mathematics

MDPI

Article

Combining Autoregressive Integrated Moving Average Model and Gaussian Process Regression to Improve Stock Price Forecast

Shiying Tu [1], Jiehu Huang [1], Huailong Mu [1], Juan Lu [1] and Ying Li [2,*]

[1] Guangxi Key Laboratory of Ocean Engineering Equipment and Technology, Beibu Gulf University, Qinzhou 535011, China; 2202011130@stu.bbgu.edu.cn (S.T.); hjh18177563137@163.com (J.H.); 2302010120@stu.bbgu.edu.cn (H.M.); lujuan3623366@163.com (J.L.)
[2] College of International Studies, Beibu Gulf University, Qinzhou 535011, China
* Correspondence: lying@bbgu.edu.cn

Abstract: Stock market performance is one key indicator of the economic condition of a country, and stock price forecasting is important for investments and financial risk management. However, the inherent nonlinearity and complexity in stock price movements imply that simple conventional modeling techniques are not adequate for stock price forecasting. In this paper, we present a hybrid model (ARIMA + GPRC) which combines the autoregressive integrated moving average (ARIMA) model and Gaussian process regression (GPR) with a combined covariance function (GPRC). The proposed hybrid model can account for both the linearity and nonlinearity in stock price movements. Based on daily data on three stocks listed on the Shanghai Stock Exchange (SSE), it is found that GPRC outperforms GPR with a single covariance function. Further, the proposed hybrid model is compared with the ARIMA model, artificial neural network (ANN), and GPRC model. Based on the forecasting trend and the statistical performance of the four models, the ARIMA + GPRC model is found to be the dominant model for stock price forecasting and can significantly improve forecasting performance.

Keywords: stock price forecast; combining model; autoregressive integrated moving average; Gaussian process regression; covariance function

MSC: 62P30

Citation: Tu, S.; Huang, J.; Mu, H.; Lu, J.; Li, Y. Combining Autoregressive Integrated Moving Average Model and Gaussian Process Regression to Improve Stock Price Forecast. *Mathematics* **2024**, *12*, 1187. https://doi.org/10.3390/math12081187

Academic Editor: Anatoliy Swishchuk

Received: 12 March 2024
Revised: 9 April 2024
Accepted: 12 April 2024
Published: 15 April 2024

1. Introduction

Stock market performance is one key indicator of the economic condition of a country [1]. A stock market is a public and open market in which stocks and derivatives of a company are traded [2]. Stock prices fluctuate in the stock market because there are many factors influencing the stock market, such as general economic conditions, political events, and traders' expectations [3]. Moreover, stock price movements are characterized by nonlinearities and high-frequent undulant components. Stock price forecasting is important for investors and stockbrokers. Thus, many researchers and financial analysts have tried to predict stock price trends with various techniques proposed over the years [4].

A time series model is often used to forecast stock prices. It is based on the past observations of the same variable, to build a model which can be used to predict the future trend. Note that it is not necessary to know the information of other variables in a time series model. One of the most efficient and widely used forecasting techniques for time series in social science is the autoregressive integrated moving average (ARIMA) model. The popularity of this model owes to its statistical properties and the famous Box–Jenkins methodology [5] in the modeling process. Recently, based on an analysis of three stock markets, Alshawarbeh and Abdulrahman [6] revealed that the hybrid ARIMA-ANN is overall superior to individual ANN and ARIMA models. Chen et al. [7] built an advanced hybrid model based on ARIMA to predict the prices of game stocks. ARIMA models

are relatively more efficient and robust than complex structural models for short-run forecasting [8]. However, there is a major limitation for ARIMA, that is, the model presumes a linear correlation structure, and the nonlinear structure cannot be captured by the ARIMA model effectively. Thus, it is not satisfactory for many complex real-world problems. To circumvent this shortcoming and to model the nonlinear relationships observed in the real world, some nonlinear models such as the autoregressive conditional heteroscedastic (ARCH) model [9] and the threshold autoregressive (TAR) model [10] have been proposed. However, these nonlinear models only apply to some specific nonlinear relationships [11], and they may not work for other nonlinear structures of time series.

On the other hand, an artificial neural network (ANN) and support vector machine (SVM) are widely used in stock price forecasting because of their ability to capture the nonlinear characteristics in the data. Stock volatility represents a crucial impact in asset pricing models, portfolio management, and trading strategies. D'Ecclesia and Clementi [12] found that the artificial neural network (ANN) was the most accurate in tracking the implied stock volatility. Based on the characteristics of the same stock in different periods and the characteristics of different stocks in the same period, an adaptive SVR was presented by Guo et al. [13]. However, these methods still have some defects. They are good at capturing the nonlinear features in the data but may neglect some linear features.

In recent years, with the development of computer technology, deep learning has become a popular algorithm, which has been applied in image recognition, unmanned driving, and other fields. It is also applied in stock forecasting. Li et al. [14] proposed a clustering-enhanced deep learning framework to improve the accuracy of stock price prediction. Agrawal et al. [15] built an evolutionary deep learning model (EDLM), which is used to identify stock price trends. The model implements the deep learning model and establishes the concept of the related tensor. Although the deep learning algorithm shows good accuracy in stock price prediction, it has many internal parameters, and the process of adjusting parameters is complex and time-consuming.

Because of the non-stationary and chaotic characteristics of the stock price data [16], the time series data of stock prices may contain both linear and nonlinear features. Therefore, the combination of linear and nonlinear models can account for the complexity in modeling the stock market and overcome the limitation of individual technology [17]. Pai et al. [18] proposed a hybrid methodology based on the SVM model and ARIMA model to predict stock price. Based on the stock prices of 10 companies, it was found that the hybrid methodology performed better than the ARIMA model or the SVM model in predicting stock prices. Hajirahimi and Khashei proposed a new parallel hybrid model to implement an integrated hybrid framework in which all pure linear and nonlinear patterns in real time series can be appropriately simulated [19]. Li et al. [20] built some combined prediction models including ARIMA + SVM and ARIMA + ANN based on the artificial intelligence technique.

Adapting to the advantages and disadvantages of the algorithm or the characteristics of the data, a combined model can overcome the limitations of the single techniques and improve the prediction effect. Although some combination models have been applied in stock price forecasting, it is difficult to find an effective model with strong adaptability for the time series of stock prices with nonlinearity, interruption, and high-frequency fluctuation. Therefore, investors and financial institutions are keen to seek more effective forecasting models.

Gaussian process regression (GPR) is a highly valuable technique in the field of machine learning, and it is widely applied in many nonlinear problems [21–24] due to its advantages of being flexible, probabilistic, and nonparametric in the nonlinear modeling process. A GPR model is completely specified by the mean function and the covariance function. Therefore, compared to the neural network and support vector machines, the number of parameters for GPR is much less, and parameter optimization and convergence for GPR are easier. However, little research has been conducted on applying the Gaussian process regression algorithm in the field of stock market prediction.

In this paper, the ARIMA model which deals with linear characteristics and the GPR model which focuses on nonlinear characteristics are combined to form a new stock price forecasting model. It is expected that both linear and nonlinear characteristics of stock price movements can be captured by the proposed model, and better forecasting results can thus be obtained.

Specifically, the ARIMA model is first used to extract the linear features of the sample data for fitting and prediction. The residual sequence of the original data is obtained based on the fitting results. Then, the GPR model is used to train the residual sequence to obtain the nonlinear characteristics of the data. Finally, the results of the ARIMA model are modified and fine-tuned to improve the accuracy of forecasting, which is important for making investment decisions.

The rest of the paper is organized as follows. The stock price forecast model framework is developed in Section 2. The methodology is described in Section 3. Section 4 presents the results and analysis. Section 5 summarizes the paper with some concluding remarks.

2. Stock Price Forecast Model Framework

Figure 1 shows the framework of the stock price forecast model consisting of four steps.

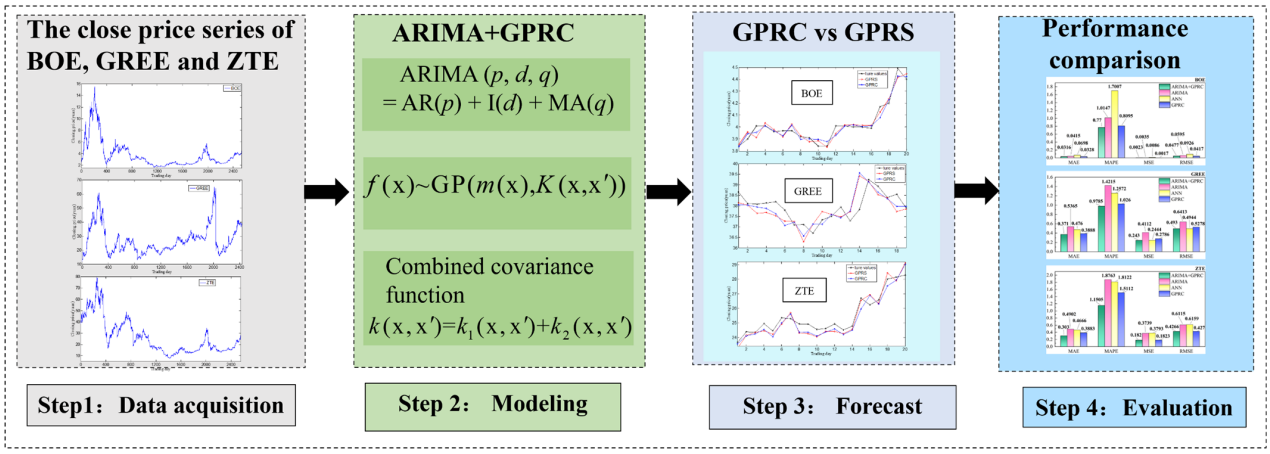


Figure 1. Stock price forecast model framework.

In Figure 1, step 1 is data acquisition. This paper selects the closing prices of Beijing Oriental Electronics Technology Group Co., Ltd. (BOE) (Beijing, China), Gree Electric Appliances (GREE) (Zhuhai, China), and Zhongxing Telecommunication Equipment Corporation (ZTE) (Shenzhen, China) on the Shanghai Stock Exchange (SSE) (Shanghai, China) from 4 January 2007 to 30 September 2017, a total of 3923 trading days for the stock price forecast, of which 3893 trading days from 4 January 2007 to 31 August 2017 are the training set, and the remaining 30 trading days in September 2017 are the test set.

Step 2 is modeling. In order to solve the complex linear and nonlinear relationship between stock price time series, the ARIMA model can preliminarily estimate the range of autoregressive p and moving average q through the diagrams of the autocorrelation function (ACF) and partial autocorrelation function (PACF), which is suitable for capturing linear features. At the same time, two different covariance functions are combined to form a GPR model to focus on nonlinear characteristics, and then a mixed model of the ARIMA model and GPRC model is constructed for the stock price forecast.

Step 3 is the forecast. The forecast performance of the GPR model with a single covariance function (GPRS) and the GPR model with a combined covariance function (GPRC) are compared based on the three stocks.

Step 4 is evaluation. The mean absolute error (MAE), mean absolute percentage error (MAPE), mean square error (MSE), and root-mean-square error (RMSE) are common indicators of the forecast model. To verify the effectiveness of the proposed hybrid model, the mixed model (ARIMA + GPRC) is compared with the ARIMA, GPRC, and ANN models.

3. Methodology

3.1. ARIMA (p, d, q)

According to the autoregressive integrated moving average model ARIMA (p, d, q), the future value of a variable can be predicted by the past observations and random errors as follows [18,20]. The following introduction to the ARIMA model and related equations are all from Refs. [18,20].

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + \theta_0 + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \quad (1)$$

where y_t denotes the actual value, and ε_t is the random error at time t; $\varphi_i (i = 1, 2, \cdots, p)$ and $\theta_j (j = 0, 1, 2, \cdots, q)$ are autoregressive coefficients and moving average coefficients, respectively. When $q = 0$, Equation (1) is an autoregressive (AR) model of order p. When $p = 0$, the model becomes a moving average (MA) model of order q.

Differencing and using logarithm are common ways to make the data become stable, and d is used to denote the time of differencing or logarithm. It is critical to determine the orders of the model, i.e., the p value and q value in ARIMA modeling. Usually, the determination of orders can be conducted by observing the correlation diagram of the autocorrelative function (ACF) and partial autocorrelation function (PACF). However, it is not accurate to determine the orders only through the trail and truncation of the ACF and PACF. Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) can be used to further determine the values of p and q.

$$\mathrm{AIC}(p) = n \ln \sigma_\varepsilon^2 + 2p \quad (2)$$

$$\mathrm{BIC}(p) = n \ln \sigma_\varepsilon^2 + p \ln n \quad (3)$$

where n denotes the number of observations, and p is the number of parameters which are to be estimated, i.e., the order of autoregression. $\sigma_\varepsilon^2 = \sum_{t=1}^{n} \varepsilon_t^2$ represents the sum of squared residuals. In (2) and (3), $\ln \sigma_\varepsilon^2$ decreases as p increases, while $2p$ and $p \ln n$ increase as p increases. Thus, the p value is considered to be an appropriate order of the model when AIC(p) or BIC(p) is minimalized. AIC(p) and BIC(p) are applied for small and large sample data, respectively. When the orders are specified, the estimation of the model parameters is straight-forward and can be conducted by the least square method.

3.2. The GPR Model Based on a Single and Combined Covariance Function

It is well known that the Gaussian process (GP) model has the advantages of a flexible modeling ability and high approximation accuracy for nonlinear problems, compared to other classes of nonlinear models. In particular, the GP model can approximate a function with high accuracy when the sample data are small. The usage and modeling properties of the Gaussian process were reviewed by Rasmussen et al. [25]. The following introduction to the GPR model and related equations are all from Ref. [25] A Gaussian process is completely specified by its mean function m(x) and covariance function $K(x, x')$; it can be represented as $f(x) \sim GP(m(x), K(x, x'))$, where $x, x' \in R^d$ are arbitrary random variables.

For the training set $D = \{(x_i, y_i) | i = 1, 2, \cdots, n\}$, observation value $\mathbf{Y} = (y_1, y_2, \cdots, y_i)^T$, the prior distribution of $\mathbf{Y}$ and the joint prior distribution between $\mathbf{Y}$ and the forecast valued of the test point y^* are obtained as follows.

$$Y \sim N(0, K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathrm{I}_n)$$

$$\begin{bmatrix} \mathbf{Y} \\ y^* \end{bmatrix} \sim N\left(0, \begin{bmatrix} \mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I}_n & \mathbf{K}(\mathbf{X}, x^*) \\ \mathbf{K}(x^*, \mathbf{X}) & \mathbf{K}(x^*, x^*) \end{bmatrix}\right)$$

where $\mathbf{I}_n$ denotes an n-dimensional identity matrix, and $\mathbf{X}$ is the training set data; x^* is defined as test point data; $\mathbf{K}(\mathbf{X}, \mathbf{X})$ is the n-th covariance matrix of $\mathbf{X}$. $\mathbf{K}(\mathbf{X}, x^*)$ represents the covariance matrix of $\mathbf{X}$, and x^*, $\mathbf{K}(x^*, x^*)$ is the x^* covariance matrix.

The covariance function is a key part in the GP model. The similarity of data points is very important in supervised learning. For example, if points a and b are similar, the corresponding function values $f(a)$ and $f(b)$ will also be similar. The covariance function is defined as the distance between the output variables corresponding to two random input variables and is used to measure the similarity of data points. There are many kinds of covariance functions in the GP model, and it is not easy to choose the most suitable one. The squared exponential covariance function is most frequently used, and it is given as

$$K_1(x, x') = \sigma_f^2 \exp(-\frac{(x - x')^2}{2l^2}) + \sigma_n^2 \delta(x, x') \tag{4}$$

where σ_f^2 represents the overall measurement of prior knowledge; l_f is a correlation measurement hyperparameter. $\delta(x, x')$ is the Kronecker operator, whose value when two data are the same is 1, otherwise its value is 0.

However, a single covariance function only captures a single feature. In order to describe different characteristics of the data and enhance the nonlinear mapping capability of the covariance function, this paper combines the squared exponential covariance function with the linear covariance function to capture the smooth and secular trend, respectively. The expression of the exponential covariance function is shown in Equation (5).

$$K_2(x, x') = xx' \tag{5}$$

The combined covariance function is

$$K(\mathrm{x},\ \mathrm{x}') = K_1(\mathrm{x},\ \mathrm{x}') + K_2(\mathrm{x},\ \mathrm{x}') \tag{6}$$

Note that the sum of two kernels is still a kernel [25]. All the hyperparameters in the GP model with the combined covariance function can be denoted as Equation (7).

$$\theta = \{l, \sigma_n, \sigma_f\} \tag{7}$$

After determining the type of covariance function, the calculation of hyperparameters θ is a crucial step in determining the GPR model. The marginal likelihood of the GPR model provides an effective method for directly solving hyperparameters using the training set, which is an advantage of GPR.

The process of solving hyperparameters is as follows: first, determine the likelihood function $L(\theta)$ as the negative logarithm of the conditional probability of the training set, and obtain the partial derivative of the hyperparameter. Then, using the conjugate gradient method to solve the hyperparameters when the likelihood function reaches its minimum value, the obtained hyperparameters are the optimal hyperparameters. The partial derivatives of the likelihood function are shown in Equation (8) [25].

$$L(\theta) = \frac{1}{2}\mathbf{Y}^T\mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_n^2\mathbf{I}_n)^{-1}\mathbf{Y} + \frac{1}{2}\lg\left|\mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_n^2\mathbf{I}_n\right| + \frac{\pi}{2}\lg 2\pi \tag{8}$$

According to the Bayesian posterior probability formula, the posterior distribution of the forecast value y^* of the test point is $y^*\left|X, Y, x^* \sim N(\mu_{y^*}, \sigma_{y^*}^2)\right.$, and the mean μ_{y^*} is taken as the predicted forecast at x^*, where the mean μ_{y^*} and variance $\sigma_{y^*}^2$ are represented by Equations (9) and (10), respectively.

$$\mu_{y^*} = \mathbf{K}(x^*, \mathbf{X})[\mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_n^2\mathbf{I}_n)^{-1}y^* \tag{9}$$

$$\sigma_{y^*}^2 = \mathbf{K}(x^*, x^*) - \mathbf{K}^T(x^*, \mathbf{X})[\mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_n^2\mathbf{I}_n)^{-1}\mathbf{K}(x^*, \mathbf{X}) \tag{10}$$

3.3. Combining Model of ARIMA and GPR (ARIMA+GPR)

The stock price time series is impacted by many factors, and it may not be simply linear or nonlinear. Thus, stock price forecasting is a complex problem. It may be inadequate to approximate the stock price time series by ARIMA models due to the assumption of the linear structure. Moreover, using GPR to handle linear and nonlinear relationships in the same time series may yield mixed results, although GPR has good nonlinear adaptability [25]. Therefore, a hybrid methodology based on ARIMA models and GPR is proposed for stock forecasting in this paper. It can account for both the linear and nonlinear modeling capabilities. When a time series has a linear autocorrelation structure and a nonlinear component, the combining methodology can be effective. The forecast of the combined methodology is as follows:

$$\hat{y}_t = \hat{L}_t + \hat{N}_t \tag{11}$$

where $\hat{L}_t$ and $\hat{N}_t$ denote the estimated value of the linear component and nonlinear component, respectively, at time t.

The implementation of the hybrid methodology consists of three steps. First, use ARIMA to model sample data, and obtain the forecast value $\hat{L}_t$ at time t. Second, calculate the deviation between the forecast value $\hat{L}_t$ of the ARIMA model and the actual value y_t, i.e.,

$$\varepsilon_t = y_t - \hat{L}_t \tag{12}$$

Third, in order to indicate the nonlinear relationships, GPR is used to model residuals which come from the ARIMA model. The GPR model for the residuals is as follows:

$$\hat{\varepsilon}_t = f(\varepsilon_{t-1}, \varepsilon_{t-2}, \cdots, \varepsilon_{t-r}) + \Delta_t \tag{13}$$

where r is the larger value between the autoregressive parameter (p) and moving average parameter (q) of the ARIMA model, Δ_t expresses the random error, and f denotes the nonlinear function determined by GPR. So $\hat{\varepsilon}_t$ represents the forecasting value of the nonlinear component at time t, namely $\hat{N}_t$ from Equation (11).

4. Results and Analysis

The Shanghai Stock Exchange (SSE) is one of the two stock exchanges in Mainland China, and it is one of the largest stock markets in the world. More and more investors from both China and abroad are attracted to the SSE for the potential to gain high returns. Thus, it is important to understand the stock price movements in the SSE, and stock price forecasting has been an important research topic. In this paper, we therefore use published stock price data from the SSE.

MATLAB R2018a software is used for testing and validation. The proposed hybrid GPR model and GPR model is implemented with the help of the GPML (Gaussian processes for machine learning) toolbox. The neural network toolbox in MATLAB R2018a is adopted for building the ANN model.

4.1. Data Set and Evaluating Indicators

In this paper, we use the historical daily close prices from the SSE, as the close prices reflect all the trading activities of the day. Three companies listed on the SSE are selected: BOE, GREE, and ZTE for this study. The sample period is from 4 January 2007 to 30 September 2017.

Figures 2–4 display the time series of the closing prices for BOE, GREE, and ZTE, respectively. Each data set is partitioned into two parts: the data from 4 January 2007 to 31 August 2017, for training and the remaining samples (September 2017) for testing. The training data set is used to build the forecasting models, and the effectiveness of the forecasting models is assessed based on the test data set. Table 1 shows the detailed information of the training and testing data set.

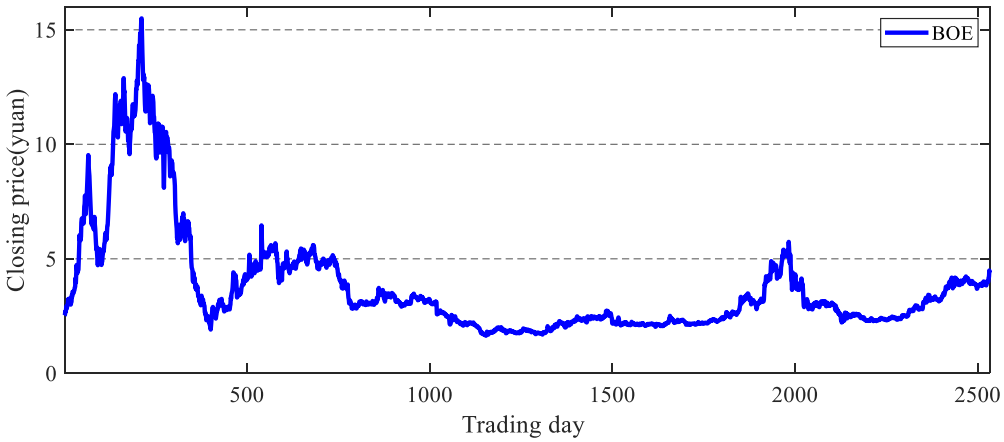


Figure 2. The close price series of BOE.

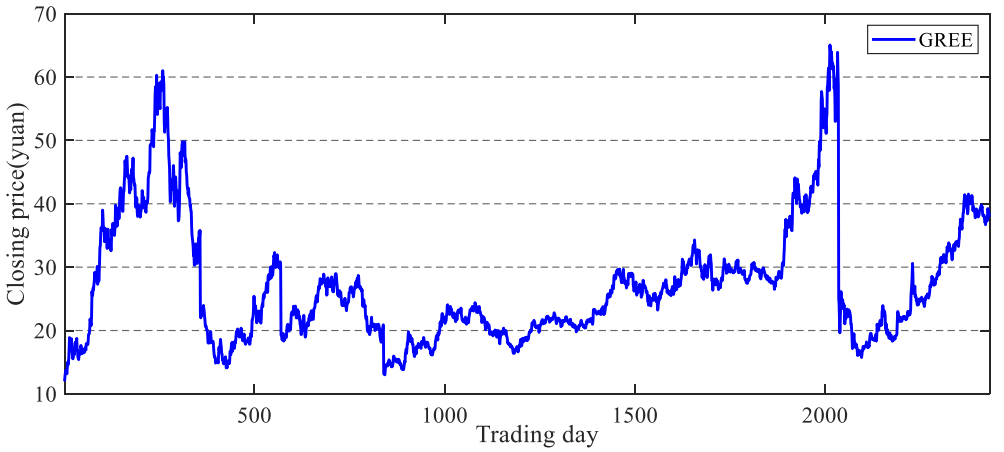


Figure 3. The close price series of GREE.

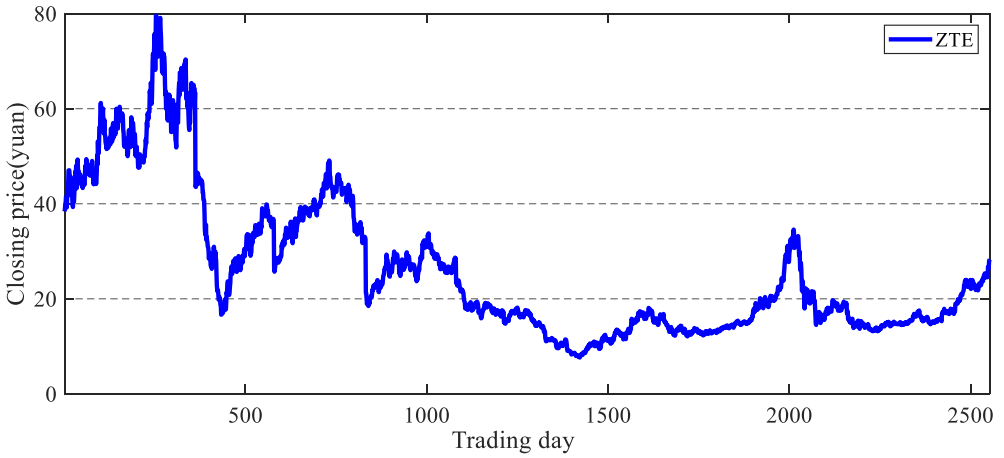


Figure 4. The close price series of ZTE.

Table 1. Descriptive statistics of three closing price data sets (CNY/day).

	Train Data Set			**Test Data Set**		
	BOE	**GREE**	**ZTE**	**BOE**	**GREE**	**ZTE**
Mean	3.7737	27.0970	26.2319	4.0220	37.8963	25.4884
Standard deviation	2.3705	9.6933	15.0992	0.1680	0.8964	1.3375
Length (days)	2513	2413	2533	20	19	20

To evaluate the performance of the hybrid model quantitatively, four performance indicators which measure the deviation between the predicted and the observed values are selected: the mean absolute error (MAE), mean absolute percentage error (MAPE), mean square error (MSE), and root-mean-square error (RMSE). Smaller values of these measures usually indicate better predictive performance. They are given as

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{14}$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left(|y_i - \hat{y}_i|/y_i * 100\%\right) \tag{15}$$

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|^2 \tag{16}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|^2} \tag{17}$$

where y_i and $\hat{y}_i$ are the observed and predicted stock price at time i, respectively; n denotes the number of predicted data.

4.2. Comparison of GPR Model with Single vs. Combined Covariance Function

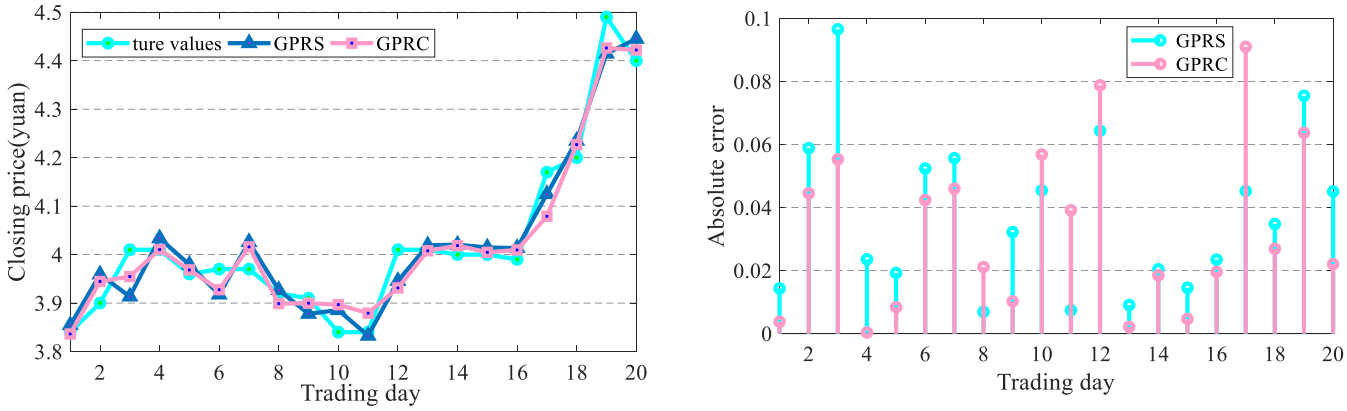
In order to select the more appropriate covariance function of the GPR model, we compare the forecasting performance of the GPR model with a single covariance function (GPRS) and that of the GPR model with a combined covariance function (GPRC). Figures 5–7 show the forecast results by a single and combined covariance function (GPRS and GPRC models) for the three selected stocks. From Figures 5a, 6a and 7a, it can be seen that both GPRC and GPRS can obtain curves with similar trends to the actual curve, but the predicted curve obtained by GPRC is closer to the actual curve than GPRS. In Figure 5b, the absolute errors obtained by GPRS and GPRC are both very small, with the maximum absolute error not exceeding 0.1. Among the 20 testing samples, 15 of the absolute error values obtained by GPRC are less than those of GPRS. Figure 6b shows the absolute error of GPRS and GPRC in obtaining the test set. Except for the error value greater than 1 in the 14th test sample, all others are less than 1. Although the error value of 11 samples obtained by GPRC is greater than that of GPRS, 8 of them are only slightly greater than that of GPRS. In Figure 7b, the absolute errors obtained by GPRS and GPRC are also very small, and the maximum absolute error does not exceed 0.1, and for the 20 samples tested, 14 of the absolute error values obtained by GPRC are less than GPRS. Therefore, it is clear from Figures 5–7 that the forecasting accuracy of GPRC is higher than that of GPRS. Tables 2–4 present the evaluation indicators of the forecasting performance. In Tables 2–4, for BOE and GREE, all indicators of GPRC are smaller than those for GPRS. For ZTE, the same holds for all indicators except MAE. Overall, the forecasting accuracy of the GPRC model is better than that of the GPRS model.

Table 2. Performance comparison between GPRS and GPRC on BOE.

Model	MAE	MAPE%	MSE	RMSE
GPRS	0.0372	0.9186	0.0020	0.0444
GPRC	0.0328	0.8095	0.0017	0.0417

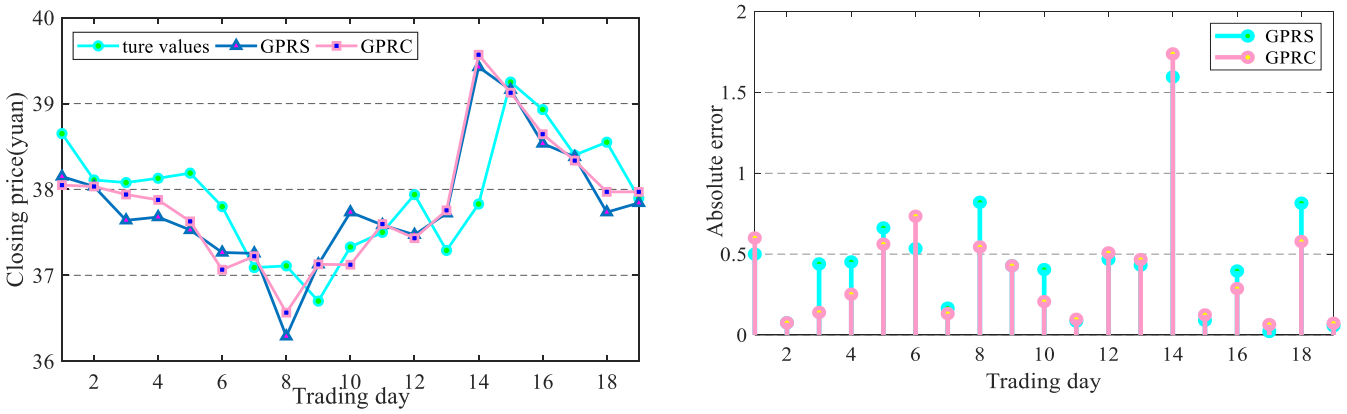
Table 3. Performance comparison between GPRS and GPRC on GREE.

Model	MAE	MAPE%	MSE	RMSE
GPRS	0.4443	1.1730	0.3270	0.5718
GPRC	0.4004	1.0570	0.3047	0.5520



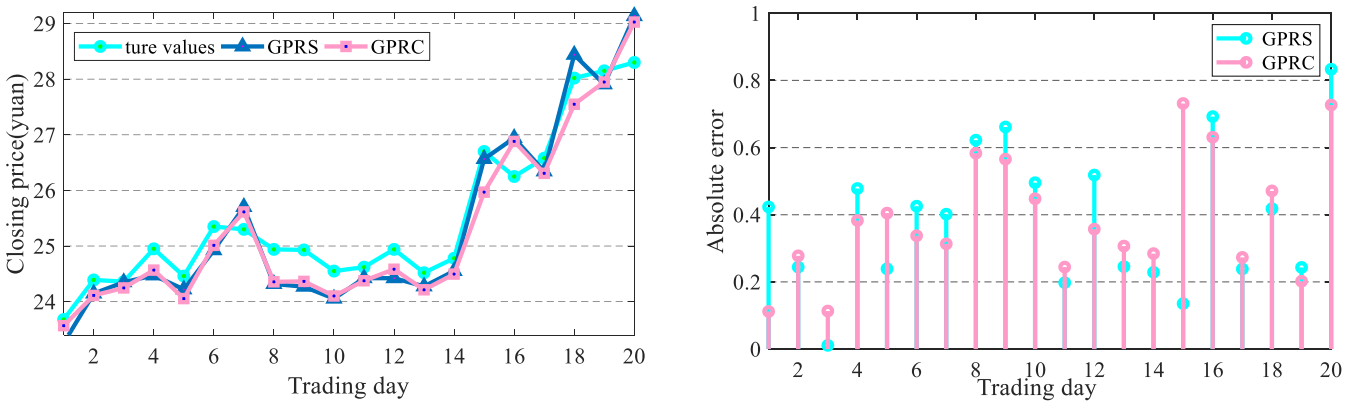
(**a**) (**b**)

Figure 5. The forecasting results of the close price series of BOE under GPRS and GPRC. Figure (**a**) shows the forecasting curves for BOE, and Figure (**b**) shows the forecasting absolute errors for BOE.



(**a**) (**b**)

Figure 6. The forecasting results of the close price series of GREE under GPRS and GPRC. Figure (**a**) shows the forecasting curves for GREE, and Figure (**b**) shows the forecasting absolute errors for GREE.



(**a**) (**b**)

Figure 7. The forecasting results of the close price series of ZTE under GPRS and GPRC. Figure (**a**) shows the forecasting curves for ZTE, and Figure (**b**) shows the forecasting absolute errors for ZTE.

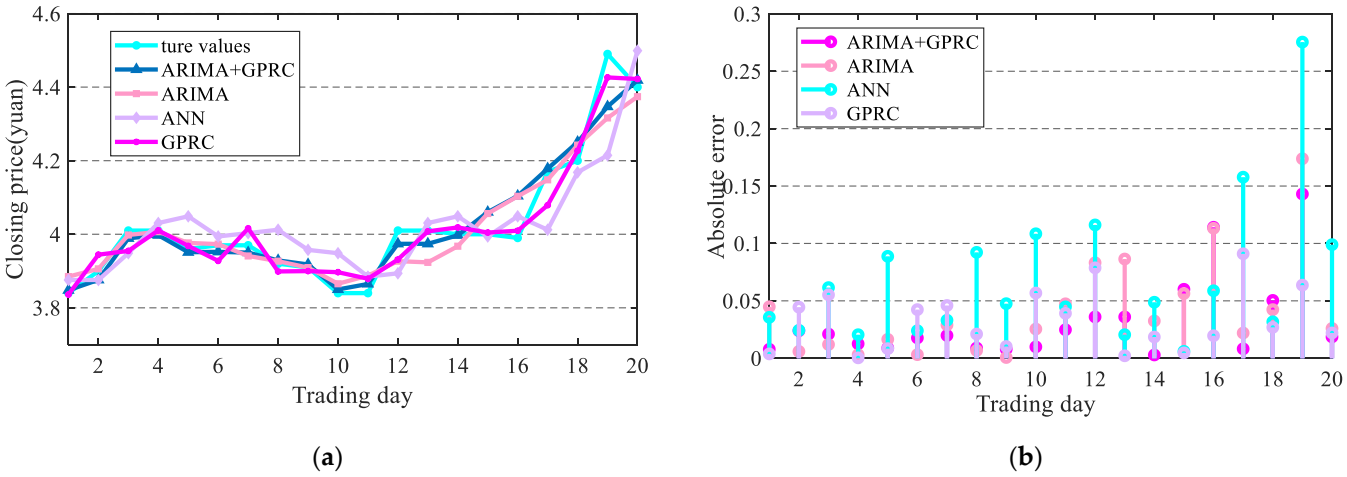
Table 4. Performance comparison between GPRS and GPRC on ZTE.

Model	MAE	MAPE%	MSE	RMSE
GPRS	0.3876	1.5152	0.1920	0.4382
GPRC	0.3883	1.5112	0.1823	0.4270

4.3. Forecasting Results of Various Models and Comparative Analysis

In order to validate the forecasting accuracy of the proposed model, we compare it with a few other models including ARIMA, GPRC, and an ANN. As shown above, the GPRC model outperforms GPRS; GPRS is not considered in this section.

For the ANN model, we employ a three-layer (one of which is a hidden layer) perceptron model, because it can approximate any continuous function in a reasonable way [26]. The number of hidden nodes is the integer number closest to log(n), where n is the number of training observations [26]. Five input variables are grouped into one vector as the input for day $t-1$. These variables are the daily high price (H_{t-1}), daily low price (L_{t-1}), the open price (O_{t-1}), daily close price (C_{t-1}), and trading volume (V_{t-1}). The output variable is the close price for day t. For the GPRC model, the input and output variables are the same as those for the ANN model. Figures 8–10 show the forecast values and absolute errors across different models (ARIMA, GPRC, ANN, ARIMA + GPRC) for the three selected stocks, respectively, and the evaluation indicator values of the four models are shown in Figures 11–13.



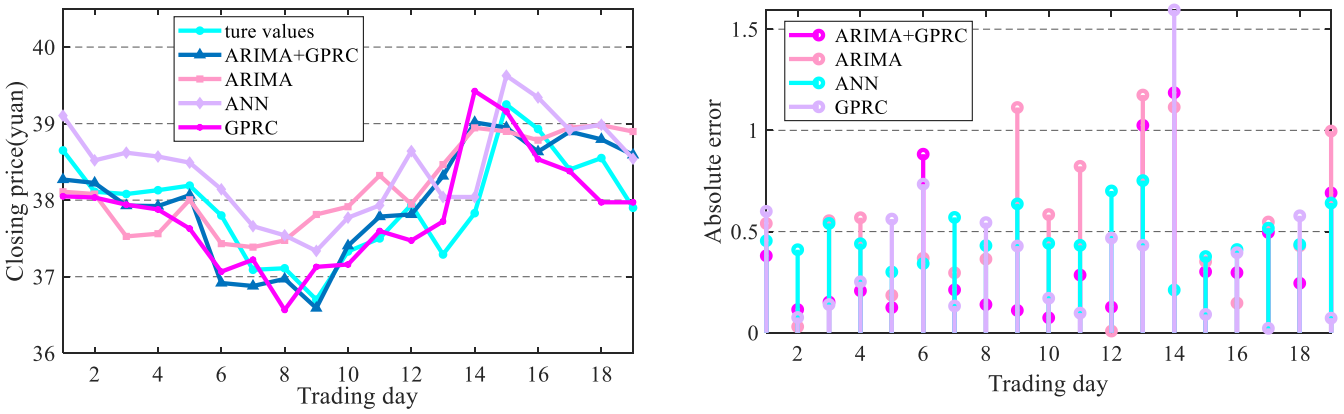
(**a**) (**b**)

Figure 8. The forecasting results of the close price series for BOE. Figure (**a**) shows the forecasting curves for BOE by four models, and Figure (**b**) shows the absolute errors for BOE by four models.

The forecasting curves and the absolute errors of four forecasting models on BOE are shown in Figure 8. In Figure 8a, all four models can obtain the same trend as the actual curve, and overall, the ARIMA + GPRC curve is closer to the actual curve. In Figure 8b, the absolute error obtained by the four models is very small, with a maximum error of no more than 0.3. Among the 20 test samples, ARIMA + GPRC has 7 samples with less error than the other three models, ARIMA has 4 samples with less error than the other three models, and GPRC has 8 samples with less error than the other three models. A further analysis of the evaluation indicators in Figure 11 shows that among the evaluation indicators obtained by ARIMA + GPRC except for REMS slightly greater than GPRC, all other indicators are smaller than models ARIMA, ANN, and GPRC. Therefore, for the closing price forecast of BOE, the ARIMA + GPRC forecasting accuracy constructed is better than the three models compared.

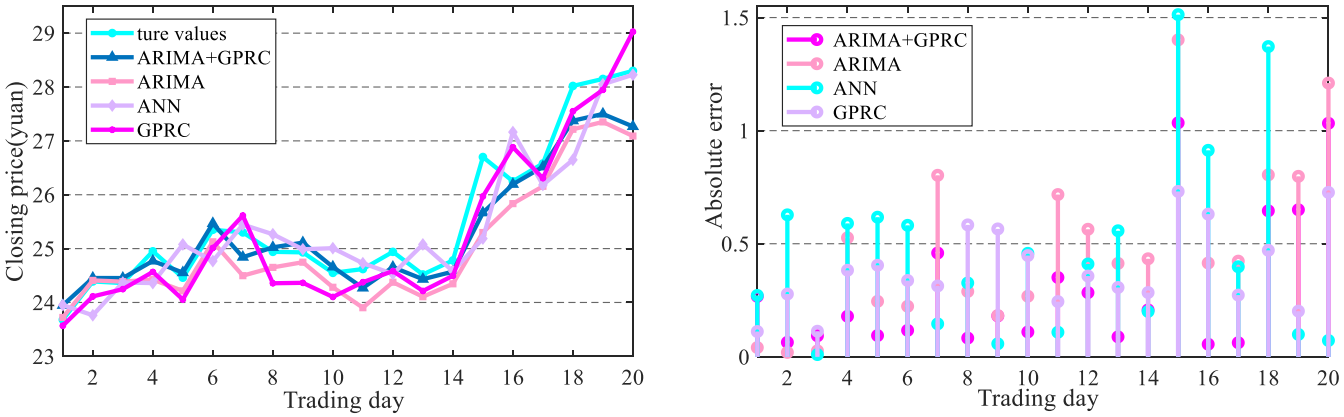
In Figure 9a, four models can obtain the same trend as the actual curve of the closing price forecast for GREE; the forecasting curves of the first 12 samples obtained by ARIMA + GPRC are the closest to the actual curves among the four models. In Figure 9b,

the absolute error obtained by the four models is small; except for samples 16, 19, and 20, the absolute errors of all samples are less than 1. Among the 19 test samples, ARIMA + GPRC has 7 samples with less error than the other three models, ARIMA has 3 samples with less error than the other three models, and GPRC has 7 samples with less error than the other three models. In Figure 11, all evaluation indicators obtained by ARIMA + GPRC are smaller than the other three models. Therefore, for the closing price forecast of GREE, the ARIMA + GPRC constructed is superior to the three models compared.



(**a**) (**b**)

Figure 9. The forecasting results of the close price series for GREE. Figure (**a**) shows the forecasting curves for GREE by four models, and Figure (**b**) shows the absolute errors for GREE by four models.



(**a**) (**b**)

Figure 10. The forecasting results of the close price series for ZTE. Figure (**a**) shows the forecasting curves for ZTE by four models, and Figure (**b**) shows the absolute errors for ZTE by four models.

From Figure 10a, it can be seen that all four models can obtain trends similar to the actual curve, except for samples 15, 19, and 20, ARIMA + GPRC can obtain the curve closest to the actual curve. In Figure 10b, the absolute error obtained by the four models is very small; except for samples 16, 18, and 20, the absolute errors of all samples are lower than 1. And among the test samples, ARIMA + GPRC, ARIMA, GPRC, and ANN have nine, two, two, and seven samples with less error than the other three models, respectively. And it is clear from Figure 13 that the ARIMA + GPRC model obtains the smallest value of all evaluation indicators among the four models. For the closing price of BOE, the MAPE obtained by ARIMA + GPRC increased by 24.1%, 54.7%, and 4.9%, respectively, by comparing to the ARIMA, ANN, and GPRC models. Calculating the forecast indicators for the GREE closing price, it was found that the MAPE of ARIMA + GPRC jumped by 31.2%,

22.2%, and 4.6%, respectively, by comparing to the models ARIMA, ANN, and GPRC. And analyzing the accuracy of the ZTE closing prices forecasted by four models, comparing to the ARIMA, ANN, and GPRC models, the MAPE for the ARIMA + GPRC model also increased by 38.7%, 36.5%, and 23.9%, respectively. Therefore, the ARIMA + GPRC model has best the forecasting accuracy among the four models.

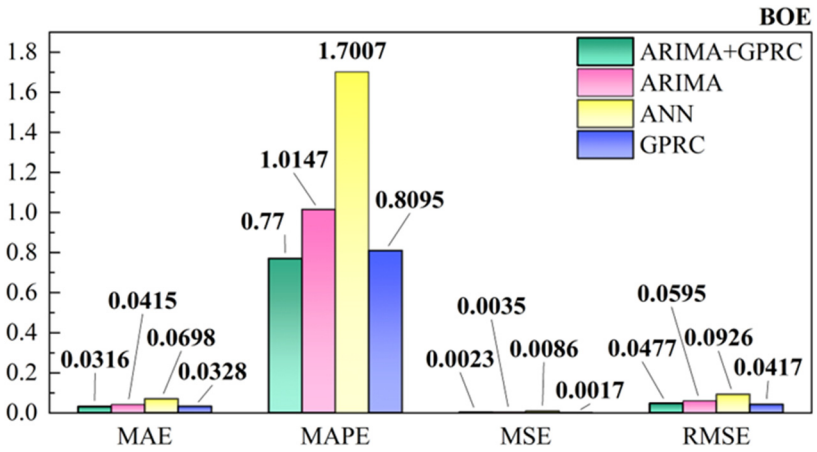


Figure 11. Performance comparison across different models on BOE.

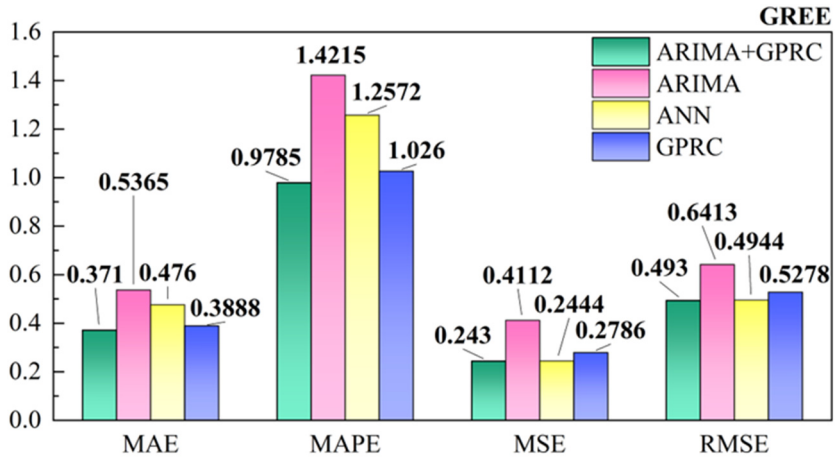


Figure 12. Performance comparison across different models on GREE.

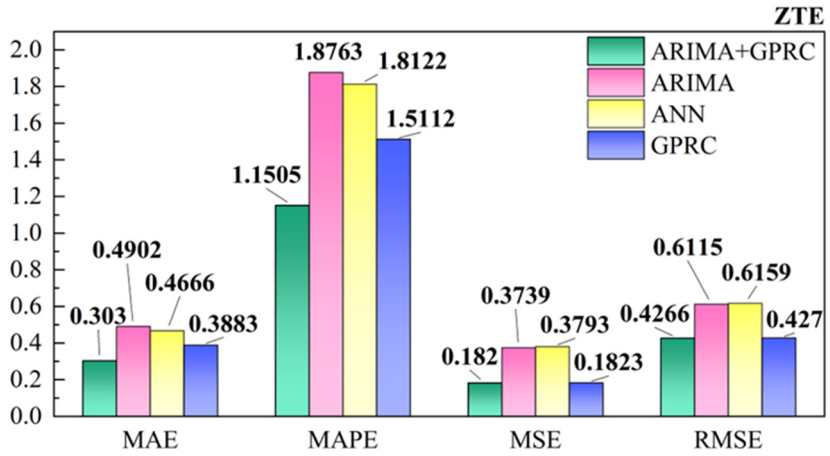


Figure 13. Performance comparison across different models on ZTE.

It can be seen from Figures 8–13 that the mixture model (ARIMA + GPRC) developed in this paper has excellent forecasting accuracy, which verifies the effectiveness and feasibility of the model built.

5. Conclusions

Over the years, a lot of research has been devoted to stock price forecasting which is critical for investment decisions. However, there are hardly any effective forecasting models. Thus, it is necessary to continue to study how to improve the effectiveness of forecasting models. In addition, it is usually difficult to predict stock prices accurately due to the complexity and volatility of the stock market.

This paper presents a new hybrid model which combines ARIMA with GPR, to improve the forecasting performance of the stock market in terms of statistical and financial terms. Meanwhile, in order to select more suitable covariance functions, the GPR model with different types of covariance functions is evaluated. It is found that GPR with a combined covariance function outperforms GPR with a single covariance function. Based on the proposed hybrid model, the ARIMA model captures the linear structure of stock prices series, and the nonlinear structure is modeled by GPRC. And using three actual data sets of the trading day price verified the validity of the ARIMA + GPRC model. The simulation results indicated that compared with ARIMA, ANN, and GPRC, in most cases, the proposed hybrid model gave the best forecasting performance in terms of MAE, MAPE, MSE, and RMSE. In summary, it can be concluded that the proposed method is an effective way to improve forecasting performance, which is beneficial to investors for investment decisions and risk management.

There are many factors that affect stock prices; in this study, only the influence of historical closing prices was considered; other influencing factors on the stock market will be taken into account in future research, which contributes to uncovering their functions and subsequent deep penetration into this area.

Author Contributions: Methodology, S.T., H.M. and J.L.; formal analysis, S.T. and Y.L.; investigation, J.H. and H.M.; writing—original draft, S.T. and Y.L.; supervision, Y.L.; funding acquisition, J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Natural Science Foundation of China (NSFC) (Grant No. 52365059), the Natural Science Foundation of Guangxi (Grant Nos. 2020JJD160004 and 2019JJB160048) and Beibu Gulf University Scientifc Research Initiation Project of Introducing High level Talents (Grant No.2022KYQD13).

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest: The authors declare that there are no conflicts of interest regarding the publication of this paper. All authors have no conflicts of interest. This article does not contain any studies with human participants performed by any of the authors.

References

1. Mohan, M.; Raja, P.K.; Velmurugan, P.; Kulothungan, A. Holt-winters algorithm to predict the stock value using recurrent neural network. *Methods* **2023**, *8*, 10. [CrossRef]
2. Buche, A.; Chandak, M.B. Stock market forecasting techniques: A survey. *J. Eng. Appl. Sci.* **2019**, *14*, 1649–1655. [CrossRef]
3. Hadavandi, E.; Shavandi, H.; Ghanbari, A. Integration of genetic fuzzy systems and artificial neural networks for stock price forecasting. *Knowl.-Based Syst.* **2010**, *23*, 800–808. [CrossRef]
4. Zhao, C.; Hu, P.; Liu, X.; Lan, X.; Zhang, H. Stock market analysis using time series relational models for stock price prediction. *Mathematics* **2023**, *11*, 1130. [CrossRef]
5. Box, G.E.; Jenkins, G.M.; Reinsel, G.C.; Ljung, G.M. *Time Series Analysis: Forecasting and Control*; John Wiley & Sons: Hoboken, NJ, USA, 2015.
6. Alshawarbeh, E.; Abdulrahman, A.T.; Hussam, E. Statistical Modeling of High Frequency Datasets Using the ARIMA-ANN Hybrid. *Mathematics* **2023**, *11*, 4594. [CrossRef]
7. Chen, Y.S.; Chou, C.L.; Lee, Y.J.; Chen, S.F.; Hsiao, W.J. Identifying Stock Prices Using an Advanced Hybrid ARIMA-Based Model: A Case of Games Catalogs. *Axioms* **2022**, *11*, 499. [CrossRef]

8. Meyler, A.; Kenny, G.; Quinn, T. *Forecasting Irish Inflation Using ARIMA Models*; MPRA: Munich, Germany, 1998.
9. Engle, R.F. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econom. J. Econom. Soc.* **1982**, *50*, 987–1007. [CrossRef]
10. Tong, H. *Threshold Models in Non-Linear Time Series Analysis*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2012.
11. De Gooijer, J.G.; Kumar, K. Some recent developments in non-linear time series modelling, testing, and forecasting. *Int. J. Forecast.* **1992**, *8*, 135–156. [CrossRef]
12. D'Ecclesia, R.L.; Clementi, D. Volatility in the stock market: ANN versus parametric models. *Ann. Oper. Res.* **2021**, *299*, 1101–1127. [CrossRef]
13. Guo, Y.; Han, S.; Shen, C.; Li, Y.; Yin, X.; Bai, Y. An adaptive SVR for high-frequency stock price forecasting. *IEEE Access* **2018**, *6*, 11397–11404. [CrossRef]
14. Li, M.; Zhu, Y.; Shen, Y.; Angelova, M. Clustering-enhanced stock price prediction using deep learning. *World Wide Web* **2022**, *26*, 207–232. [CrossRef] [PubMed]
15. Agrawal, M.; Shukla, P.K.; Nair, R.; Nayyar, A.; Masud, M. Stock prediction based on technical indicators using deep learning model. *Comput. Mater. Contin.* **2022**, *70*, 287–304. [CrossRef]
16. Hall, J.W. Adaptive Selection of US Stocks with Neural Nets. In *Trading on the Edge: Neural, Genetic, and Fuzzy Systems for Chaotic Financial Markets*; Wiley: New York, NY, USA, 1994; pp. 45–65.
17. Bahrammirzaee, A. A comparative survey of artificial intelligence applications in finance: Artificial neural networks, expert system and hybrid intelligent systems. *Neural Comput. Appl.* **2010**, *19*, 1165–1195. [CrossRef]
18. Pai, P.F.; Lin, C.S. A hybrid ARIMA and support vector machines model in stock price forecasting. *Omega* **2005**, *33*, 497–505. [CrossRef]
19. Hajirahimi, Z.; Khashei, M. A novel parallel hybrid model based on series hybrid models of ARIMA and ANN models. *Neural Process. Lett.* **2022**, *54*, 2319–2337. [CrossRef]
20. Li, Y.; Wu, C.; Liu, J.; Luo, P. A combination prediction model of stock composite index based on artificial intelligent methods and multi-agent simulation. *Int. J. Comput. Intell. Syst.* **2014**, *7*, 853–864. [CrossRef]
21. Li, X.; Zhang, Y.; Li, D.; Shum, P.P.; Huang, T. Nonlinear channel equalization using Gaussian Processes Regression in IMDD fiber link. *IEEE Photonics J.* **2022**, *14*, 1–6. [CrossRef]
22. Zheng, Y.; Zhang, W.; Liu, T.; Zhang, Y.F. Resonant responses and double-parameter multi-pulse chaotic vibrations of graphene platelets reinforced functionally graded rotating composite blade. *Chaos Solitons Fractals* **2022**, *156*, 28. [CrossRef]
23. Zheng, J.; Gong, Y.; Liu, W.; Zhou, L. Subspace Gaussian process regression model for ensemble nonlinear multivariate spectroscopic calibration. *Chemom. Intell. Lab. Syst.* **2022**, *230*, 10. [CrossRef]
24. Hong, X.; Huang, B.; Ding, Y.; Guo, F.; Chen, L.; Ren, L. Multivariate Gaussian process regression for nonlinear modelling with colored noise. *Trans. Inst. Meas. Control* **2019**, *41*, 2268–2279. [CrossRef]
25. Rasmussen, C.E.; Williams, C.K. *Gaussian Processes for Machine Learning*; MIT Press: Cambridge, MA, USA, 2006.
26. Barron, A.R. Approximation and estimation bounds for artificial neural networks. *Mach. Learn.* **1994**, *14*, 115–133. [CrossRef]