep learning model have a correction effect on series residuals, the correction mechanism is different. Through a high-dimensional mapping for data, the SVM model transforms the low-dimensional nonlinear problem into a high-dimensional linear problem. However, the correlation of high-frequency data itself will be missed during the transforming process. The LSTM model has a better ability to characterize the autocorrelation of high-frequency time series. According to the long-term and short-term memory characteristics of the data themselves, one can construct a deep learning mechanism to characterize the data themselves.

Considering that the ARIMA-LSTM model performs better in the fitting, we will verify the model's performance in the forecasting.

3.2.3 | Forecasting for high frequency data

The ARIMA modified model performs well in data fitting, and the actual scenario in which the model is applied needs to forecast the future trend more accurately. In order to verify whether the ARIMA improved model also has a good effect on the forecasting for price time series, the data here are selected from May 1, 2018, to May 1, 2019. Considering that the training for ARIMA-SVM and ARIMA-LSTM models requires a large sample size, so the data from November 1, 2017, to May 1, 2018, are added in as an additional training set. The forecasting approach is the same as the fitting one above. The 48 high-frequency data in a day are used as a standard interval for modeling by rolling fitting and the previous half-year data are set as the training set.

The forecasting curve for the price time series is shown in Figure 11, in which the left is for all the data and the right is a partial enlarged view for the last 100 data. From these figures, one can see that the forecasting effects of ARIMA-SVM and ARIMA-LSTM model are relatively satisfactory, which can accurately predict the features of the time series both in terms of trends and accuracy.

Table 5 demonstrates the evaluation criteria for forecasting results of these models. From this one can observe that the effect of the modified ARIMA model is better than the original model, and it shows a larger improvement. Moreover, the improvement effect of the modified model based on deep learning is better than that of the modified model based on machine learning, which shows that the ARIMA-LSTM model can not only accurately predict the trend of the price time series, but also has good stability. In comparison with the fitting effect of ARIMA-SVM, evaluation criteria such as MSE, RMSE, MAE, R^2, RMSPE, and RMSPE for the ARIMA-LSTM model has been improved 3.10%, 1.56%, 3.20%, 0.01‰, 3.74%, and 1.38%, respectively, which indicates the ARIMA-LSTM model possesses a much stronger generalization ability. The better performance of the ARIMA-LSTM model in forecasting can provide a reliable guarantee for its practical application. Furthermore, it can reduce the risks brought by market fluctuations, and find appropriate investment opportunities for profitable operations.

The time complexity of these models was calculated, and it was found that the parameter optimization for the best *p*-value, *d*-value, and *q*-value of ARIMA model takes an average 14.26 seconds. The average time of a pure deep learning LSTM model for training is

TABLE 4 Test for model differentiation

	ARIMA	SVM	LSTM	ARIMA_SVM	ARIMA_LSTM
ARIMA	1.00				
SVM	0.00	1.00			
LSTM	0.00	0.00	1.00		
ARIMA_SVM	0.00	0.00	0.00	1.00	
ARIMA_LSTM	0.00	0.00	0.00	0.00	1.00

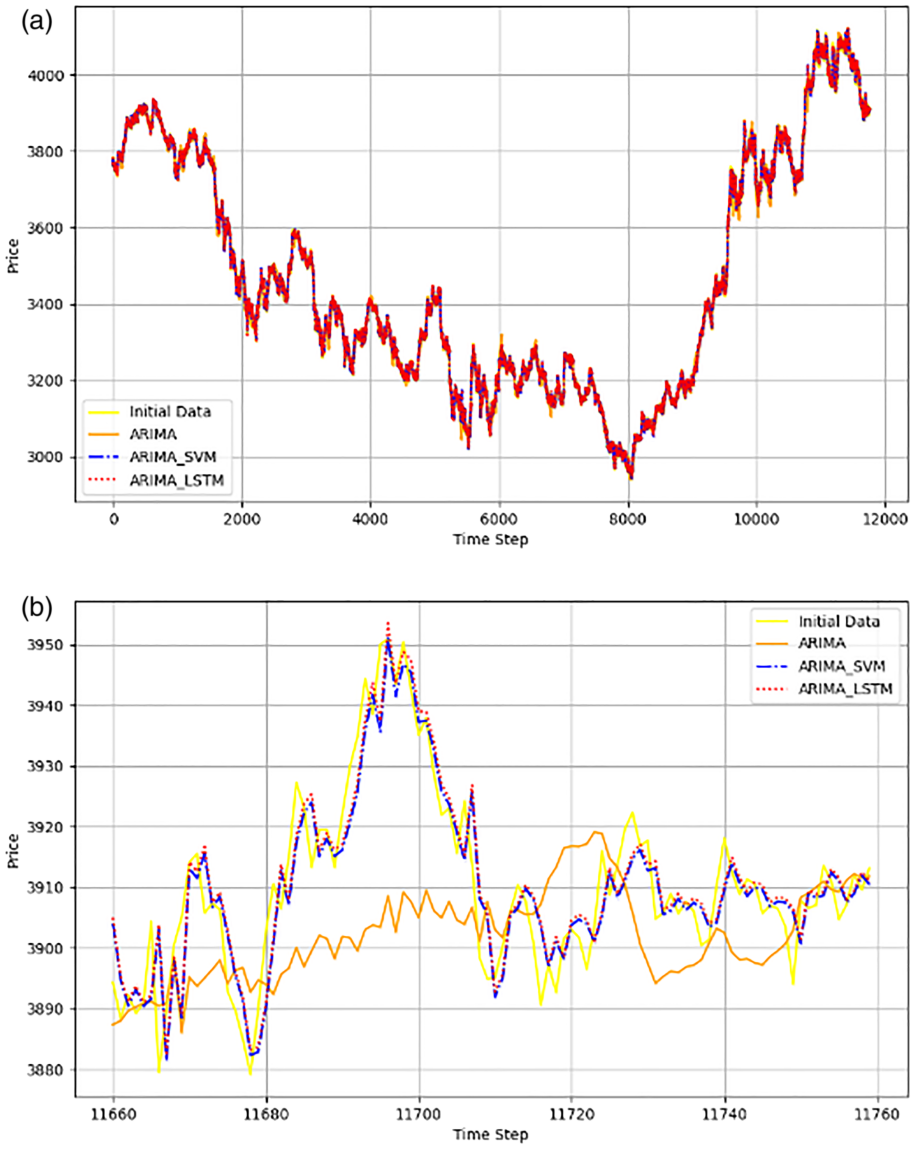

FIGURE 11 Forecasting for closing price series of CSI 300 index [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 5 Model evaluation criteria for forecasting of price series

	MSE (rank)	RMSE (rank)	MAE (rank)	R^2 (rank)	MAPE (rank)	RMSPE (rank)	Average rank
ARIMA	301.43095 (3)	17.36177 (3)	12.91725 (3)	0.99679 (3)	0.00372 (3)	0.00496 (3)	3
ARIMA_SVM	62.81497 (2)	7.92559 (2)	6.49898 (2)	0.99933 (2)	0.00188 (2)	0.00229 (2)	2
ARIMA_LSTM	24.84718 (1)	4.98470 (1)	3.59044 (1)	0.99974 (1)	0.00103 (1)	0.00143 (1)	1

2,482.68 seconds, while the training time of LSTM model for the residual series is 2,467.41 seconds, which indicates that the forecasting model for the price time series can effectively reduce its computational complexity after eliminating the linear relationship. The improvement effect is about 6.15%. The time improvement of the LSTM model for the residual series can effectively cover the time consumption of the ARIM model; that is, the ARIMA-LSTM model is better than the pure LSTM model in terms of training time.

3.2.4 | Strategy design for high frequency data

To construct the trading strategy for CSI 300 index futures based on the ARIMA-LSTM model, the window period for the backtest is selected from May 1, 2019, to November 1, 2019, and the window period for model training is selected from May 1, 2018, to May 1, 2019. The strategy is designed as follows: If the maximum value of 48 forecasting data in the future is greater than

3% according to the rolling forecasting results, the long or short operation is performed, and the position is closed when the price reaches the forecasting expectation. Taking the return of the CSI 300 index as the benchmark, the annual return of the CSI 300 index is 0.05354, and its volatility is 0.01053. In order to compare the strategic performance among these models, four criteria—the annualized return, Sharpe ratio, return volatility, and information ratio—were selected for evaluation.

From the results in Table 6, it can be observed that the performance of these models is significant compared with the return of the CSI 300 index. The annualized return of the ARIMA-LSTM model is 35.81% higher than that of the ARIMA model. Although its volatility has increased, the excess return per unit of total risk exposure is 1.6 times that of the ARIMA model. This shows that the ARIMA-LSTM model has a good practical application in strategy design. However, the strategy effect of ARIMA-SVM model is not only worse than that of the ARIMA-LSTM model, but also is far lower than that of the ARIMA model. Backtracking the forecasting data of the ARIMA-SVM model, it is found that although the ARIMA-SVM model is closer to the original value than the ARIAM model, its forecasting result was generally high and made the forecasting result too optimistic. It would be easy to cause a loss to investors if the forecasting result fails to meet expectations, reducing the benefits of the strategy. Although the ARIMA model performs worse in trend and accuracy than the ARIMA-SVM model and would miss some investment opportunities, it could avoid this risk and ensure the stability of returns. The improved ARIMA model based on deep learning not only performs better in trend and accuracy, but also most of the forecasting values are below the true value, which can control investment risks while capturing investment signals.

4 | ROBUSTNESS TEST

In this section, we will demonstrate the robustness and applicability of the improved method proposed in this paper through the returns time series of the market index and low-frequency data (e.g., monthly).

4.1 | Case 1: High-frequency returns time series

The returns time series of the market index and price time series have different properties. In order to verify the effectiveness of the ARIMA modified model, the same operation is performed on the returns time series to make a comparative analysis. Figure 12 depicts the distribution of the returns of the CSI 300 index. The returns are approximately normally distributed. However, there is asymmetry at the tail, and the probability of loss is greater than the probability of profit, which indicates that the market's "earning effect" is not obvious.

A diagram for the residual series by fitting the ARIMA model to the returns time series is depicted in Figure 13. It can be observed from the distribution of the residual sequence that the latter exhibits an approximately normal distribution, with most values near 0, which shows that the performance of the ARIMA model on the returns time series is better than that for the price time series.

Figure 14 is the fitting chart for the returns of the CSI 300. It can be seen that the performance of the ARIMA, SVM, and LSTM models are all poor, both in terms of trend and accuracy. Although the fitting effect of the ARIMA-SVM model is improved for the trend, there is

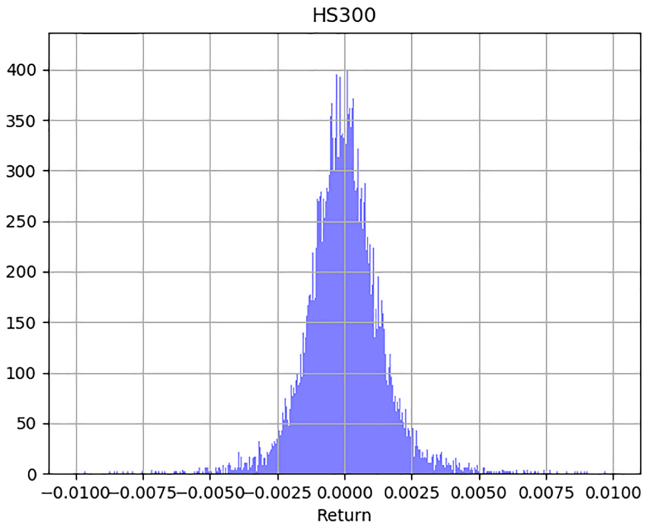

FIGURE 12 Distribution histogram for returns of CSI 300 index [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 6 Strategy evaluation criteria

	Annualized returns	Sharpe ratio	Volatility	Information ratio
ARIMA	0.56599	93.68373	0.00547	43.91174
ARIMA_SVM	0.40000	74.99134	0.00462	29.41087
ARIMA_LSTM	0.92410	150.09655	0.00580	75.30796

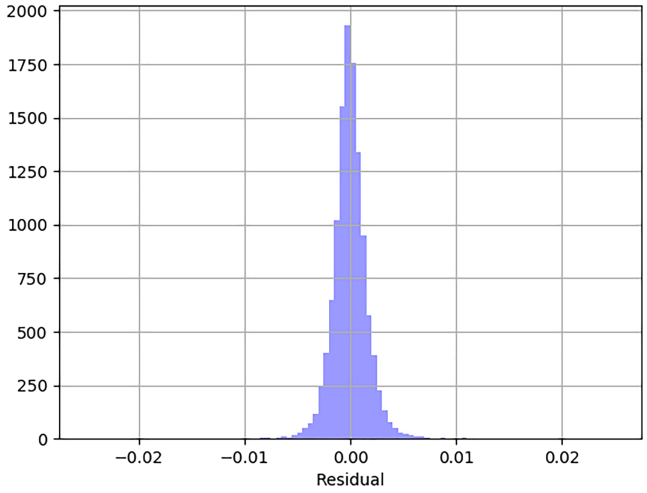

FIGURE 13 Distribution histogram for fitting residuals of returns based on ARIMA [Colour figure can be viewed at wileyonlinelibrary.com]

still a significant gap in accuracy. The ARIMA-LSTM model has significantly made both the trend and accuracy better, but it is far worse than the performance in terms of fitting effect for price time series.

Since there exists some data close to zero in the returns time series, two criteria—MAPE and RMSPE—in Table 7 are infinite, and these two criteria can be ignored here. Note that 'e-6' after MSE in Table 7 denotes $10^{(-6)}$. From the results in Table 7, one can see that the simple machine learning and deep learning model cannot fit the series trend better, and even the ARIMA-SVM model does not show good performance. Moreover, the goodness of fit for the five models on the data is less than 3%, which indicates that these models cannot be applied in practice. However, the performance of the ARIMA-LSTM model is still superior to other models, indicating that the improvement of the ARIMA model based on deep learning has practical significance.

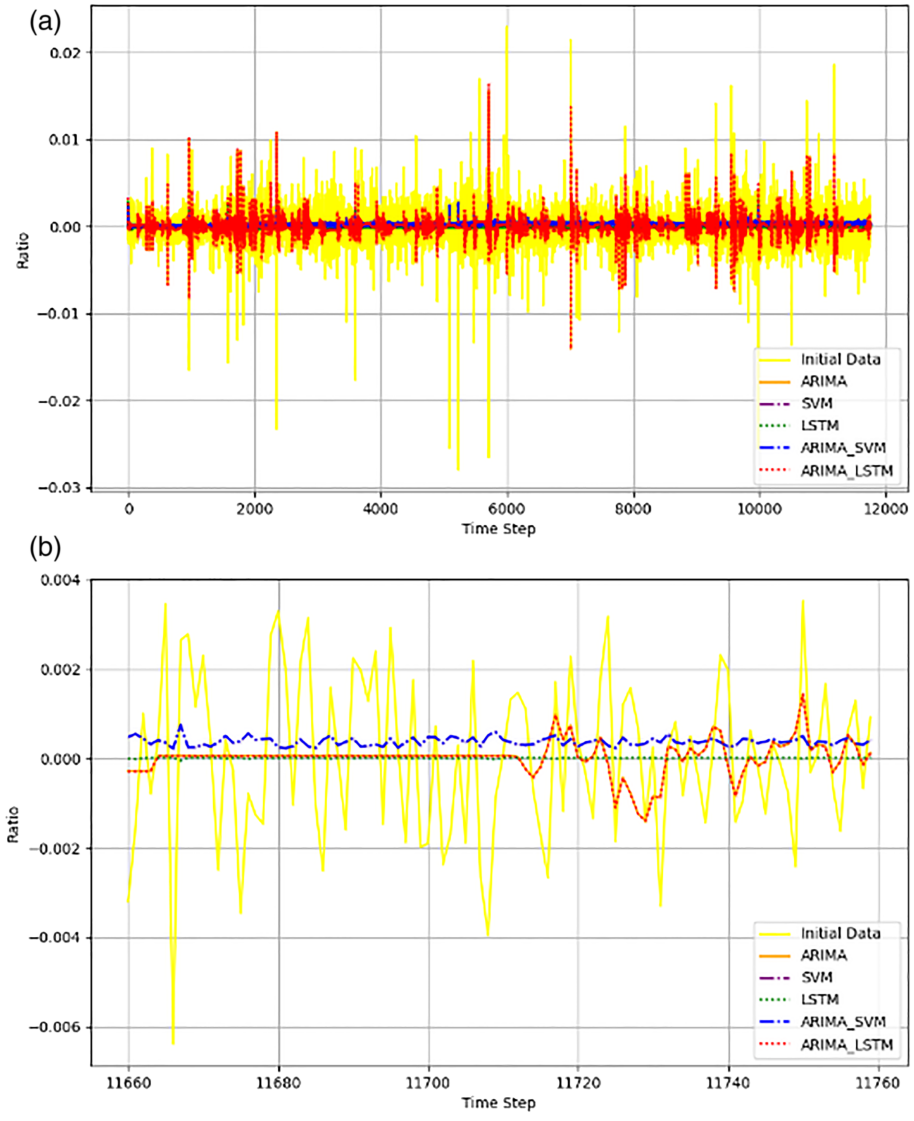

FIGURE 14 Fitting for returns series of CSI 300 index [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 7 Models evaluation criteria for fitting of returns series

	MSE(e-6) (rank)	RMSE (rank)	MAE (rank)	R^2 (rank)	MAPE (rank)	RMSPE (rank)	Average rank
ARIMA	3.24104 (3)	0.00180 (3)	0.00119 (3)	0.00234 (5)	∞	∞	3
SVM	3.31818 (5)	0.00182 (5)	0.00123 (5)	0.02619 (3)	∞	∞	5
LSTM	3.22215 (2)	0.00180 (2)	0.00119 (2)	0.02679 (2)	∞	∞	2
ARIMA_SVM	3.31011 (4)	0.00182 (4)	0.00123 (4)	0.02370 (4)	∞	∞	4
ARIMA_LSTM	3.22040 (1)	0.00180 (1)	0.00119 (1)	0.02710 (1)	∞	∞	1

TABLE 8 Model evaluation criteria

	MSE (rank)	RMSE (rank)	MAE (rank)	R^2 (rank)	MAPE (rank)	RMSPE (rank)	Average rank
ARIMA	213.30441 (3)	14.60494 (3)	12.04156 (3)	0.58358 (3)	0.00418 (3)	0.00525 (3)	3
ARIMA_SVM	45.84200 (2)	6.77067 (2)	5.12385 (2)	0.91051 (2)	0.00169 (2)	0.00222 (2)	2
ARIMA_LSTM	45.33025 (1)	6.73277 (1)	4.94918 (1)	0.91115 (1)	0.00163 (1)	0.00221 (1)	1

4.2 | Case 2: Low-frequency monthly price time series

In order to verify the broad applicability of the ARIMA-LSTM model, monthly data on the closing prices of the CSI 300 index from November 2011 to October 2019 are selected for test. One can refer to the source for the data used here in Supporting Information as Data S2. The total sample size is 96, which is low frequency data. The purpose of this verification is to show that the ARIMA modified model can be applied to low frequency data.

From the results in Table 8, we can see that the ARIMA-LSTM model also performs well on low-frequency data, and is better than the ARIMA-SVM model in various criteria. For the ARIMA model, the modified model greatly reduces the forecasting bias and also effectively improves the goodness of fit. Moreover, compared with the forecasting evaluation criteria for the high-frequency data, it is found that the performance of the ARIMA model in low-frequency data is better, which indicates that the migration of ARIMA model from low-frequency data to high-frequency data has certain limitations. However, although the performance for the modified models has been improved, the improvement for the models' evaluation criteria is relatively low compared with those for high-frequency data. This is due to the lack of sampling data; that is, the monthly data are different from the sample size of the 5-minute high frequency data by a factor of 100. For machine learning and deep learning models which rely on big data, the lack of sample size will prevent the model from acquiring data characteristics, and then they will not be able to effectively capture abnormal fluctuations and characterize future trends for the time series.

5 | CONCLUSIONS AND OTHER EXTENSIONS

With the popularity of high-frequency financial data, its time-varying and nonlinear characteristics are becoming increasingly obvious, but these characteristics have a great influence on the forecasting of time series through the traditional ARIMA model. The ARIMA model has a large deviation from the forecasting of high-frequency financial time series, which also shows that the ARIMA model has much room for improvement. Due to the gradual maturity of machine learning and deep learning techniques, the combination of machine learning model and deep learning model in the ARIMA model can be used to correct its error, which can compensate for the time-varying and nonlinear features of high-frequency sequences, and effectively improve the accuracy of the forecasting for the model. From the results of fitting and forecasting verification, the modified ARIMA model based on machine learning or deep learning has been improved significantly. In particular, the ARIMA-LSTM model possesses better performance and stability, which is also suitable for low-frequency time series. Moreover, compared with only machine learning and deep learning, after handling by the traditional model such as ARIMA, residual modeling by the LSTM model will be less computationally complex and occupy less computational time. This means that the time saved can completely cover the time consumption for the optimization of the ARIMA model.

Combining the ARIMA model with the deep learning model can provide an approach on forecasting for high-frequency financial time series and can solve the shortcomings of forecasting based on the traditional models.

Furthermore, one can also apply the improved model to forecast the individual stock for a reliable guidance of selection and timing and to prevent risks and get benefits. Moreover, one can adopt this idea to improve and optimize the forecasting model such as GARCH for volatility of high-frequency financial time series. These topics are currently under investigation. Furthermore, we will endeavor to provide the mathematical or statistical foundations for this method in the future.

ACKNOWLEDGMENTS

The authors would like to thank the editor, associate editor and two anonymous referees for their valuable suggestions, which greatly improved our paper. This research work is supported by National Natural Science Foundation of China (11901397), Ministry of Education, Humanities and Social Sciences project (18YJCZH153), National Statistical Science Research Project (2018LZ05), Youth Academic Backbone Cultivation Project of Shanghai Normal University (310-AC7031-19-003021), General Research Fund of Shanghai Normal University (SK201720) and Key Subject of Quantitative Economics of Shanghai Normal University (310-AC7031-19-004221) and Academic Innovation Team of Shanghai Normal University (310-AC7031-19-004228).

DATA AVAILABILITY STATEMENT

The data set for the empirical analysis is available as a supplementary file, which can also be obtained from Wind, a service company in mainland China providing financial data and information as Bloomberg.

REFERENCES

Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, *71* (2), 579–625. https://doi.org/10.1111/1468-0262.00418

Barrales, E. O. (2012). Lessons from the flash crash for the regulation of high-frequency traders. *Fordham Journal of Corporate and Financial Law*, *17*, 1195–1262.

Biais, B., & Foucault, T. (2014). HFT and market quality. *Bankers, Markets and Investors*, *128*(1), 5–19.

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control*. Hoboken, NJ: Wiley.

Chevallier, J., & Sévi, B. (2012). On the volatility–volume relationship in energy futures markets using intraday data. *Energy Economics*, *34*(6), 1896–1909.

de Oliveira, J. F. L., & Ludermir, T. B. (2014). A distributed PSO-ARIMA-SVR hybrid system for time series forecasting. In *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 3867–3872).

Ding, Z., Granger, C. W., & Engle, R. F. (1993). A long memory property of stock market returns and a new model. *Journal of Empirical Finance*, *1*(1), 83–106. https://doi.org/10.1016/0927-5398(93)90006-D

Giot, P., Laurent, S., & Petitjean, M. (2010). Trading activity, realized volatility and jumps. *Journal of Empirical Finance*, *17*(1), 168–175. https://doi.org/10.1016/j.jempfin.2009.07.001

Hanson, T. A., & Hall, J. (2012). Statistical arbitrage trading strategies and high frequency trading. Available at SSRN 2147012.

Hilpisch, Y. (2014). *Python for Finance: Analyze big financial data*. Farnham, UK: O'Reilly Media.

Jacquier, E., Polson, N. G., & Rossi, P. E. (2002). Bayesian analysis of stochastic volatility models. *Journal of Business and Economic Statistics*, *20*(1), 69–87. https://doi.org/10.1198/073500102753410408

Jobson, J. D., & Korkie, B. M. (1981). Performance hypothesis testing with the Sharpe and Treynor measures. *Journal of Finance*, *36*(4), 889–908. https://doi.org/10.1111/j.1540-6261.1981.tb04891.x

Kavousi-Fard, A., & Kavousi-Fard, F. (2013). A new hybrid correction method for short-term load forecasting based on ARIMA, SVR and CSA. *Journal of Experimental and Theoretical Artificial Intelligence*, *25*(4), 559–574. https://doi.org/10.1080/0952813X.2013.782351

Laughlin, G., Aguirre, A., & Grundfest, J. (2014). Information transmission between financial markets in Chicago and New York. *Financial Review*, *49*(2), 283–312. https://doi.org/10.1111/fire.12036

Menkveld, A. J. (2014). High frequency traders and market structure. *Financial Review*, *49*(2), 333–344. https://doi.org/10.1111/fire.12038

Ramos, P., Santos, N., & Rebelo, R. (2015). Performance of state space and ARIMA models for consumer retail sales forecasting. *Robotics and Computer-Integrated Manufacturing*, *34*, 151–163. https://doi.org/10.1016/j.rcim.2014.12.015

Slim, S., & Dahmene, M. (2016). Asymmetric information, volatility components and the volume–volatility relationship for the CAC40 stocks. *Global Finance Journal*, *29*, 70–84. https://doi.org/10.1016/j.gfj.2015.04.001

Van Gestel, T., Espinoza, M., Baesens, B., Suykens, J. A., Brasseur, C., & De Moor, B. (2006). A Bayesian nonlinear support vector machine error correction model. *Journal of Forecasting*, *25*(2), 77–100. https://doi.org/10.1002/for.975

Visser, M. P. (2010). Garch parameter estimation using high-frequency data. *Journal of Financial Econometrics*, *9*(1), 162–197.

Walker, G. T. (1931). On periodicity in series of related terms. *Proceedings of the Royal Society of London, Series A*, *131*(818), 518–532.

Wang, Y., Wang, J., Zhao, G., & Dong, Y. (2012). Application of residual modification approach in seasonal ARIMA for electricity demand forecasting: A case study of China. *Energy Policy*, *48*, 284–294. https://doi.org/10.1016/j.enpol.2012.05.026

Wind. (2019). Five-minute high frequency data and monthly low frequency data of the Shanghai and Shenzhen 300 Index Shanghai, China: Wind Information Co.

Yule, G. U. (1927). VII. On a method of investigating periodicities disturbed series, with special reference to Wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society of London. Series A*, *226*(636–646), 267–298.

AUTHOR BIOGRAPHIES

Zhenwei Li is a Master of Finance at the School of Finance and Business, Shanghai Normal University, China. His research interest is Financial Statistics and Modeling.

Jing Han is a Master of Insurance at the School of Finance and Management, Shanghai University of International Business and Economics, China. Her research interest is Insurance and Actuarial.

Yuping Song is an Associate Professor at the School of Finance and Business, Shanghai Normal University, China. He is an expert on analysis of high frequency financial data.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Li Z, Han J, Song Y. On the forecasting of high-frequency financial time series based on ARIMA model improved by deep learning. *Journal of Forecasting*. 2020;39: 1081–1097. https://doi.org/10.1002/for.2677