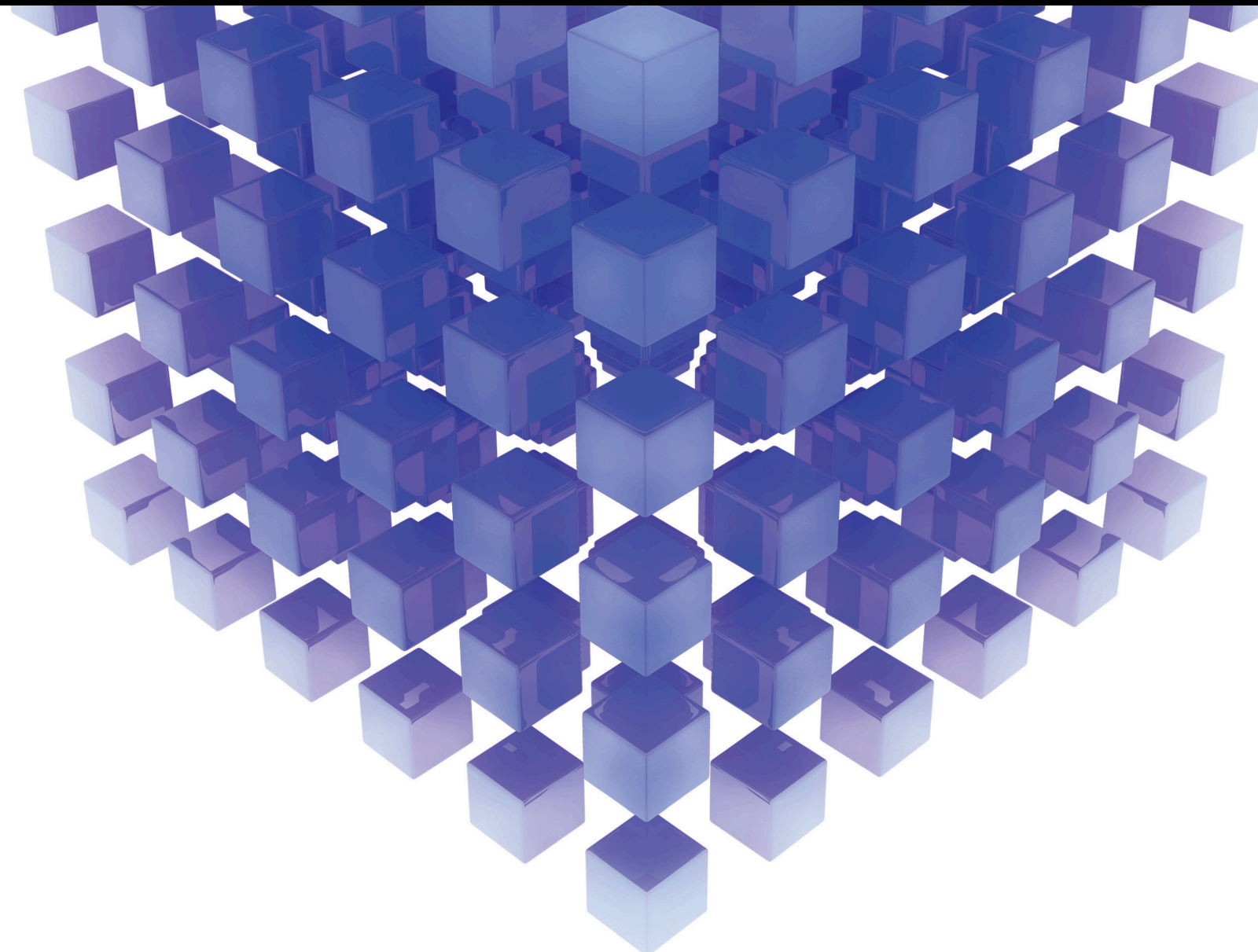# Advanced Intelligent Technologies in Energy Forecasting and Economical Applications

Lead Guest Editor: Wei-Chiang Hong
Guest Editors: Dongxiao Niu, Yinfeng Xu, Mengjie Zhang, and Pradeep Kumar Singh

Advanced Intelligent Technologies in Energy Forecasting and Economical Applications

Advanced Intelligent Technologies in Energy Forecasting and Economical Applications

Lead Guest Editor: Wei-Chiang Hong
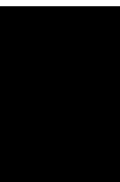Guest Editors: Dongxiao Niu, Yinfeng Xu, Mengjie Zhang, and Pradeep Kumar Singh

Rodolfo Bontempo, Italy
Alberto Borboni, Italy
Marco Bortolini, Italy
Paolo Boscariol, Italy
Daniela Boso, Italy
Guillermo Botella-Juan, Spain
Boulaïd Boulkroune, Belgium
Abdesselem Boulkroune, Algeria
Fabio Bovenga, Italy
Francesco Braghin, Italy
Ricardo Branco, Portugal
Maurizio Brocchini, Italy
Julien Bruchon, France
Matteo Bruggi, Italy
Michele Brun, Italy
Maria Elena Bruni, Italy
Vasilis Burganos, Greece
Maria Angela Butturi, Italy
Raquel Caballero-Águila, Spain
Guillermo Cabrera-Guerrero, Chile
Filippo Cacace, Italy
Pierfrancesco Cacciola, United Kingdom
Salvatore Caddemi, Italy
zuowei cai, China
Roberto Caldelli, Italy
Alberto Campagnolo, Italy
Eric Campos, Mexico
Salvatore Cannella, Italy
Francesco Cannizzaro, Italy
Maosen Cao, China
Javier Cara, Spain
Raffaele Carli, Italy
Ana Carpio, Spain
Rodrigo Carvajal, Chile
Caterina Casavola, Italy
Sara Casciati, Italy
Federica Caselli, Italy
Carmen Castillo, Spain
Inmaculada T. Castro, Spain
Miguel Castro, Portugal
Giuseppe Catalanotti, United Kingdom
Nicola Caterino, Italy
Alberto Cavallo, Italy
Gabriele Cazzulani, Italy
Luis Cea, Spain
Fatih Vehbi Celebi, Turkey
Song Cen, China

Miguel Cerrolaza, Venezuela
M. Chadli, France
Gregory Chagnon, France
Ludovic Chamoin, France
Xiaoheng Chang, China
Ching-Ter Chang, Taiwan
Kuei-Lun Chang, Taiwan
Qing Chang, USA
Dr. Prasenjit Chatterjee, India
Kacem Chehdi, France
Peter N. Cheimets, USA
Chih-Chiang Chen, Taiwan
Xiao Chen, China
He Chen, China
Zhiwen Chen, China
Chien-Ming Chen, China
Xinkai Chen, Japan
Shyi-Ming Chen, Taiwan
Kebing Chen, China
Xue-Bo Chen, China
Xizhong Chen, Ireland
Zeyang Cheng, China
Qiang Cheng, USA
Luca Chiapponi, Italy
Ryoichi Chiba, Japan
Francisco Chicano, Spain
Nicholas Chileshe, Australia
Tirivanhu Chinyoka, South Africa
Adrian Chmielewski, Poland
Seongim Choi, USA
Ioannis T. Christou, Greece
Hung-Yuan Chung, Taiwan
Simone Cinquemani, Italy
Roberto G. Citarella, Italy
Joaquim Ciurana, Spain
John D. Clayton, USA
Francesco Clementi, Italy
Piero Colajanni, Italy
Giuseppina Colicchio, Italy
Vassilios Constantoudis, Greece
Francesco Conte, Italy
Enrico Conte, Italy
Alessandro Contento, USA
Mario Cools, Belgium
Gino Cortellessa, Italy
Juan Carlos Cortés, Spain
Carlo Cosentino, Italy

Paolo Crippa, Italy
Erik Cuevas, Mexico
Guozeng Cui, China
Maria C. Cunha, Portugal
Mehmet Cunkas, Turkey
Peter Dabnichki, Australia
Luca D'Acierno, Italy
Zhifeng Dai, China
Weizhong Dai, USA
Pei Dai, China
Purushothaman Damodaran, USA
Bhabani S. Dandapat, India
Giuseppe D'Aniello, Italy
Sergey Dashkovskiy, Germany
Adiel T. de Almeida-Filho, Brazil
Fabio De Angelis, Italy
Samuele De Bartolo, Italy
Abílio De Jesus, Portugal
Pietro De Lellis, Italy
Alessandro De Luca, Italy
Stefano de Miranda, Italy
Filippo de Monte, Italy
José António Fonseca de Oliveira Correia,
Portugal
Jose Renato de Sousa, Brazil
Michael Defoort, France
Alessandro Della Corte, Italy
Laurent Dewasme, Belgium
Sanku Dey, India
Gianpaolo Di Bona, Italy
Angelo Di Egidio, Italy
Roberta Di Pace, Italy
Francesca Di Puccio, Italy
Ramón I. Diego, Spain
Yannis Dimakopoulos, Greece
Rossana Dimitri, Italy
Alexandre B. Dolgui, France
José M. Domínguez, Spain
Georgios Dounias, Greece
Bo Du, China
Z. Du, China
George S. Dulikravich, USA
Emil Dumic, Croatia
Bogdan Dumitrescu, Romania
Saeed Eftekhar Azam, USA
Antonio Elipe, Spain
Anders Eriksson, Sweden

R. Emre Erkmen, Canada
Francisco Periago Esparza, Spain
Gilberto Espinosa-Paredes, Mexico
Leandro F. F. Miguel, Brazil
Andrea L. Facci, Italy
Giovanni Falsone, Italy
Hua Fan, China
Nicholas Fantuzzi, Italy
Muhammad Shahid Farid, Pakistan
Hamed Faroqi, Iran
Mohammad Fattahi, Iran
Yann Favennec, France
Fiorenzo A. Fazzolari, United Kingdom
Giuseppe Fedele, Italy
Roberto Fedele, Italy
Zhongyang Fei, China
Mohammad Ferdows, Bangladesh
Arturo J. Fernández, Spain
Jesus M. Fernandez Oro, Spain
Massimiliano Ferraioli, Italy
Massimiliano Ferrara, Italy
Francesco Ferrise, Italy
Constantin Fetecau, Romania
Eric Feulvarch, France
Iztok Fister Jr., Slovenia
Thierry Floquet, France
Eric Florentin, France
Gerardo Flores, Mexico
Alessandro Formisano, Italy
FRANCESCO FOTI, Italy
Francesco Franco, Italy
Elisa Francomano, Italy
Juan Frausto-Solis, Mexico
Shujun Fu, China
Juan C. G. Prada, Spain
Matteo Gaeta, Italy
Mauro Gaggero, Italy
Zoran Gajic, USA
Jaime Gallardo-Alvarado, Mexico
Mosè Gallo, Italy
Akemi Gálvez, Spain
Rita Gamberini, Italy
Maria L. Gandarias, Spain
Xingbao Gao, China
Yan Gao, China
Hao Gao, Hong Kong
Shangce Gao, Japan

Zhong-Ke Gao, China
Zhiwei Gao, United Kingdom
Giovanni Garcea, Italy
José García, Chile
Luis Rodolfo Garcia Carrillo, USA
Jose M. Garcia-Aznar, Spain
Akhil Garg, China
Harish Garg, India
Alessandro Gasparetto, Italy
Gianluca Gatti, Italy
Oleg V. Gendelman, Israel
Stylianos Georgantzinos, Greece
Fotios Georgiades, India
Parviz Ghadimi, Iran
Georgios I. Giannopoulos, Greece
Agathoklis Giaralis, United Kingdom
Pablo Gil, Spain
Anna M. Gil-Lafuente, Spain
Ivan Giorgio, Italy
Gaetano Giunta, Luxembourg
Alessio Gizzi, Italy
Jefferson L.M.A. Gomes, United Kingdom
HECTOR GOMEZ, Chile
José Francisco Gómez Aguilar, Mexico
Emilio Gómez-Déniz, Spain
Antonio M. Gonçalves de Lima, Brazil
David González, Spain
Chris Goodrich, USA
Rama S. R. Gorla, USA
Veena Goswami, India
Xunjie Gou, Spain
Jakub Grabski, Poland
Antoine Grall, France
George A. Gravvanis, Greece
Fabrizio Greco, Italy
David Greiner, Spain
Jason Gu, Canada
Federico Guarracino, Italy
Michele Guida, Italy
Muhammet Gul, Turkey
Hu Guo, China
Jian-Ping Guo, China
Zhaoxia Guo, China
Dong-Sheng Guo, China
Quang Phuc Ha, Australia
Li Haitao, China
Petr Hájek, Czech Republic

Muhammad Hamid, United Kingdom
Shigeyuki Hamori, Japan
Xingsi Han, China
Zhen-Lai Han, China
Weimin Han, USA
Renke Han, United Kingdom
Thomas Hanne, Switzerland
Xinan Hao, China
Mohammad A. Hariri-Ardebili, USA
Khalid Hattaf, Morocco
Xiao-Qiao He, China
Yu-Ling He, China
Defeng He, China
Fu-Qiang He, China
salim HEDDAM, Algeria
Ramdane Hedjar, Saudi Arabia
Jude Hemanth, India
Reza Hemmati, Iran
Nicolae Herisanu, Romania
Alfredo G. Hernández-Diaz, Spain
M.I. Herreros, Spain
Eckhard Hitzer, Japan
Paul Honeine, France
Jaromir Horacek, Czech Republic
S. Hassan Hosseinnia, The Netherlands
Yingkun Hou, China
Xiaorong Hou, China
Lei Hou, China
Yunfeng Hu, China
Gordon Huang, Canada
Can Huang, China
Sajid Hussain, Canada
Asier Ibeas, Spain
Wubshet Ibrahim, Ethiopia
Orest V. Iftime, The Netherlands
Przemyslaw Ignaciuk, Poland
Muhammad Imran, Pakistan
Giacomo Innocenti, Italy
Emilio Insfran Pelozo, Spain
Alessio Ishizaka, France
Nazrul Islam, USA
Benoit Iung, France
Benjamin Ivorra, Spain
Breno Jacob, Brazil
Tushar Jain, India
Amin Jajarmi, Iran
Payman Jalali, Finland

Lingfan Zhang, China
Yifan Zhao, United Kingdom
Yongmin Zhong, Australia
Zebo Zhou, China
Debao Zhou, USA
Jian G. Zhou, United Kingdom
Zhe Zhou, China
Quanxin Zhu, China
Wu-Le Zhu, China
Gaetano Zizzo, Italy
Zhixiang Zou, China

Contents

Time Series Prediction of Electricity Demand Using Adaptive Neuro-Fuzzy Inference Systems
Amevi Acakpovi ⓘ, Alfred Tettey Ternor, Nana Yaw Asabere, Patrick Adjei, and Abdul-Shakud Iddrisu

Editorial

Advanced Intelligent Technologies in Energy Forecasting and Economical Applications

Wei-Chiang Hong ⓘ,[1] **Dongxiao Niu,**[2] **Yinfeng Xu,**[3] **Mengjie Zhang,**[4] **and Pradeep Kumar Singh**[5]

[1]*Asia Eastern University of Science and Technology, New Taipei City, Taiwan*
[2]*North China Electric Power University, Beijing, China*
[3]*Donghua University, Shanghai, China*
[4]*Victoria University of Wellington, Wellington, New Zealand*
[5]*KIET Group of Institutions, Delhi-NCR, Ghaziabad, U.P., India*

Correspondence should be addressed to Wei-Chiang Hong; samuelsonhong@gmail.com

Accurate energy forecasting is essential to achieve greater efficiency and reliability in power system operation and security, energy pricing problems, and scheduling and planning of energy supply systems, among other areas. Particularly, in the Big Data era, forecasting models are always based on a complex function combination, and energy data are always complicated; examples include seasonality, cyclicity, fluctuation, and dynamic nonlinearity. During the past few decades, many energy forecasting models have been proposed, including traditional statistical models and artificial intelligent models. However, most of these models often possess theoretical drawbacks which limit their forecasting performance. These forecasting models have resulted in an over-reliance on the use of informal judgments and higher expenses when lacking the ability to determine data characteristics and patterns.

Recently, due to the development of advanced intelligent computing technologies, many novel technologies hybridised or combined with the energy forecasting and economic planning models mentioned previously have received much attention. The hybridization of optimization methods and superior evolutionary algorithms can provide important improvements via well parameter determinations in the optimization process, which is of great assistance to actions taken by energy decision-makers. It is important to explore the development of the modeling methodology by applying these advanced intelligent technologies.

The aim of this special issue is to collate original research articles with a focus on the applications of advanced intelligent technologies in economical modeling and energy forecasting. After a rigorous round of double-blind peer review process, finally 10 papers are accepted for publication in this special issue, which all display a broad range of cutting edge topics in the advanced intelligent technologies. The editors of this special issue believe that the advanced intelligent technologies and computation techniques will play more important role in energy forecasting accuracy improvements to overcome some critical shortcomings of a single model or directly improve the shortcoming by theoretical innovative arrangements.

The first paper is "Time Series Prediction of Electricity Demand Using Adaptive Neuro-Fuzzy Inference Systems" by Acakpovi et al.; they concerned with the reliable prediction of electricity demands by using the adaptive neuro-fuzzy inference system (ANFIS). Based on the experimental analysis and the comparison results, the performance of the avalanche photodiode is around 11% better than the pin diode.

The second paper is "A Comparison of Hour-Ahead Solar Irradiance Forecasting Models Based on LSTM

Network" by Huang et al.; they investigated the Model I-A and Model II-B based on traditional long short-term memory (LSTM) and the effects of different parameters. With real data over 5 years, the experimental results demonstrate that Model II-BD shows the best performance because it considers the weather information of the next moment, the root mean square error (RMSE) is 62.1618 W/m^2, the normalized root mean square error (nRMSE) is 32.2702%, and the forecast skill (FS) is 0.4477, which is 19.19% more accurate than the backpropagation neural network (BPNN) in terms of RMSE.

The third paper is "Optimal Control and Simulation for Enterprise Financial Risk in Industry Environment" by Liang et al.; they developed a technique for enterprise financial risk optimal control with exponential decay rate and simulation in industry environment. The factors of industry environment risks to enterprise financial activities are considered; seven kinds of industry environment risks influencing enterprise financial activities are chosen as state variables. Numerical simulation results illustrate the effectiveness of the proposed technique.

The fourth paper is "Optimizing the Procurement of IaaS Reservation Contracts via Workload Predicting and Integer Programming" by Zhu et al.; they investigated the problem of how to minimize IaaS rental cost associated with hosting web applications, while meeting the demand in the future business cycle, by using integer liner program model and a long short-term memory (LSTM). The experimental prediction results show the LSTM-based algorithm gains an advantage over several popular models, such as the Holter–Winters, the seasonal autoregressive integrated moving average (SARIMA), and the support vector regression (SVR).

The fifth paper is "CO_2 Emissions, Energy Consumption, and Economic Growth Nexus: Evidence from 30 Provinces in China" by Zou and Zhang; they established a spatial Durbin model including economic growth, energy consumption equation, and CO_2 emissions and studied the dynamic relationship and spatial spillover among economic growth, energy consumption, and CO_2 emissions effects. The results show that the economic growth can significantly improve carbon dioxide emissions, and China's economic growth level has become a positive driving force for carbon dioxide emissions. However, economic growth will not be significantly affected by the reduction of carbon dioxide emissions, and energy consumption and carbon emissions are interrelated, which has a negative spatial spillover effect on the carbon dioxide emissions of the surrounding provinces and cities.

The sixth paper is "The Improved Least Square Support Vector Machine Based on Wolf Pack Algorithm and Data Inconsistency Rate for Cost Prediction of Substation Projects" by Wang et al.; they proposed a forecasting model based on the improved least squares support vector machine (ILSSVM) optimized by wolf pack algorithm (WPA) to improve the accuracy and stability of the cost forecasting of substation projects. They selected 88 substation projects in different regions from 2015 to 2017 to conduct the training tests to verify the validity of the model. The results indicate

that the new hybrid WPA-DIR-ILSSVM model presents better accuracy, robustness, and generality in cost forecasting of substation projects.

The seventh paper is "A Novel Hybrid Approach Based on BAT Algorithm with Artificial Neural Network to Forecast Iran's Oil Consumption" by Bahmani et al.; they developed a function of population, GDP, import, and export by applying a hybrid bat algorithm (BAT) and artificial neural network (ANN) to forecast oil consumption in Iran. The result of the model shows that the findings are in close agreement with the observed data, and the performance of the method was excellent. Authors demonstrate that its efficiency could be a helpful and reliable tool for monitoring oil consumption.

The eighth paper is "Nonlinear Effect Analysis of Electricity Price on Household Electricity Consumption" by Zhang and Wen; they investigated the scenario effect of per capita income and regional differences in urbanization development on the relationship between electricity sales price and urban household electricity consumption. They discussed the effect of electricity sales price on urban household electricity consumption from the perspective of regional difference in income and urbanization, which provides the decision-making basis and empirical support for developing regional electricity price policy and household energy consumption policy.

The ninth paper is "The Impact and Stability Analysis of Commercial Banks' Risk Preference on SMEs' Credit Financing Based on DSGE Model" by Gao et al.; they analyzed the impact of financial intermediary departments' risk preference on corporate finance. Under the revised DSGE framework, they discussed the impact and stability analysis of commercial banks' risk preferences on SMEs' financing. The results showed that positive interest rate shocks inhibit commercial banks' credit to SMEs, and with the increasing weight of commercial banks' risk preference for default rate, the trend of credit repression will be intensified.

The tenth paper is "Evolutionary Framework with Bidirectional Long Short-Term Memory Network for Stock Price Prediction" by Zheng et al.; they proposed a novel bidirectional long short-term memory network (BiLSTM) framework called evolutionary BiLSTM (EBiLSTM) for the prediction of stock price. Experiments on several stock market indexes demonstrate that EBiLSTM can achieve better prediction performance than others without the evolutionary operator.

Conflicts of Interest

The editors declare that they have no conflicts of interest regarding the publication of this special issue.

Wei-Chiang Hong
Dongxiao Niu
Yinfeng Xu
Mengjie Zhang
Pradeep Kumar Singh

Research Article

Evolutionary Framework with Bidirectional Long Short-Term Memory Network for Stock Price Prediction

Hongying Zheng,[1] Hongyu Wang,[1,2] and Jianyong Chen [iD] [2]

[1]*Sino-German Robotics School, Shenzhen Institute of Information Technology, Shenzhen 518172, China*
[2]*Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen University, Shenzhen 518060, China*

Correspondence should be addressed to Jianyong Chen; jychen@szu.edu.cn

As an important part of the social economy, stock market plays an important role in economic development, and accurate prediction of stock price is important as it can lower the risk of investment decision-making. However, the task of predicting future stock price is very difficult. This difficulty arises from stocks with nonstationary behavior and without any explicit form. In this paper, we propose a novel bidirectional Long Short-Term Memory Network (BiLSTM) framework called evolutionary BiLSTM (EBiLSTM) for the prediction of stock price. In the framework, three independent BiLSTMs correspond to different objective functions and act as mutation individuals, then their respective losses for evolution are calculated, and finally, the optimal objective function is identified by the minimum of loss. Since BiLSTM is effective in the prediction of time series and the evolutionary framework can get an optimal solution for multiple objectives, their combination well adapts to the nonstationary behavior of stock prices. Experiments on several stock market indexes demonstrate that EBiLSTM can achieve better prediction performance than others without the evolutionary operator.

1. Introduction

It is essential for investors to forecast the future price of a stock because the risk of decision-making can be mitigated by appropriately determining its future movement. The topic has attracted many researchers from various academic fields. However, it is a challenging task to predict accurately. In the early stage, most of them used moving average [1], linear regression [2], hidden Markov model (HMM) [3], autoregressive integrated moving average (ARIMA) [4], and prophet model [5] to predict stock prices and their trend.

Currently, neural networks based on deep learning are dominant so far in time series prediction as surveyed, especially Long Short-Term Memory (LSTM) [6]. Recurrent Neural Network (RNN) is also widely used to predict stock prices [7], which applies a decision-making method based on an estimate of the zero-crossing rate to enhance the ability of prediction. Relative insensitivity to gap length is an advantage of LSTM over CNN, RNN, HMM, and other

learning methods in numerous applications [8–10]. In some of the initial researches, scholars used only raw financial data for price prediction, which utilized LSTM to predict high and low prices of soybean futures using the data set from the Dalian Commodity Exchange [11]. Later some researchers found that preprocessing of the original data can improve the accuracy of the prediction [12], which proposed a movement trend-based data prediction method to preprocess the trend indicator.

Deep neural networks (DNN) have also been widely applied in stock price prediction to identify trends and patterns. Go and Hong [13] firstly trained DNN by the data of financial time series and then tested and confirmed the predictability of their model. Fluctuation of the stock price was predicted by DNN with 715 novel input features [14]. The performance of their model was also compared with the other models with simple price-based input features. To predict the stock market behavior, the performance of DNN was examined in which high-frequency intraday stock

returns were considered as the input [15]. In the model, the predictability of principal component analysis (PCA), autoencoder, and restricted Boltzmann machine (RBM) was analyzed. The results showed that DNN has good predictability with the information received from the residuals of the autoregressive mode. Moreover, it has been found that financial news may be one of the key factors to produce fluctuations in stock prices. Several sentiment analysis studies have tried to point out the relationship between reaction from investors and news [16], which utilized a novel two-stream gated recurrent unit network to predict the directions of stock prices by using Stock2Vec.

However, due to the complexity of stock data, it is difficult to obtain satisfactory accuracy by only using simple preprocessing or a single LSTM model. By both the complex preprocessing and a hybrid model, the prediction accuracy can be significantly improved [16, 17]. In [17], a large-scale deep learning model was proposed to predict price movements from data of Limit Order Book (LOB) [18]. The architecture utilized CNN to learn the spatial feature of the LOBs and LSTM to remember longer time dependencies series. Framework with multiple LSTMs was also studied extensively to improve the performance of prediction, in which different LSTM can capture different features of data [19, 20].

On the other hand, evolutionary algorithms are inspired by biological evolution mechanisms and simulating evolutionary processes such as reproduction, mutation, genetic recombination, and natural selection, for evolutionary calculation of candidate solutions to optimization problems. Since the evolutionary algorithm is a highly robust and widely applicable global optimization algorithm [21], many scholars have begun to use it to optimize various complex models.

The combination of evolutionary algorithms and neural networks can further improve network performance. In recent years, there have been many achievements in practical applications in multiple fields [22–29]. To minimize human participation in designing deep learning algorithms and automatically discover such configurations, there have been some attempts to optimize deep learning hyperparameters through an evolutionary search [22, 23]. For network optimization, Generative Adversarial Network (GAN) [24] can generate attacked images by one-pixel adversarial perturbation based on differential evolution (DE), that is, blackbox attack, which only required a little adversarial information and can fool many types of networks due to the inherent features of DE. Moreover, GAN can also make the generated image more stable through an evolutionary algorithm [25], in which the adversarial game was composed of a population of generators and acted as the discriminator, thereby improving the generative performance. In reinforcement learning, the network topology can also be optimized by combining differential evolution and metaheuristic algorithms [26]. In [27], the transfer learning was used as agents and embedded in the neural network to determine which parts of the network can be reused for a new task. During learning, a tournament selection genetic algorithm was used to select pathways through the neural

networks. LSTM combined with evolutionary algorithms has been reported in the prediction of time series [28, 29]. In [28], the gradient descent method in LSTM was combined with the particle swarm optimization (PSO) algorithm to update the weights of network. In [29], an evolutionary attention-based LSTM was proposed for multivariate time series prediction, which can refrain from being trapped into partial optimization like traditional gradient-based methods.

In this paper, EBiLSTM is proposed for stock price prediction, which takes the BiLSTM training procedure as an evolutionary problem. Specifically, the training process of each BiLSTM has its adaptive loss functions and is independent. A population of BiLSTMs evolves corresponding to the training process of multi-BiLSTMs. The BiLSTM is trained for predicting the stock price of the next day during each training (or evolutionary) iteration.

In summary, contributions in this paper are listed as follows:

(i) A framework named EBiLSTM is proposed which integrates BiLSTM and evolutionary algorithm to effectively predict stock price. As far as we know, it is the first report on the approach.

(ii) An evolution strategy is proposed which uses multiobjective functions (square loss, abs loss, and Huber loss) to optimize BiLSTM.

(iii) Performances are evaluated with several stock market indexes and the results demonstrate that the proposed EBiLSTM can get more accuracy of prediction than others.

The rest of this paper is organized as follows: in Section 2, EBiLSTM together with its training process is proposed. Section 3 provides the experimental validation of the method. Finally, the conclusion is presented in Section 4.

2. Method

2.1. Framework. In contrast to conventional BiLSTM which utilizes a single BiLSTM to train the stock data of the real world, an evolutionary algorithm is used that evolves a population of BiLSTM (s), that is, {BiLSTM}, in the training process. In this population, each individual stands for a possible solution in the weights space of the BiLSTM. During the evolutionary process, mutation operations (different objective functions are chosen dynamically) are used to generate different offspring individuals (the weights of different BiLSTM). As shown in Figure 1, there are three substages in each step of evolution:

(i) Mutation: Given an individual $BiLSTM_{\theta}$ in the population, the variation operators are used to produce its offspring $\{BiLSTM_{\theta 1}, BiLSTM_{\theta 2}, \ldots\}$. Each individual creates some copies and mutations are used to modify each of them. The modified copies are taken as children.

(ii) Evaluation: BiLSTM training process is taken as fitness function $F(\cdot)$ which is used to evaluate the
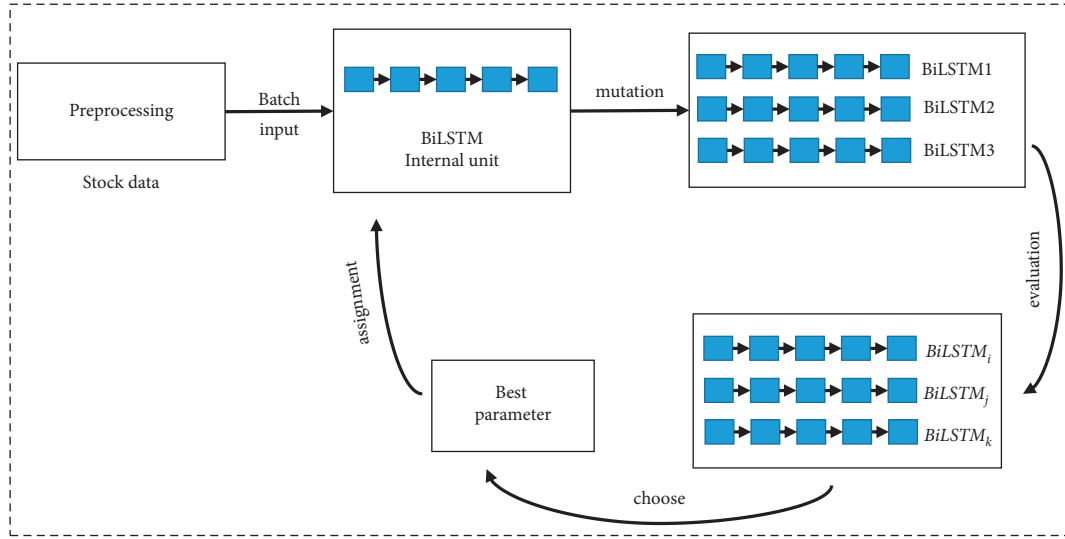
FIGURE 1: Framework of EBiLSTM.

performance of each child. It can be represented by fitness score.

(iii) Selection: According to the fitness score, all children are sorted from high to low. Some of the lowest ones, that is, the worst ones, are deleted. The remaining children are kept and act as parents to be further evolved at the next iteration.

After each evolutionary circle, the BiLSTM is updated to predict the price of the next day; that is,

$$L_{\text{BiLSTM}} = F_{\text{BiLSTM}} \,(\text{square, abs or huber}). \tag{1}$$

Here, we take a simple example to show the process of evolution. Data of stock A is split into various batches to train BiLSTM. The first batch, that is, data_1, is input to BiLSTM. According to the difference between the outputs and the corresponding labels, various losses, such as square, abs, and Huber, are calculated. Among them, the least one is selected and the weights of BiLSTM are updated. Then, the second batch, that is, data_2, is input to BiLSTM. The training process continues until the loss becomes small enough. Thus, the adaptive losses from the evolution of the BiLSTM population can produce optimal solutions.

2.2. Mutation. Sexual reproduction with different mutations is employed to produce the next BiLSTM's individuals (i.e., children). Specifically, these mutation operations are taken as different training objectives, which try to narrow the distance between the predicted value and the real stock price. In this section, the mutation is presented in detail. To analyze the properties of these mutations, we assume that the optimal BiLSTM model can be got from each evolutionary circle.

(1) Abs mutation (L1 loss): The abs mutation represents the abs objective function in the original BiLSTM:

$$M_{\text{abs}} = |y - f(x)|. \tag{2}$$

The abs aims to minimize the absolute value between the prediction value and the real close price (label). If there are outliers in the training set which may corrupt the training process, it is necessary to use L1 loss. However, L1 loss has a serious problem. Its gradient is kept the same throughout. When L1 loss is small, the gradient will be large which may impede learning.

(2) Square mutation (L2 loss): The square mutation represents the square objective function in the original BiLSTM:

$$M_{\text{square}} = (y - f(x))^2. \tag{3}$$

Gradients of the square mutation can be used for BiLSTM training. When L2 loss approaches zero, it means that the prediction accuracy of BiLSTM is very high (i.e., $L_{\text{BiLSTM}} \longrightarrow 0$). While when L2 loss is close to infinity, it means that the training of BiLSTM is not effective. Because L2 loss is square of the error $(y - f(x) = e)$, the error (e) increases a lot when $e > 1$. Once there is an outlier in our data, e may be high and e^2 may be $\gg |e|$. Due to the outliers, the weights of a model will be affected more seriously by L2 loss than by L1 loss.

(3) Huber mutation: the Huber mutation represents the Huber objective function in the original BiLSTM:

$$M_{\text{huber}} = \begin{cases} \dfrac{1}{2}[y - f(x)]^2 & |y - f(x)| \le \delta, \\\\ \delta|y - f(x)| - \dfrac{1}{2\delta^2} & |y - f(x)| > \delta. \end{cases} \tag{4}$$

2.3. Evaluation. For evolutionary algorithm, fitness function (i.e., evaluation operation) is used to measure the quality of an individual. In this paper, the loss function is taken as the

fitness function which focuses on minimal loss. As shown in equation (1), the smaller the loss value is, the better the fitness is.

Note that BiLSTM is constantly updated to get the optimal solution in the training process. If a BiLSTM has a relatively high fitness value, its prediction result can get better performance.

2.4. Selection.

In an evolutionary algorithm, the selection is the counterpart of the mutation operators. In the proposed EBiLSTM, a simple yet useful survivor selection strategy is used to determine the next generation based on the fitness score of existing individuals.

Particularly, the BiLSTMs (i.e., population) are optimized in a dynamic procedure. Thus, the fitness function is changeable and we can evaluate the fitness score of each BiLSTM from the corresponding training process in the same evolutionary generation. It indicates that we cannot compare the fitness scores evaluated in different generations with each other. Moreover, because the mutation operators of the proposed EBiLSTM actually represent selection from different BiLSTM training objectives, the desired offspring represents the effective training strategies. Taking into account both the fitness function and the mutation operators, the selection mechanism of EBiLSTM is taken as the comma selection, i.e., (μ, λ)-selection. Specifically, after the current offspring population $\{BiLSTM_i\}_{i=1}^{\lambda}$ is sorted according to their fitness scores F_i, the μ – best individuals are selected as population of the next generation.

2.5. Data Preprocessing.

Stock data have variables with multiple dimensions, such as opening price, closing price, the highest price, the lowest price, and trading rate. Among these prices, the closing price is always the most concern by investors. Therefore, we use the closing price as the input variable. Our experiments show that the results are basically the same when other prices are used as input variables.

As shown in Figure 2, the overall data is divided into a training set and a test set, respectively. A rolling window is used to segment data. We use the way of $N+1$ (i.e., the closing price of the previous N days is used to predict the closing price of the next day) to train EBiLSTM continuously. After the train finishes, the data of the last N days in the training set are used as input data to predict the first day of the test set. Then the rolling window of input data is shifted forward one day and the last day of the input data becomes vacant. In this case, the first day of the test data is used as the last day of the input data to predict the second day of the test data and so forth. Once the last day of the test set is predicted, the test process is over.

To reduce the noise of stock data and benefit the detection of stock price pattern, it is necessary to smooth and normalize the stock price data since every stock may have its specific domain and scale. Data normalization is defined as adjusting values measured on different scales to a uniform scale [30], while data smoothing is defined as transforming stock prices into variations of daily change. The data smoothing is shown in the following equation:
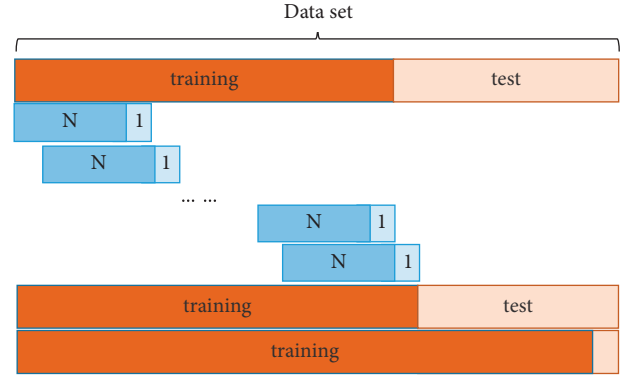
FIGURE 2: Schematic diagram of the rolling window.

$$x_{(s,t)} = \begin{cases} 0, & t = 1, \\ \dfrac{x_t - x_{t-1}}{x_{t-1}}, & t > 1, \end{cases} \tag{5}$$

where $x_{(s,t)}$ denotes the result of data smoothing at the tth day. Here, when $t = 1$, we set $x_{(s,t)} = 0$.

The learning of BiLSTM is in fact to get stock patterns which can be magnified by "min–max" normalization of the data set. The "min–max" normalization method is shown as follows:

$$x_{(n,t)} = \dfrac{x_{(s,t)} - x_{(s,\min)}}{x_{(s,\max)} - x_{(s,\min)}}, \tag{6}$$

where $x_{(n,t)}$ denotes the data after normalization, $x_{(s,t)}$ is original data, $x_{(s,\min)}$ is the minimum among the data set, and $x_{(s,\max)}$ is the maximum.

Accordingly, denormalization and desmoothness are required at the end of the prediction process to get the original price, which are given by

$$\widehat{x}_{(s,t)} = \widehat{x}_{(n,t)}\left[x_{(s,\max)} - x_{(s,\min)}\right] + x_{(s,\min)}, \tag{7}$$

$$\widehat{x}_t = \widehat{x}_{(s,t)}x_{t-1} + x_{t-1}, \tag{8}$$

where $\widehat{x}_{(n,t)}$ denotes the predicted data, $\widehat{x}_{(s,t)}$ denotes the predicted data after denormalization, and $\widehat{x}_t$ denotes the predicted data after both denormalization and desmoothness.

2.6. BiLSTM.

As a variant of LSTM [6], BiLSTM can capture context information more comprehensively and the correlations between contexts. Two LSTM networks, one is with a forward direction and the other is with a backward direction, are connected to the same output layer. Both of them are trained with the same sequence of data. There are three gates, that is, input gate, forget gate, and output gate, in a unit of LSTM. Equations (9)–(14) show the calculation processes:

$$i_t = \sigma\left(w_i[h_{t-1}, x_t] + b_i\right), \tag{9}$$

$$f_t = \sigma\left(w_f[h_{t-1}, x_t] + b_f\right), \tag{10}$$

Input: population size $P = N$, the number of mutations n_m, the batch size m, batch data D, and initial weight ω_0,
output: close price of the next day
(1) $\omega = \omega_0$
(2) Initializes model parameter ω_0:
(3) **for** $i = 1$ **to** $m/(Nn_m)$
(4) $param \longleftarrow \omega$ save model parameters
(5)   **for** $j = 1$ **to** N
(6)   **for** $k = 1$ **to** n_m
(7) $M(param)$ assign parameters to the model
(8)   get a batch D as input x_i of EBiLSTM;
(9)   **switch(k)**
(10)   **case1**: $loss_{square}, param_{square} \longleftarrow M(x_i, square, param)$
(11)   **case2**: $loss_{abs}, param_{abs} \longleftarrow M(x_i, abs, param)$
(12)   **case3**: $loss_{huber}, param_{huber} \longleftarrow M(x_i, huber, param)$
(13)   **end switch**
(14)   **if** $k = n_m$
(15) $loss_{min} \longleftarrow \min(loss_{square}, loss_{abs}, loss_{huber})$
(16) $param_{new} \longleftarrow (loss_{min}, param_{square}, param_{abs}, param_{huber})$
(17) $\omega \longleftarrow param_{new}$
(18)   **end for**
(19)   **end for**
(20) **end for**

ALGORITHM 1:

$$o_t = \sigma(w_o[h_{t-1}, x_t] + b_o), \tag{11}$$

$$\widetilde{c}_t = \tanh(w_c[h_{t-1}, x_t] + b_c), \tag{12}$$

$$c_t = f_t * c_{t-1} + i_t * \widetilde{c}_t, \tag{13}$$

$$h_t = o_t * \tanh(c_t). \tag{14}$$

Here, w_i, w_f, and w_o are the weights of LSTM and b_i, b_f, and b_o are the biases. i_t is input gate, f_t is forget gate, and o_t is output gate. The input vector is x_t and the output vector is h_t. c_t is the cell state and $t * \widetilde{c}_t$ means the candidate of the cell state. For the forward LSTM, it can be presented as $\overrightarrow{h}_t = \text{LSTM}(x_t, \overrightarrow{h}_{t-1})$. Accordingly, the backward LSTM is with $\overleftarrow{h}_t = \text{LSTM}(x_t, h_{t+1})$. Both $\overrightarrow{h}$ and $\overleftarrow{h}$ are the output of BiLSTM at time t,

$$h_t = \left[\overrightarrow{h}_t; \overleftarrow{h}_t\right]. \tag{15}$$

2.7. EBiLSTM. The complete process of EBiLSTM can be shown in Algorithm 1. At each evolutionary circle, BiLSTMs are updated with different mutations (or objectives). Among children of the next generation, only well-performing ones will survive and participate in the next rotation of training, following the principle of "survival of the fittest." Unlike a single BiLSTM with a fixed and static training objective, EBiLSTM integrates the advantages of different training objectives and selects the best solution. Therefore, during training, EBiLSTM can not only largely suppress the limitations (local optimal, etc.) of individual training objectives but also harness their advantages to find a better solution.

3. Experiment

3.1. Implementation Details. In order to evaluate the proposed EBiLSTM, experiments on several stock prediction tasks are run and their prediction results are presented in this section. Comparing with previous BiLSTM models, we show that the proposed EBiLSTM can achieve better stock prediction. Configurations from Table 1 are used in all the following experiments.

We evaluate EBiLSTM with three stock market indexes, as shown in Table 2. Every index includes data of 4750 days which are large enough to train EBiLSTM effectively.

3.2. Evaluation Metrics. To evaluate the proposed EBiLSTM, we use Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Mean Square Error (MSE) as quantitative metrics. They are shown as follows:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |\widehat{x}_i - x_i|, \tag{16}$$

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \left|\frac{\widehat{x}_i - x_i}{x_i}\right| \times 100, \tag{17}$$

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (\widehat{x}_i - x_i)^2. \tag{18}$$

For MAE, the error is calculated as an average of absolute differences between the target values and the predictions. It is a linear score and always nonnegative which means that all the individual differences are weighted equally on average [31]. The closer to 0 its value is, the higher the accuracy is. MAPE is measured by calculating the absolute error in each

TABLE 1: The structure of EBiLSTM and hyperparameter.

Input: real closing price (20 days)
[Layer1] internal cells (32) (BiLSTM | LSTM | GRU | RNN)
Output: predicted price in the next day
Hyperparameter: lr = 0.00001, epoch = 150, and timestep = 21
Survived parents number $\mu = 1$

TABLE 2: Data period of stock market indexes.

Index	Date
Shanghai Securities Composite Index	From Jan 2, 2001, to Jan 23, 2020
A-Share Index	From Jan 2, 2001, to Jan 23, 2020
Standard & Poor's 500 index	From Sep 8, 2000, to Sep 6, 2019

period, dividing this by actual value for that period, and finding the average of absolute percentage errors. MSE is basically based on the average squared error of our predictions. For each point, it calculates the square difference between the predictions and the target and then averages those values. The higher this value is, the worse the model is.

3.3. Effectiveness.
To evaluate the effectiveness of the EBiLSTM, it is compared with BiLSTMs under different loss functions, that is, BiLSTM-square, BiLSTM-abs, and BiLSTM-Huber. After training, closing prices of 50 days are predicted. For each model, simulations are taken 5 times independently and their average results of metrics are shown in Tables 3–5. Three stock market indexes shown in Table 1 are used in simulations, respectively. From these tables, it is evident that EBiLSTM achieves the best performance among the four models. The three models with a single objective function always get worse results because their objective functions cannot keep optimal during training. It may be easier for them to be suffered from local minima of parameters.

3.4. Training Stability.
To further examine the performance of prediction for different length of days, EBiLSTM as well as BiLSTM-square with a single objective is simulated with different days of prediction. The results with Shanghai Securities Composite Index are shown in Table 6. In this table, EBiLSTM can always get the best performance at different days of prediction.

3.5. Generality.
The architecture of EBiLSTM is general which can integrate different deep learning algorithms and keeps good performance. To demonstrate the generality, LSTM, GRU, and RNN are used to replace BiLSTM in the architecture, which are named ELSTM, EGRU, and ERNN, respectively. Simulation results with Shanghai Securities Composite Index are shown in Tables 7–9. In Table 7, ELSTM is compared with its corresponding algorithms with a single objective, that is, LSTM-square, LSTM-abs, and

TABLE 3: Standard & Poor's 500.

Metrics	EBiLSTM	BiLSTM-square	BiLSTM-abs	BiLSTM-Huber
MSE	**945**	1253	1107	1191
MAPE	**0.82**	0.91	0.89	0.90
MAE	**23.98**	26.63	25.88	26.20

TABLE 4: Shanghai Securities Composite Index.

Metrics	EBiLSTM	BiLSTM-square	BiLSTM-abs	BiLSTM-Huber
MSE	**608**	804	713	765
MAPE	**0.61**	0.74	0.68	0.71
MAE	**18.28**	22.09	20.42	21.22

TABLE 5: A-Share Index.

Metrics	EBiLSTM	BiLSTM-square	BiLSTM-abs	BiLSTM-Huber
MSE	**721**	899	785	901
MAPE	**0.66**	0.73	0.68	0.73
MAE	**20.78**	22.95	21.33	22.79

TABLE 6: Shanghai Securities Composite Index.

Days	Algorithms	MSE	MAPE	MAE
50	EBiLSTM	**608**	**0.61**	**18.28**
	BiLSTM-square	804	0.74	22.09
100	EBiLSTM	**534**	**0.62**	**18.44**
	BiLSTM-square	701	0.71	21.06
150	EBiLSTM	**637**	**0.66**	**19.57**
	BiLSTM-square	830	0.78	22.94
200	EBiLSTM	**951**	**0.75**	**22.26**
	BiLSTM-square	1231	0.88	26.00

TABLE 7: Shanghai Securities Composite Index.

Metrics	ELSTM	LSTM-square [34]	LSTM-abs	LSTM-Huber
MSE	**653**	706	649	681
MAPE	**0.65**	0.68	0.65	0.67
MAE	**19.53**	20.38	19.44	20.16

TABLE 8: Shanghai Securities Composite Index.

Metrics	EGRU	GRU-square	GRU-abs	GRU-Huber
MSE	**584**	844	812	801
MAPE	**0.63**	0.75	0.74	0.73
MAE	**18.95**	22.60	22.11	21.97

TABLE 9: Shanghai Securities Composite Index.

Metrics	ERNN	RNN -square	RNN -abs	RNN-Huber
MSE	**574**	979	1044	989
MAPE	**0.64**	0.83	0.85	0.83
MAE	**19.17**	25.00	25.49	24.92

LSTM-Huber. The results show that ELSTM is the best among them. Similar results can be got for EGRU and ERNN which are shown in Tables 8 and 9, respectively.

4. Conclusion

Stock market exchanges have become popular, encouraging researchers to find predictions using new technologies or methods. Proper predictive techniques can help investors get higher profits from the stock market. However, it is difficult to improve the prediction only by using neural networks because gradient descent in the training process is easy to fall into local optimal. To overcome it, we propose an evolutionary framework (EBiLSTM) for stock prediction. In the framework, we propose an evolutionary algorithm to evolve a population of BiLSTM. In contrast to conventional BiLSTM, the evolution of EBiLSTM selects three different BiLSTM objective functions as mutated individuals, then calculates their respective losses for evaluation, and finally selects the optimal objective function through the minimum loss. Experiments show that EBiLSTM can improve the training stability of BiLSTM and achieves more accuracy of prediction than others in various stock market indexes. For further investigation, there are still some promising directions in the future. For example, the BiLSTM with attention mechanism might have more potential to get better performance. In forecast-based trading, it is interesting to design a portfolio allocation to improve the performance of BiLSTM.

Data Availability

The data are available at https://tushare.pro/.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

References

[1] S. Lauren and S. D. Harlili, "Stock trend prediction using simple moving average supported by news classification," in *Proceedings of the 2014 International Conference of Advanced Informatics: Concept, Theory and Application (ICAICTA)*, Bandung, Indonesia, August 2014.

[2] A. Izzah, Y. A. Sari, R. Widyastuti, and T. A. Cinderatama, "Mobile app for stock prediction using improved multiple linear regression," in *Proceedings of the 2017 International Conference on Sustainable Information Engineering and Technology (SIET)*, Batu, Java, Indonesia, November 2017.

[3] P. Somani, S. Talele, and S. Sawant, "Stock market prediction using hidden markov model," in *Proceedings of the 2014 IEEE 7th Joint International Information Technology and Artificial Intelligence Conference*, Chongqing, China, December 2014.

[4] L. Xiong and Y. Lu, "Hybrid ARIMA-BPNN model for time series prediction of the Chinese stock market," in *Proceedings of the 2017 3rd International Conference on Information Management (ICIM)*, Chengdu, China, April 2017.

[5] B. L. Pooja, S. Kanakaraddi, and M. M. Raikar, "Sentiment based stock market prediction," in *Proceedings of the 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, Belgaum, India, December 2018.

[6] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[7] Y. F. Lin, Y. L. Ueng, W. H. Chung, and T. M. Huang, "Stock price range forecast via a recurrent neural network based on the zero-crossing rate approach," in *Proceedings of the IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)*, Shenzhen, China, May 2019.

[8] Y. Xing, C. Lv, and D. Cao, "Personalized vehicle trajectory prediction based on joint time-series modeling for connected vehicles," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 2, pp. 1341–1352, 2020.

[9] M. Hwang, B. Thananjeyan, and S. Paradis, "Efficiently calibrating cable-driven surgical robots with RGBD sensing, temporal windowing, and linear and recurrent neural network compensation," 2020, https://arxiv.org/abs/2003.08520v4.

[10] M. Ravanelli, P. Brakel, M. Omologo, and Y. Bengio, "Light gated recurrent units for speech recognition," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 92–102, 2018.

[11] C. Wang and Q. Gao, "High and low prices prediction of soybean futures with LSTM neural network," in *Proceedings of the IEEE International Conference on Software Engineering and Service Science (ICSESS)*, Beijing, China, November 2018.

[12] Y. Liu and X. Liu, "A trend-based stock index forecasting model with gated recurrent neural network," in *Proceedings of the IEEE International Conference on Progress in Informatics and Computing (PIC)*, Suzhou, China, December 2018.

[13] Y. H. Go and J. K. Hong, "Prediction of stock value using pattern matching algorithm based on deep learning," *International Journal of Recent Technology and Engineering*, vol. 8, no. 2, pp. 31–35, 2019.

[14] Y. Song, J. W. Lee, and J. Lee, "A study on novel filtering and relationship between input-features and target-vectors in a deep learning model for stock price prediction," *Applied Intelligence*, vol. 49, no. 3, pp. 897–911, 2019.

[15] E. Chong, C. Han, and F. C. Park, "Deep learning networks for stock market analysis and prediction: methodology, data representations, and case studies," *Expert Systems with Applications*, vol. 83, pp. 187–205, 2017.

[16] D. Lien Minh, A. Sadeghi-Niaraki, H. D. Huy, K. Min, and H. Moon, "Deep learning approach for short-term stock trends prediction based on two-stream gated recurrent unit network," *IEEE Access*, vol. 6, pp. 55392–55404, 2018.

[17] Z. Zhang, S. Zohren, and S. Roberts, "DeepLOB: deep convolutional neural networks for limit order books," *IEEE Transactions on Signal Processing*, vol. 67, no. 11, pp. 3001–3012, 2019.

[18] C. A. Parlour and D. J. Seppi, "Limit order markets: a survey," *Handbook of Financial Intermediation and Banking*, Elsevier, vol. 5, pp. 63–96, Amsterdam, Netherlands, 2008.

[19] X. Shao, D. Ma, Y. Liu, and Q. Yin, "Short-term forecast of stock price of multi-branch LSTM based on K-means," in *Proceedings of the 2017 4th International Conference on*

Systems and Informatics (ICSAI), Hangzhou, China, November 2017.

[20] O. Orojo, J. Tepper, T. M. McGinnity, and M. Mahmud, "A multi-recurrent network for crude oil price prediction," in *Proceedings of the 2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, Xiamen, China, December 2019.

[21] F. G. Mohammadi, M. H Amini, and H. R. Arabnia, "Evolutionary computation, optimization and learning algorithms for data science," 2019, https://arxiv.org/abs/1908.08006v1.

[22] R. Miikkulainen, J. Liang, and E. Meyerson, "Evolving deep neural networks," 2017, https://arxiv.org/abs/1703.00548v2.

[23] S. R. Young, D. C. Rose, and T. P. Karnowski, "Optimizing deep learning hyper-parameters through an evolutionary algorithm," in *Proceedings of the Workshop on Machine Learning in High-Performance Computing Environments*, Austin, Texas, November 2015.

[24] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 5, pp. 828–841, 2019.

[25] C. Wang, C. Xu, X. Yao, and D. Tao, "Evolutionary generative adversarial networks," *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 6, pp. 921–934, 2019.

[26] A. Banerjee, D. Ghosh, and S. Das, "Evolving network topology in policy gradient reinforcement learning algorithms," in *Proceedings of the 2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*, Gangtok, India, February 2019.

[27] C. Fernando, D. Banarse, and C. Blundell, "PathNet: evolution channels gradient descent in super neural networks," 2017, https://arxiv.org/abs/1701.08734.

[28] Y. Hu, X. Sun, X. Nie, Y. Li, and L. Liu, "An enhanced LSTM for trend following of time series," *IEEE Access*, vol. 7, pp. 34020–34030, 2019.

[29] Y. R. Li, Z. F. Zhu, and D. Q. Kong, "EA-LSTM: evolutionary attention-based LSTM for time series prediction," 2018, https://arxiv.org/abs/1811.03760.

[30] R. Hafezi, J. Shahrabi, and E. Hadavandi, "A bat-neural network multi-agent system (BNNMAS) for stock price prediction: case study of DAX stock price," *Applied Soft Computing*, vol. 29, pp. 196–210, 2015.

[31] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? —arguments against avoiding RMSE in the literature," *Geoscientific Model Development*, vol. 7, no. 3, pp. 1247–1250, 2014.

Research Article

The Impact and Stability Analysis of Commercial Banks' Risk Preference on SMEs' Credit Financing Based on DSGE Model

Di Gao [ID],[1] **Zhaohui Hao** [ID],[2] **Jiangming Ma** [ID],[3] **Huanyu He** [ID],[1] and **Meng Li** [ID][4]

[1]*School of International Business, Southwestern University of Finance and Economics, Chengdu, Sichuan 611130, China*
[2]*School of Finance, Southwestern University of Finance and Economics, Chengdu, Sichuan 611130, China*
[3]*School of Economics, Xihua University, Chengdu, Sichuan 610039, China*
[4]*School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan 611731, China*

Correspondence should be addressed to Zhaohui Hao; haozh1005@163.com

As an important driving force for China's economic transformation and upgrading, the problems of financing difficulties and expensive financing for SMEs have become increasingly prominent. The main objective of this paper was to analyze the impact of financial intermediary departments' risk preference on corporate finance. Under the revised DSGE framework, this paper discusses the impact and stability analysis of commercial banks' risk preferences on SMEs' financing. The results show that positive interest rate shocks inhibit commercial banks' credit to SMEs, and with the increasing weight of commercial banks' risk preference for default rate, the trend of credit repression will be intensified.

1. Introduction

The essence of finance is to serve the real economy (Report to the 19th National Congress). With the transformation of China's economy from high-speed development to high-quality development, maintaining financial system stability has become the focus of China's economic work. After the subprime mortgage crisis, many economists began to incorporate financial stability into the monetary policy framework [1, 2]. Under the monetary policy goal of stabilizing finance, commercial banks increase the demand for the security of loan funds and reduce the supply of credit to high-risk enterprises. Based on the information asymmetry theory, small and medium-sized enterprises (SMEs) are more likely to produce adverse selection and moral hazard problems, resulting in a decline in the banks nonperforming loan rate and fund security. SMEs are a vital force for national economic and social development, and they play an important and irreplaceable role in stabilizing growth, increasing employment, promoting innovation, and improving people's livelihood. According to data released by the China Association of Small and Medium Enterprises in 2019,

SMEs (including individual industrial and commercial households) account for 94.15 % of the total number of enterprises, the value of the final products and services created is equivalent to 60% of the total GDP, and the tax payment account for 50% of the total national tax. Under the backdrop of world economic growth slowing down, the external financing problem of SMEs is becoming increasingly prominent, and SMEs are troubled by financing constraints. According to data from the China Household Finance Survey, in 2011, the scale of commercial bank credit available to SMEs accounted for 12.2% of the total credit scale; by 2017, the proportion dropped to 9.8%. The purpose of this paper was to analyze how financial intermediary departments' risk preference affect SMEs financing under the revised DSGE framework.

The existence of financing constraints in SMEs has always been the focus of theoretical attention. On the one hand, information asymmetry believes that the difficulty of financing for SMEs is due to the information asymmetry between banks and enterprises. Stiglitz and Weiss [3] pay attention to the credit decision problem of commercial banks for the first time. In the credit market, information

asymmetry between banks and enterprises is likely to induce adverse selection and moral hazards. Compared with large enterprises, SMEs are prone to lose their credit qualification due to the lack of qualified collateral in the process of qualification examination before obtaining bank credit. In addition, in order to ensure the safety of credit, banks need to pay high supervision costs; otherwise, SMEs will lose their credit qualification due to the lack of qualified collateral. The moral hazard of business operators will damage the interests of commercial banks. In order to reduce operating costs and avoid their own operational risks, banks will tend to implement the policy of reluctant loans to SMEs [4, 5]. Therefore, in the context of mortgage loans as the main loan business of commercial banks, the financing difficulty of SMEs is largely due to the lack of qualified collateral. Casey et al. [5] use Euro area firm-level data since the recent financial crisis, as opposed to credit financing, and smaller, self-rationing borrowers are more likely to apply for grant finance. Farinha and Félix [6] examine the importance of credit demand and credit-supply-related factors in explaining the evolution of credit granted to Portuguese SMEs. The results suggest that credit supply mostly depends on the firms' ability to generate cash flows and reimburse their debt, and on the amount of collateral.

On the other hand, based on risk-taking theory, SMEs have a weak ability to bear risk shocks. When the external economy fluctuates, the income of SMEs will have a high degree of uncertainty, and there will be a high risk of default on bank credit. The results of Blumberg and Letterie [7] show that the rejection of loans depend largely on corporate commitments and signals of loan repayment, as well as the chance of success of investment projects. Ioannidou et al. [8], Bonfim and Soares [9] studied from the perspective of bank risk-taking and found that when interest rates were low, Banks would lend more to risky companies in pursuit of more returns. This shows that under certain conditions, the change of risk preference of commercial Banks can have an impact on the credit level of SMEs. At the same time, Chinese researchers incorporate the problem of corporate default risk into the credit decision-making mechanism when they focus on commercial banks' credit decision-making. Pang [10] discusses the credit decision-making model and mechanism of commercial banks under the default risk framework. Dai et al. [11] study the relationship between credit level and bank loans of listed companies based on China's "laolian" data. The results show that the financing cost of high-risk enterprises is determined by the high default rate of enterprises.

Based on existing literature, the main reasons of financing difficulty and high cost for SMEs could be summarized as follows: (1) The lack of loan collateral provided by SMEs is easy to cause financing difficulties; (2) SMEs have a high risk of default after they successfully obtain loans, which forces commercial banks to raise the loan interest rate to make up for the losses caused by default. In that case, the problem of expensive financing is formed.

Although the problem of credit financing for SMEs has always been a hot issue in the theoretical field, it is mostly limited to microscopic empirical studies of available data

[12–14]. Using the revised DSGE framework, this paper studies the credit decision-making mechanism of commercial banks on SMEs, and examines the impact of different risk preference weights on SMEs' credit financing. With the introduction of the financial accelerator mechanism into the DSGE framework by Bernanke et al. [15], many scholars have introduced a separate financial intermediary sector into the model [16–19]. In this paper, the DSGE model including the banking system is used as the basic analysis tool, and we reconstruct the enterprise sector of the model to depict the influence of risk tolerance of commercial banks on SMEs credit financing. In order to analyze this problem accurately, this paper not only describes the loan behavior of commercial banks to SMEs but also considers its role in the macroeconomy as well as the influence and feedback of monetary policy transmission.

This paper makes several contributions to related literature. First, previous studies assume that commercial banks are fully competitive, so the debt contract between commercial banks and enterprises depends on the return of enterprises [20, 21]. This paper assumes that commercial banks are monopolistic competition, and the debt contract between SMEs and commercial banks depends on the loan interest rate elasticity of commercial banks, which is introduced into the risk-preference analysis framework of commercial banks for SMEs. Second, considering the state-owned nature, China's commercial banks pay more attention to the safety of credit funds, that is, the fact of default risk than profit maximization. In this paper, corporate default is included in the lending standard, and different credit default risk preference of commercial banks for SMEs is investigated. Third, based on the fact that there is a devaluation gap of mortgage assets before and after default, this paper examines the impact of different asset devaluation of SMEs on commercial bank lending.

The paper is organized as follows. In Section 2, the basic DSGE model is described, including heterogeneous economic sectors such as enterprise sector, household sector, final product sector, and central bank. In Section 3, the debt contract and the stability analysis between enterprises and commercial banks are described in detail. In Section 4, numerical simulation is introduced. In Section 5, the main conclusions are summarized.

2. Statement of Background

At present, the mainstream monetary policy model adopted by western developed countries is the new Keynesian DSGE model. It is based on the core of RBC model and constructed by introducing various Keynesian economic assumptions. Some scholars also call it new neoclassical synthesis [22]. Since the research published by Kydland and Prescott [23] and Prescott [24], Real Business Cycle (RBC) theory provides a standard research framework for economic cycle analysis and modern macroeconomic theory research. The new Keynesian model is established and developed on the basis of the RBC model; therefore, the many similarities between the New Keynesian model and RBC model are as follows: First, the representative family lives indefinitely and

chooses consumption and labor to pursue utility maximization under the condition of intertemporal budget constraints. Second, there are a large number of enterprises with the same technology, and they are all impacted by external technology. Third, consistent with the RBC theory, the equilibrium state of the model is represented as a random process of endogenous variables consistent with the most intertemporal behaviors of households and firms and with market clearing conditions.

However, there are many differences between the new Keynesian model and the RBC model in theoretical assumptions, the most important of which can be summarized as follows: First is the monopoly competition hypothesis. In contrast to the market clearing hypothesis of the neoclassical Walras auctioneer, the new Keynesian theory holds that the prices of goods and factors of production are determined by the optimal behavior of private sector agents. Second is the nominal viscosity hypothesis. Firms face costs when adjusting output prices and workers when adjusting wages, so nominal variables cannot immediately adjust to market clearing levels. Third is the short-term, nonneutral assumption of monetary policy. Due to the existence of the nominal stickiness hypothesis, short-term nominal price changes will not lead to a 1: 1 change in expected inflation, which enables monetary policy to have an impact on the real interest rate in the economy, thus leading to corresponding fluctuations in consumption, investment, output, and employment. But, in the long run, all wages and prices will adjust and the economy will eventually return to equilibrium.

In the current study, there are two main methods of empirical research on the effectiveness of monetary policy. One is the vector autoregression (VAR) model represented by Sims [25], Bernake and Blinder [26]. As soon as the model was put forward, it was praised by many economists for its simple structure and good predictive ability, and it was widely used. However, the effectiveness of the VAR model has also been criticized—some researchers hold that the VAR model lacks strict theoretical basis, while others consider that there is only a relatively limited amount of information in the model due to the limited data conditions [27, 28]. The other one, which is widely used in monetary policy analysis and adopted by many central Banks, is the DSGE model. With the continuous improvement of economists, the matching degree of the DSGE model with the real economy has been greatly enhanced. For example, Christiano et al. [29] introduced a large number of frictions including sticky wage and price, habit preference, investment adjustment cost, and capacity utilization rate. A large number of studies have proved that the DSGE model can not only fit macroeconomic data well but also perform well in out of sample prediction, even better than traditional econometric models [16]. Based on the research of Christiano et al. [16] and Qiu and Zhou [30], and combined with the theoretical model constructed in our research, this part proposed the following research hypothesis, and all the other assumptions are basically the same as those in the traditional DSGE model:

Hypothesis 1. In order to characterize limited banks' risk preference for SMEs, this paper divides enterprises into two types: large enterprises (proportion $1 - \eta$) and SMEs (proportion η).

Hypothesis 2. In order to match the two types of enterprises, this paper assumes that commercial banks are divided into two types: A and B. Type A commercial banks provide loans to large enterprises, and type B commercial banks provide loans to SMEs.

Hypothesis 3. The traditional DSGE model introduces capacity utilization to investigate the debt contract between capital production departments and banks. Considering that small and medium-sized enterprises (especially individual industrial and commercial households) have weak capital production capacity and are mostly engaged in the production of intermediate goods, which is different from the traditional DSGE model assumption, this paper uses the Douglas production function to investigate the relationship between intermediate goods sector and interbank credit activities.

In order to explain the basic model constructed in our study more clearly, the setting logic of the main departments in the theoretical model is shown in Figure 1:

2.1. The Household Sector.
Assuming that there are a large number of families with homogeneous and indefinite periods in the economy, each family chooses to consume, provide labor, and save under budget constraints to maximize intertemporal utility. The lifetime utility function faced by a family is as follows:

$$\max E_0 \sum_{t=0}^{\infty} \beta_h^t U(c_t, n_t), \tag{1}$$

where E_t is an expectation operator; $\beta_h \epsilon(0, 1)$ represents subjective discount factor; c_t and n_t are the actual household consumption and the labor provided during the period t, respectively; and $U(\cdot)$ is the immediate utility function of the family. Without loss of generality, the immediate utility function is

$$U = \frac{(c_t - bc_{t-1})^{1-\sigma}}{1 - \sigma} - \frac{n_t^{1+\varphi}}{1 - \varphi}, \tag{2}$$

where σ and φ are the reciprocal of the intertemporal substitution elasticity of household consumption and the elasticity of labor supply, respectively.

The family is bound by

$$i_t + c_t + d_t = w_t n_t + r_t^k k_{t-1} + \frac{r_{t-1}^d d_{t-1}}{\pi_t} + \Pi_t, \tag{3}$$

where i_t is the investment (or savings); d_t is the financial assets for residents (simplified as bank deposits); w_t is the wage; r_t^k is the capital rent; k_t is the ending capital stock; r_t^d is the nominal deposit rate; π_t is the inflation rate; Π_t is the profits from companies and banks in the period t.

Capital accumulation equation is
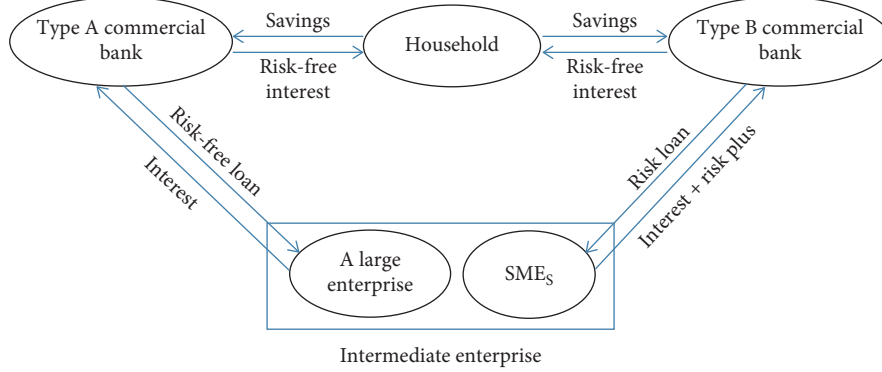
Figure 1: Logical framework.

$$k_t = (1 - \delta)k_{t-1} + i_t, \tag{4}$$

where δ is the capital depreciation rate. The optimal behavior decision of the household sector is the optimal choice of consumption c_t, labor n_t, capital stock k_t, and savings d_t under the constraints of household budget and capital accumulation equations. Therefore, the first-order condition of this optimization problem is as follows:

$$\frac{(c_t - bc_{t-1})^{-\sigma}}{(1 - b)^{-\sigma}} - \beta b \frac{(c_{t+1} - bc_t)^{-\sigma}}{(1 - b)^{-\sigma}} = \lambda_t,$$

$$\lambda_t w_t = n_t^{\varphi},$$

$$E_t \left(\beta \frac{\lambda_{t+1}}{\pi_{t+1}} r_t^d \right) = \lambda_t, \tag{5}$$

$$E_t \left\{ \beta \lambda_{t+1} \left[(1 - \delta) + r_{t+1}^k \right] \right\} = \lambda_t.$$

where λ_t, is, respectively, the Lagrange multiplier corresponding to the constraints of household budget, while λ_t is actually the marginal utility of wealth.

2.2. The Business Sector

2.2.1. SMEs.
Let the production function of SMEs be the Cobb–Douglas form:

$$y_t^{xw}(s) = A_t^{xw} \left[k_{t-1}^{xw}(s) \right]^{\alpha} \left[n_t^{xw}(s) \right]^{1-\alpha}, \tag{6}$$

, where $y_t^{xw}(s)$ is the s_{th} type intermediate product; $k_{t-1}^{xw}(s)$ is the capital used to produce intermediate products; $n_t^{xw}(s)$ is the labor used to produce intermediate products; α and $1 - \alpha$ are the elasticity of output with respect to capital and labor, respectively; A_t^{xw} represents the technological level, that is, the productivity of SMEs, which satisfies the condition that technological shocks obey the process $AR(1)$.

At the beginning of the period t, the SMEs determined the scale of production, and obtained loans from banks. Then, SMEs employ labor $n_t^{xw}(s)$ and lease capital $k_t^{xw}(s)$ with own net assets and loans for productions. The impact capability of SEMs is relatively weak, assuming that the returns realized by SMEs during each period would encounter a heterogeneous and uncertain impact ω_t^{xw}, and $\ln(\omega_t^{xw}) \sim N(-(\sigma^2/2), \sigma^2)$. This shock makes SMEs' future earnings uncertain, thus converting the original earnings into effective earnings $\omega_t^{xw} y_t^{xw}(t)$; at the end of the period t, after the SMEs sell the current products to obtain production returns, the bank principal and interest are repaid.

Suppose a standard debt contract is signed between the SME and the bank, that is, a contract specifying the loan interest rate and loan amount: $(r_t^{l,xw}, l_t^{xw})$, where r_t^l is the SMEs' loan interest rate; l_t^{xw} is the SMEs' loan amount. At this time, the total investment of SMEs for production is

$$w_t n_t^{xw}(s) + r_t^k k_{t-1}^{xw}(s) = l_t^{xw} + N_t^{xw}. \tag{7}$$

The left side of (7) indicates the production scale determined by the SME at the beginning of the period, that is, the production input (cost); $l_t^{xw} = k_t^{xw}[w_t n_t^{xw}(s) + r_t^k k_{t-1}^{xw}(s)]$ represents the loan received by the SMEs from the bank at the beginning of the period t, where k_t^{xw} represents the proportion of bank loans in the production investment of SMEs, which can be regarded as the leverage ratio of SMEs' external financing; $N_t^{xw} = (1 - k_t^{xw}/k_t^{xw})l_t^{xw}$ represents the net assets of SMEs at the beginning of the period t.

The return on production of SMEs can be expressed as

$$1 + R_t^{k,xw}(s) = \frac{\omega_t^{xw} y_t^{xw}(s)}{w_t n_t^{xw}(s) + r_t^k k_{t-1}^{xw}(s)}, \tag{8}$$

where $Rk_t^{xw}(s)$ indicates the input-output ratio of the s_{th} SME.

Intermediate product manufacturing companies determine the proportion of factor demand through cost minimization. Given the final demand, by choosing $n_t^{xw}(s)$ and $k_t^{xw}(s)$ to minimize the cost in the production process, namely,

$$\min_{n_t^{xw}(s), k_{t-1}^{xw}(s)} \left[w_t \left(1 + \psi_n^{xw} r_t^{l,xw} \right) n_t^{xw}(s) + r_t^k \left(1 + \psi_k^{xw} r_t^{l,xw} \right) k_{t-1}^{xw}(s) \right], \tag{9}$$

where ψ_n^{xw} represents the proportion of wages which the enterprise has to pay before production; ψ_k^{xw} represents the proportion of rent that the enterprise has to pay. The capital-labor ratio of all intermediate goods, enterprises of SMEs obtained from the first-order conditions is

$$\frac{k_{t-1}^{xw}}{n_t^{xw}} = \frac{a}{1 - a} \frac{w_t}{r_t^k} \frac{\left(1 + \psi_n^{xw} r_t^{l,xw} \right)}{\left(1 + \psi_k^{xw} r_t^{l,xw} \right)}. \tag{10}$$

Thus, it can be further obtained that the actual marginal cost of per unit product produced by an intermediate product enterprise is

$$
\begin{aligned}
mc_t^{xw} = &\left[r_t^k \left(1 + 1 + \psi_k^{xw} r_t^{l,xw} \right) k_{t-1}^{xw}(s) \right]^a \\
&\cdot \left[w_t \left(1 + \psi_n^{xw} r_t^{l,xw} \right) \right]^{1-a} a^{-a} (1-a)^{-(1-a)} \left(A_t^{xw} \right)^{-1}.
\end{aligned}
\tag{11}
$$

2.2.2. Large Enterprises. Similar to small and medium-sized enterprises, at the beginning of the period t, large enterprises determine the scale of production, obtain loans from banks, which is used to hire workers and purchase the capital required for production; compared with SMEs, large and medium-sized enterprises have a certain amount of assets that can be used for bank loan mortgages. Even in the event of corporate default, banks can protect their income by processing mortgaged assets. Thus, this article assumes that there is no default for large and medium-sized enterprises. Therefore, the scale of loans available to large enterprises is

$$
w_t n_t^{dz}(s) + r_t^k k_{t-1}^{dz}(s) = l_t^{dz} + N_t^{dz}, \triangleleft
\tag{12}
$$

where the left side of (12) is the total production input of the large enterprises at the beginning of the period t; the first item on the right side of (12) is the loan of the large enterprises at the beginning of the period t, $l_t^{dz} = k_t^{dz} [w_t n_t^{dz}(s) + r_t^k k_{t-1}^{dz}(s)]$, where k_t^{dz} represents the proportion of bank loans in the production investment of large enterprises and $N_t^{dz} = (1 - k_t^{dz}/k_t^{dz}) l_t^{dz}$ represents the net assets of large enterprises at the beginning of the period t.

Given the final demand, by minimizing the cost, the capital-labor ratio of all intermediate goods enterprises for large enterprises can be obtained as

$$
\frac{k_t^{dz}}{n_t^{dz}} = \frac{a}{1-a} \frac{w_t}{r_t^k} \frac{\left(1 + \psi_n^d r_t^{l,d} \right)}{\left(1 + \psi_k^d r_t^{l,d} \right)}.
\tag{13}
$$

From (13), it can be further obtained that the actual marginal cost of per unit product produced by an intermediate product enterprise is

$$
\begin{aligned}
mc_t^{dz} = &\left[r_t^k \left(1 + \psi_k^{dz} r_t^{l,dz} \right) \right]^a \left[w_t \left(1 + \psi_n^{dz} r_t^{l,dz} \right) \right]^{1-a} a^{-a} \\
&\cdot (1-a)^{-(1-a)} \left(A_t^{dz} \right)^{-1}.
\end{aligned}
\tag{14}
$$

Regarding the debt contract between large enterprises and banks, based on the high default rate of SME credit, this article assumes that there is no default situation for large and medium-sized enterprises. At this time, the conditions for large and medium-sized enterprises to meet the loan are

$$
\begin{aligned}
\left(1 + R_t^{k,dz} \right) & \left[w_t n_t^{dz}(s) + r_t^k k_{t-1}^{dz}(s) \right] - \left(1 + r_t^{l,dz} \right) l_t^{dz} \\
&= (1 + \lambda_{dz}) \left(1 + r_t^d \right) N_t^{dz},
\end{aligned}
\tag{15}
$$

where λ_{dz} represents the extra profit required by large companies compared to the profit margin obtained by depositing net assets in banks; the left side of (15) is the gains from large and medium-sized companies' production in the period t; the right side of (15) is the opportunity cost of large and medium-sized companies' production.

2.3. Determination of the Optimal Price for the Enterprise. Since the market for intermediate products is in a state of monopolistic competition, manufacturers of intermediate products have certain pricing power for intermediate products under demand constraints. Suppose that the $1 - \theta$ percent of intermediate goods manufacturers adjust prices; at this time, under the condition of demand constraints, the intermediate goods company can maximize profit by selecting the optimal intermediate goods price, and its behavior equation can be expressed as

$$
\max_{p_t^{*,j}(s)} E_t \left\{ \sum_{k=0}^{\infty} \left[(\beta\theta)^k v_{t+k} \left(p_{t+k}^{a,j}(s) - mc_{t+k}^j \right) y_{t+k}^j(s) \right] \right\},
\tag{16}
$$

where β is the discount rate of intermediate product manufacturers. Since the manufacturers produce on behalf of residents, the discount rate is equal to the discount rate used by residents; v_t is the marginal utility of wealth; j takes dz and xw separately, indicating large enterprises and SMEs. $p_t^{a,j}(s) = p_t^{*,j}(s) (\pi_t^j, \ldots, \pi_{t+k-1}^j / \pi_{t+1}, \ldots, \pi_{t+k})$ represents the relative price of the product of the s_{th} category enterprise; $y_{t+k}^j(s) = |(p_t^{a,j}(s)/p_t^j)|^{-(1+\varepsilon/\varepsilon)}$ represents the s_{th} type of intermediate products produced by the enterprise; y_{t+k}^j represents the final product produced by the enterprise; $p_t^{*,j}(s)$ represents the optimal solution. By solving (16), the optimal pricing of intermediate goods manufacturers is

$$
E_t \sum_{k=0}^{\infty} \left\{ \frac{(\beta\theta)^k v_{t+k} y_{t+k}^j(s)}{\varepsilon p_t^{a,j}(s)} \left| p_t^{*,j}(s) \frac{\pi_t^j, \ldots, \pi_{t+k-1}^j}{\pi_{t+1}, \ldots, \pi_{t+k}} - (1+\varepsilon) mc_{t+k}^j \right| \right\} = 0.
\tag{17}
$$

Since $1 - \theta$ percent of intermediate goods manufacturers adjust prices in each period, the corresponding enterprise price index can be expressed as

$$
p_t^j = \left| \theta \left(p_{t-1}^j \right)^{-(1/\varepsilon)} + (1-\theta) \left(p_t^{*,j} \right)^{-(1/\varepsilon)} \right|^{-\varepsilon}.
\tag{18}
$$

From (18), the new Keynesian Phillips curve (logarithmic linearized form) is derived:

$$
\hat{\pi}_t^j = \beta E_t \hat{\pi}_{t+1}^j + \frac{(1-\theta)(1-\beta\theta)}{\theta^j} \left(\widehat{mc}_t^{dz} - p_t^j \right),
\tag{19}
$$

where $\widehat{\pi}_t^j$ and $\widehat{mc}_t^j$ denote the deviation of inflation and real marginal cost of SMEs from their steady states, respectively; p_t^j represents the relative price of SME product prices relative to final consumer goods. Then, we get

$$\widehat{\pi}_t^j - \widehat{\pi}_t = p_t^j - p_{t-1}^j. \tag{20}$$

2.4. End Product Manufacturer.

The manufacturer that produces the final product processes the intermediate product into a final product and provides it to other economic entities. Assuming that the intermediate products are continuously distributed in the interval $[0, 1]$, the manufacturer that produces the final product uses the following production technology:

$$Y_t = \left| \eta^{(1/vy)} \left(Y_t^{xw} \right)^{1+(1/vy)} + (1 - \eta)^{(1/vy)} \left(Y_t^{dz} \right)^{1+(1/vy)} \right|^{(vy/vy+1)}, \tag{21}$$

where vy represents the elasticity of products substitution between SMEs and large and medium-sized enterprise in the final product synthesis; η indicates the proportion of SMEs and large and medium-sized enterprises in the synthesis of products. From the profit maximization condition, the demand function of the final product manufacturer is

$$Y_t^{xw}(s) = \eta \left(P_t^{xw} \right)^{vy} Y_t,$$
$$Y_t^{dz}(s) = (1 - \eta) \left(P_t^{dz} \right)^{vy} Y_t. \tag{22}$$

Since the deal of final product manufacturer in a perfectly competitive market, its profit is zero and thus we can obtain

$$1 = \left[\eta \left(P_t^{xw} \right)^{1+vy} + (1 - \eta) \left(P_t^{dz} \right)^{1+vy} \right]^{(1/vy+1)} \tag{23}$$

2.5. Central Bank and Market Clearing.

Central banks usually focus on inflation and output gaps. The equations of monetary policy that follow the Taylor rule are

$$r_t^d = \left(r_{t-1}^d \right)^{\rho m} \left[\overline{r}^d \left(\frac{\pi_{t+1}}{\overline{\pi}} \right)^{\xi_\pi} \left(\frac{y_t}{\overline{y}} \right)^{\xi_y} \right]^{(1-\rho m)} u_t,$$
$$u_t = \phi_m u_{t-1} + \varepsilon_t, \tag{24}$$

where r_t^d, $\overline{\pi}$, and $\overline{y}$ represent the steady-state nominal interest rate, inflation, and output, respectively; ρm reflects the smoothness of the interest rate; and the bigger ρm is, the greater the persistence of an external shock to interest rates will be; ξ_π and ξ_y are response coefficients of inflation and output gap to the rate of monetary growth, respectively; $u_{m,t}$ expresses monetary policy shocks.

When the model is in equilibrium, all markets are cleared. Market equilibrium includes product market, capital market, labor market, and credit market equilibrium. According to Walras' rule, only three of these market equilibriums are required.

Product market clearing: $y_t = c_t + i_t$;

Credit market clearing: $l_t^{dz} + l_t^{xw} = l_t^b$, that is, corporate credit demand equals bank credit supply;

Labor market clearing: $n_t^{xw} + n_t^{dz} = n_t$, that is, the labor demand of an enterprise is equal to the labor supply of residents.

There are three external shocks in the model economy: monetary policy shocks and production technology shocks for large enterprises and SMEs. They satisfy the standard first-order autoregressive equations. The specific form is

$$\ln z_t = (1 - \rho_z) \ln z + \rho_z \ln z_{t-1} + e_{i,t}, \tag{25}$$

where ρ_i reflects the degree of impact inertia and $e_{i,t}$ reflects external shocks, satisfying $N(0, \sigma_i^2)$ and $i \in (u, a_{xw}, a_{dz})$.

3. Main Results

Following Dib [31], Peng and Fang [32], there are two types of homogeneous commercial banks, namely, A and B in the model, which provide loans to large enterprises and SMEs, respectively. There is no default risk when commercial banks lend to large enterprises. But, there is default risk when they provide short-term loans to SMEs.

When the SMEs' credit defaults, the commercial bank has the right to dispose of the remaining property of the SME. Because of information asymmetry, commercial banks do not know the real state of capital gains. Therefore, there is a cost to identify the state. This paper assumes that the regulatory cost that financial institutions needs to pay is μ percent of the total profit ratio paid by enterprises to commercial banks as the proportion of assets lost by banks after bankruptcy and liquidation. When the monitoring cost μ increases, the mortgage rate ω rises. Different from previous research, this paper sets σ_a^{xw} to measure the depreciation of corporate assets after default. When the σ_a^{xw} increases, the mortgage rate decreases.

In reality, large enterprises with large asset scale and good credit often prioritize the credit resource allocation of commercial Banks. However, SMEs are more likely to default when their production activities are impacted by external risks; as a result, the security of bank credit funds decline, resulting in the reluctance to lend to SMEs. In order to describe commercial banks' consideration about SMEs' default rate and net assets in loans, this paper sets λ_f and λ_n to measure commercial banks' sensitivity to SMEs' default rate and net assets risk preference, respectively. When the λ_f is high, it indicates that commercial Banks pay more attention to the default rate of SMEs in loans. When the parameter λ_n is high, it indicates that commercial banks pay more attention to the net assets of SMEs in lending.

3.1. Debt Contract between Enterprise and Commercial Banks.

Due to information asymmetry, banks do not fully grasp the heterogeneous and uncertain impacts, production conditions, and investment returns faced by SMEs, so it is costly to identify these situations. Suppose that the bank only knows the distribution of ω_t^{xw} in advance; the actual value of the audit ω_t^{xw} needs to pay the monitoring cost equivalent to the

μ percent of the total revenue for the enterprise, that is, $\mu\omega_t^{xw}[w_t(1 + \psi_n^{xw}r_t^{l,xw})n_t^{xw}(s) + r_t^k(1 + \psi_k^{xw}r_t^{l,xw})k_{t-1}^{xw}(s)]$. The arrangements for debt contracts signed by banks and entrepreneurs are as follows: the nondefault interest rate for entrepreneur loans l_t^{xw} is $r_t^{l,xw}$, it depends on the critical value $\bar{\omega}_t^{xw}$ of the random variable ω_t^{xw}, namely,

$$\bar{\omega}_t^{xw}(1 + R_t^{k,xw})[w_t(1 + \psi_n^{xw}r_t^{l,xw})n_t^{xw} + r_t^k(1 + \psi_k^{xw}r_t^{l,xw})k_{t-1}^{xw}]$$
$$= (1 + r_t^{l,xw})l_t^{xw}. \tag{26}$$

When $\omega_t^{xw} \geq \bar{\omega}_t^{xw}$, SMEs repay principal and interest of commercial banks $(1 + r_t^{l,xw})l_t^{xw}$, and earn income $\omega_t^{xw}(1 +$

$R_t^{k,xw})[w_t(1 + \psi_n^{xw}r_t^{l,xw})n_t^{xw} + r_t^k(1 + \psi_k^{xw}r_t^{l,xw})k_{t-1}^{xw}] - (1 + r_t^{l,xw})l_t^{xw}$. When $\omega_t^{xw} \leq \bar{\omega}_t^{xw}$, SMEs cannot repay the principal and interest, and their income is zero. At this time, the commercial bank reserves the right to dispose of the manufacturers' property. In this article, after the enterprise fails, the business implies a certain percentage of the property when the business implies the disposal of the enterprises property. Meanwhile, that is, the actual available property of the bank's enterprises is $(1 - \mu)\,\omega_t^{xw}\sigma_a^{xw}(1 + R_t^{k,xw})[w_t(1 + \psi_n^{xw}r_t^{l,}\,xw)n_t^{xw} + r_t^k(1 + \psi_k^{xw}r_t^{l,xw})k_{t-1}^{xw}]$, where σ_a^{xw} measure the degree of property devaluation after a business failure. Then, the returns of commercial banks are

$$[1 - F(\bar{\omega}_t^{xw})](1 + r_t^{l,xw})l_t^{xw}(s) + (1 - \mu)\sigma_a^{xw}\int_{\bar{\omega}_t^{xw}}^{\infty}\omega_t^{xw}dF(\bar{\omega}_t^{xw})\omega_t^{xw}(1 + R_t^{k,xw})$$
$$[w_t(1 + \psi_n^{xw}r_t^{l,xw})n_t^{xw}(s) + r_t^k(1 + \psi_k^{xw}r_t^{l,xw})k_{t-1}^{xw}(s)]. \tag{27}$$

Different from the assumption of SMEs, large companies do not have credit defaults. When loans are due, the principal and interest of commercial banks can be guaranteed. Therefore, commercial banks can obtain from large enterprise loans as $(1 + r_t^{l,dz})l_t^{dz}(s)$.

Unlike Bernanke et al. [15] and other scholars who set up complete competition for commercial banks, this article is based on the character of China's bank-led financial system, in which commercial banks have a high degree of monopoly and draw on the design of Qiu and Zhou [30]. In the model, suppose that commercial banks have the ability to monopolize bargaining power in the process of SMEs' crediting,

and therefore $\bar{\omega}_t^{xw}$ are endogenous, depending on the commercial banks' decision to maximize profits. Therefore, suppose that the commercial banks are in a monopolistic competitive market. Then, the commercial Banks in the model are in the monopolistic competition market, and the aggregate loan demand each type of bank faces is $l_t^b = |\int_0^1 l_t^b(s)^{(\vartheta - 1/\vartheta)}ds|^{(\vartheta/\vartheta - 1)}$, where $\vartheta(\vartheta > 1)$ is the elasticity of substitution between commercial bank loans. Let loan interest rate index be $r_t^l = |\int_0^1 (r_t^l(s))^{1-\vartheta}ds|^{(1/1-\vartheta)}$, then the loan demand of commercial bank is $l_t^b = |(r_t^l(s)/r_t^l)|^{-\vartheta}l_t^b$.

So the decision problem of the s_{th} commercial bank is

$$\max_{r_t^{l,xw},\bar{\omega}_t^{xw},r_t^{l,dz}}[1 - F(\bar{\omega}_t^{xw})](1 + r_t^{l,xw})l_t^{xw}(s) + (1 - \mu)\sigma_a^{xw}\int_{\bar{\omega}_t^{xw}}^{\infty}\omega_t^{xw}dF(\bar{\omega}_t^{xw})\omega_t^{xw}(1 + R_t^{k,xw})$$

$$[w_t(1 + \psi_n^{xw}r_t^{l,xw})n_t^{xw}(s) + r_t^k(1 + \psi_k^{xw}r_t^{l,xw})k_{t-1}^{xw}(s)] + (1 + r_t^{l,dz})l_t^{dz}(s) - (1 + r_t^d)d_t^{xw}, \tag{28}$$

$$\text{s.t. } l_t^{b,j}(s) = \left[\frac{r_t^{l,j}(s)}{r_t^{l,j}}\right]^{-\vartheta}l_t^{b,j},$$

where j gets xw and dz, representing two types of SMEs and large enterprises, respectively.

Commercial banks choose the loan interest rates of large enterprises and SMEs, and SME mortgage rates to maximize their profits, so that they can get

$$1 + r_t^{l,xw} = \frac{\vartheta_t^{xw}(1 + r_t^d)}{(\vartheta_t^{xw} - 1)[1 - F(\bar{\omega}_t^{xw})]}, \tag{29}$$

$$\frac{[1 - F(\bar{\omega}_t^{xw})]}{\vartheta_t^{xw}} - [1 - \sigma_a^{xw}(1 - \mu)]\omega_t^{xw}F'(\bar{\omega}_t^{xw}) = 0, \tag{30}$$

$$1 + r_t^{l,dz} = \frac{\vartheta_t^{xw}}{(\vartheta_t^{xw} - 1)}(1 + r_t^d). \tag{31}$$

Equation (30) represents loan contracts between SMEs and commercial banks, which is different from Bernanke et al. [15] and Christiano et al. [16]. From (30), we find that ω_t^{xw} depends on the endogenous variable loan interest rate elasticity ϑ_t^{xw}, the parameter σ_a^{xw} of the depreciation for assets after the default of SMEs, and the commercial banks' size of the monitoring cost parameter μ for corporate loans. From the first-order conditions of the SME loan interest rate, when commercial banks provide loans to SMEs, they not only focus on the net assets of SMEs but also care for the

TABLE 1: Parameter calibration.

Parameters	Description	Values
α^{xw}	SMEs' capital output elasticity	0.45
α^{dz}	Elasticity of capital output of large enterprises	0.6
β	Discount factor	0.9936
δ	Depreciation rate	0.025
σ	Reciprocal household consumption elasticity	2
φ	Reciprocal of labor supply elasticity	1
θ	Price stickiness	0.8
η	SMEs' intermediate products account for the proportion of final products	0.28
ε^{xw}	SME middleware substitution elasticity	3
ε^{dz}	Intermediate substitution elasticity of large enterprises	1
μ	Banks' proportion of credit monitoring costs	0.12
ϑ^{xw}	SME credit interest rate flexibility	132.7
ϑ^{dz}	Large enterprise credit interest rate flexibility	137.5
ρ_m	Monetary policy smoothing factor	0.38
ρ_z	Autoregressive coefficient	0.95

elasticity of the SME loan interest rate and the critical value of the external risk. This article describes the risk preference of commercial banks for SMEs' default rate and net asset status through the following formula:

$$\vartheta_t^{xw} = \vartheta^{xw} \exp\left[\lambda_n\left(\frac{N_t^{xw}}{N^{xw}} - 1\right) - \lambda_f\left(\frac{F(\varpi_t^{xw})}{F(\varpi^{xw})} - 1\right)\right], \quad (32)$$

where λ_n reflects the sensitivity between SMEs' own net assets and loan demand elasticity; λ_f reflects the sensitivity between SMEs' default rate and loan demand elasticity; $\lambda_n > 0$ and $\lambda_f > 0$.

3.2. Stability Analysis. Based on the description of each economic sector in the model, the behavior of each major economic entity is analyzed from the perspective of optimization. The decision equation of the loan contract between SMEs and commercial banks is

$$\frac{[1 - F(\varpi_t^{xw})]}{\vartheta_t^{xw}} - [1 - \sigma_a^{xw}(1 - \mu)]\omega_t^{xw} F\prime(\varpi_t^{xw}) = 0. \quad (33)$$

From the calculation of MATLAB, it can be seen that the mortgage rate of commercial banks to SMEs is negatively correlated with the elasticity of commercial banks' loan interest rates to SMEs. That is $(\partial\varpi_t^{xw}/\partial\vartheta_t^{xw}) < 0$, as the elasticity of bank loan interest rates ϑ_t^{xw} increases, the mortgage rate ϖ_t^{xw} of SMEs' loans will become smaller. In addition, the elasticity of loan interest rates for commercial banks to SMEs is positively related to the depreciation of corporate net assets when SMEs fail. That is $(\partial\vartheta_t^{xw}/\partial\sigma_a^{xw}) > 0$ as the elasticity of bank loan interest σ_a^{xw} increases, the mortgage rate of SMEs' loans will become bigger.

In addition, the decision equation of commercial banks' loan interest rates for SMEs is:

$$1 + r_t^{l,xw} = \frac{\vartheta_t^{xw}(1 + r_t^d)}{(\vartheta_t^{xw} - 1)[1 - F(\varpi_t^{xw})]}. \quad (34)$$

From (34), the interest rate of commercial bank loans to SMEs is also negatively related to the elasticity of commercial

bank loan rates to SMEs, that is $(\partial r_t^{l,xw}/\partial\vartheta_t^{xw}) > 0$, as the elasticity of bank loan interest rates increases, the interest rate of commercial banks' loans to SMEs decreases.

In addition, this paper introduces the mechanism of the impact of commercial banks' default on SMEs' net assets and loans, that is, by changing the elasticity of loan demand, the loan interest rate, and mortgage interest rate of commercial banks will be affected.

The risk preference equation of a commercial bank is

$$\vartheta_t^{xw} = \vartheta^{xw} \exp\left[\lambda_n\left(\frac{N_t^{xw}}{xw}\right) - \lambda_f\left(\frac{F(\varpi_t^{xw})}{F(\varpi^{xw})} - 1\right)\right]. \quad (35)$$

From (35), the correlation between the elasticity of loan interest rates for commercial banks to SMEs and the net assets of enterprises depends on the parameters λ_n. When $\lambda_n > 0$, $(\partial\vartheta_t^{xw}/\partial N_t^{xw}) > 0$, the larger the net assets of the enterprise, the greater the elasticity of the loan interest rate of commercial banks. The correlation between the elasticity of loan interest rates of commercial banks to SMEs and the default rate of enterprises depends on the parameters λ_f. When $\lambda_f > 0$, $(\partial\vartheta_t^{xw}/\partial F(\varpi_t^{xw})) > 0$, the higher the corporate default rate, the greater the elasticity of commercial bank loan interest rates.

Based on the risk preference of commercial banks for corporate net assets and their aversion to corporate defaults when lending to enterprises, this paper sets $\lambda_n > 0$ and $\lambda_f > 0$. Through the above static comparison analysis among the interest rate determination equation, loan contract and risk preference equation of commercial bank, corollaries are derived as follows:

Corollary 1. *The smaller the tolerance of commercial banks for default of SMEs, the smaller the size of loans available to SMEs.*

Corollary 2. *The greater the sensitivity of commercial banks to the size of SMEs' net assets, the larger the scale of loans available to SMEs.*

TABLE 2: Model of the steady-state analysis of main economic variables at different levels of parameters σ_a^{xw}.

	l^{xw}	lev^{xw}	k^{xw}	ϖ^{xw}	$F(\varpi^{xw})$	$R^{k,xw}$	$\varpi^{xw}v^{xw}$
100	0.4403	0.8975	0.3293	0.3238	0.0100	1.0414	1.3925
99	0.4344	0.8963	0.3208	0.3196	0.0092	1.0272	1.3910
98	0.4291	0.8952	0.3132	0.3157	0.0085	1.0144	1.3897
97	0.4241	0.8942	0.3062	0.3121	0.0079	1.0026	1.3884
96	0.4195	0.8933	0.2999	0.3089	0.0074	0.9918	1.3873
95	0.4152	0.8924	0.2941	0.3058	0.0070	0.9818	1.3862
94	0.4112	0.8915	0.2887	0.3030	0.0066	0.9725	1.3852
93	0.4075	0.8907	0.2837	0.3003	0.0062	0.9638	1.3842
92	0.4040	0.8899	0.2791	0.2979	0.0059	0.9557	1.3833
91	0.4007	0.8892	0.2748	0.2955	0.0056	0.9480	1.3824
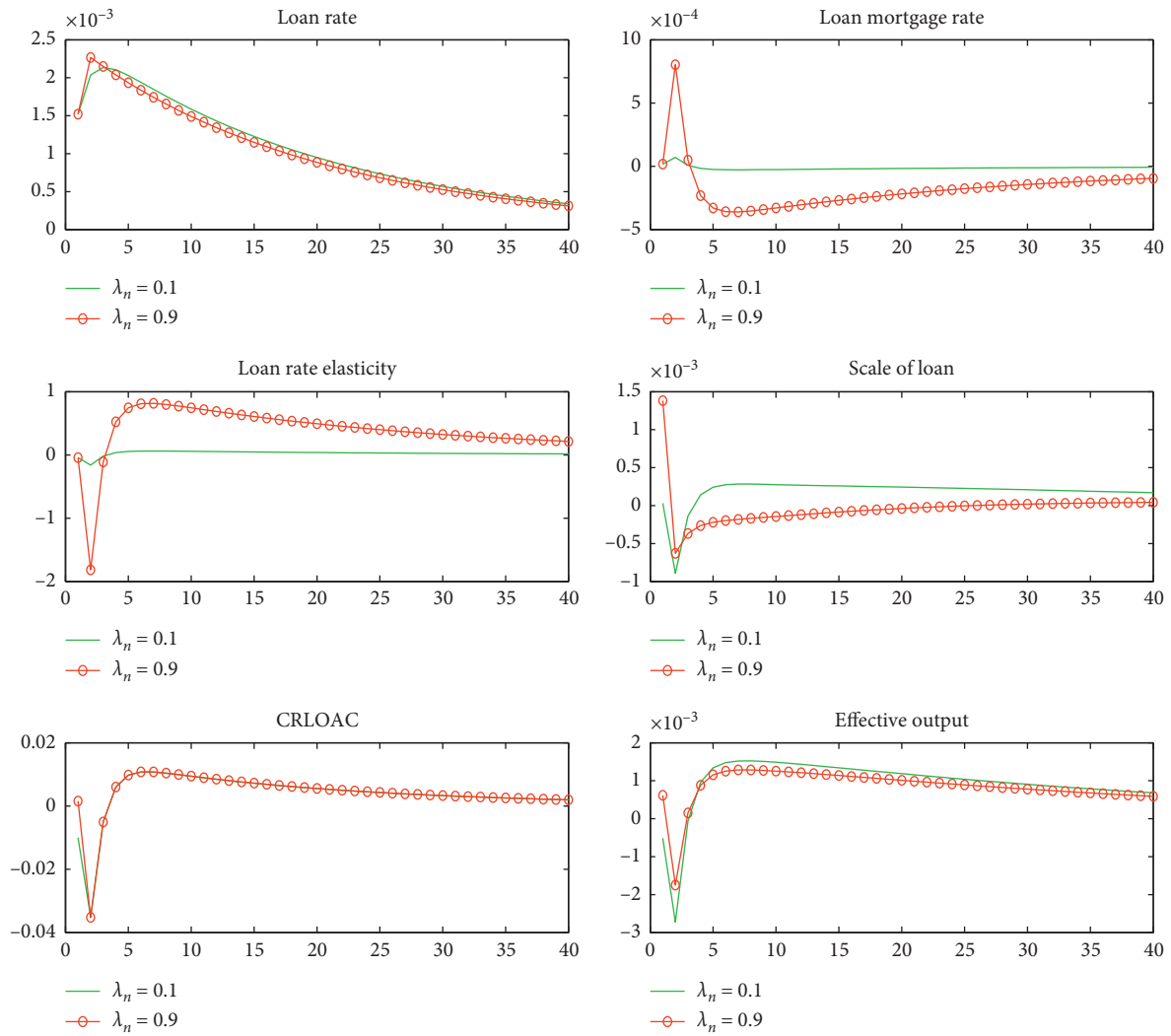
FIGURE 2: Simulation results of commercial banks' preference mechanism for the net asset value of SMEs.

Corollary 3. *The higher the quality of SMEs' net assets, the larger the scale of SMEs' bank loans.*

4. Numerical Simulation Analysis

In this part, this article performs numerical simulation calculation on the above DSGE model. In order to examine the credit allocation of SMEs under different risk preferences of commercial banks, we use the calibration method to obtain model-related parameters and then use the model steady-state equation and numerical simulation to analyze the changes in the dynamic adjustment process when the model steady-state and endogenous variables face the impact of exogenous monetary policy under different parameter settings.
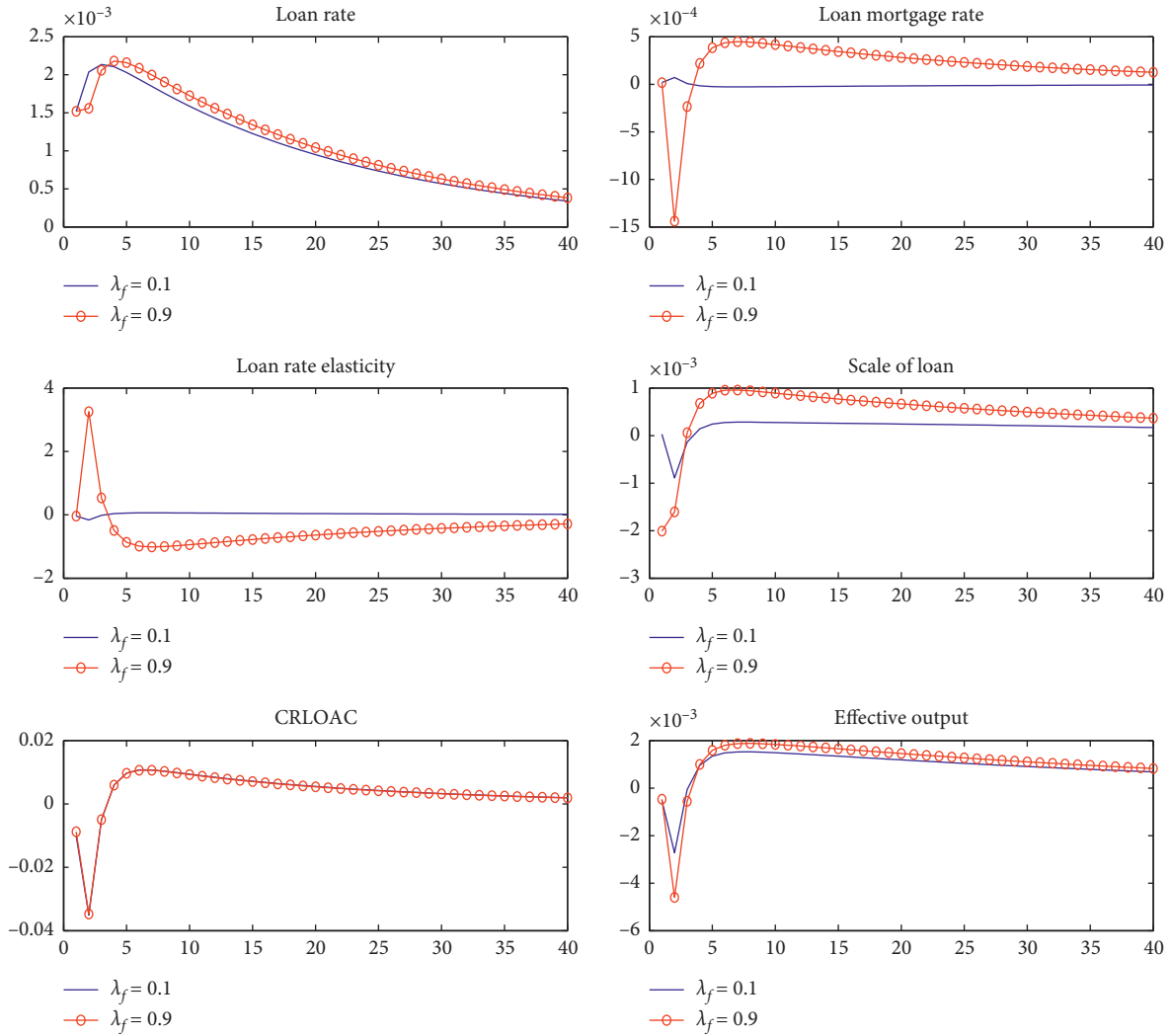
FIGURE 3: Simulation results of commercial banks' preference mechanism for the default risk of SMEs.

4.1. Model Parameter Calibration. This paper uses the macro- and microeconomic data of the National Bureau of Statistics and the steady-state equation of the model to obtain the model parameter values. The model calibration process mainly uses the average value of macro- and microeconomic data that can be observed in reality to calculate the model parameter value through the steady-state equation of the nonlinear model. The sample period of macro and micro is 1998Q1-2018Q4. The economic variables involved in this paper, such as real GDP, real m2, CPI, and so on, and the data of SME loans and large-scale enterprise loans are all from Wind database.

Table 1 shows the specific information on parameter calibration. From the average value of the one-year deposit interest rate of 2.57%, the calibration r^d will be 1.0064, and the calibration result of the subjective discount factor β for residents is 0.9936. Referenced Peng and Fang [32], the reciprocal elasticity of household consumption substitution σ, labor supply elasticity φ, and the price stickiness parameter are set 2 , 1, and 0.8, respectively. According to estimates by Lin and Yang [33], the elasticity of capital output for large enterprises and SMEs are 0.6 and 0.45, respectively. The cost μ

of bank monitoring for SMEs' loans is 0.12, which is commonly used in literature. The common reference value of capital depreciation rate is 0.025. The substitution elasticity of intermediate goods for large enterprises and SMEs are set to 1 and 3. The large enterprise loan interest rate adopts the 1-year loan interest rate announced by the Statistics Bureau. According to the 2019 one-year loan interest rate issued by the People's Bank of China was 6.07%. The steady-state values of corporate loan interest rate elasticity ϑ^{dz} are 137.5 and 132.7, respectively. According to estimates by Qiu and Zhou [30], the smoothing factor ρ_m for monetary policy is 0.38. The autoregressive coefficient ρ_z is taken as 0.95. Calculated from the SME loan and large enterprise loan data in the Wind database, the combined proportion of the final product synthesis of large enterprise intermediate products and SME intermediate products is 0.28 and 0.72, respectively.

4.2. Numerical Simulation and Impulse Response. SMEs obtain loans from commercial banks through mortgaged net assets. When external risk shocks exceed the capacity of SMEs, they close down. At this time, commercial banks have
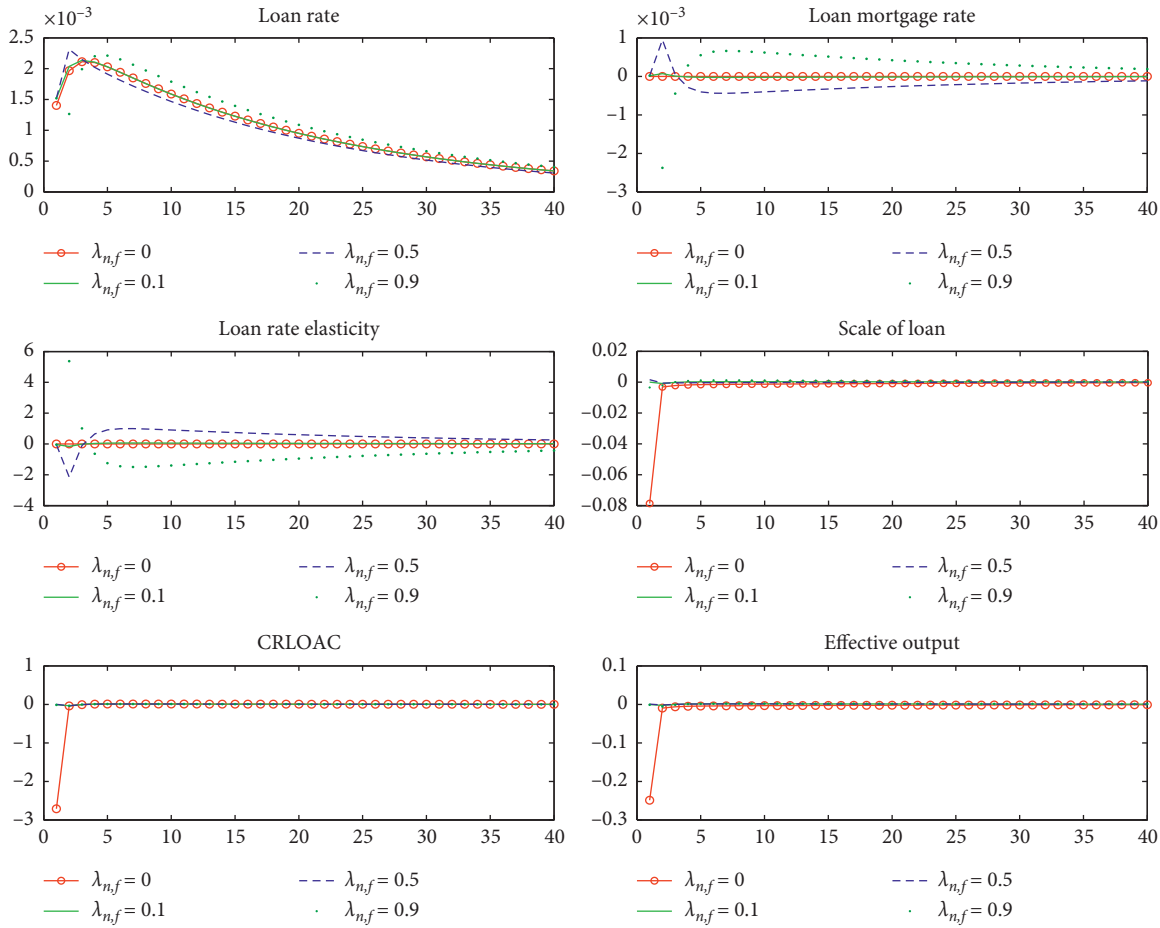
FIGURE 4: Simulation results of commercial banks' risk neutrality and different overall comparison.

the right to dispose for SMEs, thereby making up for credit losses. The reality is that when SMEs go bankrupt, when commercial banks auction the remaining net assets of SMEs, the remaining assets of SMEs after credit default often have a gap with the net assets pledged before the credit default. The larger the gap, the remaining will be the net assets after SMEs default. The stronger the depreciation of assets, the greater the credit losses of commercial banks. This article uses σ_a^{xw} to indicate the depreciation of the remaining net assets for SMEs after default. The smaller the σ_a^{xw}, the more severe the depreciation. From the steady-state loan contract $\{1 - ([1 - F(\varpi_t^{xw})]/\vartheta_t^{xw})/[\varpi_t^{xw} F'(\varpi_t^{xw})]\}/(1 - \mu) = \sigma_a^{xw}$, when σ_a^{xw} decrease, ϖ_t^{xw} increases. The fewer loans a unit of net assets can use for mortgages, the external credit financing leverage ratio of SMEs decreases, and the effective output decreases. SMEs have difficulty in financing, but the high default rate of SMEs is also an important factor in their financing difficulties. Table 2 reports the steady-state results of the major economic variables at different levels of parameters σ_a^{xw}. The results show that as σ_a^{xw} decrease, the steady-state values of the major economic variables all show a downward trend.

4.3. Influence of Different Parameter Settings on Impulse Response Function of Endogenous Variables. Under the

framework of tight monetary policy and the heterogeneity of enterprises, this article examines the impact of different risk preference of banks on SMEs' credit financing. The main conclusion is: with the increase of interest rate by central bank, the credit financing of SMEs is suppressed. And, different risk preferences of commercial banks lead to different credit financing conditions of SMEs.

Figure 2 analyzes the impact of the different weights for commercial banks' risk preference on SMEs' credit financing and effective output under the random impact of a positive interest rate of 10%. In the tight monetary policy cycle, the credit supply of commercial banks is tightened, corporate production activities are reduced, output is reduced, and corporate default risks are increased. Meanwhile, banks increase the inspection weight of corporate mortgage assets. From the risk preference equation of commercial banks, we get that when the assets of SMEs deviate from the steady-state value downward, the elasticity ϑ_t^{xw} of commercial banks' credit interest rates on SMEs decrease. From the perspective of credit interest rate channels: when the ϑ_t^{xw} decreases, the credit interest rate of SMEs increase and the cost of credit rise. From the perspective of risk-bearing channels: as the ϑ_t^{xw} decreases, the mortgage rate ω_t^{xw} increases. The high mortgage rate
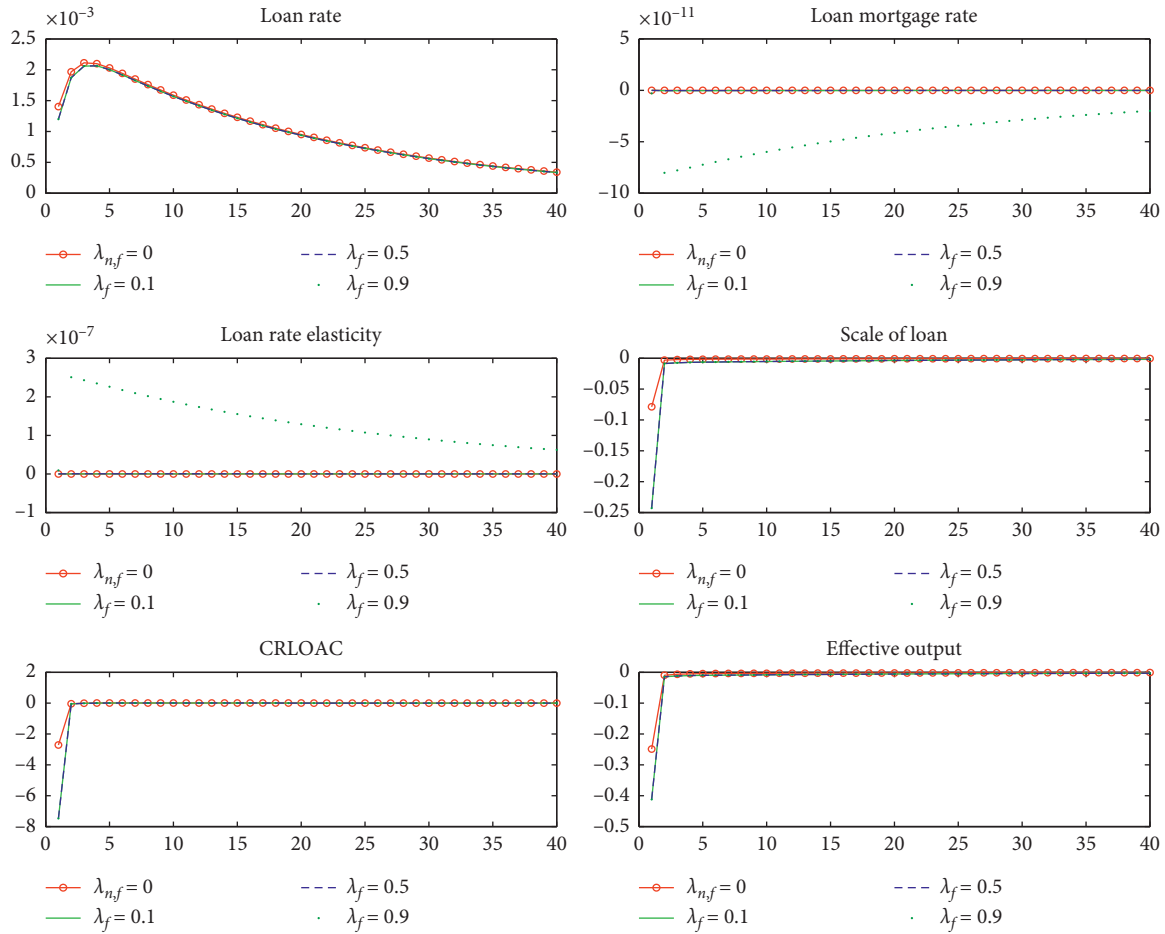
FIGURE 5: Simulation results of different comparison of commercial banks' risk preference based on the perspective of net asset value.

indicates that the enterprise's sample assets can obtain more loans from the bank, and the cost of borrowing is reduced. From the pulse effect results of Figure 2, with central bank interest rate hikes, the rising cost of borrowing due to rising interest rates exceeds the falling cost of borrowing due to rising mortgage rates, and SME credit financing is suppressed, and the effective output declines. However, as the weight of commercial banks' risk preference for SMEs' net assets increased, the effect of lowering the cost of borrowing from the risk-taking channel exceeds the effect of the rising cost of borrowing from the interest rate channel. The downward deviation from the steady-state trend is suppressed.

Figure 3 analyzes the SMEs' credit financing and effective output impact of the different weights of commercial banks on the default risk of SMEs under a positive interest rate impact. The high default rate of SMEs on bank credit has become an important reason restricting their credit. In the process of central bank interest rate hikes, the scale of bank credit has decreased, and the default rate has decreased. From the risk preference equation of commercial banks, it can be seen that as the default rate decreases, the ϑ_t^{xw} increases. From the perspective of interest rate channels: since the loan interest rate of SMEs is negatively correlated with the elasticity of credit interest

rates, when the ϑ_t^{xw} increases, the loan interest rate decreases, and the cost of obtaining credit for SMEs decreases. From the perspective of risk-bearing channels: an increase in credit interest rates ϑ_t^{xw} leads to a reduction in the mortgage rate and an increase in the borrowing costs of SMEs. The impulse response of Figure 3 shows that in the tight monetary policy cycle, as commercial banks attach importance to SME defaults, the increase in parameters λ_f has led to further increase in SME credit financing constraints, and effective output has further deviated from the steady-state trend.

For the sake of robustness, this paper tests the effectiveness of the risk preference mechanism of commercial banks from two aspects, and compares the simulation results. Specifically, it is obtained by assuming that the commercial banks are neutral ($\lambda_{n,f} = 0$) to the net asset value and default risk preference of SMEs and that the risk preference of commercial banks is different ($\lambda_{n,f} = 0.1$, 0.5, 0.9). Figure 4 shows the simulation results when the parameters of asset value and default risk preference of commercial banks for SMEs change at the same time. Compared with that when commercial banks are neutral to the net asset value and default risk preference of SMEs, when commercial banks have risk preference, economic variables such as loan rate, loan mortgage rate, and loan rate elasticity
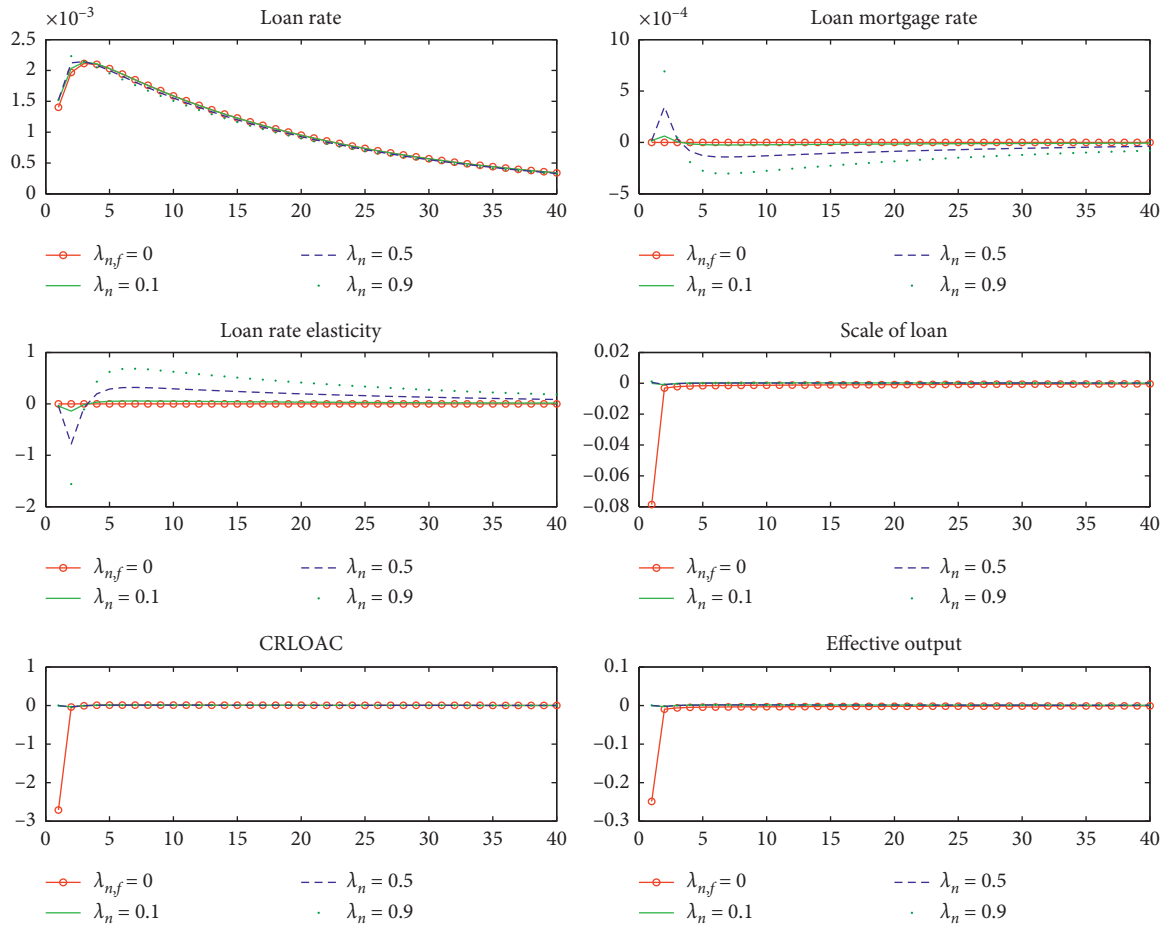
Figure 6: Simulation results of different comparison of risk preference of commercial banks based on default perspective.

will fluctuate in the short term. When $\lambda_{n,f} = 0.5$ and $\lambda_{n,f} = 0.9$, the simulation results are contrary, which may be caused by the substitution effect of monetary policy interest rate transmission channel and bank risk-taking transmission channel. However, it does not affect the effectiveness of the risk preference mechanism of commercial banks. Simulation results of different comparisons of commercial banks' risk preference based on the perspective of net asset value as shown in Figure 5. Simulation results of different comparison of risk preference of commercial banks based on default perspective as shown in Figure 6. The results show that, when commercial banks have risk preference, the economic variables of SMEs will fluctuate significantly in the short term, and the direction is opposite, which is consistent with the simulation results. This shows the effectiveness of the risk preference mechanism of commercial banks. Thus, the risk preference mechanism of commercial banks constructed in this paper can explain the credit financing problems of SMEs.

For the sake of robustness, the cases for $\lambda_n = 0$ and $\lambda_f = 0$ are investigated. When both λ_n and λ_f are zero, commercial banks only consider the effect of the interest rate channel. The results show that there is no significant relationship between the dynamic process of other variables and the risk taking channel, except the mortgage rate. With the

introduction of the risk-bearing channel, the mortgage rate will take on different states with the different values of λ_n and λ_f, and the commercial banks have different preferences for corporate net assets and default risk, thus affecting the financing constraints of SMEs.

5. Conclusion and Implication

By constructing a DSGE model with commercial banks' credit risk preferences, this paper examines the impact of different risk preferences of commercial banks on SME credit financing, and draws the following main conclusions: Firstly, from the perspective of high default rate of SMEs, this paper analyzes the depreciation gap of mortgage assets before and after default. The comparative static analysis of the theoretical model shows that with the increase of the devaluation of mortgage assets, the critical value of risk-taking of SMEs is reduced, and the borrowing cost is increased, so the loan scale and leverage ratio of SMEs are reduced, and the effective output is reduced. Secondly, this paper conducts numerical simulation of the theoretical model under the framework of the central bank's interest rate hike and the assumption that the potential output of SMEs can be used as collateral. It is found that with the increase of the weight of commercial banks' risk preference

for corporate assets, the loan interest rate elasticity of commercial banks to SMEs decreases, and the reduction of the loan interest rate elasticity will lead to the increase of mortgage rate, the reduction of borrowing costs, the increase of loan scale and leverage ratio, and the increase of the effective output of SMEs. Thirdly, the study found that under the tightening monetary policy background, the increase in the weight of commercial banks' default preference for SMEs leads to an increase in the elasticity of loan interest rates, a reduction in mortgage rates, a rise in borrowing costs, and reduced loan scale and effective output of SMEs. Therefore, reducing the information asymmetry between SMEs and commercial banks, improving the quality of net assets of enterprises, or expanding the scope of corporate collateral can increase the availability of credit financing for SMEs.

The research has the following implications for solving the problem of financing difficulty and high cost for SMEs: on the one hand, as the direct supplier of external financing for enterprises, financial intermediaries have different perceptions of asset quality and loan risk, which will directly affect their lending behaviors. Therefore, when the central bank has the policy tendency of directional lending to SMEs, commercial banks should moderately increase the tolerance of SMEs' credit risk. On the other hand, financial innovation on loan collateral for SMEs should be carried out, thereby improving the grade of loan collateral or expanding the scope of loan collateral should be paid attention. However, it can only be used as an unconventional means of lending in a special period, which will increase the loan risk of commercial banks in the long run.

This study mainly investigates the transmission and implementation effect of commercial banks' monetary policy on SME financing. In the future research, direct financing channels in capital market can be brought into the framework. Meanwhile, this study focuses on the theoretical research at the macro level, and further empirical studies can be conducted at the micro level to investigate the impact of commercial banks' risk preference on SMEs' financing.

Appendix

A. An Overview of the Model

In the process of constructing the dynamic stochastic general equilibrium (DSGE) model, based on the rational man hypothesis, this paper models the economic behaviors of household sector, enterprise sector, financial intermediary sector, and central bank sector in the economy. At the same time, the household sector should be satisfied with the maximization of lifetime utility, the enterprise sector should be satisfied with the maximization of corporate profits, and the corporate default rate should be considered under the profit-maximization framework when the financial sector signs debt contracts with enterprises. In addition, the monetary policy of the central bank follows the Taylor rule. After the above economic models are given, in order to obtain the optimal behavior of each economic sector, the first-order derivative of each economic sector behavior model is carried out. The following are the first-order

conditions (FOC) of the main equations involved in each economic sector in the model, specifically:

A.1. The Household Sector

Combined with the household utility function model (1) and model (2), household utility constraint model (3), and capital accumulation equation model (4), the first-order conditions of consumption c_t, labor n_t, deposit d_t, and capital k_t can be obtained as follows:

(1) FOC on consumption

$$\frac{(c_t - bc_{t-1})^{-\sigma}}{(1-b)^{-\sigma}} - \beta b \frac{(c_{t+1} - bc_t)^{-\sigma}}{(1-b)^{-\sigma}} = \lambda_t; \qquad \text{(A.1)}$$

(2) Wage pricing equation

$$\lambda_t w_t = n_t^{\varphi}; \qquad \text{(A.2)}$$

(3) First-order terms for deposits

$$\left(\beta \frac{\lambda_{t+1}}{\pi_{t+1}} r_t^d\right) = \lambda_t; \qquad \text{(A.3)}$$

(4) FOC for investment

$$\beta \lambda_{t+1}\left[(1-\delta) + r_{t+1}^k\right] = \lambda_t. \qquad \text{(A.4)}$$

A.2. The Enterprise Sector

Based on the intermediate product production function, model (6), cost constraint condition model (7) or (12), and under the minimum cost condition model (9), first-order conditions of input factors and actual marginal cost can be obtained by differentiating n_t and k_t. According to the optimal price determining behavior of enterprises, with the optimal price determining equation (16) of enterprises, by taking the first derivative of the optimal price, the new Keynesian Phillips curve can be obtained. In addition, with the final manufacturer's production function model (21) being given, the first-order conditions of enterprise input factors and the total price determination equation can be obtained. The details are as follows:

(1) Capital factor input of SMEs

$$\left(1 + \Psi_k^{xw} r_t^{l,xw}\right) r_t^k k_{t-1}^{xw} = a m c_t^{xw} y_t^{xw}; \qquad \text{(A.5)}$$

(2) Labor factor input of SMEs

$$\left(1 + \Psi_n^{xw} r_t^{l,xw}\right) w_t n_t^{xw} = (1-a) m c_t^{xw} y_t^{xw}; \qquad \text{(A.6)}$$

(3) Actual marginal cost of SMEs

$$mc_t^{xw} = \left[r_t^k\left(1 + \Psi_k^{xw} r_t^{l,xw}\right)\right]^a \left[w_t\left(1 + \Psi_n^{xw} r_t^{l,xw}\right)\right]^{1-a} a^{-a}$$
$$\cdot (1-a)^{-(1-a)} \left(A_t^{xw}\right)^{-1}; \qquad \text{(A.7)}$$

(4) Return on production of SMEs

$$1 + R_t^{k,xw} = \frac{\omega_t^{xw} y_t^{xw}}{w_t n_t^{xw} + r_t^k k_{t-1}^{xw}}; \qquad (A.8)$$

(5) Capital factor input of large enterprises

$$\left(1 + \Psi_k^d r_t^{l,d}\right) r_t^k k_t^{dz} = \text{amc}_t^{dz} y_t^{dz}; \qquad (A.9)$$

(6) Labor factor input in large enterprises

$$\left(1 + \Psi_k^d r_t^{l,d}\right) r_t^k k_t^{dz} = \text{amc}_t^{dz} y_t^{dz}; \qquad (A.10)$$

(7) Actual marginal cost of large enterprises

$$mc_t^{dz} = \left[r_t^k\left(1 + \Psi_k^{dz} r_t^{l,dz}\right)\right]^a \left[w_t\left(1 + \Psi_n^{dz} r_t^{l,dz}\right)\right]^{1-a} a^{-a}$$
$$\cdot (1-a)^{-(1-a)} \left(A_t^{dz}\right)^{-1}; \qquad (A.11)$$

(8) Return on production of large enterprises

$$1 + R_t^{k,dz} = \frac{y_t^{dz}}{w_t n_t^{dz} + r_t^k k_{t-1}^{dz}}; \qquad (A.12)$$

(9) The New Keynes Phillips curve

$$\widehat{\pi}_t^j = \beta E_t \widehat{\pi}_{t+1}^j + \frac{(1-\theta)(1-\beta\theta)}{\theta^j}\left(\widehat{mc}_t^j - p_t^j\right);$$

$$\widehat{\pi}_t^j - \widehat{\pi}_t = p_t^j - p_{t-1}^j; \qquad (A.13)$$

(10) The equation of total price

$$1 = \left[\eta\left(P_t^{xw}\right)^{1+v_y} + (1-\eta)\left(P_t^{dz}\right)^{1+v_y}\right]^{\left(1/1+v_y\right)}; \qquad (A.14)$$

(11) The input of SMEs in final product synthesis

$$Y_t^{xw}(s) = \eta\left(P_t^{xw}\right)^{v_y} Y_t; \qquad (A.15)$$

(12) The input of large enterprises in final product synthesis

$$Y_t^{dz}(s) = (1-\eta)\left(P_t^{dz}\right)^{v_y} Y_t. \qquad (A.16)$$

A.3. The Debt Covenants between Enterprises and Commercial Banks

Under the condition of considering the risk of corporate default, with the behavior equation of commercial Banks in the paper (28) being given, combined with the constraint condition (27), mortgage rates in the commercial bank ω_t^{xw} and loan interest rates $r_t^{l,j}$ can be firstly differentiated, and the following FOC can be obtained:

(1) Marginal default conditions for SMEs

$$\omega_t^{xw}\left(1 + R_t^{k,xw}\right)\left[w_t\left(1 + \Psi_n^{xw} r_t^{l,xw}\right)n_t^{xw} + r_t^k\left(1 + \Psi_k^{xw} r_t^{l,xw}\right)k_{t-1}^{xw}\right] = \left(1 + r_t^{l,xw}\right)l_t^{xw}; \qquad (A.17)$$

(2) The condition of loans of large enterprises

$$\left(1 + R_t^{k,dz}\right)\left[w_t n_t^{dz}(s) + r_t^k k_{t-1}^{dz}(s)\right] - \left(1 + r_t^{l,dz}\right)l_t^{dz}$$
$$= (1 + \lambda_{dz})\left(1 + r_t^d\right)N_t^{dz}; \qquad (A.18)$$

(3) The equation of loan interest of SMEs

$$1 + r_t^{l,xw} = \frac{\vartheta_t^{xw}\left(1 + r_t^d\right)}{\left(\vartheta_t^{xw} - 1\right)\left[1 - F\left(\omega_t^{xw}\right)\right]}; \qquad (A.19)$$

(4) The contracts of loan of SMEs

$$\frac{\left[1 - F\left(\omega_t^{xw}\right)\right]}{\vartheta_t^{xw}} - \left[1 - \sigma_a^{xw}(1-\mu)\right]\omega_t^{xw} F\prime\left(\omega_t^{xw}\right); \qquad (A.20)$$

(5) The equation of loan interest of large enterprises

$$1 + r_t^{l,dz} = \frac{\vartheta_t^{dz}}{\left(\vartheta_t^{dz} - 1\right)}\left(1 + r_t^d\right); \qquad (A.21)$$

(6) The equation of loan interest rate elastic of SMEs

$$\vartheta_t^{xw} = \vartheta^{xw} \exp\left[\lambda_n\left(\frac{N_t^{xw}}{N^{xw}} - 1\right) - \lambda_f\left(\frac{F\left(\omega_t^{xw}\right)}{F\left(\omega^{xw}\right)} - 1\right)\right]; \qquad (A.22)$$

(7) Central bank monetary policy

$$r_t^d = \left(r_{t-1}^d\right)^{\rho_m}\left[\overline{r}^d\left(\frac{\pi_{t+1}}{\overline{\pi}}\right)^{\xi_\pi}\left(\frac{y_t}{\overline{y}}\right)^{\xi_y}\right]^{\left(1-\rho_m\right)} u_t. \qquad (A.23)$$

A.4. Other Conditions

Based on the equilibrium and clearing of labor market, capital market, and financial market in the paper, the following equations can be obtained:

(1) Total resource constraint

$$y_t = c_t + i_t; \tag{A.24}$$

(2) Capital accumulation equation

$$k_t = (1 - \delta)k_{t-1} + i_t; \tag{A.25}$$

(3) Capital market clearing

$$k_t^{xw} + k_t^{dz} = k_t; \tag{A.26}$$

(4) Labor market clearing

$$n_t^{xw} + n_t^{dz} = n_t, \tag{A.27}$$

(5) Total credit scale

$$l_t^{xw} + l_t^{dz} = l_t^{b}. \tag{A.28}$$

Given the first-order conditions in the above paper, the model parameters can be calibrated by combining the existing relevant literature and the economic data over the years in China. After the model parameters are obtained, Matlab and Dynare software can be directly used to carry out editing operations so as to obtain the corresponding simulation results in the paper. Finally, the main data and programming code are given.

Data Availability

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.

Acknowledgments

References

[1] C. Goodhart, "The boundary problem in financial regulation," *National Institute Economic Review*, vol. 206, no. 1, pp. 48–55, 2008.

[2] D. L. Kohn, "Monetary policy and asset prices revisited," *The American Economic Review*, vol. 29, no. 31, 2009.

[3] J. E. Stiglitz and A. Weiss, "Credit rationing in markets with imperfect information," *The American Economic Review*, vol. 71, no. 3, pp. 393–410, 1981.

[4] V. Vermoesen, M. Deloof, and E. Laveren, "Long-term debt maturity and financing constraints of SMEs during the global financial crisis," *International Small Business Economics*, vol. 41, no. 2, pp. 433–448, 2013.

[5] E. Casey and C. M. O'Toole, "Bank lending constraints, trade credit and alternative financing during the financial crisis: evidence from European SMEs," *Journal of Corporate Finance*, vol. 27, pp. 173–193, 2014.

[6] L. Farinha and S. Félix, "Credit rationing for Portuguese SMEs," *Finance Research Letters*, vol. 14, pp. 167–177, 2015.

[7] B. F. Blumberg and W. A. Letterie, "Business starters and credit rationing," *Small Business Economics*, vol. 30, no. 2, pp. 187–200, 2008.

[8] V. Ioannidou, S. Ongena, and J. Peydr, "Monetary policy, risk-taking and pricing: evidence from a Quasi-natural experiment," *Review of Finance*, vol. 19, no. 1, pp. 95–144, 2014.

[9] D. Bonfim and C. Soares, "The risk-taking channel of monetary policy: exploring all avenues," *Journal of Money, Credit and Banking*, vol. 50, no. 7, pp. 1507–1541, 2018.

[10] S. Pang, "Credit decision model and mechanisms under default risk," *Journal of Management Sciences in china*, vol. 15, no. 4, pp. 58–70, 2012.

[11] Y. Y. Dai, P. D. Zhang, and Y. Pan, "The role of honesty in public firms bank loans:evidence from China," *Journal of Financial Research*, vol. 8, pp. 77–95, 2019.

[12] G. McGuinness and T. Hogan, "Bank credit and trade credit: evidence from SMEs over the financial crisis," *International Small Business Journal*, vol. 34, no. 4, pp. 412–445, 2016.

[13] A. Ferrando, A. Popov, and G. F. Udell, "Sovereign stress and SMEs' access to finance: evidence from the ECB's SAFE survey," *Journal of Banking & Finance*, vol. 81, pp. 65–80, 2017.

[14] L. Wu and L. Xu, "The role of venture capital in SME loans in China," *Research in International Business and Finance*, vol. 51, 2020.

[15] B. Bernanke, M. Gertler, and S. Gilchrist, "The financial accelerator in a quantitative business cycle framework?," NBER Working Paper No. w6455, 1998.

[16] L. J. Christiano, R. Motto, and M. Rostagno, "Financial factors in economic fluctuations," ECB Working Paper Series, European Central Bank (ECB), 2010.

[17] R. Meeks and B. Nelson, "Shadow banks and macroeconomic instability," Bank of Italy Temi di Discussione (Working Paper), 2013.

[18] L. Dedola, P. Karadi, and G. Lombardo, "Global implications of national unconventional policies," *Journal of Monetary Economics*, vol. 60, no. 1, pp. 66–85, 2013.

[19] F. Smets and R. Wouters, "Shocks and frictions in U.S. business cycles: a Bayesian DSGE approach," *American Economic Review*, vol. 93, no. 3, pp. 586–606, 2007.

[20] S. Fukuda, "Bank's capital requirements in the business cycle: a DSGE analysis with a news shock," *Journal of Applied Finance Banking*, vol. 6, no. 3, pp. 1–8, 2016.

[21] B. Jonathan, J. Benchimol, F. André, and A. Fourans, "Central bank losses and monetary policy rules: a DSGE investigation," *International Review of Economics Finance*, vol. 61, pp. 289–303, 2019.

[22] M. Goodfriend and R. King, "The new neoclassical synthesis and the role of monetary policy," Federal Reserve Bank of Richmond Working Paper, 1998.

[23] F. Kydland and E. Prescott, "Time to build and aggregate fluctuations," *Econometrica*, vol. 50, no. 6, pp. 1345–1370, 1982.

[24] E. Prescott, "Theory ahead of business cycle measurement," *Federal Reserve Bank of Minneapolis Quarterly Review*, vol. 10, no. 4, pp. 9–22, 1986.

[25] C. Sims, "Interpreting the macroeconomic time series facts: the effects of monetary policy," *European Economic Review*, vol. 36, no. 5, pp. 975–1000, 1992.

[26] B. Bernanke and A. Blinder, "The federal funds rate and the channels of monetary transmission," *American Economic Review*, vol. 82, no. 4, pp. 901–921, 1992.

[27] B. Bernanke and I. Mihov, "Measuring monetary policy," *Quarterly Journal of Ecnomics*, vol. 113, no. 3, pp. 869–902, 1998.

[28] B. Bernanke, J. Boivin, and P. Eliasz, "Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach," *Quarterly Journal of Ecnomics*, vol. 120, no. 1, pp. 387–422, 2005.

[29] L. Christiano, M. Eichenbaum, and C. Evans, "Nominal rigidities and the dynamic effects of a shock to monetary policy," *Journal of Political Economy*, vol. 113, no. 1, pp. 1–45, 2005.

[30] X. Qiu and Q. L. Zhou, "Shadow banking and monetary policy transition," *Economic Research Journal*, vol. 5, pp. 91–105, 2014.

[31] A. Dib, "Banks, credit market frictions, and business cycles?," Bank of Canada Working Paper, 2010.

[32] Y. C. Peng and Y. Fang, "Structural monetary policy, industrial structure upgrade and economic stability," *Economic Research Journal*, vol. 7, pp. 29–42, 2016.

[33] R. W. Lin and Y. Yang, "Chinas market-oriented reform and the evolution of the effectiveness of the monetary policy: an analysis based on the DSGE model," *Management World*, vol. 6, pp. 39–52, 2014.

Research Article

Nonlinear Effect Analysis of Electricity Price on Household Electricity Consumption

Lianwei Zhang [ID] **and Xiaoni Wen**

School of Economics & Management, Xidian University, Xi'an, China

Correspondence should be addressed to Lianwei Zhang; lwzhang2017@163.com

The household energy consumption has been a hot field in the study of household energy consumption in recent years. With the increase of residents' income level and the pushing of urbanization, there is a complex nonlinear relationship between energy price and energy consumption. The purpose of this paper is to investigate the scenario effect of per capita income and regional differences in urbanization development on the relationship between electricity sales price and urban household electricity consumption. To this direction, based on the regional characteristics of economic development in China, with the residents' disposable income and the urbanization level as the conversion variables and the electricity sales price as the core explanatory variable, the panel smooth transition regression (PSTR) model of electricity sales price and urban household electricity consumption from the perspectives of income level and urbanization has been constructed in this paper. The empirical results show the following: (1) Under the consideration of regional difference of residents' income level, with the increase of residents' disposable income level, there is a significant negative correlation between electricity sales price and urban household electricity consumption in the whole country, the eastern region, and the central region, while such correlation is significantly positive in the western region. (2) Under the consideration of the difference of urbanization development level, the national regional electricity sales price and the urbanization level are positively related to the urban household electricity consumption, and the urbanization level in the western region plays the biggest role in promoting the urban household electricity consumption, followed by the eastern region and then the central region which plays the smallest role. This paper discusses the effect of electricity sales price on urban household electricity consumption from the perspective of regional difference in income and urbanization, which provides the decision-making basis and empirical support for developing regional electricity price policy and household energy consumption policy.

1. Introduction

With the rapid development of urbanization and industrialization, China's economy has entered a new normal stage, and its economic growth has turned into an important historical phase shifting from high-speed to high-quality development. As a strong and basic support of economy development, energy consumption has been widely paid attention to and discussed. In recent years, the residents' energy consumption has become an important economic sector for the growth of total energy consumption in China. Meanwhile, with the annually increasing economic level and urbanization, China's urban household energy consumption is also increasing; in particular, the growth of energy demand for electricity has become a significant source of household energy consumption. The report of the 19th National Congress of the Communist Party of China points out that the market direction of China's energy consumption is to "drive the revolution of energy production and consumption and build a clean, low-carbon, safe and efficient energy system". Therefore, in the new normal stage of China's economic development, studying energy consumption characteristics and influencing factors of urban households is helpful for policy makers and governments to realize the green and sustainable development.

On the other hand, the geographic variation and the stage characteristics of economic development facing Chinese residents are also the influencing factors causing objective differences in energy consumption behavior of urban households in China. First, there are differences in temperature, production habits, cultural habits, and other social perspectives in the east-west and north-south regions of China, which leads to regional differences in energy consumption. Since the reform and opening-up, China's economic development has gradually formed a pattern of developed coastal regions and relatively undeveloped central and western regions, and this is reflected in regional income level (i.e., higher income and consumption level in the eastern and coastal regions and lower ones in the western regions). Meanwhile, the Chinese urban-rural dual economic structure also makes the issues on household energy consumption face more complex economic background and human characteristics.

The influencing factors of household energy consumption mainly include the climate characteristics, the household income, and the regional cultural traditions [1, 2]. The *Report on China's Household Energy Consumption* (2016) (hereinafter referred to as the report) [3] indicates that per capita energy consumption quite varies from south to north. Per capita energy consumption in the south is 291.35 kg standard coal/year, and that in the north is 1.68 times the per capita level in the south. Since 2000, energy consumption in household sector has grown rapidly, from 467 million tons of standard coal in 2000 to 501 million tons of standard coal in 2015, with an annual average growth rate of 7.6%, while per capita energy consumption at the same time has also increased from 126.44 kg of standard coal in 2000 to 365.4 kg of standard coal in 2015, with an annual average growth rate of 7.3%. However, compared to developed countries, such as UK and USA, China's household energy consumption only corresponds to one-third of that in the USA and one-half of that in the UK.

Therefore, it is necessary to explore specific influencing factors of China's household energy consumption. This research can provide valuable reference for further improving China's regional energy governance policies and realizing the objective of energy conservation and emission reduction. The organization of this paper is as follows: Section 2 summarizes the relevant literature and presents the research basis of this paper. Section 3 designs the research, including model construction, research variables, and data source. Section 4 provides empirical analysis and results. Section 5 makes conclusion discussion and policy meaning analysis. Section 6 summarizes this study.

2. Literature Review

In recent years, these are growing interest and discussions about the issues of household energy consumption [4]. The demand of electricity consumption, as one of the clean energy sources, has been growing rapidly, partly because of its substitutability for other unclean energy sources. China's "*West-to-East Power Transmission Project*" reflects the imbalance in power demand. In the following section, we will review previous research on spatial characteristics, influencing factors, and the guarantee measures of energy demand.

First, the energy demand and the economic growth are important socioeconomic indicators for social and economic operation. There is dynamic change process of regional difference for the relationship between energy consumption and economic development. Due to the difference of resource endowment in different regions, China's energy supply pattern varies from region to region, characterized by regional difference and shifting of center of energy production and consumption, as well as rate change [5]. Second, in terms of the drivers of energy consumption, regional economic development plays a greater role in affecting energy consumption in household sector, and this is mainly related to the change of China's economic center [6]. Zhang et al. [7] study the factors that influences the shift of the center of China's oil production and consumption and find that the infrastructure and demand have a great impact on oil consumption in different regions over different periods, so that there has been the change of the center of natural gas consumption. Finally, the urbanization, industrialization, residents' income, cultural characteristics, and regional climate are all the social influencing factors at the household energy consumption level [8, 9]. All the studies above have reflected the impact of urbanization, industrialization, and regional differences on energy consumption only under group studying with certain classification criteria, but they have not made study from unified perspective.

For the difference in intraregional growth of electricity consumption in China, Liu et al. [10] point out that the annual average growth of electricity consumption in Northeast China is the lowest, while that in Northwest China is the highest. Therefore, the dynamic change process and characteristics of the center of power demand and consumption are important considerations in the energy infrastructure. Of course, intraregional economic scale and investment scale are also important to drive the growth of electricity consumption. Due to the difference of power consumption intensity and investment efficiency, the drivers of energy consumption are different across regions. Therefore, the study of regional differential management based on influencing factors is important theoretical support and reference for energy policy making.

The rebound effect of electricity consumption is an important influencing aspect of residents' electricity consumption. Su [11] explores the maintenance level and influencing factors of rebound effect from a micro perspective and finds that the improvement of power consumption efficiency will raise the expected power saving quantity by 43.51%. The author points out that the rebound effect of electricity consumption is both closely related to per capita income and significantly related to gender difference. For this, the relationship of the rebound effect of residents' income level and electricity consumption is shaped by an inverted "V," and that of male consumers is higher than that of female consumers. From the analysis of indirect rebound effects of main energy consumption, including electricity, gasoline, and natural gas, the rebound effects of these three

kinds of energy consumption are all below 10% over the period 1997–2012, but the indirect rebound effect of energy consumption has a difference between energy type and urban and rural areas. In these three types of energy consumption, the indirect rebound effect of electricity consumption is the largest [12]. The rebound effect of energy consumption determines the limitation of energy consumption reduction caused by energy efficiency improvement [13]. Given the experience of the developed countries, the rebound effect of energy consumption makes it difficult to predict the relationship between energy efficiency and energy saving effect. The rebound effect would partially offset the energy saving achieved for increased energy efficiency [14].

Energy price, income, urbanization, and industrialization are very important factors in investigating energy consumption in household sector. Liu et al. [15] have demonstrated the key effect of economic measures on residents' willingness to save energy based on Beijing, Hangzhou, Guangzhou, and Guiyang residents' household electricity consumption behavior and the economic intervention effect analysis, while the change of electricity price played a significant role in reducing residents' electricity consumption. There is difference in energy saving consciousness among different income groups. With the improvement of income level, the electricity consumption of low-grade consumption residents has shown an increasing trend. Yang et al. [16] studied the effect of residents' income and urbanization level on urban residents' appliances consumption and found that the income level and urbanization rate had positive impact on residents' appliances consumption, which led to the increase of household energy consumption. In recent years, the rapid promotion of urbanization has become an important driving factor for growth of energy consumption in household sector. It is found that the economic development, economic structure, urbanization, and income all have positively affected total electricity consumption, and the contribution rate of urbanization to electric power growth is the most significant [17].

Demand policy of electricity consumption in China relates to the electricity price system, the electricity sales price subsidy, and other aspects. Since the implementation of the stepped incremental electricity price policy for household electricity consumption in July 2012, the distribution and utilization of electricity resources have become more reasonable, and this has improved the efficiency of electricity consumption and guaranteed the limited range of equity and other multiple goals [18]. However, under the background of the improvement of overall income level of residents, there is a significant difference in the energy saving effect and difference between low-income and high-income groups [19]. With the development of the market-oriented reform of electricity price, the elimination of hidden cross-subsidization that distorts electricity price has become the necessary goal of the market-oriented reform of electricity price. The goal is to make the income of electric power enterprises more reasonable without damaging the ability of residents to pay for consumption, and simultaneously to

significantly reduce the emission of electric power industry. Therefore, the objects of electricity price marketization reform are more concentrated on low-income residents, and this is an efficient and fair market-oriented electricity consumption subsidization mechanism [20].

China's electricity consumption has been characterized by periodic path evolution [21]. With the rise of energy economics in the 1970s, the studies on power economics phenomenon, especially on the relationship between power demand and economic growth in developing countries, have gradually become the focus of the academic circles at home and abroad. The existing studies have mainly formed the following views: first, the power consumption has become a one-way Granger cause of economic growth [22]; second, the economic growth has become a one-way Granger cause of power consumption [23]; third, there has been a Granger cause relationship between power consumption and economic growth [24]; fourth, there has been difference in the causal relationship between them in different situations [25, 26]. Since the income level and life style of urban and rural residents are quite different, the characteristics and patterns of household electricity consumption in Northwest China are also significantly different. The electric consumption has obvious stepped change characteristics from rural areas to small cities, medium cities, and big cities. The demand level of rural household electricity is relatively lower, and it is mainly used for the most basic lighting, cooking, and entertainment, while that of urban household electricity is relatively higher, and it is especially used for the pursuit of life convenience, cleanliness, and comfort. Zeng et al. [27] pointed out that electricity consumption between urban and rural and among regions in China has reached a fair level after the exploration on the fairness of electricity consumption. Meanwhile, due to the difference in China's regional economic development, population scale, and other aspects, it is necessary to develop relevant electricity policies suitable for the actual situation of local regions. Ma et al. [28] explored the two behavioral mechanisms of rational decision-making and inertia decision-making in the process of residents' power consumption based on the perspective of residents' rational consumption. It has been found that income affects significantly residents' electricity consumption, and living habits have an important impact on consumption behavior [29]. However, economic development makes residents' income increase with stage characteristics; these features of changes in income have not been considered adequately in research with respect of energy consumption behavior.

In conclusion, the existing studies on household energy consumption are relatively systematic. Accordingly, many results can been summarized: (1) The household energy has regional characteristics (obviously seen in eastern, middle, and western China). (2) The impacts of influencing factors of household energy consumption and their degree vary from region to region. (3) The impact of economic development, such as industrialization and urbanization, on household energy consumption plays a role of external regulation and changes the environment of the study of the relationship between subjects. Overall, there is still necessity for further

expansion in the existing studies, and it could be reflected in the following aspects: (1) the way to put the study of unified influence relationship in different regions and develop the evolution characteristics of variable relationship; (2) considering the same external influence factors to study the characteristics of the relationship between variables. Therefore, based on the core relationship between the price and quantity of household energy consumption, this paper investigates the external action mechanism of the change of household income and the level of urbanization promotion on their relationship, and then a systematic analysis of urban household (electric) energy consumption in China is made.

3. Study Design

3.1. Model. The IPAT equation, presented firstly in the 1970s, identified three factors that determine the human environmental impact by using the equation formula $I = P \times A \times T$, with impact (I), population (P), affluence (A), and technology (T) [30]. IPAT method provides a theoretical basis for analysis of energy and environmental problems. For the STIRPAT model developed subsequently, the study on expansibility has also been made based on the idea of IPAT model. Among them, I, P, A, and T represent the observed environmental impact, population scale, wealth, and technological progress, respectively. Owing to the different effects of household income, urbanization, and other factors on residents' energy consumption in different regions, it may be nonlinear and there is a threshold value. The transformation of traditional STR model in handling of related problems is nonsmooth. The panel smooth transition regression (PSTR) model is improved based on the traditional model, which can both describe the heterogeneity of model regression coefficient on the section and realize the smooth transition among different regions, so that the nonlinear characteristics of variables can be captured well, and it is more in line with the performance of the relationship between real economic variables. The biggest difference between panel smooth transformation regression (PSTR) model and panel threshold regression model lies in the following: (1) With panel threshold regression analysis, different groups are distinguished according to the observed values, and the boundaries between groups are obvious and discontinuous. However, this strict limitation does not always fit the real-world situation. (2) The panel smooth transformation model relaxes this restriction. The boundary is a function of the threshold variable and can fluctuate within a certain range. Panel smooth transformation model is a generalization of panel threshold regression model, which is more in line with economic and social reality.

The traditional panel fixation effect and the random effect model have difficulties in accurately measuring individual difference between explanatory variable and explained variable. For this, Hansen (1999) [31] proposed a panel threshold regression (PTR) model by introducing threshold variable q_{it} and transition function.

$$y_{it} = \alpha_i + \beta_0' x_{it} I(q_{it} \le c) + \beta_1' x_{it} I(q_{it} \succ c) + u_{it}. \tag{1}$$

Because transition function is an indicative function, there are only two conditions for valuing regression coefficient vector, β_0 or β_1, and this depends on whether the value of the threshold variable q_{it} is greater than the critical value c. This means that even if some individual q_{it} values are very close, there will be significant differences in the corresponding regression coefficient values (either β_0 or β_1) for their distribution on both sides of c. This jumpy change reduces the applicability of PTR model.

To avoid this, González et al. [32] set the transition function in the form of Logistic function and proposed the panel smooth transition regression (PSTR) model. The model and its transition function are as follows:

$$y_{it} = \alpha_i + \beta_0' x_{it} + \beta_1' x_{it} g(q_{it}; \gamma, c) + u_{it}, \tag{2}$$

$$g(q_{it}; \gamma, c) = \left[1 + \exp\left(-\gamma \prod_{j=1}^{m} (q_{it} - c_j) \right) \right]^{-1}, \tag{3}$$

$$\gamma \succ 0, c_1 \le c_1 \le \cdots \le c_1,$$

where $g(q_{it}, \gamma, c)$ is transition function; q_{it} is threshold variable, and its value range is [0, 1]; γ is smoothing coefficient, which determines transition speed; c is critical value; and m is the number of critical values. Since $g(q_{it}, \gamma, c)$ is a continuous function of q_{it}, and the regression coefficient vector $\beta_0 + \beta_1 (q_{it}, \gamma, c)$ is also a continuous function of threshold variable, there is no jumpy change. PSTR model consists of a linear part ($\alpha_i + \beta_0' x_{it}$) and a nonlinear part $\beta_1' x_{it} (q_{it}, \gamma, c)$. The linear part is the first partition of the model. Each transition function corresponds to a new system, and (2) is a PSTR model containing two systems. The general expression of PSTR model is

$$y_{it} = \alpha_i + \beta_0' x_{it} + \sum_{j=1}^{r} \beta_j' x_{it} g_i(q_{it}^{(j)}; \gamma_j, c_j) + u_{it}. \tag{4}$$

The general expression (4) of PSTR model includes $r + 1$ partitions. Then, the PSTR model corresponding to (4) is equivalent to the PTR model of Hansen (1999). When $\gamma \longrightarrow 0$, the PSTR model corresponding to (4) is equivalent to the panel fixation effect regression model. Finally, in (3), $\gamma > 0$, $C_1 \le C_2 \le \cdots \le C_m$ is a constraint to estimate the regression coefficient.

In this paper, the PSTR model has been used for identifying the nonlinear characteristics between the urban household electricity consumption and electricity sales price. As a threshold model, PSTR meets the nonlinear relationship between the total electricity demand and the price, and the specific form of PSTR model in the study is as follows:

$$y_{it} = \alpha_{it} + b_1 x_{it} + b_2 x_{it} g(q_{it}; \gamma, c) + \varepsilon_{it}, \tag{5}$$

where α_{it} is the intercept; ε_{it} is the residual term; and γ is the transition speed, which reflects the speed from "0" state to "1" state; when the value tends to be 0, PSTR model will degenerate into the traditional linear regression model, and the nonlinear relationship between the variables is not observed. This parameter affects the smoothness of the model. c is smooth parameter value, which represents the

turning point of state change. q_{it} is the transition variable, and it can be any explanatory variable, combination form, or any other exogenous variable. The specific form of the transition function in the model is

$$g(q_{it}; \gamma, c) = \left\{ 1 + \exp\left[-\gamma \prod_{j=1}^{m} (q_{it} - c_j) \right] \right\}^{-1}, \quad (6)$$

where $\gamma > o$, $c_1 \le c_2 \le \cdots \le c_m$, $0 \le g(q_{it}; \gamma, c) \le 1$, and the function value of x_{it} is smooth between b_1 and $b_1 + b_2$. Usually, $m = 1$ or $m = 2$. When $m = 1$, $g(q_{it}; \gamma, c) = [1 + \exp[-\gamma(q_{it} - c)]]^{-1}$, and $g(q_{it}; \gamma, c)$ has a location parameter; at this time, $\lim_{q_{it} \rightarrow -\infty} g(q_{it}; \gamma, c) = 0$ (and $\lim_{q_{it} \rightarrow +\infty} g(q_{it}; \gamma, c) = 1$). When the transition function value is 0, the model is called low mechanism; the model is called high mechanism if transition function value is 1. When the transition function value is between 0 and 1, the corresponding model is smoothly converted between these two mechanisms. When $m = 2$, there are two location parameters of $g(q_{it}; \gamma, c)$; at this time, $g(q_{it}; \gamma, c_1, c_2) = [1 + \exp[-\gamma(q_{it} - c_1)(q_{it} - c_2)]]^{-1}$, the conversion function has the minimum value in $((c_1 + c_2)/2)$, and its corresponding mechanism is the intermediate mechanism, while $\lim_{q_{it} \rightarrow +\infty} g(q_{it}; \gamma, c_1, c_2) = 1$. Combined with the theoretical analysis of the impact of the electricity price of urban households on electricity consumption, the following analysis model is constructed:

$$\text{elec} = \mu_{1t} + \sum \beta_{1t} \times x_{1t} + \beta'_{1t} g(\text{income}; \gamma_1, c_1) \times x_{1t} + u_{1t},$$

$$\text{elec} = \mu_{2t} + \sum \beta_{2t} \times x_{2t} + \beta'_{2t} g(\text{urbanization}; \gamma_1, c_1)$$

$$\times x_{2t} + u_{2t},$$

$$(7)$$

where $x_{1t} = (\text{ecp}, \text{income}, \text{CPI}, \text{ids}, \text{rain}, \text{temp}, \text{pca})$ is the explanatory variable matrix consisting of the core explanatory variable, the transition variable, and the control variable. $\beta_{1t} = (\beta_{11}, \beta_{12}, \beta_{13}, \beta_{14}, \beta_{15}, \beta_{16}, \beta_{17})$ is the model linear estimation coefficient of the explanatory variable in the PSTR model. $\beta'_{1t} = (\beta'_{11}, \beta'_{12}, \beta'_{13}, \beta'_{14}, \beta'_{15}, \beta'_{16}, \beta'_{17})$ is the model nonlinear estimation coefficient of the explanatory variable in the PSTR model. $x_{2t} = (\text{ecp}, \text{income}, \text{CPI}, \text{ids}, \text{rain}, \text{temp}, \text{pca})$ is an explanatory variable matrix consisting of core explanatory variable, transition variable, and control variable. $\beta_{2t} = (\beta_{21}, \beta_{22}, \beta_{23}, \beta_{24}, \beta_{25}, \beta_{26}, \beta_{27})$ is the model linear estimation coefficient of explanatory variable in PSTR model, and $\beta_{2t}' = (\beta'_{21}, \beta'_{22}, \beta'_{23}, \beta'_{24}, \beta'_{25}, \beta'_{26}, \beta'_{27})$ is the model nonlinear estimation coefficient of explanatory variable in PSTR model. Among them, the value range of urbanization rate of transition variable is between (0, 1), so the logarithm is mapped to $[-\infty, +\infty]$ in the process of empirical analysis to facilitate the discussion of results. The parameter estimation of PSTR model can effectively overcome the problem of parameter heterogeneity and obtain relatively reliable and stable estimation results. For PSTR model, the estimate of parameter is obtained by nonlinear least square method, and the initial slope coefficient and location parameters of the transition function are obtained

by seeking high-precision analog degradation, while the NLS method is used to estimate the nonlinear parameters of the model.

3.2. Variable Selection

3.2.1. Explained Variable. In this paper, the electricity demand of household energy consumption of urban residents is taken as the main indicator to measure the energy consumption level of urban residents, and it is denoted as *elec*. The electricity consumption of urban residents is selected from 30 provincial capitals and key cities (such as Shenzhen and Ningbo). In fact, in the process of urbanization development, electricity consumption has become the main type of household energy consumption, such as cooking and heating. On the other hand, rural household energy consumption is also gradually shifting from biomass (firewood) to clean energy consumption such as electricity. Therefore, electricity consumption becomes the main behavioral characteristic of household energy consumption.

3.2.2. Core Explanatory Variable. The core explanatory variable in this paper is electricity sales price, denoted as *ecp*. The electricity sales price is expressed by the prevailing sales price of each province in the country. The relationship between electricity price and demand is complex due to the rigid characteristics of power demand. The increase of residents' income makes energy more affordable and accessible. In addition, urbanization development changes energy behavior, to some extent, in respect of alternative energy sources, especially the increase in electricity consumption, to replace unclean energy sources. This in turn makes the relationship between electricity price and its consumption much more complex. In this paper, we try to investigate how the price drives electricity demands in household sector, at different scenarios.

3.2.3. Transition Variable. In this paper, the per capita disposable income (*income*) and urbanization rate (*urbanization*) are selected as the transition variables in PSTR model. Because of the typical urban-rural duality of China's economic development, there are obvious regional characteristics in terms of income and urbanization at a national scale. Per capita disposable income of residents in the eastern coastal regions is generally higher than that in the central and western regions. The urbanization is still developed slowly in the western region, especially in the remote underdeveloped areas. Therefore, to comprehensively analyze the regional differences in the impact of electricity sales price on the electricity consumption of urban residents in China, it is necessary to analyze the state characteristics of the relationship between the electricity consumption and electricity sales price of urban residents under different incomes and different urbanization levels.

3.2.4. Control Variables. The electricity consumption demand is closely interconnected with the economy, society,

and environment. Among them, in respect of economic factors, the growth of income level and expenditure level has greatly driven the consumption, so that the power consumption demand has also been increased. The optimization and adjustment of industrial structure reflect the level and quality of social and economic development. The adjustment of industrial structure and the transformation of national economy to service industry will play a role in promoting the reduction of power consumption to a certain extent. In respect of environmental factors, the unique temperature difference between the north and the south and the rainfall difference between the east and the west have a significant seasonal impact on the annual electricity consumption of urban residents. To fully investigate the impact of these factors on the household electricity consumption of urban residents in China, this paper has selected CPI, industrial structure (ids), urban annual average rainfall (rain), urban annual average temperature (temp), and per capita housing construction area of major cities (PCA) as the control variables of PSTR model analysis.

3.3. Data Sources and Descriptive Statistics.

Table 1 presents statistical characteristics of variables selected in this study. The explained variable is urban household electricity consumption. The core explanatory variable is the electricity sales price, and the transition variables are urban residents' per capita disposable income and urbanization rate, while control variables include industrial structure (proportion of secondary industry to total GDP), housing construction area, CPI (consumer price index), annual average temperature, and annual average rainfall. All data are from relevant statistical yearbooks of 30 provincial capitals of China (including the statistical yearbooks of cities and the statistical yearbooks of economic development of different provinces, excluding Taiwan, Hong Kong, Macao, and Tibet).

4. Empirical Test and Analysis

4.1. Linear Test and Residual Nonlinear Test.

The regression analysis of PSTR model is required to meet the three-step assumptions. First, test the establishment of the linear and nonlinear relationship of the model and the number of transition functions. Second, determine the position parameters of the PSTR model. Third, calculate the initial values of the smoothing parameters and the position parameters based on the above. Table 2 shows the verification for basic hypothesis of the model herein. The test results show that the null hypothesis H0: $r = 0$ passes the significance test at a significant level of 1%, while the null hypothesis H1: $r = 1$ is not significant at the 5% level. Therefore, the models are nonlinear. In addition, the number of transition functions is 1; that is, $r = 1$. At the same time, the values of AIC and BIC of all models $m = 1$ are less than the test value of $m = 2$, so the constructed model has a position parameter.

4.2. Effect of Electricity Sales Price on Urban Household Electricity Consumption under Residents' per Capita

Disposable Income. Table 3 shows the results. Models 1-4 show regression results of the models with regard to influence of electricity sales price on urban household electricity consumption nationwide and in eastern, central, and western regions in the context of urban residents' per capita disposable income between different regions. It can be seen that most of the factors influencing urban household electricity consumption pass the significance test. Herein, the influence of electricity sales price on urban household electricity consumption is explained by regions from the perspective of regional per capita income differences.

From a nationwide perspective, the position parameter for nonlinear conversion of PSTR model is RMB 9.91 thousand, indicating that nonlinear income threshold value of urban household electricity consumption based on the influence of electricity sales price is RMB 9.91 thousand. Therefore, there are two transition mechanisms in the model. When per capita disposable income is less than RMB 9.91 thousand, the model is in low mechanism, and the coefficient of influence of electricity sales price on urban household electricity consumption is 0.42. When per capita disposable income is higher than RMB 9.91 thousand, the model is in high mechanism, and the coefficient is −2.38. On the one hand, the correlation between electricity sales price and electricity consumption has turned from positive to negative as the urban residents' per capita income level increases. On the other hand, the influence of urban residents' per capita disposable income level on electricity consumption has also changed from positive to negative with the increase of income level. The coefficient of influence of urban residents' per capita income on electricity consumption is 0.64 when the model is in low mechanism and −0.27 when the model is in high mechanism. Therefore, urban residents' per capita disposable income has an effect on the selectivity of household electricity consumption. Herein, it is believed that the increase in per capita disposable income has enhanced the diversity of selections of urban household energy consumption, and there are alternative energy products for electricity consumption.

For eastern region, the position parameter for the nonlinear conversion of the PSTR model is RMB 9.49 thousand, indicating that the nonlinear income threshold value of the electricity sales price that affects urban household electricity consumption in the eastern region is RMB 9.49 thousand. Similar to the national situation, there are two transition mechanisms in the eastern region model. When the per capita disposable income is lower than RMB 9.49 thousand, the model is in low mechanism, and there is a negative linear correlation between electricity sales price and urban household electricity consumption (−0.23). When the per capita disposable income is higher than RMB 9.49 thousand, there is a positive nonlinear correlation between electricity sales price and urban household electricity consumption (1.96). For residents' per capita disposable income under the two mechanisms, urban household electricity consumption decreases as the residents' disposable income level increases. The influence coefficient of residents' disposable income on urban household electricity consumption is −0.24 when the model is in low mechanism and −0.92 with

Table 1: Descriptive statistical characteristics of variables.

Variable description	Obs.	Mean	Std. Dev.	Min	Max
Household energy consumption	420	428098	401371	21293	2100000
Electricity sales price	420	502.361	66.676	334.000	694.560
Income	420	25740.840	11664.100	8397	61172
Urbanization	420	52.646	14.255	26.870	89.600
Industrial structure	420	42.919	8.530	18.100	60.100
Per capita housing construction area	420	1101.602	1591.397	45.570	13755.900
CPI	420	102.611	1.747	97.510	108.400
Annual average temperature	420	14.582	5.074	4.283	25.358
Annual average rainfall	420	79.162	44.339	9.363	244.975

Table 2: Linear test and residual nonlinear test results.

Model	$H_0: r=0$, $H_1: r=1$			$H_1: r=1$, $H_2: r=2$			AIC		BIC	
	LM	LMF	LRT	LM	LMF	LRT	$m=1$	$m=2$	$m=1$	$m=2$
M1	85.32 (0.00)	4.48 (0.00)	95.38 (0.00)	6.21 (0.51)	0.79 (0.60)	6.26 (0.51)	3.54	3.68	4.23	4.28
M2	79.83 (0.00)	6.89 (0.00)	118.23 (0.00)	13.77 (0.05)	1.70 (0.12)	14.50 (0.43)	4.21	4.49	4.98	5.20
M3	55.42 (0.00)	3.87 (0.00)	76.49 (0.00)	21.86 (0.00)	2.88 (0.01)	24.32 (0.00)	4.21	4.37	4.29	4.41
M4	53.72 (0.00)	3.04 (0.00)	65.08 (0.00)	20.47 (0.00)	2.67 (0.01)	21.87 (0.00)	5.45	5.87	6.82	7.01
M5	71.76 (0.00)	3.62 (0.00)	78.77 (0.00)	13.35 (0.06)	1.73 (0.10)	13.57 (0.06)	4.01	4.65	5.21	5.45
M6	78.64 (0.00)	6.65 (0.00)	115.48 (0.00)	25.19 (0.00)	3.42 (0.00)	27.78 (0.00)	4.23	4.58	4.88	4.98
M7	73.52 (0.00)	7.55 (0.00)	119.65 (0.00)	27.65 (0.00)	3.89 (0.00)	31.76 (0.00)	4.30	4.53	5.32	5.67
M8	60.52 (0.00)	3.68 (0.00)	75.12 (0.00)	7.07 (0.42)	0.84 (0.55)	7.23 (0.41)	5.61	5.81	6.01	6.42

Table 3: Model parameter estimation results (in the context of urban residents' disposable income change).

	M1	M2	M3	M4
β_{11}	0.42*(1.41)	−0.23*(−1.71)	−39.05***−−2.72)	0.62***(7.07)
β_{12}	0.64***(2.34)	−0.24***(−2.36)	10.24*(1.38)	0.71***(2.29)
β_{13}	0.25*(1.53)	0.02 (0.69)	−7.09***(−2.32)	0.04 (0.23)
β_{14}	0.12***(4.37)	0.83 (0.85)	0.95 (1.03)	0.06*(1.28)
β_{15}	−1.44***(−2.48)	−0.01 (−0.02)	−96.91***(−2.96)	−0.61 (−0.92)
β_{16}	−0.27***(−2.36)	0.12 (0.89)	−3.07***(−2.22)	−0.05 (−0.38)
β_{17}	−0.20***(−3.41)	1.56 (1.12)	0.31 (0.21)	−0.07*(−1.42)
β'_{11}	−2.80*(−1.94)	2.19**(2.13)	7.82*(1.93)	0.67 (0.37)
β'_{12}	−0.91 (−1.18)	−0.68***(−2.83)	−17.12 (−1.26)	16.05*(1.65)
β'_{13}	−1.46***(−6.23)	−0.16*(−1.73)	11.38**(2.09)	−3.02***(−2.56)
β'_{14}	−0.05 (−1.07)	−2.04*(−1.85)	−1.60 (−0.97)	1.00 (1.54)
β'_{15}	2.37 (2.87)	0.81 (0.86)	177.60***(2.98)	−23.26 (−1.38)
β'_{16}	0.67***(4.76)	−0.01 (−0.34)	5.73**(2.33)	1.13*(1.81)
β'_{17}	0.29***(2.78)	−0.51 (−0.22)	−0.58 (−0.23)	0.57 (1.21)
γ_1	1.77	1.51	0.17	10.52
c_1	9.91	9.49	8.62	10.60

Note: ***, **, *significant at the levels of 1%, 5%, and 10%, respectively. The t values for parameter estimates are in parentheses.

the model in high mechanism. Herein, it is believed that the positive relationship between electricity sales price and urban household electricity consumption may be partly due to the insensitivity of step tariff to electricity demand under the conditions of high income, and high-income levels have stimulated the potential of urban households in energy consumption at specific stages of income levels.

For the central region, the threshold value of the influence of urban residents' per capita disposable income is RMB 8.62 thousand. There are two transition mechanisms for the models with regard to the influence of electricity sales price on urban household electricity consumption in central region. The model is in low mechanism when the residents' per capita disposable income is less than RMB 8.62 thousand and in high mechanism when the residents' per capita disposable income is greater than RMB 8.62 thousand. In central region, the slope changes gently as the conversion degree is small (the value of conversion degree is 0.17). When the residents' disposable income level is low, the relationship between the electricity sales price and the urban household electricity consumption is negative, but the degree of this negative relationship gradually decreases as per

capita disposable income increases. Under both mechanisms, the relationship between electricity sales price and urban household electricity consumption in central region is linearly negative and nonlinearly negative. According to the separate investigation for the influence of the residents' disposable income on urban household electricity consumption, residents' disposable income and urban household electricity consumption are positively correlated when the model is in low mechanism. The nonlinear relationship between them shows a negative correlation (coefficient is −6.88) when the model is in high mechanism.

For western region, the threshold value of effect of urban residents' per capita disposable income is RMB 10.52 thousand. The threshold value of the electricity sales price that affects the urban household electricity consumption in this region is RMB 10.60 thousand. There are two transition mechanisms. When per capita disposable income is less than RMB 10.60 thousand, the model is in low-mechanism state, and at this point, the linear coefficient of the effect of electricity sales price on the urban household electricity consumption is 0.62. When the per capita disposable income is greater than RMB 10.60 thousand, the model is in high-mechanism state, and at this point, the linear coefficient of the effect of electricity sales price on the urban household electricity consumption is 1.29. There is a positive correlation between electricity sales price and urban household electricity consumption under the two mechanisms. Urban household electricity consumption has gradually increased in western region as the income level increases. In addition, it can be found that the influence of residents' disposable income level is larger than that of the electricity sales price on household electricity consumption under both mechanisms. Therefore, it can be predicted that income level in western region is still the main factor that determines urban household energy consumption with the implementation of energy conservation, emission reduction, and new energy policies, and the price sensitivity of energy consumption is generally low.

Figure 1 shows the transition functions of regression results, corresponding to the models in Table 3. Model 1 to Model 4 are transition functions of regression results in terms of the influence of electricity sales price on urban household electricity consumption throughout the country and eastern region, central region, and western region considering the urban residents' per capita disposable income among different regions. From the perspective of regional analysis, some observations of the models are between the low and high mechanisms of the models. Therefore, there are two mechanisms in these models with regard to the influence of the electricity sales price on urban household electricity consumption in the context of urban households' per capita disposable income. From the transition function graph shown in Figure 1, it is found that the degrees of the model transition throughout the country and in eastern region and central region are lower than those in the western region, which are 1.77, 1.51, and 0.17, respectively. The slope of the transition function changes more gently. In the western region, the degree of transition is 10.52, the slope of the transition function changes

drastically, and most of the observations are in the low-mechanism state of the model.

4.3. Effect of Electricity Sales Price on Urban Household Electricity Consumption under Urbanization Rate.
Table 4 shows the estimate results of the effects of electricity sales price on urban household electricity consumption. Models 5–8 are the regression results about the relationship between electricity sales price and urban household electricity consumption throughout the country and in eastern region, central region, and western region under urbanization development in different regions. We find that most of the factors influencing urban household electricity consumption have passed the significance test. This will explain the influence of electricity sales price on urban household electricity consumption in regions from the perspective of progress differences in regional urbanization development.

The effect of electricity sales price on urban household electricity consumption is seen from the background of urbanization development nationwide. There are two transition mechanisms for Model 5. The threshold value of urbanization is $urbanization = e^{-3.79}$. The influencing coefficient of electricity sales price on the urban household electricity consumption is 0.39 when it is less than the threshold value and 1.28 when it is greater than the threshold value. This shows that the total urban household electricity consumption has continued to increase with urbanization development. It is found that urbanization development positively affects urban household electricity consumption (coefficient is 1.71) in the low mechanisms of the model. The higher the level of urbanization is, the greater the growth in urban household electricity consumption is. In the high mechanisms of the model, both have nonlinear positive correlation (coefficient is 1.02). It can be seen that there is significant nonlinear relationship between electricity sales price and urban household electricity consumption nationwide. In addition, urbanization positively drives the urban household electricity consumption and promotes the growth of total electricity consumption.

The regression results show that there are two transition mechanisms for Model 6. The threshold value of urbanization in the eastern region is $urbanization = e^{-4.05}$. There is always a positive correlation between the electricity sales price and the urban household electricity consumption. Under the two mechanisms, the coefficient of the influence of electricity sales price on urban household electricity consumption increased from 2.33 to 2.95, indicating that the effect has gradually increased as the urbanization rate rises. In eastern region, the influence of urbanization development on urban household electricity consumption has changed from negative (coefficient is −0.54) to positive (coefficient is 1.94) in the low mechanism. This shows that urbanization development has promoted the growth of urban household electricity consumption, and it has become an important driving force of growth in urban household energy consumption.

The regression results also show that there are two transition mechanisms in Model 7. The threshold value of
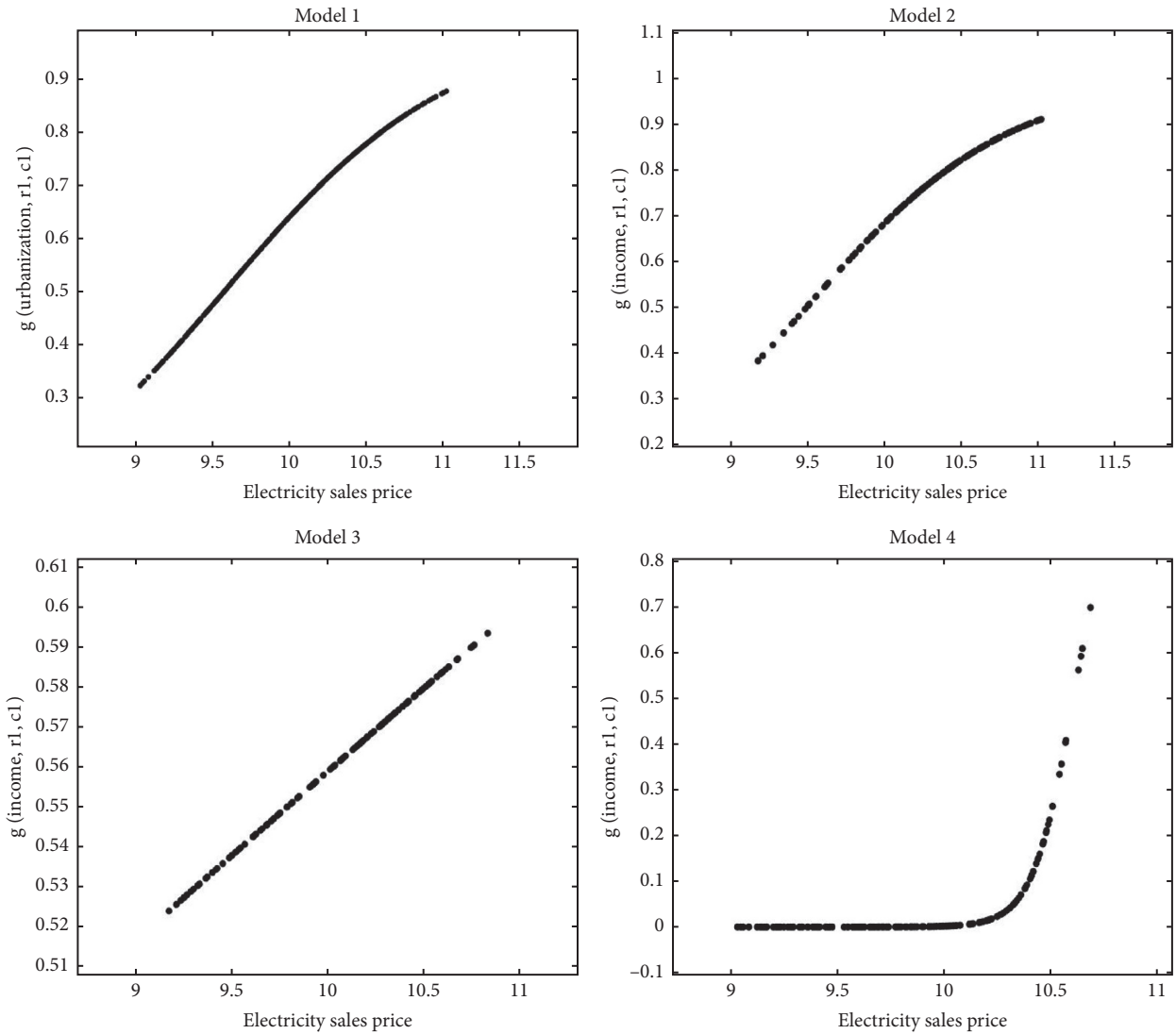
FIGURE 1: Transition functions of different model regression results in the context of urban households' per capita disposable income.

urbanization in central region is urbanization $= e^{-3.69}$. At different stages of urbanization development, the electricity sales price positively affects the urban household energy consumption. With continuous improvement of urbanization rate, the total urban household electricity consumption has continued to increase. Regression results show that the coefficient of the effect of electricity sales price on urban household electricity consumption is 1.69 under a high-mechanism state. From the fact that urbanization affects the total urban household electricity consumption, it can be seen that urbanization development has promoted the growth of electricity consumption. The coefficient is 2.18 when the model is in low mechanism and 1.27 when the model is in high mechanism. Therefore, electricity sales price and urbanization level are positively correlated with urban household electricity consumption in central region whether it is a low-mechanism linear relationship or a high-mechanism nonlinear relationship.

The regression results show that there are two transition mechanisms for Model 8, and the threshold value of

urbanization in western region is urbanization $= e^{-3.91}$. In western region, the coefficient of influence of electricity sales price on urban household electricity consumption is 1.79 when the model is in low mechanism and 0.29 when the model is high mechanism. It can be seen that the electricity sales price remains positively correlated to the urban household electricity consumption with urbanization development. However, its degrees of influence are different in the two states. The role of urbanization in promoting household electricity consumption growth has gradually increased as the urbanization rate rises. It is found that urbanization development has a positive effect on the growth of urban household electricity consumption through both mechanisms. On the one hand, urbanization development has changed the energy consumption structure of rural households after they became urban households. Under the guidance of policies that promote clean energy consumption, electricity has become a rigid energy source for urban household in several choices. On the other hand, the improvement of urbanization has increased the total demand

TABLE 4: Model parameter estimates (in the context of urbanization change).

	M5	M6	M7	M8
β_{21}	0.39*(1.52)	2.33***(5.34)	0.37*(1.73)	1.79***(5.92)
β_{22}	1.71***(8.30)	−0.54** (−2.07)	2.18***(7.87)	0.27 (0.87)
β_{23}	−0.95*** (−3.42)	−3.26*** (−8.59)	−1.53*** (−3.94)	0.18***(4.07)
β_{24}	0.26***(4.66)	0.07 (1.04)	0.58***(6.37)	0.14***(3.88)
β_{25}	0.70 (1.24)	4.55***(4.97)	−0.47 (−0.48)	−0.26 (−0.38)
β_{26}	−0.54*** (−3.40)	−0.28 (−0.52)	2.21***(3.26)	0.18 (1.02)
β_{27}	−0.08 (−1.24)	−0.04 (−0.23)	0.01 (0.08)	−0.18*** (−3.05)
β'_{21}	0.89***(3.22)	0.62 (1.05)	1.32***(4.11)	−1.50*** (−3.66)
β'_{22}	-0.69** (−2.83)	2.48***(5.55)	−0.91**(-2.18)	2.55***(5.67)
β'_{23}	−0.14 (−0.51)	2.56***(5.97)	1.43***(3.21)	−1.74*** (−5.24)
β'_{24}	−0.12** (−2.12)	−0.20*(−2.12)	−0.39*** (−4.17)	−0.12** (−2.01)
β'_{25}	0.09 (0.25)	−6.39*** (−8.35)	0.86*(1.53)	−0.60 (−1.00)
β'_{26}	0.59***(4.78)	0.88* (1.57)	−2.22 (−3.43)	−0.19*(−1.15)
β'_{27}	0.09*(1.44)	0.09 (0.51)	−0.02 (−0.15)	0.32***(3.80)
γ_2	70.27	15.52	57.85	19.10
c_2	3.79	4.05	3.69	3.91

Note. ***, **, * significant at the levels of 1%, 5%, and 10%, respectively. The t values for parameter estimates are in parentheses.

for energy consumption in the original urban households. Electricity is the main source of energy for urban households. The increase in urbanization level, the convenience in energy acquisition, the further improvement of quality of life, and the diversity of household electrical appliances have promoted the further increase of the total urban household electricity consumption.

Figure 2 shows the transition function of the regression results of the models in Table 4. Models 5–8 are transition functions of regression results in terms of the influence of electricity sales price on the urban household electricity consumption nationwide and in eastern, central, and western regions considering the urbanization development levels among different regions. From regional perspective, some observations of the models are between the low and high mechanisms. Therefore, there are two mechanisms with regard to the influence of the electricity sales price on urban household electricity consumption under different levels of urbanization development. From the transition function graph shown in Figure 2, it can be seen that the degree of model transition nationwide and in eastern and central regions is lower than that in the western regions, which are 70.27, 15.52, 57.85, and 19.1, respectively. In comparison, the slopes of the transition functions change gently in the eastern and western regions, while those nationwide and in the central region change sharply. The influence of the background of urbanization development on the relationship between electricity sales prices and urban household electricity consumption nationwide and in the central region

is more significant than those in the eastern and western regions.

5. Empirical Discussion and Policy Analysis

Through empirical analysis, this paper argues that there is a complex nonlinear relationship between electricity sales prices and urban household electricity consumption. This complex nonlinear relationship shows different influencing directions and degrees under different income levels and urbanization development. In the following, the specific reasons for the formation of research conclusions on electricity sales prices and urban household electricity consumption and the implications for policies are discussed in detail from perspectives of regional differences in both residents' disposable income and urbanization development.

Models 1–4 are for investigating the relationship between the electricity sales price, residents' disposable income, and urban household electricity consumption under changes in the residents' disposable income. Nationwide, electricity sales prices and residents' income level are negatively correlated to the urban household electricity consumption. Given the difference in residents' income, electricity sales price is positively correlated to the urban household electricity consumption in eastern and western regions but negatively correlated to urban household electricity consumption in central region. These results can be explained from the following perspectives. First, environmental regulatory policies cause regional differences. In recent years, the production and consumption of clean energy have been advocated nationwide. As a clean energy, electricity is an alternative to most energy categories. Especially in urban living environments, electricity consumption, as a major source of household energy consumption, is relatively insensitive to electricity sale price. Electricity becomes a "rigid-demanded" consumer product. Second, from the development trend of electricity demand, the rise in production costs has led to increase in electricity sales price. The promotion effect of cost is greater than that of growth in demand for electricity consumption.

In addition, income is a very important influential factor for household electricity consumption. As is well known, Chinese economic development holds very obvious regional differences. Eastern and southern coastal regions are considered traditional developed areas. Central part of China is defined as the less developed area, and west region, especially for northwest part of China, is undeveloped area. In fact, from the perspective of the resource endowment of regional development, the differences in economic development between the east, the central, and the west regions are fully reflected in the level of per capita income. Regional differences, to some extent, determine residents' consumption behavior. This paper shows the practical significance of income difference for consumption relationship. Therefore, household energy policies should be flexible enough to promote equity in energy access between regions.

Models 5–8 are for investigating the relationship between electricity sales prices, urbanization levels, and urban household electricity consumption in the context of
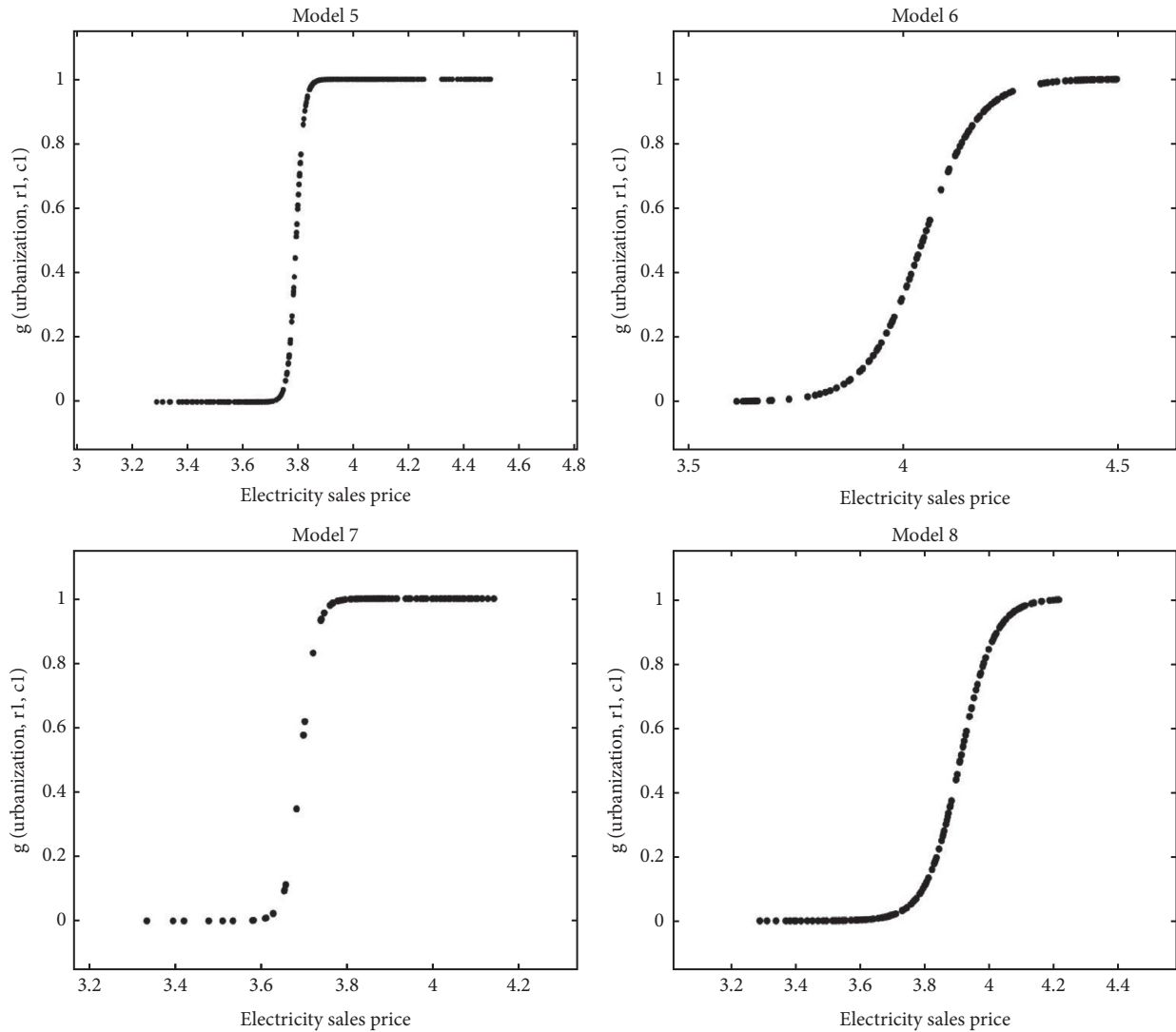
FIGURE 2: Transition functions of regression results of different models in the context of urbanization development.

urbanization development. The regression results of the models show that electricity sales price remains positively correlated to the urban household electricity consumption nationwide in the context of regional differences in urbanization. The coefficients of influence of electricity sales price on urban household electricity consumption are the highest in eastern region and the lowest in western region. However, the positive influence of urbanization development on urban household electricity consumption is the strongest in western region. The reasons for the positive relationship between the electricity sales price and the urban household electricity consumption in the context of urbanization may be as follows: First, the process of urbanization development has changed the role of the original residents in life, which is represented by typical demolition and urban village reconstruction. The change of role has reduced the habitual household energy option and choice range of the original residents. Second, urbanization development has changed the characteristics of overall growth in urban household energy consumption. The advancement

of urbanization as a whole has improved the quality of life of residents and promoted the general increase in the absolute amount of energy demanded by urban households. Furthermore, the incremental demand for household electricity-consuming products has further promoted the sustainable growth in total electricity consumption.

Furthermore, China's urbanization development still has a strong momentum. On the one hand, urbanization development has changed the residents' energy consumption habits and energy structure, especially for the shifting from high-carbon energy sources to cleaner ones. On the other hand, the development of urbanization is a manifestation of economic development. However, regional urbanization shows certain stage characteristic due to the different development resource endowment. This in turn plays a dominant role in affecting household energy consumption. Therefore, one very important suggestive finding is that the effectiveness of household energy policies should be fully considered at different stages of economic development and their phase characteristics.

In short, per capita disposable income and the level of urbanization are the key factors affecting urban household electricity consumption. The relationship between electricity sales price and total urban household electricity consumption is generally affected by regional differences. The increase in per capita disposable income has improved the actual affordability of urban household electricity consumption and further stimulated the potential of the growth of household energy consumption. The process of urbanization has changed not only the residents "energy consumption habits, but also the range of residents" energy choices. Based on the above analysis, the special manifestations of the relationship between reference variables at different regional levels need to be considered in the research of policies on household energy consumption. Starting from the basic variable relationship, comprehensive consideration is given to the context factors for the variable relationship. This paper mainly considers the regional differences characterized by the two factors of residents' disposable income and urbanization level.

6. Conclusions

This paper analyzes nonlinear relationship between electricity sales prices and urban household electricity consumption using PSTR approach. Considering the regional differences in the relationship between variables studied, this paper investigates systematically the complex relationship between the two and their specific forms under residents' disposable income and urbanization rate. To this direction, two sets of panel are constructed for smooth transition of regression models to explore the nonlinear relationship, with residents' disposable income and urbanization rates as converted variables. The research concludes that the degrees and models of influence of electricity sales price on urban household electricity consumption are different in the eastern, central, and western areas of China due to differences in income and urbanization. Such different relationship between variables provides an important research reference and decision basis for the Chinese government to implement energy policy and promote the synergy of policy benefits between regions.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

References

[1] M. Zhang and C. Bai, "Exploring the influencing factors and decoupling state of residential energy consumption in Shandong," *Journal of Cleaner Production*, vol. 194, pp. 253–262, 2018.

[2] M. Zhang and F. Guo, "Analysis of rural residential commercial energy consumption in China," *Energy*, vol. 52, pp. 222–229, 2013.

[3] X. Y. Zheng, *Research Report on Chinese Household Energy Consumption*, Science Press, Beijing, China, 2016, in Chinese.

[4] M. Zhang and B. Su, "Assessing China's rural household energy sustainable development using improved grouped principal component method," *Energy*, vol. 113, pp. 509–514, 2016.

[5] H. Xiao, Z. Ma, Z. Mi et al., "Spatio-temporal simulation of energy consumption in China's provinces based on satellite night-time light data," *Applied Energy*, vol. 231, pp. 1070–1078, 2018.

[6] S. K. Sun, J. F. Song, F. F. Wang, P. T. Wu, and Y. B. Wang, "Evaluating the impacts of socio-economic factors on regional grain virtual water flows in China using a structural equation modeling approach," *Journal of Hydrology*, vol. 571, pp. 132–141, 2019.

[7] Y. Zhang, J. Zhang, Z. Yang, and J. Li, "Analysis of the distribution and evolution of energy supply and demand centers of gravity in China," *Energy Policy*, vol. 49, pp. 695–706, 2012.

[8] L. F. Jing, Y. J. Zhang, and B. Wang, "The impact of urbanization on residential energy consumption in China: an aggregated and disaggregated analysis," *Renewable and Sustainable Energy Reviews*, vol. 75, pp. 220–233, 2017.

[9] Y. J. Guo, Q. Zhang, K. K. Lai, Y. Q. Zhang, S. B. Wang, and W. L. Zhang, "The impact of urban transportation infrastructure on air quality," *Sustainability*, vol. 12, no. 14, pp. 1–23, 2020.

[10] G.-X. Liu, M. Wu, F.-R. Jia, and Q. Wang, "Material flow analysis and spatial pattern analysis of petroleum products consumption and petroleum-related CO2 emissions in China during 1995-2017," *Journal of Cleaner Production*, vol. 209, pp. 40–52, 2019.

[11] Y.-W. Su, "Residential electricity demand in Taiwan: consumption behavior and rebound effect," *Energy Policy*, vol. 124, pp. 36–45, 2019.

[12] Z. Yan, X. Ouyang, and K. Du, "Economy-wide estimates of energy rebound effect: evidence from China's provinces," *Energy Economics*, vol. 83, pp. 389–401, 2019.

[13] X. Wang, X. Wen, and C. Xie, "An evaluation of technical progress and energy rebound effects in China's iron & steel industry," *Energy Policy*, vol. 123, pp. 259–265, 2018.

[14] Y. Wu, J. Su, K. Li, and C. Sun, "Comparative study on power efficiency of China's provincial steel industry and its influencing factors," *Energy*, vol. 175, pp. 1009–1020, 2019.

[15] G. Liu, X. Dong, Q. Jiang, C. Dong, and J. Li, "Natural gas consumption of urban households in China and corresponding influencing factors," *Energy Policy*, vol. 122, pp. 17–26, 2018.

[16] Y. Yang, J. Liu, Y. Lin, and Q. Li, "The impact of urbanization on China's residential energy consumption," *Structural Change and Economic Dynamics*, vol. 49, pp. 170–182, 2019.

[17] Z. Wang, X. Z. Deng, X. B. Li, Q. Zhou, and H. M. Yan, "Impact analysis of government investment on water projects in the arid Gansu Province of China," *Physics and Chemistry of the Earth, Parts A/B/C*, vol. 79, pp. 54–66, 2015.

[18] J. Ma, "On-grid electricity tariffs in China: development, reform and prospects," *Energy Policy*, vol. 39, no. 5, pp. 2633–2645, 2011.

[19] Y. Wu and L. Zhang, "Evaluation of energy saving effects of tiered electricity pricing and investigation of the energy saving willingness of residents," *Energy Policy*, vol. 109, pp. 208–217, 2017.

[20] P. R. Hartley, K. B. Medlock, and O. Jankovska, "Electricity reform and retail pricing in Texas," *Energy Economics*, vol. 80, pp. 1–11, 2019.

[21] Y. Yang and G. Meng, "A bibliometric analysis of comparative research on the evolution of international and Chinese ecological footprint research hotspots and frontiers since 2000," *Ecological Indicators*, vol. 102, pp. 650–665, 2019.

[22] J. Yuan, C. Zhao, S. Yu, and Z. Hu, "Electricity consumption and economic growth in China: cointegration and co-feature analysis," *Energy Economics*, vol. 29, no. 6, pp. 1179–1191, 2007.

[23] C. Deng, K. Li, C. Peng, and F. Han, "Analysis of technological progress and input prices on electricity consumption: evidence from China," *Journal of Cleaner Production*, vol. 196, pp. 1390–1406, 2018.

[24] B. Lin and C. Liu, "Why is electricity consumption inconsistent with economic growth in China?" *Energy Policy*, vol. 88, pp. 310–316, 2016.

[25] C. Zhang, K. Zhou, S. Yang, and Z. Shao, "Exploring the transformation and upgrading of China's economy using electricity consumption data: a VAR-VEC based model," *Physica A: Statistical Mechanics and Its Applications*, vol. 473, pp. 144–155, 2017.

[26] M. Li, R. Shan, M. Hernandez, V. Mallampalli, and D. Patiño-Echeverri, "Effects of population, urbanization, household size, and income on electric appliance adoption in the Chinese residential sector towards 2050," *Applied Energy*, vol. 236, pp. 293–306, 2019.

[27] S. Zeng, Z.-M. Chen, A. Alsaedi, and T. Hayat, "Price elasticity, block tariffs, and equity of natural gas demand in China: investigation based on household-level survey data," *Journal of Cleaner Production*, vol. 179, pp. 441–449, 2018.

[28] B. Ma, Y. Yu, and F. Urban, "Green transition of energy systems in rural China: national survey evidence of households' discrete choices on water heaters," *Energy Policy*, vol. 113, pp. 559–570, 2018.

[29] S. Wang, W. Li, H. Dincer, and S. Yuksel, "Recognitive approach to the energy policies and investments in renewable energy resources via the fuzzy hybrid models," *Energies*, vol. 12, p. 3645, 2019.

[30] M. R. Chertow, "The IPAT equation and its variants; changing views of technology and environmental impact," *Journal of Industrial Ecology*, vol. 4, no. 4, pp. 13–29, 2008.

[31] B. E. Hansen, "Threshold effects in non-dynamic panels: estimation, testing, and inference," *Journal of Econometrics*, vol. 93, no. 2, pp. 345–368, 1999.

[32] A. González, T. Teräsvirta, and D. Dijk, "Panel smooth transition regression models," *SSE/EFI Working Paper Series in Economics and Finance*, Econometric Institute, Erasmus University Rotterdam, Rotterdam, Netherlands, 2005.

Research Article

A Novel Hybrid Approach Based on BAT Algorithm with Artificial Neural Network to Forecast Iran's Oil Consumption

Mojtaba Bahmani [ID],[1] **Mehdi Nejati** [ID],[1] **Amin GhasemiNejad** [ID],[1] **Fateme Nazari Robati** [ID],[1] **Mehrdad Lashkary** [ID],[2] **and Naeeme Amani Zarin** [ID][1]

[1]*Shahid Bahonar University of Kerman, Kerman, Iran*
[2]*Allameh Tabataba'i University of Tehran, Tehran, Iran*

Correspondence should be addressed to Amin GhasemiNejad; ghaseminejad@aem.uk.ac.ir

In this paper, we develop a function of population, GDP, import, and export by applying a hybrid bat algorithm (BAT) and artificial neural network (ANN). We apply these methods to forecast oil consumption in Iran. For this purpose, an improved artificial neural network (ANN) structure, which is optimized by the BAT is proposed. The variables between 1980 and 2017 were used, partly for installing and testing the method. This method would be helpful in forecasting oil consumption and would provide a level playing field for checking the energy policy authority impacts on the structure of the energy sector in an economy such as Iran with high economic interventionism by the government. The result of the model shows that the findings are in close agreement with the observed data, and the performance of the method was excellent. We demonstrate that its efficiency could be a helpful and reliable tool for monitoring oil consumption.

1. Introduction

The efficient application of energy is a factor which can considerably affect the sustainable development of countries. Energy is a foundation element of the modern industrial economy. Over the past few decades, energy has become a very critical element in industries and also a fundamental product and factor in the growth of the economy in different world regions. The ever-increasing dependence of human life on energy has made this factor play a critically crucial role either potentially or actively in the functions of various economic sectors of countries [1]. Given the vital applications of energy in different economic sectors and also the rise of the population over past years, energy has turned the center of attention as one of the most significant factors of production [2]. Oil, natural gas, electric power, solar, wood, animal, and plant waste could be recognized as Iran's basic energy resource [3]. The oil industry is a fundamental factor in the economy of Iran and the government's annual budget. It affects foreign trade, the national capital, and developments in nonpetroleum exports [4].

Therefore, recognizing the factors that affect the needed energy in a country is required in managing the energy supplement. Based on the reality that the energy demand procedure and factors that affect it go through vague and complicated patterns, determining efficient instruments for appropriate energy use is necessary. Due to the increasing demand for energy, the assessment of it is necessary. Different optimization techniques, e.g., birds, bats, and fireflies, are appropriate to forecast these models [1]. Different models are presented to determine the possible approach ever in modeling energy issues as a function of various indicators. Given the high dependency of the Iranian economy on oil incomes, acquiring data by employing precious and qualified methods in this field will result in more efficient planning. Hence, the studies in this field have a considerable degree of importance. We examine the relationship between oil consumption in Iran and diverse independent variables, based on socioeconomic factors, using the BANN (hybrid bat algorithm and artificial neural network) technique. Results arising from the present study provide important reference information for the utility companies in pursuing patterns of

oil consumption and selecting a more accurate approach to estimate future oil demand.

A few studies that propose several models for energy demand policy management with different techniques are introduced. Pourdaryaei et al. [5, 6] in two studies worked on short-run electricity price forecasting using hybrid algorithms. Using a hybrid method for electricity price forecasting via artificial neural network and artificial cooperative search algorithm and using the hybrid backtracking algorithm and the adaptive neuro-fuzzy inference system (ANFIS) approach have improved the accuracy and effectiveness. Kaboli and Alqallaf [7] conducted a study on solving the problem of sending nonconvex economic load via an artificial algorithm for cooperative research. Thus, they used a method that is to interfere and work with feasible solutions throughout the optimization. Hlal et al. [8] conducted an optimization study based on NSGA-II and MOPSO for sizing a hybrid PV/wind/battery energy storage system. Kıran et al. [9] applied artificial bee colony (ABC) and particle swarm optimization (PSO) techniques to forecast electricity energy demand in Turkey in two forms (linear and quadratic) by using selected socioeconomic and demographic variables that included gross domestic product (GDP), population, import, and export. Behrang et al. [4] presented GSA (gravitational search algorithm) oil demand estimation models to be used in the Iran context. In another study, Behrang et al. [10] predicted the total energy demand in Iran from 2006 to 2030, using the bees algorithm technique. Ceylan et al. [11] developed the harmony search (HS) technique to forecast Turkish transport energy demand in tree forms (linear, exponential, and quadratic) based on three factors including population, GDP, and vehicle kilometers. The energy consumption in Turkey is also determined by using the ant colony (ACO) by Toksari [12] with independent variables such as population, GDP, import, and export.

According to what has been presented, using nonclassic methods for the identification and prediction of complex systems-related problems has been expanded [13]. Hence, in this study, a novel approach for oil consumption forecasting is presented. Heuristic methods are used for the better optimal solutions and to overcome the complexity of the nonlinearities [14]. The authors' main goal, therefore, is to employ artificial methods, because of their accuracy, and present new methods, such as the proposed hybrid, in the field of oil and gas, to acquire more precise information and data.

This paper is organized as follows. First, a brief introduction is presented. In the next section, we explain the specification of the BANN (hybrid bat algorithm and artificial neural network). Section 3 presents optimization and method validation. In Section 4, the model estimations are done, and finally, in the last section, the conclusion is given.

2. Methods and Material

2.1. Bat Algorithm (BA). The BAT algorithm was introduced by Yang [15]. This algorithm is an evolutionary algorithm inspired by the behavior of natural bats in locating their foods. It is used to solve various problems. BA starts with an initial population of bats. It employs a homologous manner by performing an echolocation method for updating its position

[16]. Bats send a signal with the loudness of frequency 20 kHz to 500 kHz. This signal is deflected back after striking the object to bat as an echo signal. The echo signal is used to calculate the bat's distance to any object which is the destination of the bat [17, 18]. Bats fly towards the object or prey, and they reduce their pulse rate when they reach the closer object. Bats continue to do so till the distance becomes zero [18, 19]. Also, they can distinguish between insects and obstacles to hunt prey based on the echo [20]. The BA is summarized in the following three main principal rules [15–21]:

(1) Each bat utilizes echolocation characteristics to estimate the distance and to distinguish between prey and obstacles.

(2) Each bat flies randomly with position x_i, velocity v_i with constant frequency f_{min}, updated wavelength λ, and loudness L_0 to seek for prey. It automatically changes the frequency of its pulse emission and regulates the rate of pulse release r in a range [0, 1] based on the proximity of its target.

(3) Frequency, loudness, and pulse released rate of each bat is varied. The loudness is adjusted from a large value (positive) L_0 to a minimum constant value L_{min}.

In the simulations, each bat is associated with a velocity v_i^t and a location x_i^t with iteration t, in a d-dimensional search or solution space. Therefore, the above three rules can be translated into the updating equations for x_i^t and velocities v_i^t:

$$
\begin{aligned}
f_i &= f_{min} + (f_{max} - f_{min})\beta, \\
v_i^t &= v_i^{t-1} + (x_i^t - x^*)f_i, \\
x_i^t &= x_i^{t-1} + v_i^t,
\end{aligned} \tag{1}
$$

where $\beta\varepsilon[0, 1]$ is a random vector outcome from a uniform distribution. x^* is the best current universal position, achieved after comparing the locations of all then bats. The frequency f during a range $[f_{min}, f_{max}]$ corresponds to a range of wavelengths $[\lambda_{min}, \lambda_{max}]$. Indeed, the frequency just varies while the wavelength f is constant, and it is assumed that $f \in [0, max]$ in implementation [22].

For the local search, a solution is selected among the best current solutions, and a new solution for each bat is produced using a random walk.

$$
x_{new} = x_{old} + \varepsilon L^t, \tag{2}
$$

where $\varepsilon\varepsilon[-1, 1]$ is a random number while L^t is the average loudness of all bats. The loudness normally reduces when a bat finds its prey while the rate of the pulse emission increases. The loudness may be chosen as any value of convenience. It can be chosen as any value of convenience between L_{min} and L_{max} assuming $L_{min} = 0$ which means that a bat has just found the prey and provisionally stops emitting sound. It is shown in the following:

$$
\begin{aligned}
L_i^{t+1} &= \alpha L_i^t, \quad \forall 0 \le \alpha \le 1, \\
r_i^{t+1} &= r_i^0 \left[1 - e^{-\gamma t}\right], \quad \forall \gamma > 0,
\end{aligned} \tag{3}
$$
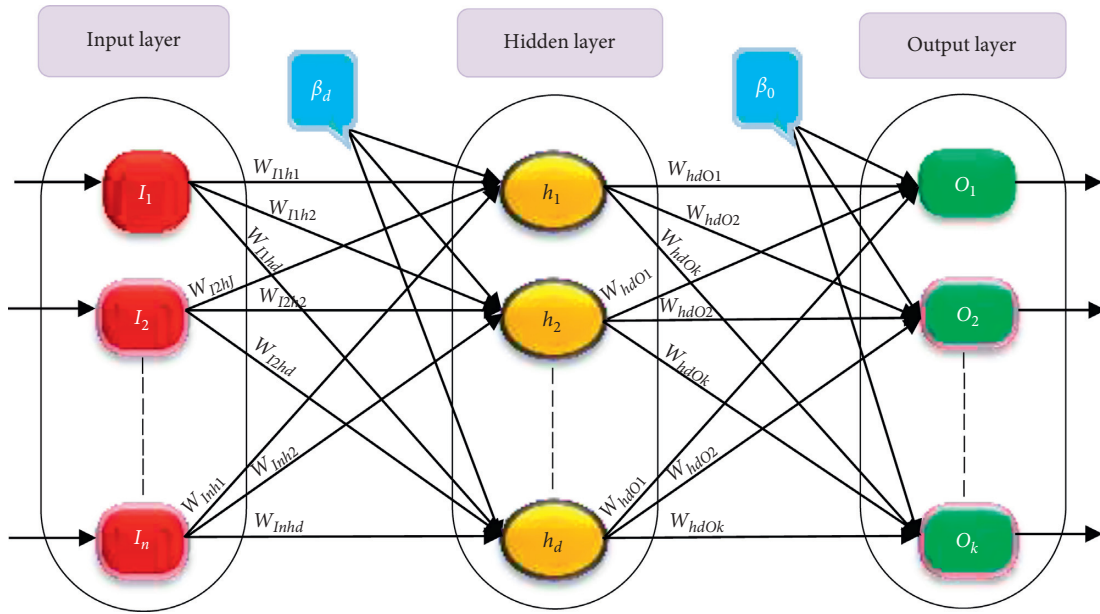
FIGURE 1: A three-layer artificial neural network.
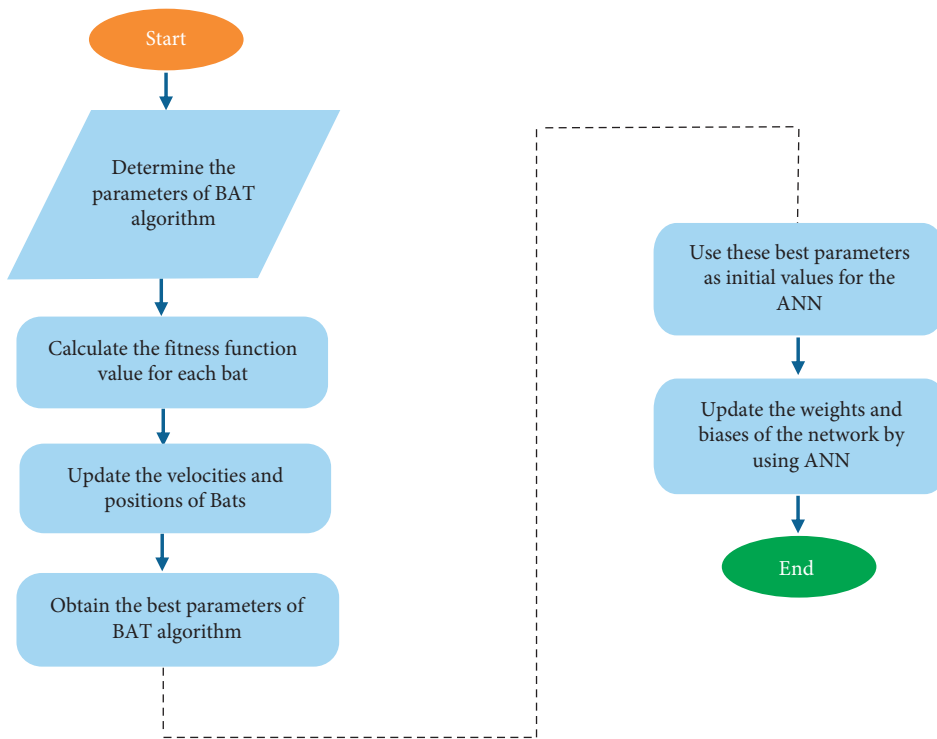
FIGURE 2: Flowchart of the BANN.

TABLE 1: Minimum and maximum values of research variables for normalization.

Variable	Maximum	Minimum
Per capita GDP (constant 2010 US$)	6952.444	3640.315
Population, total (million persons)	81	39
Import (a thousand barrels a day)	213	6.9
Export (a thousand barrels a day)	2867	911
Oil consumption (mboe)	99.889	28.434
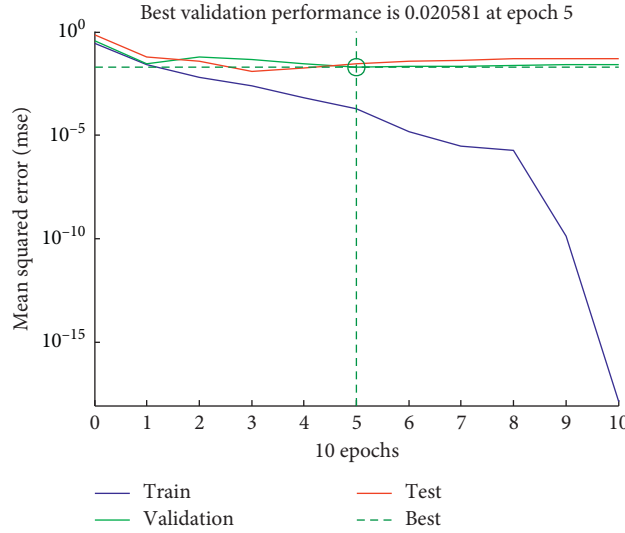
FIGURE 3: Validation performance graph.

where t is the iteration number during the optimization process. As time moves towards infinity, the loudness L_i^t tends to be zero; then, $r_i^t = r_i^0$.

2.2. Artificial Neural Network (ANN).

Artificial neural network is a famous artificial intelligent method that simulates the work of the human mind mechanism. It is a method that handles information coming from different nodes in this model known as neurons. These nodes are embedded in different layers which will work together to solve a complicated problem [23].

Artificial neural network is a mathematical approximation to the human nervous system that have been widely used to solve various nonlinear problems. The structure of ANN contains the elements of the input layer, hidden layer, and output layer. A network can contain multiple different hidden layers enabling the network to have computational and processing abilities. The number of layers is dependent on the problem complexity. In general, a neural network is a set of connected input and output units that each connection has an associated weight [13].

An artificial neural network (ANN) contains the elements of the input layer, hidden layer, and output layer. The ANN can contain multiple different hidden layers enabling the network to have computational and processing abilities.

Figure 1 shows an artificial neural network (ANN) with a hidden layer, which contains some weights connecting between layers. The output values are going to be calculated through the subsequent steps.

First, the sum of weights is calculated as follows:

$$S_j = \sum_{i=1}^{n} w_{ij} I_i + \beta_i, \tag{4}$$

where I_i is the input variable, w_{ij} is the weight between the input variable I_i and neuron j, and β_i is the input variable's bias term.

Second, neurons' output values in the hidden layers are produced from the received values of weighted summation

(equation (4)) by using an activation function. A popular choice of such a function is a sigmoid function as follows:

$$f_j(I) = \frac{1}{1 + e^{-S_j}}, \tag{5}$$

where f_j is the sigmoid function for neuron j and S_j is the sum of weights.

Finally, the output of neuron j is calculated as follows:

$$O_j = \sum_{i=1}^{k} w_{ij} f_j + \beta_j, \tag{6}$$

where O_j is the output of neuron j, w_{ij} is the weight between the output variable O_i and neuron j, f_j is the activation function for neuron j, and β_i is the output variable's bias term [24].

2.3. Proposed Method.

The goal of employing a bat algorithm and artificial neural network simultaneously is to present a higher performance estimation and forecast. When the bat algorithm is embedded into an ANN to optimize its initial weights and biases, it will play a well-suited role in addressing a number of the deficiencies caused by the randomness of the initial weights and biases of ANN. The main goal of using the BAT algorithm to optimize the ANN is to combine them, using the initial connection weights between ANN layers and the initial biases between neural nodes, to optimize the distribution, execute global searches within the solution space, and find the optimal initial weights and biases of the ANN at a rapid convergence rate. Subsequently, the initial weights and biases obtained by the ANN are often used for training and testing the sample set. The ANN with the optimal weights and biases is formed and trained to predict the output. The flowchart of the BANN algorithm is shown in Figure 2.

3. Optimization and Method Validation

Optimal tuning of PI controllers for load frequency controller design was proposed by Abd-Elazim and Ali
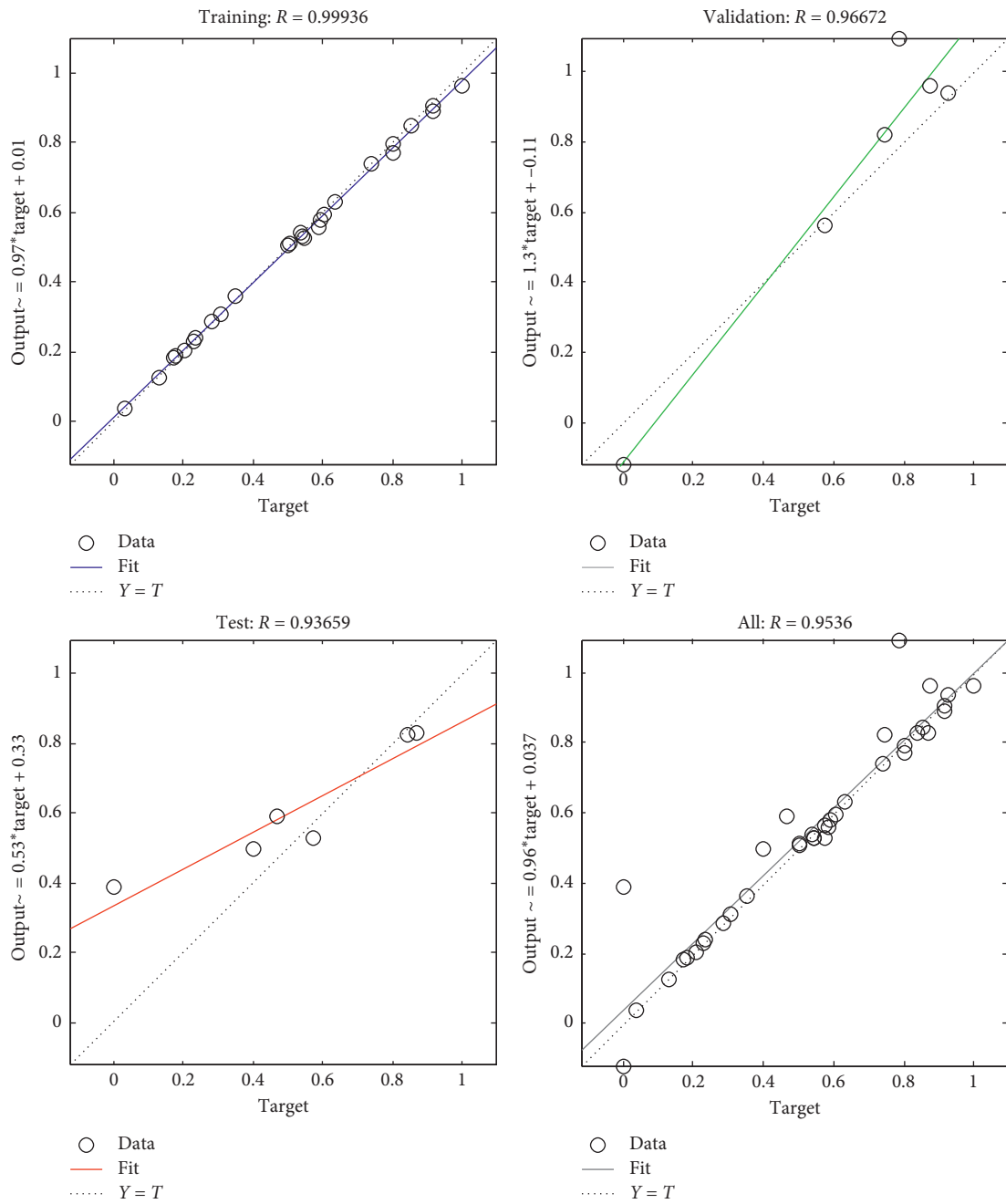
FIGURE 4: Regression plot.
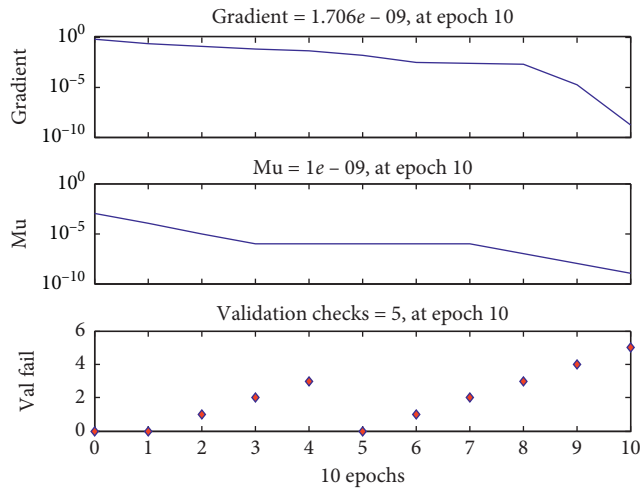
Figure 5: Optimized neural network output and model performance criteria for data.

Table 2: Neural network operation.

Neural network process	Data usage (%)	MSE	R
Training	70	0.0001973	0.999356
Validation	15	0.0205805	0.966715
Testing	15	0.0292405	0.936587
Total	100	0.0080010	0.953600

Table 3: Performance evaluation of neural network outputs.

MSE	RMSE	Std error
0.008001	0.089451	0.080401235
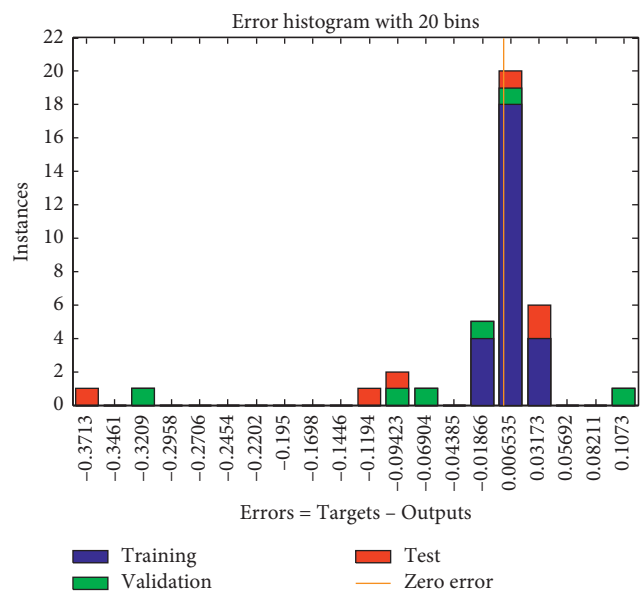
Figure 6: Error histogram for BANN outputs.
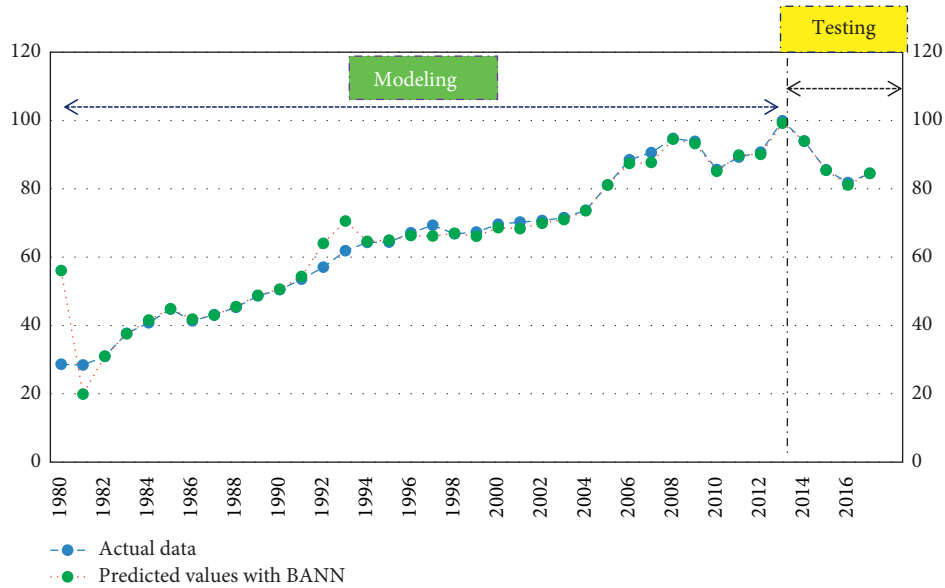
Figure 7: Comparison of actual oil consumption values and their predicted values with BANN.

Table 4: BANN operation.

Years	Actual data	BANN − output$_{predicted}$	Relative error %
2013	99.889	99.15401	0.007358
2014	93.947	93.93771	9.89E-05
2015	85.579	85.37274	0.00241
2016	81.795	81.12205	0.008227
2017	84.496	84.56049	0.000763
Average	–	–	0.003771546

Table 5: Comparison of various models introduced in the introduction and present study[a].

Source	Method	Target (country)	Average relative errors (%)
Toksarı [12]	Ant colony algorithm	Total energy (Turkey)	1.07
	Harmony search	Total energy (Turkey)	21.74
Ceylan et al. [11]	Harmony search	Total energy (Turkey)	13.41
	Harmony search	Total energy (Turkey)	39.32
	Genetic algorithm	Oil (Iran)	2.83
	Genetic algorithm	Oil (Iran)	1.72
Assareh et al. [3]	Particle swarm optimization	Oil (Iran)	1.4
	Particle swarm optimization	Oil (Iran)	1.36
	Gravitational search algorithm	Oil (Iran)	1.14
	Gravitational search algorithm	Oil (Iran)	1.52
Behrang et al. [4]	Gravitational search algorithm	Oil (Iran)	1.43
	Gravitational search algorithm	Oil (Iran)	3.32
	Gravitational search algorithm	Oil (Iran)	1.33
	Particle swarm optimization	Electricity (Turkey)	3.99
Kıran et al. [9]	Particle swarm optimization	Electricity (Turkey)	4.406
	Artificial bee colony	Electricity (Turkey)	3.20
	Artificial bee colony	Electricity (Turkey)	4.47
Present study	Hybrid bat algorithm with artificial neural network (BANN)	Oil (Iran)	0.0037

The average relative errors are separately based on the testing period of each model.

[25]. Their results of numerical simulation demonstrated the superiority of BA in comparison with simulated annealing in PI controller optimization. Bat algorithm which is a heuristic method is used for selecting features.

It is a metaheuristic algorithm. The process of optimization uses it. Function minimization or maximization is a major problem in optimization. Randomization is used for computing the best solution in the best optimum way

[26]. Different measures are applied to examine the estimation accuracy and forecasting ability of different estimators. The most popular ones are mean squared error (MSE) or root mean squared error (RMSE) [13] and correlation coefficient (R).

4. Empirical Results

To obtain oil consumption data, it is necessary to normalize in the first step. Equation (7) was used in this regard.

$$X_N = \frac{(X_R - X_{\min})}{(X_{\max} - X_{\min})},\qquad(7)$$

X_N: normalized value, X_R: the value to be normalized, $X_{\min}$: the minimum value in all the values for the related variable, and $X_{\max}$: the maximum value in all the values for the related variable. $X_{\min}$ and $X_{\max}$ values for each variable are selected between 1980 and 2012 and are shown in Table 1.

The input parameters of the BANN included population, gross domestic product, import, and export, and the oil consumption was considered as the BANN output parameter [1]. The data on these parameters were divided into training, testing, and data validation. 70% of these data were used for training, 15% of data were used for validation, and the rest were used for testing.

Figures 3–5 show the best validation performance graph and regression plot between actual and predicted data in the BANN method. Tables 2 and 3 and Figure 6 show the performance evaluation of BANN outputs.

Figure 7 and Table 4 for the modeling and the testing data show the performance of the BANN method.

Table 5 shows the comparison of different models introduced in the introduction and present study.

5. Conclusion

Demonstration of the significance of employing hybrid estimation strategies in the energy sector is the point of the present study. The relationship between economic development and energy consumption in developing countries is something essential, and it needs proper calculation for a wide variety of economic, social, and technological features. Therefore, in this study, the BANN (hybrid bat algorithm and artificial neural network) has been successfully used to estimate Iran's oil consumption based on the structure of Iran's socioeconomic conditions. Oil consumption is estimated based on population, GDP, import, and export. The proposed method anticipated oil consumption in terms of relative errors and RMSE.

Given the high dependency of the Iranian economy on oil incomes, acquiring data by employing precious and qualified methods in this field will result in more efficient planning. According to Table 5, the empirical results of Iran's data exhibit that the accuracy of the BANN method was more precise than that of the other methods. Regarding the results, oil consumption is influenced by population, GDP, import, and export. Hence, the findings proved that the recommended model was an appropriate tool for effective oil consumption prediction in Iran. It will provide a level playing field for checking the energy policy authority impacts on the structure of Iran's energy with high economic interventionism by the government.

The BANN success in such a study suggests that it could be applied as a practical instrument for energy-economic analysis in various areas, such as designing energy systems with more theoretical specifications complexity. Forecasting oil consumption can also be studied by other metaheuristics including harmony search and simulated annealing. The results of different methods could be put into analogy with the BANN method.

Data Availability

All data are available in the British Petroleum Company plc and BP Amoco plc [27].

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] S. A. Jalaee, A. GhasemiNejad, M. Lashkary, and M. Rezaee Jafari, "Forecasting Iran's energy demand using cuckoo optimization algorithm," *Mathematical Problems in Engineering*, vol. 2019, Article ID 2041756, 8 pages, 2019.

[2] M. Balat, "Energy consumption and economic growth in Turkey during the past two decades," *Energy Policy*, vol. 36, no. 1, pp. 118–127, 2008.

[3] E. Assareh, M. A. Behrang, M. R. Assari, and A. Ghanbarzadeh, "Application of PSO (particle swarm optimization) and GA (genetic algorithm) techniques on demand estimation of oil in Iran," *Energy*, vol. 35, no. 12, pp. 5223–5229, 2010.

[4] M. A. Behrang, E. Assareh, M. Ghalambaz, M. R. Assari, and A. R. Noghrehabadi, "Forecasting future oil demand in Iran using GSA (gravitational search algorithm)," *Energy*, vol. 36, no. 9, pp. 5649–5654, 2011.

[5] A. Pourdaryaei, H. Mokhlis, H. A. Illias, S. H. A. Kaboli, S. Ahmad, and S. P. Ang, "Hybrid ANN and artificial cooperative search algorithm to forecast short-term electricity price in de-regulated electricity market," *IEEE Access*, vol. 7, pp. 125369–125386, 2019.

[6] A. Pourdaryaei, H. Mokhlis, H. A. Illias, S. H. A. Kaboli, and S. Ahmad, "Short-term electricity price forecasting via hybrid backtracking search algorithm and ANFIS approach," *IEEE Access*, vol. 7, pp. 77674–77691, 2019.

[7] S. H. A. Kaboli and A. K. Alqallaf, "Solving non-convex economic load dispatch problem via artificial cooperative search algorithm," *Expert Systems with Applications*, vol. 128, pp. 14–27, 2019.

[8] M. I. Hlal, V. K. Ramachandaramurthya, S. Padmanaban et al., "NSGA-II and MOPSO based optimization for sizing of hybrid PV/wind/battery energy storage system," *International Journal of Power Electronics and Drive Systems (IJPEDS)*, vol. 10, no. 1, pp. 463–478, 2019.

[9] M. S. Kıran, E. ÖZceylan, M. GüNdüZ, and T. Paksoy, "Swarm intelligence approaches to estimate electricity energy demand in Turkey," *Knowledge-Based Systems*, vol. 36, pp. 93–103, 2012.

[10] M. A. Behrang, E. Assareh, M. R. Assari, and A. Ghanbarzadeh, "Total energy demand estimation in Iran using bees algorithm," *Energy Sources, Part B: Economics, Planning, and Policy*, vol. 6, no. 3, pp. 294–303, 2011.

[11] H. Ceylan, H. Ceylan, S. Haldenbilen, and O. Baskan, "Transport energy modeling with meta-heuristic harmony search algorithm, an application to Turkey," *Energy Policy*, vol. 36, no. 7, pp. 2527–2535, 2008.

[12] M. D. Toksarı, "Ant colony optimization approach to estimate energy demand of Turkey," *Energy Policy*, vol. 35, no. 8, pp. 3984–3990, 2007.

[13] S. A. Jalaee, M. Lashkary, and A. GhasemiNejad, "The Phillips curve in Iran: econometric versus artificial neural networks," *Heliyon*, vol. 5, no. 8, Article ID e02344, 2019.

[14] M. Modiri-Delshad, S. H. A. Kaboli, E. Taslimi, J. Selvaraj, and N. A. Rahim, "An iterated-based optimization method for economic dispatch in power system," in *Proceedings of the 2013 IEEE Conference on Clean Energy and Technology (CEAT)*, pp. 88–92, IEEE, Langkawi, Malaysia, November 2013.

[15] X.-S. Yang, "A new metaheuristic bat-inspired algorithm," *Nature Inspired Cooperative Strategies for Optimization (NICSO 2010)*, vol. 284, pp. 65–74, 2010.

[16] A. S. Oshaba, E. S. Ali, and S. M. Abd Elazim, "MPPT control design of PV system supplied SRM using BAT search algorithm," *Sustainable Energy, Grids and Networks*, vol. 2, pp. 51–60, 2015.

[17] M. A. Lones, "Metaheuristics in nature-inspired algorithms," in *Proceedings of the 2014 Conference Companion on Genetic and Evolutionary Computation Companion*, pp. 1419–1422, New York, NY, USA, July 2014.

[18] S. Mishra, K. Shaw, and D. Mishra, "A new meta-heuristic bat inspired classification approach for microarray data," *Procedia Technology*, vol. 4, pp. 802–806, 2012.

[19] G. Komarasamy and A. Wahi, "An optimized *K*-means clustering technique using bat algorithm," *European Journal of Scientific Research*, vol. 84, pp. 26–273, 2012.

[20] M. Bahmani, A. GhasemiNejad, F. N. Robati, and N. A. Zarin, "A novel approach to forecast global CO_2 emission using Bat and Cuckoo optimization algorithms," *MethodsX*, vol. 7, Article ID 100986, 2020.

[21] Y. Xin-She, "Bat algorithm for multi-objective optimization," *International Journal of Bio-Inspired Computation*, vol. 3, pp. 267–274, 2011.

[22] X. S. Yang and A. Hossein Gandomi, "Bat algorithm: a novel approach for global engineering optimization," *Engineering Computations*, vol. 29, no. 5, pp. 464–483, 2012.

[23] X. S. Yang and X. He, "Bat algorithm: literature review and applications," *International Journal of Bio-Inspired Computation*, vol. 5, no. 3, pp. 141–149, 2013.

[24] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The Bulletin of Mathematical Biophysics*, vol. 5, no. 4, pp. 115–133, 1943.

[25] S. M. Abd-Elazim and E. S. Ali, "Load frequency controller design via BAT algorithm for nonlinear interconnected power system," *International Journal of Electrical Power & Energy Systems*, vol. 77, pp. 166–177, 2016.

[26] E. S. Ali, "Optimization of power system stabilizers using BAT search algorithm," *International Journal of Electrical Power & Energy Systems*, vol. 61, pp. 683–690, 2014.

Research Article

The Improved Least Square Support Vector Machine Based on Wolf Pack Algorithm and Data Inconsistency Rate for Cost Prediction of Substation Projects

Haichao Wang,[1] Yi Liang ⬡,[2,3] Wei Ding,[1] Dongxiao Niu,[4] Si Li,[4] and Fenghua Wang[5]

[1]*Long Yuan (Beijing) Wind Power Engineering & Consulting Co., Ltd., Beijing 100034, China*
[2]*School of Management, Hebei Geo University, Shijiazhuang 050031, China*
[3]*Strategy and Management Base of Mineral Resources in Hebei Province, Hebei GEO University, Shijiazhuang 050031, China*
[4]*School of Economics and Management, North China Electric Power University, Beijing 102206, China*
[5]*State Grid Zhejiang Electric Power Company, Hangzhou 310007, China*

Correspondence should be addressed to Yi Liang; louisliang@hgu.edu.cn

Accurate and stable cost forecasting of substation projects is of great significance to ensure the economic construction and sustainable operation of power engineering projects. In this paper, a forecasting model based on the improved least squares support vector machine (ILSSVM) optimized by wolf pack algorithm (WPA) is proposed to improve the accuracy and stability of the cost forecasting of substation projects. Firstly, the optimal features are selected through the data inconsistency rate (DIR), which helps reduce redundant input vectors. Secondly, the wolf pack algorithm is used to optimize the parameters of the improved least square support vector machine. Lastly, the cost forecasting method of WPA-DIR-ILSSVM is established. In this paper, 88 substation projects in different regions from 2015 to 2017 are chosen to conduct the training tests to verify the validity of the model. The results indicate that the new hybrid WPA-DIR-ILSSVM model presents better accuracy, robustness, and generality in cost forecasting of substation projects.

1. Introduction

Poor control over the cost of substation projects easily leads to the high cost, which seriously affects the economics and sustainability of power engineering projects [1]. The forecasting of cost level is an important part of the cost control of substation projects, and it also has important guiding significance for the cost saving of substation projects. However, the attributes of historical data indicators are numerous due to the influence of factors such as the overall planning of the power grid, total capacity, topographical features, design and construction level, and the comprehensive economic level of the construction area. Simultaneously, the number of construction projects in the same period is limited, and it is impossible to collect more comparable engineering projects in a short time, which leads to less sample data and higher

difficulty for the cost forecasting of substation projects [2]. Therefore, the construction of cost forecasting model and the realization of accurate cost forecasting results of substation projects are of great significance to the sustainability of power engineering investment.

At present, few scholars have studied the cost forecasting of substation projects, but many have studied the cost forecasting of other engineering projects. The forecasting methods are mainly divided into two categories: one is the traditional forecasting method and the other is the modern intelligent forecasting method. The traditional forecasting methods mainly include time series prediction [3], regression analysis [4], Bayesian model [5], and fuzzy prediction [6]. A time series method for cost forecasting of projects based on the bill of quantities pricing model was built in [3], which could control the forecasting error within 5%. In [4], a

forecasting model based on the integral linear regression of multiple structures according to the features of building project cost was established, and the principal component factor method was introduced to solve the problem of multicollinearity among variables. In [5], a Bayesian project cost forecasting model that adaptively integrates preproject cost risk assessment and actual performance data into a range of possible project costs at a chosen confidence level is proposed. Gao et al. [6] utilized the significant cost theory to determine the significant cost factors affecting the cost and adopted the fuzzy c-means clustering method to predict the project cost. Basic theory and the verification way of this method are relatively mature and perfect, and the calculation process is also relatively simple. However, the method is often suitable for a single object, and the forecasting accuracy is not ideal.

Therefore, in the background of the rapid development of artificial intelligence technology, the use of intelligent forecasting methods to predict the cost of substation projects has a more important significance. Intelligent forecasting methods mainly include artificial neural networks (ANNs) and support vector machines (SVMs) [7]. The artificial neural networks, such as back propagation neural network (BPNN), radial basis function neural network (RBFNN), and general regression neural network (GRNN), are mainly used in the forecasting field [8]. A combination forecasting method for power grid engineering technical transformation projects was proposed in [9], which was based on Data Envelopment Analysis (DEA) and genetic BP neural network model. These factors were used as input variables of BP neural network to establish a 3-layer structure, and the forecasting model was proved by training the actual engineering data. In [10], a cost forecasting model of engineering projects based on BPNN and RBFNN was constructed, in which the radial basis function (RBF) neural network was proved to have higher forecasting accuracy on the basis of the training test with sample data from 55 engineering projects. In [11], a forecasting model for transmission line ice coating based on generalized regression neural network and fruit fly optimization algorithm is proposed and good forecasting results are achieved. However, we can find that the artificial neural network has the disadvantage of slow convergence rate and it is easy to fall into local optimization through the above research. As a result, the forecasting accuracy is greatly reduced. In [12], it is proposed that the SVM model could avoid the neural network structure selection and local minima problems. Therefore, the SVM model has been widely used in the research of projects cost forecasting. The influencing factors of substation projects cost were determined through literature review and other methods and the cost forecasting model of substation projects was established based on the theory of SVM in [13]. In [14], parameters of the SVM forecasting model were optimized by the improved firefly algorithm, which achieved a higher accuracy of the forecasting effect. Compared with the neural network, a better forecasting performance of substation projects cost can be obtained by applying the SVM model. According to the research mentioned above, it can be noted that there are some shortcomings in the

traditional support vector machine. When the traditional support vector machine is used for cost forecasting, the solution process needs to be converted into a quadratic programming process through the kernel function, which reduces the efficiency and the convergence precision is not high. However, the least squares support vector machine (LSSVM) avoids the quadratic programming process by using the least squares linear system as the loss function. Meanwhile, the forecasting problem is transformed into the equation sets and the inequality constraints are transformed into equality constraints by means of using the kernel function, which increases the accuracy and speed of the forecasting [15, 16]. Therefore, some scholars have tried to use LSSVM to carry out the cost forecasting of engineering projects [17, 18]. In [17], the variable importance in projection (VIP) was used to extract the features of multiple factors that affect the project cost. And, the LSSVM was used as the nonlinear approximator to establish the cost forecasting model of engineering projects. Liu [18] proposed a cost forecasting model of projects based on chaotic theory and least squares support vector machine, which aimed at the time-varying and chaotic of the cost change.

Although LSSVM shows the better forecasting performance than SVM in cost forecasting, there is still the problem of blind selection of the penalty parameter and the kernel parameter [19]. Therefore, it is necessary to select appropriate intelligent algorithms to optimize the parameters. The common intelligent algorithms mainly include genetic algorithm [20], ant colony algorithm [21], fruit fly optimization algorithm [22], and particle swarm algorithm [23]. Although the above algorithms all have their own advantages, there are still some corresponding flaws. The genetic algorithm has the problems of easy precocity, complicated calculation, small processing scale, difficult to deal with nonlinear constraints, poor stability, and so on. The ant colony algorithm and the firefly algorithm cannot guarantee convergence to the best point, and it is easy to fall into a local optimum, leading to a decrease in the forecasting accuracy. And, the particle swarm algorithm cannot fully satisfy the optimization of parameters in the LSSVM due to the insufficient local search accuracy. Based on the above analysis, the wolf pack algorithm (WPA) was applied to optimize the parameters of LSSVM in this paper. The performance of WPA will not be affected by a small change in parameters and the selection of parameters is relatively easy, so it has a good global convergence and computational robustness, which is suitable for solving the high-dimensional, multipeak complex functions, especially [24].

Otherwise, the number of influencing factors of substation projects cost is very large. If all the influencing factors are used as the input indicators of the forecasting model, then a large amount of redundant data appears, so feature selection is also important. The feature selection refers to the identification and selection of appropriate input vectors in the forecasting model to reduce redundant data and improve the computational efficiency [25]. The DIR model refers that the feature set is divided into several subsets, and the minimum inconsistency rate is calculated by the feature

subsets to determine the optimal feature subset and complete the feature selection [26]. Both [27, 28] adopted the DIR model for feature selection and obtained effective forecasting results. The use of the DIR model for feature selection can eliminate redundant features based on the data set inconsistency, and the characteristics of the correlation between features are also taken into account at the same time. The selected optimal feature can represent all data information perfectly because the relationship between features is not ignored. Therefore, the data inconsistency rate (DIR) was chosen for the feature selection in this paper.

According to the above research, an ILSSVM model integrated DIR with WPA is proposed. It is the first time to combine these three models in cost forecasting and several comparing methods are utilized to validate the effectiveness of the proposed hybrid model. The paper is organized as follows: Section 2 introduces the implementation process of DIR and ILSSVM optimized by FOA. Section 3 provides a case to validate the proposed model. Section 4 obtains the conclusion in this paper.

There are three innovations in this paper: (1) DIR is applied to the selection of influencing factors of substation project cost prediction; (2) LSSVM model is improved by applying the wavelet kernel function to replace the traditional radial basis kernel function; (3) LSSVM parameters are optimized by WPA.

2. Basic Theory

2.1. Wolf Pack Algorithm. Wolf pack algorithm (WPA) is derived from the study of wolf hunting behaviors. During the hunting process, different divisions of labor are assigned to wolves according to their respective roles. Wolves in the pack are divided into three types: the leader wolf, the scout wolf, and the ferocious wolf. And, these three wolves work together to complete the hunting process. The wandering behavior, summoning behavior and siege behavior of WPA are the three main behaviors, which are bionic based on the behavioral features of the wolves in the hunting process. Simultaneously, the generation of leader wolf and the replacement of wolf pack are followed by the rules of "survival of the fittest" and "winner-take-all," respectively [29]. The optimal solution is obtained through continuous iterations of the bionic behaviors. The bionic model of wolf pack algorithm is as shown in Figure 1.

The principle and steps of wolf pack algorithm are as follows:

(1) *Initialization of the Wolf Pack.* Suppose that there are N artificial wolves in the D-dimensional space and the location of the ith wolf is shown as follows:

$$X_i = (x_{i1}, x_{i2}, \ldots, x_{id}), \quad 1 \le i \le N, 1 \le d \le D. \quad (1)$$

The initial position is generated by the following equation:

$$x_{id} = x_{\min} + \text{rand} \times (x_{\max} - x_{\min}), \quad (2)$$

where rand is a uniformly distributed random number in [0, 1] and $x_{\max}$ and $x_{\min}$ are the upper and lower limits of the search space, respectively.

(2) *Generation of the Leader Wolf.* The wolf at the optimal position of the target function value is chosen as the leader wolf. The leader wolf neither updates the position of the hunting process nor participates in hunting activities, but directly iterates. If $Y_{\text{lead}} < Y_i$, then $Y_{\text{lead}} = Y_i$, namely, the scout wolf becomes a leader wolf at this time. Conversely, if the scout wolf i migrates towards h directions until the maximum H is obtained or location cannot be further optimized, the search stops at this moment. Among the h positions searched by the wolf i, the position of the jth point at the dth dimension is shown as follows:

$$y_{ijd} = y_{id} + \text{rand} \times \text{step}_a. \quad (3)$$

(3) *Close to the Prey.* The location update of the wolf pack is promoted by the summoning behavior of the leader wolf. Driven by the summoning behavior, the new location is obtained by wolves based on the summons of the leader wolf. The updated position of wolf i at the dth dimension is shown as follows:

$$z_{id} = x_{id} + \text{rand} \times \text{step}_b \times (x_{id} - x_{lid}), \quad (4)$$

where step_a is the length of the wolf's stride in the search process, step_b is the length of stride when the wolf moves to the target, x_{id} is the current position of the wolf i at the dth dimension, and x_{lid} is the position of the leader wolf at the dth dimension.

(4) *Encirclement of the Prey.* After discovering the prey, the surrounding wolves will complete the reclamation of the target prey according to the signals issued by the leader wolf. Equations of the reclamation and the reclamation steps are as follows:

$$X_i^{t+1} = \begin{cases} X_i^t, & r_m < \theta, \\ X_i + \text{rand} \times ra, & r_m > \theta, \end{cases} \quad ra(t) = ra_{\min}$$

$$\times (x_{\max} - x_{\min}) \times e^{((\ln(ra_{\min}/ra_{\max}))/\max t)}, \quad (5)$$

where t is the number of iterations, ra is the length of the wolf's stride during the reclamation, X_i is the position of the leader that issued the signals, and X_i^t is the position of the wolf i in the tth iteration.

(5) *Update Mechanism of Wolves Competition.* Wolves will be eliminated without food in the process of reclamation hunting; instead, the wolves that survived are the first to obtain food. Weak wolves that without food will be eliminated while an equal
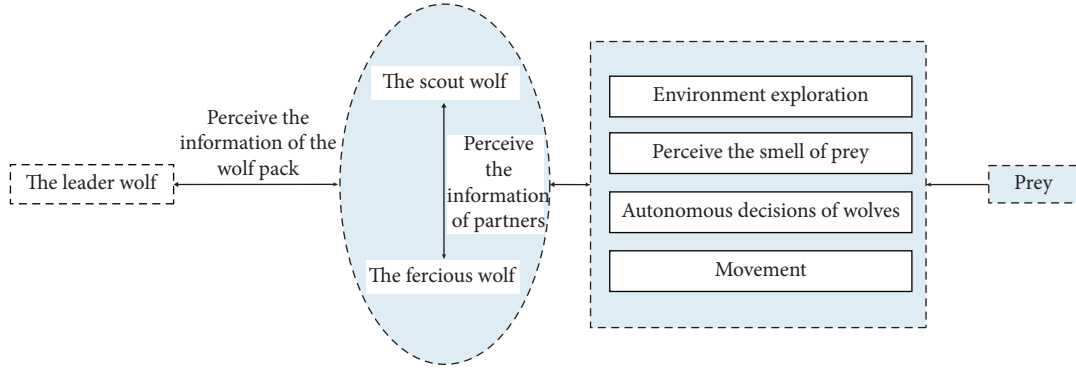
Figure 1: Bionic graph of Wolf pack algorithm.

number of new wolves that are randomly generated will be added into the wolf pack.

2.2. Data Inconsistency Rate (DIR).

The feature selection of many historical cost data of substation projects takes aim at distinguishing the most relevant data features, which makes the input vector of the power projects cost forecasting model with a strong pertinence and reduces the redundancy of the input information in order to improve the accuracy of the cost forecasting of the substation projects. The inconsistency rate of data can accurately describe the discrete characteristics of the input features. Different feature characteristics can be obtained by different division modes and different division modes can obtain different frequency distributions. The discrimination ability of data categories can be distinguished by the inconsistency rate of calculation. The higher the inconsistency rate of data, the stronger the classification ability of feature vectors.

It is necessary to know the specific calculation formula of the inconsistency rate for selecting the features by means of the inconsistency rate method. Therefore, it is assumed that the collected cost data of substation projects have g features, such as main transformer capacity, floor space, and main transformer unit price. which are, respectively, represented by the values of $G1, G2, \ldots, Gg$. Γ is the feature set and L is a subset of the feature set. Next, set the standard class M with c categories and N data instances. The feature value corresponding to the feature Fi is represented as z_{ji}. λ_i is the class value of M, then the data instance can be represented as $[z_j, \lambda_i]$, where $z_j = [z_{j1}, z_{j2}, z_{j3}, \ldots, z_{jg}]$. The calculation formula of the inconsistency rate is shown as follows:

$$\tau = \frac{\sum_{k=1}^{p} \left(\sum_{l=1}^{c} f_{kl} - \max_l \{f_{kl}\} \right)}{N}, \tag{6}$$

where f_{kl} is the number of data instances that centrally belongs to the feature subset of the x_k mode. x_k is the data set with a total of P feature partition modes, where $(k = 1, 2, \ldots, p$ and $p \leq N)$. The feature selection on the basis of the inconsistency rate is shown as follows:

(1) Initialize the optimal feature subset Γ to an empty set.

(2) Calculate the inconsistency rate of datasets $G1, G2, \ldots, Gg$ which belongs to the subset mode that consists of the subset Γ and each remaining feature.

(3) Calculate the inconsistency rate statistical table of the feature subsets and rank the inconsistent data from small to large.

(4) Select the feature subset L with the smallest number of features. If $\tau_L \approx \tau_\Gamma$ or $\tau_{L'}/\tau_L$ is the minimum of all adjacent inconsistency rate, then the optimal feature subset is L, where L' is the last feature subset adjacent to L.

By calculating the inconsistency rate, on the one hand, redundant features can be eliminated based on the inconsistency of data set. On the other hand, it can also take into account the correlation between features in the selection process, which has a better presentation of all data information through the selected optimal features.

2.3. Improved Least Squares Support Vector Machine (LSSVM)

2.3.1. Least Squares Support Vector Machine.

Least squares support vector machine (LSSVM) is an extension of support vector machine (SVM). It constructs the optimal decision surface by projecting the input vector into the high-dimensional space nonlinearly. Next, the inequalities of SVM are inverted into the equation sets by applying the risk minimization principle, which reduces complexity and speeds up the rate of calculation [30].

Suppose that $T = \{(x_i, y_i)\}_{i=1}^{N}$ is the given sample set, N is the total number of samples. The regression model of the sample is shown as follows:

$$y(x) = w^T \cdot \phi(x) + b, \tag{7}$$

where $\phi(*)$ is the function that projects training samples into a high-dimensional space, w is the weighted vector, and b is the offset parameter.

The optimization problem of LSSVM can be converted into the following function to solve [31]:

$$\min \frac{1}{2} w^T w + \frac{1}{2} \gamma \sum_{i=1}^{N} \xi_i^2 \tag{8}$$

$$\text{s.t.} \quad y_i = w^T \phi(x_i) + b + \xi_i, \quad i = 1, 2, 3, \ldots, N,$$

where γ is the penalty coefficient, which is used to balance the complexity and accuracy of the model, and ξ_i is the estimation error. The above equations can be solved by converting them into the Lagrangian function, which is detailed in 2.3.3.

2.3.2. Improved Method of LSSVM

(1) Horizontal Weighting of Input Vectors. The cost forecasting of substation projects is mostly a multiinput and single-output model. The values of the input vectors are horizontally distributed with the item serial number. And, the influence of the actual value of the substation projects cost influencing factors on the final forecasting value can be reflected by means of the weighted processing. Therefore, the weighted processing of input vectors is shown as follows:

$$\widehat{x}_i = x_{ki} \cdot \delta (1 - \delta)^{n-i}, \quad k = 1, 2, \ldots, l, \tag{9}$$

where $\widehat{x}_i$ is the weighted input vector, x_{ki} is the original input vector, k is the dimension number of input vector, and δ is a constant.

(2) Longitudinal Weighting of the Training Sample Sets. The cost forecasting value is not only related to the elements in the input vectors but also has a certain correlation with the sample groups, which means that the close sample has a greater influence on the forecasting value and the long-range sample has less influence on the forecasting value. Therefore, it is necessary to reduce the influence of close samples on the forecasting model by assigning different degrees of subordination to the influencing factors of the current substation projects cost and also enhance the influence of the long-range samples on the forecasting model at the same time. The linear membership degree μ_i is used to calculate the degree of the given membership and the equation is shown as follows [32]:

$$\mu_i = \frac{\beta + i(1 - \beta)}{N}, \quad 0 \le \mu_i \le 1, \tag{10}$$

where μ_i is the degree of membership and β is a constant between [0, 1] and $i = 1, 2, \ldots, N$.

Then the input sample set can be changed as follows:

$$T = \{(x_1, y_1, \mu_1) \ (x_2, y_2, \mu_2) \ \cdots \ (x_N, y_N, \mu_N)\}. \tag{11}$$

The determination of β affects the fitting performance of LSSVM directly [33]. The value of β can be obtained through the gray correlation coefficient, and the calculation formulas are shown as follows:

$$r(x_0(k), x_i(k)) = \frac{\Delta_{ki}(\min) + \rho \Delta_{ki}(\max)}{\Delta_{ki} + \rho \Delta_{ki}(\max)},$$

$$\Delta_{ki} = |x_0(k) - x_i(k)|, \quad \rho \in [0, 1], \tag{12}$$

$$\beta_i = \sum_{k=1}^{N} r(x_0(k), x_i(k)).$$

In this paper, $x_0 = Y, Y = \{y_1, y_2, \ldots, y_N\}$ since the forecasting model for cost prediction of substation projects is usually a multiinput and single-output model.

2.3.3. Weighted Least Squares Support Vector Machine.
The improvement in the above section is applied to the LSSVM to get the weighted least squares support vector machine. The objective function is described as follows:

$$\min \frac{1}{2} w^T w + \frac{1}{2} \gamma \sum_{i=1}^{N} \mu_i \xi_i^2 \tag{13}$$

$$\text{s.t.} \quad y_i = w^T \phi(x_i) + b + \xi_i, \quad i = 1, 2, 3, \ldots, N.$$

To solve the above problem, the Lagrangian function is established as follows:

$$L(w, b, \xi_i, \alpha_i) = \frac{1}{2} w^T w + \frac{1}{2} \gamma \sum_{i=1}^{N} \mu_i \xi_i^2$$

$$- \sum_{i=1}^{N} \alpha_i [w^T \phi(x_i) + b + \xi_i - y_i], \tag{14}$$

where α_i is the Lagrange multiplier. The variables of the function are deduced and the derivation is equivalent to zero. The specific calculation is shown as follows:

$$\begin{cases} \dfrac{\partial L}{\partial w} = 0 \longrightarrow w = \sum_{i=1}^{N} \alpha_i \phi(x_i), \\[2ex] \dfrac{\partial L}{\partial b} = 0 \longrightarrow \sum_{i=1}^{N} \alpha_i = 0, \\[2ex] \dfrac{\partial L}{\partial \xi} = 0 \longrightarrow \alpha_i = \gamma \mu_i \xi_i, \\[2ex] \dfrac{\partial L}{\partial \alpha} = 0 \longrightarrow w^T + b + \xi_i - y_i = 0. \end{cases} \tag{15}$$

Convert the equations to the following problem by eliminating w and ξ_i:

$$\begin{bmatrix} 0 & e_n^T \\ e_n & \Omega + \gamma^{-1} \mu^{-1} \cdot I \end{bmatrix} \cdot \begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix}, \tag{16}$$

where $\Omega = \phi^T(x_i)\phi(x_i)$, $e_n = [1, 1, \ldots, 1]^T$, $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_n]$, and $y = [y_1, y_2, \ldots, y_n]^T$.
The equation (17) is obtained as follows:

$$y(x) = \sum_{i=1}^{N} \alpha_i K(x_i, x) + b, \quad (17)$$

where $K(x_i, x)$ is the kernel function. The wavelet kernel function $K(x_i, x) = \prod_{i=1}^{N} \Psi((x_i - x_i')/\sigma_i)$ is selected to replace the Gaussian kernel function in the standard least squares support vector machine and the construction of the wavelet kernel function will be detailed in the next section. The wavelet kernel function is brought into $y(x)$ as follows:

$$y(x) = \sum_{i=1}^{N} \alpha_i \prod_{i=1}^{N} \Psi\left(\frac{x_i - x_i'}{\sigma_i}\right) + b,$$

$$(18)$$

$$\psi(x) = \cos(1.75x) \cdot \exp\left(\frac{-x^2}{2}\right).$$

Finally, the regression model of the weighted least squares support vector machine is shown as follows:

$$y(x) = \sum_{i=1}^{N} \alpha_i \prod_{i=1}^{N} \left\{ \cos\left[\frac{1.75(x_i - x_i')}{\sigma_i}\right] \cdot \exp\left[\frac{-(x_i - x_i')^2}{2}\right] \right\} + b.$$

$$(19)$$

In this paper, the wavelet kernel function was used to replace the traditional radial basis kernel function, which is mainly based on the following considerations. (a) The wavelet kernel function has the excellent specialty of describing the data information step by step, and the LSSVM that uses the wavelet kernel function as the kernel function can simulate any function with high precision. However, the traditional Gaussian function is relatively less effective. (b) The wavelet kernel function is orthogonal or approximately orthogonal, while the traditional Gaussian kernel function is related or even redundant. (c) The wavelet kernel function can analyze and process the multiresolution of wavelet signals. Therefore, the nonlinear processing ability of the wavelet kernel function is better than the Gaussian kernel function, which can improve the generalization ability and robustness of the LSSVM regression model.

2.3.4. Construction of the Wavelet Kernel Function. The kernel function of LSSVM is the inner product of two input spatial data points in a spatial feature, and it has two obvious features. Firstly, $k(x, x') = k(x', x)$ is the symmetric function of inner product kernel variables. Secondly, the sum of the kernel functions on the same plane is a constant. In short, only when the kernel function satisfies the following two theorems can it become the kernel function of the least squares support vector machine [34, 35].

(1) Mercer's Theorem. $k(x, x')$ is the continuous symmetric core that can be extended to the following form:

$$k(x, x') = \sum_{i=1}^{\infty} \lambda_i g_i(x) g_i(x'), \quad (20)$$

where λ_i is a positive value. The following necessary and sufficient conditions need to be met to ensure complete convergence of the above extensions.

$$\iint k(x, x') g(x) g(x) dx dx' \geq 0, \quad x, x' \in R^n, \quad (21)$$

For all the functions of $g(*)$, the condition that

$$\begin{cases} g(*) \neq 0 \\ \int_{R^n} g^2(\xi) d\xi < \infty \end{cases} \text{ needs to be satisfied.}$$

Where $g(x_i)$ is the feature function, λ_i is the eigenvalue, and all of them are positive. Therefore, it is known that the kernel function $k(x, x')$ is a positive definite function.

(2) Smola and Scholkopf Theorem. When the kernel function satisfies Mercer's theorem, $k(x, x')$ can be used as the kernel equation of the least square support vector machine when it is proved to satisfy the following equation (22):

$$F(x)(\omega) = (2\pi)^{-(n/2)} \int_{R^n} \exp(-J(\omega \cdot x)) k(x) dx \geq 0, \quad x \in R^n.$$

$$(22)$$

(3) Construction of the Wavelet Kernel Function. When the wavelet kernel equation satisfies conditions that $\psi(x) \in L^2(R) \cap L'(R)$ and $\widehat{\psi}(x) = 0$, $\widehat{\psi}(x)$ is the Fourier transform of $\psi(x)$. Then, $\psi(x)$ can be defined as follows [36]:

$$\psi_{\sigma,m}(x) = (\sigma)^{(1/2)} \psi\left(\frac{x - m}{\sigma}\right), \quad x \in R, \quad (23)$$

where σ is the shrinkage coefficient, m is the horizontal float coefficient, and $\sigma > 0, m \in R$.

When $f(x)$ satisfies $f(x) \in L^2(R)$, $f(x)$ can be wavelet transformed as follows:

$$W(\sigma, m) = \sigma^{(1/2)} \int_{-\infty}^{+\infty} f(x) \psi^*\left(\frac{x - m}{\sigma}\right) dx, \quad (24)$$

where $\psi^*(x)$ is the complex conjugate function of $\psi(x)$. The wavelet transform function $W(\sigma, m)$ is reversible and can be used for reconstructing the original signals, and then the following formula can be obtained:

$$f(x) = C_\varphi^{-1} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} W(\sigma, m) \psi_{\sigma,m}(x) \frac{d\sigma}{\sigma^2} dm, \quad (25)$$

where $C_\varphi = \int_{-\infty}^{+\infty} (|\widehat{\psi}(w)|^2/|w|) dw < \infty$ and $\widehat{\psi}(w) = \int \psi(x) \exp(-Jwx) dx$. where C_φ is a constant. The wavelet decomposition theory is an infinite approximation to a set of functions based on linear combination of the wavelet functions. Suppose that $\psi(x)$ is a one-dimensional function, then the multidimensional wavelet equation can be described according to the tensor theory as follows:

$$\psi_l(x) = \prod_{i=1}^{l} \psi(x_i), x \in R^{l \times d}, \quad x_i \in R^d. \quad (26)$$

In this way, the horizontal floating kernel function is constructed as follows:

$$k(x, x') = \prod_{i=1}^{l} \psi\left(\frac{x_i - x_i'}{\sigma_i}\right), \quad \sigma_i > 0. \tag{27}$$

The kernel equation needs to satisfy the Fourier transform in the least squares support vector machine. Therefore, the wavelet kernel function can be used as the kernel function of LSSVM when it satisfies the Fourier transform. It is necessary to prove that the following equation is established:

$$F(k)(w) = (2\pi)^{(-l/2)} \int_{R^l} \exp(-J(wx))k(x)\mathrm{d}x \geq 0. \tag{28}$$

In this paper, the Morlet wavelet mother function is chosen to prove the above equation, which ensures the generalization of the wavelet kernel function, namely,

$$\psi(x) = \cos(1.75x)\exp\left(-\frac{x^2}{2}\right). \tag{29}$$

And, $k(x, x')$ can be expressed as follows:

$$k(x, x') = \prod_{i=1}^{l} \left\{ \cos\left[1.75\left(\frac{\|x_i - x_i'\|^2}{2\sigma^2}\right)\right] \exp\left[-\frac{\|x_i - x_i'\|^2}{2\sigma^2}\right] \right\}$$

$$= \prod_{i=1}^{l} \cos\left(\frac{1.75X_i}{\sigma}\right) \exp\left(-\frac{\|x_i\|^2}{2\sigma^2}\right), \tag{30}$$

where σ can be obtained through the sample fitting. $x \in R^N$ and $\sigma, x_i \in R^N$. It can be seen from the above equation that the multidimensional wavelet function can be used as a kernel function of the multidimensional least squares support vector machine. And, the mathematical proof process is shown as follows:

$$\int_{R^l} \exp(-Jwx)k(x)\mathrm{d}x$$

$$= \int_{R^l} \exp(-Jwx) \cdot \left(\prod_{i=1}^{l}\left\{\cos\left[1.75\left(\frac{x_i - x_i'}{\sigma}\right)\right]\exp\left(-\frac{\|x_i - x_i'\|^2}{2\sigma^2}\right)\right\}\mathrm{d}x\right)$$

$$= \prod_{i=1}^{l} \int_{R^l} \exp(-Jw_ix_i) \cdot \left(\frac{\exp(j1.75x_i/\sigma) + \exp(-1.75x_i/\sigma)}{2}\right) \cdot \exp\left(-\frac{\|x_i\|^2}{2\sigma^2}\right)\mathrm{d}x_i \tag{31}$$

$$= \prod_{i=1}^{l} \frac{1}{2}\int_{R^l} \left\{\exp\left[-\frac{\|x_i\|^2}{2\sigma^2} + \left(\frac{1.75j}{\sigma} - jw_i\sigma\right) \cdot x_i\right] + \exp\left[-\frac{\|x_i\|^2}{2\sigma^2} - \left(\frac{1.75j}{\sigma} - jw_i\sigma\right) \cdot x_i\right]\right\}\mathrm{d}x_i$$

$$= \prod_{i=1}^{l} \frac{|\sigma|\sqrt{2\pi}}{2}\left\{\exp\left(-\frac{(1.75 - w_i\sigma)^2}{2}\right) + \exp\left(-\frac{(1.75 + w_i\sigma)^2}{2}\right)\right\}.$$

Finally, the equation is obtained as follows:

$$F(x)(w) = \prod_{i=1}^{l} \left(\frac{|\sigma|}{2}\right)\left(\exp\left(-\frac{(1.75 - w_i\sigma)^2}{2}\right)\right.$$

$$\left. + \exp\left(-\frac{(1.75 + w_i\sigma)^2}{2}\right)\right), \tag{32}$$

where $|\sigma| \neq 0$, so $F(x)(w) \geq 0$. In addition, it can be concluded that the wavelet kernel function can be used as the kernel function of least squares support vector machine.

3. DIR-ILSSVM Optimized by the WPA Algorithm

The cost forecasting model of substation projects combining WPA, DIR, and ILSSVM is constructed as illustrated in Figure 2. As is shown from the figure, the cost forecasting model of substation projects proposed in this paper mainly includes three sections. The first section is the feature selection based on the inconsistency rate. The second section is the sample training based on the ILSSVM model. The third section is the cost forecasting based on the ILSSVM model. When the established feature subset L cannot satisfy stopping conditions of the algorithm, the program will continue to loop until it reaches the desired precision, and an optimal feature subset is output. Therefore, in the cost forecasting model of the substation projects constructed in this paper, the purpose of the first section is to find the optimal feature subset and the optimal regression model parameters by iterative calculations. The purpose of the second section is to calculate the forecasting accuracy of the training sample in each iteration, so that the fitness function value can be calculated. The third section will make use of the optimal feature subsets and parameters obtained in the above two sections. And, finally, the substation projects cost of the test sample can be forecasted through retraining of the ILSSVM regression model.
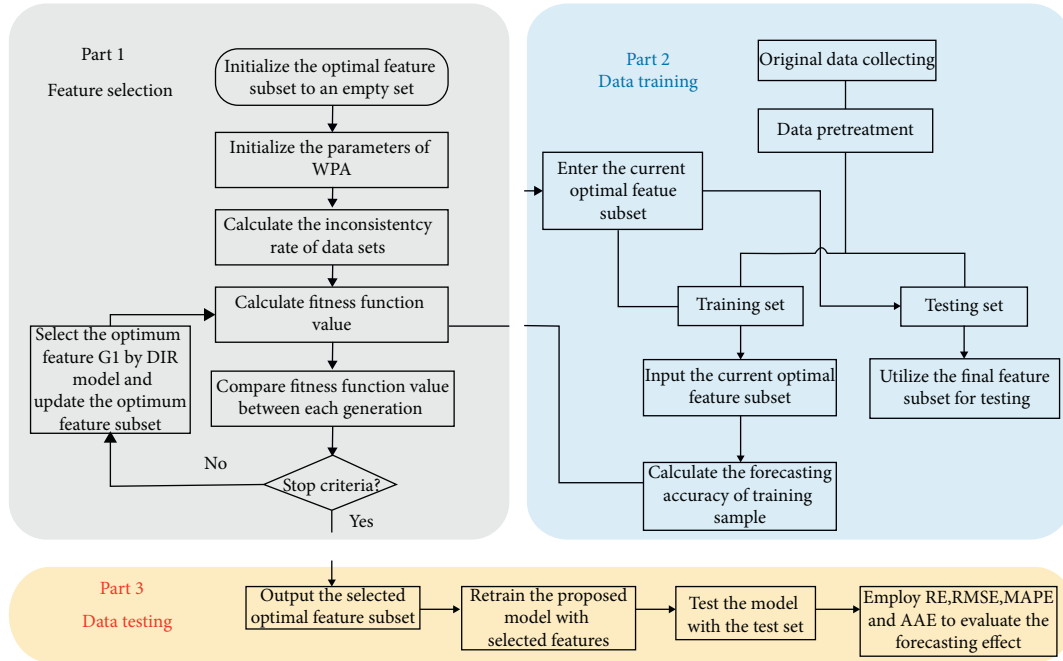
Figure 2: The flow chart of WPA-DIR-ILSSVM.

The specific steps for cost forecasting of substation projects are listed as follows:

(1) *Determine the Initial Candidate Feature Values.* By combing the related references [37–39], in this paper, the candidate features of the influencing factors of substation projects cost are selected as follows, including the floor space, construction properties, substation voltage level, main transformer capacity, number of outlets on the high-voltage side, number of outlets on the low-voltage side, topography, duration, substation type, number of transformers, economic development level of the construction area, inflation rate, transformer single unit price, high-side circuit breaker unit price, high-side circuit breaker unit, low-voltage capacitor quantity, high-voltage fuse price, current transformer price, power capacitor price, reactor price, power bus price, arrester price, measuring instrument price, relay protection device price, signal system price, automatic device price, site leveling fee, foundation treatment fee, designer's skill level, number of accidents, deviation rate of project volume, construction progress level, and days of rain and snow. In the IR algorithm, the optimal feature subset needs to be initialized to an empty set, namely, $\Gamma =$.

(2) *Initialize the Parameters of WPA.* Set the total number of wolf $N = 50$, the number of iterations $t = 100$, $\text{step}_a = 1.5$, $\text{step}_b = 0.8$, $q = 6$, and $h = 5$.

(3) *Calculate the Inconsistency Rate.* After step (1) and step (2) are completed, the candidate features are gradually substituted into the IR feature selection model. Calculate the inconsistency rate of datasets $G_1, G_2, \cdots, G_g$ which belongs to the subset mode that consists of the subset Γ and each remaining feature. The feature G_i corresponding to the minimum inconsistency rate is selected as the optimal feature. $\Gamma = \{\Gamma, G_i\}$ is the updated optimal feature set.

(4) *Determine the Optimal Feature Subset and Parameters of the Optimal Regression Model.* The current feature subset is substituted into the ILSSVM model to calculate the forecasting accuracy $r(j)$ in the learning process of the current cycle training sample, and the fitness function value Fitness(j) of each cycle can be obtained. The optimal feature subset can be determined through the comparison of fitness function values between each generation. Determine whether each iteration reaches the stopping condition of the algorithm. If it is not satisfied, the new feature subset is reinitialized and the new cycle is entered until the optimal subset of the global optimal feature is obtained. It is important to note that the parameters of the ILSSVM model also need to be optimized and the initial value of γ and σ will be allocated randomly. The fitness function based on double factors of forecasting accuracy and feature selection quantity is constructed as follows:

$$\text{fitness}(j) = -\left(a + r(j) + b \times \frac{1}{\text{Numfeature}(j)}\right),$$
(33)

where Numfeature(x_i) is the optimal number of features selected for each iteration and a and b are constants between [0, 1]. The optimal feature quantity and the fitness function value are proportional to each other in each iteration, and the cost

forecasting accuracy of substation projects is inversely proportional to the fitness function value.

(5) *Stop Optimization and Start the Cost Forecasting.* Circulation ends at the maximum number of iteration. Here, the optimal feature subset and the best value of γ and σ can be substituted into ILSSVM model for cost forecasting of substation projects.

4. Case Study

4.1. Collection and Processing of Data. The relevant cost data of substation projects from 2015 to 2017 in different regions is collected in this paper, including 88 voltage level substation projects. The cost levels and the influencing factors of the first 66 substation projects are used as the training set, and the last 22 data sets are used as the test set. The collected original data of the substation projects cost are shown in Table 1.

The explanation of the processing of input indicators is shown as follows. The data of the construction properties of substation projects are mainly divided into three categories, in which the assignment of new substation is 1, the main transformer is 2, and t the interval project is 3. The data of the substation types are mainly divided into three categories, in which the assignment of the indoor type is 1, the half indoor type is 2, and the outdoor type is 3. The topography and landform are mainly divided into the following eight situations, in which the assignment of hillock is 1, slope is 2, plain is 3, flat ground is 4, paddy is 5, dry land is 6, mountain is 7, and muddy land is 8. The level of economic development in the construction area is based on the data of the local GNP and the technical level of the designers is based on the proportion of employees with bachelor degree or above in this project. The difference between the actual progress of the construction and the scheduled progress plan is selected to represent the construction progress level. And, the data are normalized according to the following formula (34):

$$Y = \{y_i\} = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}, \quad i = 1, 2, 3, \ldots, n, \quad (34)$$

where x_i is the actual value, $x_{\min}$ and $x_{\max}$ are the minimum and maximum values of the sample data, and y_i is the normalized value.

4.2. Evaluation Indicators of the Forecasting Results. Evaluation indicators of the cost forecasting results of substation projects adopted in this paper are shown as follows.

(1) Relative error (RE):

$$\text{RE} = \frac{x_i - \widehat{x}_i}{x_i} \times 100\%. \quad (35)$$

(2) Root-mean-square error (RMSE):

TABLE 1: The collected original data of the substation project cost (unit: yuan/kV·A).

Serial number	Cost
1	299.3
2	285.6
3	305.5
4	310.3
5	405.6
6	358.3
7	256.9
8	305.8
9	356.9
10	1058.6
11	501.2
12	208.6
13	356.2
14	401.5
15	378.6
16	369.5
17	301.6
18	325.8
19	337.1
20	368.9
21	370.2
22	450.1
23	980.6
24	286.8
25	279.5
26	308.6
27	312.8
28	315.9
29	364.2
30	372.5
31	383.9
32	295.6
33	270.2
34	260.8
35	239.3
36	381.7
37	406.9
38	315.6
39	285.5
40	333.7
41	336.3
42	309.0
43	341.6
44	244.2
45	346.8
46	349.5
47	352.1
48	394.7
49	405.6
50	428.2
51	449.3
52	470.4
53	491.5
54	212.6
55	353.7
56	254.8
57	375.9
58	397.0
59	418.1

Serial number	Cost
60	335.3
61	326.2
62	317.1
63	308.0
64	298.9
65	289.9
66	280.8
67	271.7
68	262.6
69	253.5
70	244.5
71	235.4
72	326.3
73	217.2
74	208.1
75	199.1
76	390.0
77	280.9
78	285.1
79	289.4
80	293.7
81	297.9
82	402.2
83	306.4
84	310.7
85	274.9
86	319.2
87	283.4
88	369.5

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \widehat{x}_i}{x_i} \right)^2}. \tag{36}$$

(3) Mean absolute percentage error (MAPE):

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{(x_i - \widehat{x}_i)}{x_i} \right| \cdot 100\%. \tag{37}$$

(4) Average absolute error (AAE):

$$\text{AAE} = \frac{1}{n} \frac{\left(\sum_{i=1}^{n} |x_i - \widehat{x}_i| \right)}{\left((1/n) \sum_{i=1}^{n} x_i \right)}. \tag{38}$$

In equations (35)–(38), x is the actual value of the substation projects cost, $\widehat{x}$ is the forecasting value of the substation projects cost, and n is the number of data groups. The smaller the value of above indicators, the higher the accuracy of cost forecasting.

4.3. Feature Selection. The main content of this section is the selection of the optimal feature subset based on the DIR model, which helps determine the input indicators of the forecasting model. In this paper, Matlab R2014b is used to carry on the programming computation. The test platform

environment is based on Intel Core i5-6300U, 4G memory, and Windows 10 Pro edition system.

The iterative process diagram of the extraction of training sample features based on the WPA-DIR-ILSSVM model is shown as Figure 3. The accuracy curve shown in the figure describes the forecasting accuracy of ILSSVM for training samples in different iterations. The fitness curve describes the fitness function value calculated during each iteration. The selected number is the optimal number of features calculated by the DIR model in the process of convergence. And, the reduced number of features refers to the number of features eliminated by the WPA algorithm during the convergence process.

From the figure above, it can be seen that the WPA algorithm converges when the number of iterations is 39 and the optimal fitness function value is −0.91 at this time. The forecasting accuracy of the training sample is 98.9% at the 39th iteration, which shows that the fitting ability of ILSSVM is enhanced through the learning and training of the algorithm and the forecasting accuracy of the training sample is the highest. Meanwhile, the number of selected features also tends to be stable when WPA reaches the 39th time. The algorithm eliminates 26 redundant features from the 33 candidate features, and the final selected features are construction properties, substation voltage level, main transformer capacity, substation type, number of transformers, main transformer single unit price, and floor space.

4.4. Cost Forecasting of Substation Projects and Result Analysis. The input vectors are substituted into the ILSSVM model for training and testing after obtaining all the optimal features of the sample data. In this paper, the self-written program is used to run and calculate in Matlab software. It is worth that the wavelet kernel function is chosen as the kernel function of the ILSSVM regression model in this paper. And, the important parameters of the model are obtained through optimization based on WPA algorithm to ensure the accuracy of ILSSVM. The parameter settings of WPA have been given in Section 3, and parameters of ILSSVM calculated by running the program are $\gamma = 43.0126$ and $\sigma = 19.0382$.

The unoptimized ILSSVM, LSSVM, SVM, and BP neural networks (BPNNs) are also selected to conduct the cost forecasting of substation projects, which helps prove the forecasting performance of the cost forecasting model proposed in this paper. The topological structure of the BPNN model is 7-5-1, and the transfer functions of the hidden layer and the output layer adopt the tansig function and purelin function, respectively. The maximum number of training is 200, the minimum error of the training target is 0.0001, the training rate is 0.1, and the initial weight and threshold are obtained by its own training. In the SVM model, the penalty parameter c is 10.276, the kernel function parameter σ is 0.0013 and the loss function parameter ε is 2.4375. In the LSSVM model, c is 10.108, σ is 0.0026, and ε is 0.0018. And, in the ILLSVM model, c is 16.263, σ is 0.0012, and ε is 0.0015.

(a)

(b)

(c)

(d)
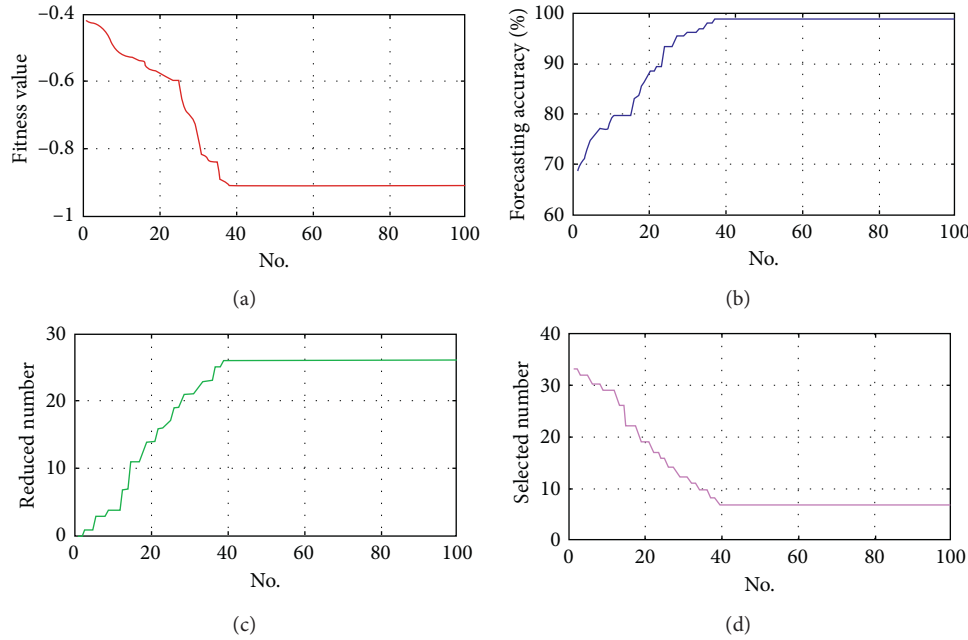
FIGURE 3: The curve of convergence for feature selection.

The forecasting results of BPNN, SVM, LSSVM, ILSSVM, and the model constructed in this paper for cost forecasting of the test set are shown in Table 2.

The forecasting results in Table 2 are plotted as shown in Figure 4 to facilitate the analysis more intuitively. The RE of each forecasting model is shown in Figure 5. It can be seen from Table 2 and Figures 4 and 5 that the forecasting error of all WPA-DIR-ILSSVM models is within [−3%, 3%], and the minimum value of the absolute relative errors is 0.06% and the maximum value is 1.30%. Among them, the number of errors outside [−1%, 1%] is 5, which are substation projects with serial numbers 67, 70, 72, 79, and 83, with relative errors of −1.14% and 1.13%, 1.30%, 1.28%, and 1.08%, respectively. The relative errors of the 11 sample points in the ILSSVM forecasting model is controlled within the range of [−3%, 3%], and the error range of 3 sample points is [−2%, 2%], with the serial number of 70, 77, and 83, respectively. For the substation projects above, the relative errors are −1.94%, 1.68%, and −0.89%, respectively, and the minimum value of the absolute relative error is 0.89% and the maximum is 3.66%. The relative errors of the 4 sample points in the LSSVM forecasting model is controlled within the range of [−3%, 3%], which are substation projects with serial numbers 68, 76, 72, 80, and 86, with relative errors of −2.78%, 2.19%, −2.74%, and 1.36%, respectively. However, all of them are outside the range of [−1%, 1%], the minimum value of the absolute relative error is 1.36%, and the maximum is 6.27%. The minimum value of the absolute relative error in the SVM forecasting model is 1.44% and the maximum is 7.10%, in which the errors of most sample points are between [−6%, −4%] and [4%, 6%]. The minimum value of the absolute relative error in the BPNN forecasting model is 2.57% and the maximum is 7.71%, in which the errors of most sample points are between [−7%, −5%] and

[5%, 7%]. Meanwhile, the fluctuation range of errors is relatively large in the BPNN model. According to the results of the absolute relative error, the WPA-DIR-ILSSVM model has the highest forecasting accuracy, next are the ILSSVM, LSSVM, and SVM models, and the worst is the BNPP model. Therefore, the forecasting accuracy and stability of WPA-DIR-ILSSVM model are the best, which reflects that the WPA algorithm can effectively enhance the training and learning process so as to avoid falling into a local optimum and improve the global searching ability of ILSSVM. Simultaneously, satisfactory forecasting results can be obtained by using the input vectors based on the DIR model. Additionally, ILSSVM presents more satisfactory performance than LSSVM, SVM, and BPNN. This result indicates that the more accurate forecasting results can be achieved by means of the improvement of LSSVM.

The RMSE, MAPE, and AAE of BPNN, SVM, LSSVM, ILSSVM, and WPA-DIR-ILSSVM are shown in Figure 6. From Figure 6, we can conclude that the RMSE, MAPE, and AAE of the proposed model are 0.8025%, 0.7159%, and 0.7157%, which are all the smallest among the above five models. In addition, the RMSE, MAPE, and AAE of the ILSSVM model are 2.6858%, 2.7961%, and 2.7956%, respectively. The RMSE, MAPE, and AAE of the LSSVM model are 4.5163%, 4.3614%, and 4.2778%, respectively. The RMSE, MAPE, and AAE of the SVM model are 4.5558%, 4.3895%, and 4.3203%. The RMSE, MAPE, and AAE of the BPNN model are 5.5044%, 5.2589%, and 5.1402%, respectively. The overall forecasting error of model the degree of error dispersion can be reflected through these indicators. Therefore, the overall forecasting effect of ILSSVM is better than that of LSSVM, SVM, and BPNN, and the overall forecasting effect of LSSVM is better than that of SVM and BPNN, which indicates that the overall forecasting

TABLE 2: The forecasting values and actual values of the test set (unit: yuan/kV·A).

Serial number	Actual value	BPNN	SVM	LSSVM	ILSSVM	WPA-DIR-ILSSVM
67	271.7	286.4	287.1	284.6	280.5	268.6
68	262.6	251.3	273.1	255.3	269.6	264.2
69	253.5	270.0	239.6	264.0	259.6	255.9
70	244.5	225.6	254.5	259.5	239.7	247.2
71	235.4	251.4	245.9	246.5	243.1	236.7
72	326.3	335.6	335.2	341.1	335.7	330.5
73	217.2	208.3	201.8	203.6	211.6	215.6
74	208.1	219.0	216.5	218.0	213.9	209.5
75	199.1	210.2	195.2	208.6	204.9	200.6
76	390.0	376.9	395.6	398.5	381.6	387.6
77	280.9	296.1	294.8	269.6	285.6	280.5
78	285.1	302.8	269.6	296.9	295.0	288.0
79	289.4	271.9	302.1	303.2	280.2	293.1
80	293.7	310.5	281.0	285.6	302.8	291.3
81	297.9	313.7	312.3	311.1	307.6	298.1
82	402.2	412.5	389.5	385.6	413.6	405.9
83	306.4	325.3	321.6	322.6	303.7	309.7
84	310.7	298.7	327.0	292.7	320.2	309.1
85	274.9	290.3	287.2	288.6	283.2	276.7
86	319.2	338.1	302.9	323.5	307.5	318.6
87	283.4	300.1	295.7	270.6	293.2	283.6
88	369.5	350.2	385.9	387.1	380.9	372.0
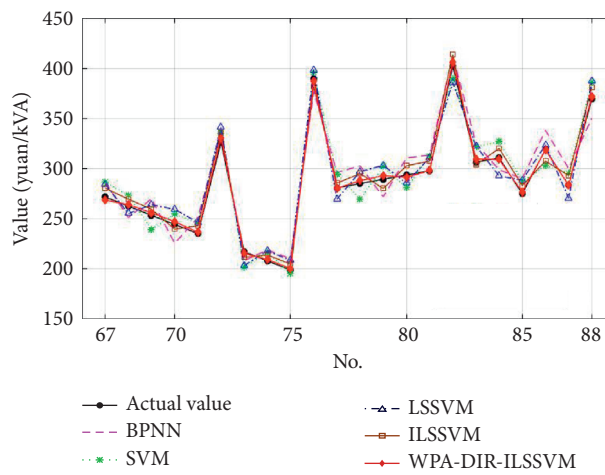
FIGURE 4: Comparison of forecasting results of the test sets.
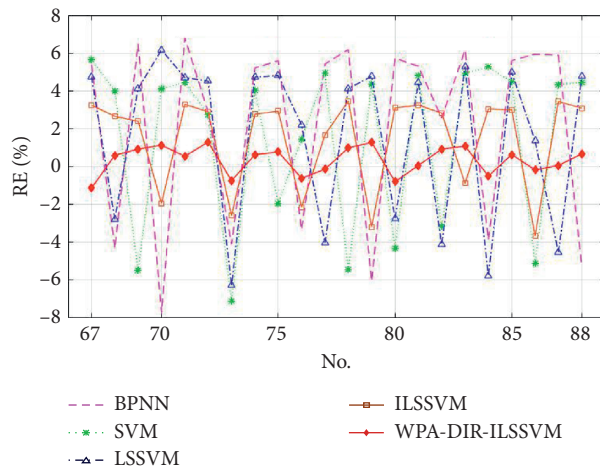
FIGURE 5: RE of the forecasting results of the test set.
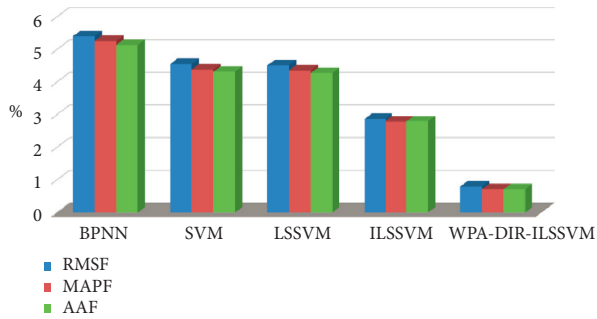
Figure 6: RMSE, MAPE, and AAE of the forecasting results.

performance of LSSVM is significantly improved after the weighted improvement. The forecasting accuracy of WPA-DIR-ILSSVM is better than that of ILSSVM, which proves that the parameters γ and σ of ILSSVM selected by the WPA algorithm have a good optimization effect. And, the DIR model ensures the integrity of the input information while reducing redundant data, in which the ideal forecasting results are achieved. In addition, K-fold is utilized in this paper to improve generalization performance, which advances the accuracy of prediction.

5. Conclusions

This paper presents a hybrid cost forecasting model that combines DIR with ILSSVM optimized by WPA. First, in order to forecast the substation projects cost, the DIR combined with the WPA is employed to select the input feature. Furthermore, the WPA is also adopted to optimize the parameters of the ILSSVM. Finally, after obtaining the optimized input subset and the best value of γ and σ, the proposed model is utilized for the cost forecasting of substation projects. Several conclusions based on the studies can be obtained as follows: (a) by the utilization of DIR, the influence of unrelated noises can be reduced and the forecasting performance can be effectively improved. (b) The optimization algorithm WPA adds the model with strong global searching capability and the ILLSVM model optimized by WPA shows good performance. (c) Based on the error valuation criteria, compared with LSSVM, a better forecasting result can be achieved based on ILSSVM, which shows that the method of improving LSSVM by replacing the traditional radial basis kernel function with wavelet kernel function is effective. (d) Through the example verification of substation projects in different regions, different voltage grades, and different scales, an ideal forecasting effect is obtained, which shows that the model proposed in this paper is more adaptable and stable. Hence, the proposed cost forecasting method of WPA-DIR-ILSSVM is effective and feasible, and it may be an effective alternative for the cost forecasting in the electric-power industry. However, more sample data are needed for verification. At the same time, adopting more intelligent models to forecast the substation projects cost is also our next work.

Data Availability

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 article.

Acknowledgments

References

[1] A. S. Sidhu, M. G. Pollitt, and K. L. Anaya, "A social cost benefit analysis of grid-scale electrical energy storage projects: a case study," *Applied Energy*, vol. 212, no. 15, pp. 881–894, 2018.

[2] Z. Zhang, Z. Yu, W. Li et al., "Research on the whole life cycle cost management and maintenance decision of secondary equipment based on physical ID," in *Proceedings of the 2019 IEEE 3rd Conference on Energy Internet and Energy System Integration (EI2)*, IEEE, Changsha, China, June 2020.

[3] L. Liu and G. Jiang, "LCC prediction model of substation in high latitude and cold area based on SSA-LS-SVM," *Smart Power*, vol. 48, no. 6, pp. 54–60, 2020.

[4] W. Li, "A model of building project cost estimation based on multiple structure integral linear regression," *Archit Technologies*, vol. 46, pp. 846–849, 2015.

[5] B.-C. Kim, "Integrating risk assessment and actual performance for probabilistic project cost forecasting: a second moment bayesian model," *IEEE Transactions on Engineering Management*, vol. 62, no. 2, pp. 158–170, 2015.

[6] L. Gao and Y. Niu, "Intelligent estimation of high-speed railway cost based om fuzzy C means clustering-fuzzy inference system," *Journal of Shijiazhuang Tiedao University (Social Science Edition)*, vol. 14, no. 2, pp. 36–43, 2020.

[7] S. Dong, Y. Zhang, Z. He, N. Deng, X. Yu, and S. Yao, "Investigation of support vector machine and back propagation artificial neural network for performance prediction of the organic rankine cycle system," *Energy*, vol. 144, pp. 851–864, 2018.

[8] D. Keles, J. Scelle, F. Paraschiv, and W. Fichtner, "Extended forecast methods for day-ahead electricity spot prices applying artificial neural networks," *Applied Energy*, vol. 162, pp. 218–230, 2016.

[9] Z. Dong, Y. Q. Wang, Y. Wang, X. Zhao, C. Y. Rong, and J. Nie, "Cost forcast of power grid transformation project based on DEA and genetic BP neural network," *Hebei Electric Power*, vol. 39, no. 1, pp. 26–29, 2020.

[10] Y. P. Ling, P. F. Yan, C. Z. Han, and C. G. Yang, "BP neural network based cost prediction model for transmission projects," *Electric Power*, vol. 45, pp. 95–99, 2012.

[11] D. X. Niu, H. C. Wang, H. Y. Chen, and Y. Liang, "The general regression neural network based on the fruit fly optimization algorithm and the data inconsistency rate for transmission line icing prediction," *Energies*, vol. 10, p. 2066, 2017.

[12] R. C. Deo, X. Wen, and F. Qi, "A wavelet-coupled support vector machine model for forecasting global incident solar

radiation using limited meteorological dataset," *Applied Energy*, vol. 168, pp. 568–593, 2016.

[13] N. N. Wang, F. Wang, Y. T. Yin, H. Li, and Y. Hou, "Research on cost predicting of power transformation projects based on SVM," *Construction Economics & Finance*, vol. 37, pp. 48–52, 2016.

[14] Z. Y. Song, D. X. Niu, X. L. Xiao, and L. Zhu, "Substation engineering cost forecasting method based on modified firefly algorithm and support vector machine," *Electric Power*, vol. 50, pp. 168–173, 2017.

[15] X. Yuan, Q. Tan, X. Lei, Y. Yuan, and X. Wu, "Wind power prediction using hybrid autoregressive fractionally integrated moving average and least square support vector machine," *Energy*, vol. 129, pp. 122–137, 2017.

[16] Z. Tian, "Short-term wind speed prediction based on LMD and improved FA optimized combined kernel function LSSVM," *Engineering Applications of Artificial Intelligence*, vol. 91, Article ID 103573, 2020.

[17] W. T. Huang, P. Zhou, and J. X. Cheng, "An estimation method of engineering cost based on adaboost and variable selection with LSSVM," *Journal of Chongqing Jiaotong University (Natural Sciences)*, vol. 35, pp. 54–57, 2016.

[18] M. Liu, "Prediction model of project cost of based on chaotic theory and least square support vector machine," *Journal of Inner Mongolia Normal University (Natural Science)*, vol. 44, pp. 333–338, 2015.

[19] Z. Tian, "Kernel principal component analysis-based least squares support vector machine optimized by improved grey wolf optimization algorithm and application in dynamic liquid level forecasting of beam pump," *Transactions of the Institute of Measurement and Control*, vol. 42, no. 6, pp. 1135–1150, 2020.

[20] G. R. Ruiz, C. F. Bandera, G. A. Temes, and S. O. Gutierrez, "Genetic algorithm for building envelope calibration," *Applied Energy*, vol. 168, pp. 691–705, 2016.

[21] G. Li, L. Boukhatem, and J. Wu, "Adaptive quality-of-service-based routing for vehicular ad hoc networks with ant colony optimization," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 4, pp. 3249–3264, 2017.

[22] D. Niu, Y. Liang, and W.-C. Hong, "Wind speed forecasting based on EMD and GRNN optimized by FOA," *Energies*, vol. 10, no. 12, p. 2001, 2017.

[23] M. Azaza and F. Wallin, "Multi objective particle swarm optimization of hybrid micro-grid system: a case study in Sweden," *Energy*, vol. 123, pp. 108–118, 2017.

[24] Y.-C. Li and P.-D. Xu, "Improved wolf pack algorithm for optimum design of truss structures," *Civil Engineering Journal*, vol. 6, no. 8, pp. 1411–1427, 2020.

[25] X. Zhang, C. Mei, D. Chen, and J. Li, "Feature selection in mixed data: a method using a novel fuzzy rough set-based information entropy," *Pattern Recognition*, vol. 56, pp. 1–15, 2016.

[26] J. Liu, J. Ma, Y. Zhang et al., "Discriminative feature representation to improve projection data inconsistency for low dose CT imaging," *IEEE Transactions on Medical Imaging*, vol. 36, no. 12, pp. 2499–2509, 2018.

[27] T. N. Ma, D. X. Niu, Y. L. Huang, and Z. D. Du, "Short-term load forecasting for distributed energy system based on spark platform and multi-variable L2-boosting regression model," *Power System Technology*, vol. 40, pp. 1642–1649, 2016.

[28] J.-P. Liu and C.-L. Li, "The short-term power load forecasting based on sperm whale algorithm and wavelet least square support vector machine with DWT-IR for feature selection," *Sustainability*, vol. 9, no. 7, p. 1188, 2017.

[29] J. J. Xue, Y. Wang, H. Li, and J. Y. Xiao, "A smart wolf pack algorithm and its convergence analysis," *Control and Decision*, vol. 31, pp. 2131–2139, 2016.

[30] Y. Lv, F. Hong, T. Yang, F. Fang, and J. Liu, "A dynamic model for the bed temperature prediction of circulating fluidized bed boilers based on least squares support vector machine with real operational data," *Energy*, vol. 124, pp. 284–294, 2017.

[31] A. Barati-Harooni, A. Najafi-Marghmaleki, M. Arabloo, and A. H. Mohammadi, "An accurate CSA-LSSVM model for estimation of densities of ionic liquids," *Journal of Molecular Liquids*, vol. 224, pp. 954–964, 2016.

[32] M. Malvoni, M. G. De Giorgi, and P. M. Congedo, "Photovoltaic forecast based on hybrid PCA-LSSVM using dimensionality reduced data," *Neurocomputing*, vol. 211, pp. 72–83, 2016.

[33] R. Dong, J. Xu, and B. Lin, "ROI-based study on impact factors of distributed PV projects by LSSVM-PSO," *Energy*, vol. 124, pp. 336–349, 2017.

[34] Q. Wu and C. Peng, "Wind power grid connected capacity prediction using LSSVM optimized by the bat algorithm," *Energies*, vol. 8, no. 12, pp. 14346–14360, 2016.

[35] X. Song, J. J. Zhao, J. Song et al., "Local demagnetization fault recognition of permanent magnet synchronous linear motor based on S-transform and PSO-LSSVM," *IEEE Transactions on Power Electronics*, vol. 35, no. 8, pp. 7816–7825, 2020.

[36] Z. Zhao, "Backtracking search optimization algorithm-based least square support vector machine and its applications," *Engineering Applications of Artificial Intelligence*, vol. 94, Article ID 103801, 2020.

[37] X. Zhang, "Analysis on the influencing factors of project cost in substation construction stage," *Low Carbon World*, vol. 33, pp. 64-65, 2017.

[38] S. Montoya-Bueno, J. Muñoz-Hernandez, and L. Baringo, "A benders' decomposition approach for renewable generation investment in distribution systems," *Energies*, vol. 13, no. 5, p. 1225, 2020.

[39] X. Yi, X. Chuanyu, W. Wu et al., "Life cycle cost prediction of substation based on advanced PSO and least squares support vector machine," *IOP Conference Series Earth and Environmental Ence*, vol. 461, Article ID 012083, 2020.

Research Article

CO$_2$ Emissions, Energy Consumption, and Economic Growth Nexus: Evidence from 30 Provinces in China

Shaohui Zou[1,2] **and Tian Zhang** [1]

[1]*School of Energy, Xi'an University of Science and Technology, Xi'an 710054, China*
[2]*Energy Economy and Management Research Center, Xi'an University of Science and Technology, Xi'an 710054, China*

Correspondence should be addressed to Tian Zhang; 2274540847@qq.com

Under the situation of global low-carbon development, the contradiction among energy consumption, economic growth, and CO$_2$ emissions is increasingly prominent. Considering the possible two-way feedback among the three, based on the panel data of 30 regions in China from 2000 to 2017, this paper establishes a spatial Durbin model including economic growth, energy consumption equation, and CO$_2$ emissions and studies the dynamic relationship and spatial spillover among economic growth, energy consumption, and CO$_2$ emissions effects. The results show that the economic growth can significantly improve carbon dioxide emissions, and China's economic growth level has become a positive driving force for carbon dioxide emissions. However, economic growth will not be significantly affected by the reduction of carbon dioxide emissions. There is a two-way relationship between energy consumption (ENC) and carbon dioxide emissions (CO$_2$). Energy consumption and carbon emissions are interrelated, which has a negative spatial spillover effect on the carbon dioxide emissions of the surrounding provinces and cities.

1. Introduction

The process of global industrialization has brought about the rapid economic growth of all countries. With the continuous improvement of energy consumption, the rapid economic growth leads to the rise of carbon dioxide emissions. The problem of global climate change caused by the aggravation of greenhouse gas emissions is threatening the survival and development of human beings, which become a worldwide concern. Since the 1960s, environmental problems have erupted all over the world, prompting governments to pay more and more attention to environmental problems. The world has successively held "Rio Conference," "Kyoto Conference," and "Copenhagen Conference" and signed the Kyoto Protocol. At present, with the high energy consumption mode of economic growth, China has become the largest fossil energy consumption country in the world and China's per capita carbon emissions are also increasing. Facing such problems, our government makes great efforts to take measures to reduce carbon emissions and advocate

the development concept of "Green GDP." In the 13th five-year plan, China proposes the object of reducing CO$_2$ emissions by 40%–45% in 2020 and 60%–65% in 2030 compared with 2005. Therefore, it is very important to develop low-carbon economy and study the nexus among energy, environment, and economy. In the past 30 years, China's economy has maintained a sustained growth, with an average annual GDP growth rate of 7%. With economic growing, energy consumption and CO$_2$ emissions are also increasing. Under the existing technical conditions, whether reducing emissions and energy consumption will affect economic development, whether economic development will react on energy consumption and carbon dioxide emissions, and what kind of relationship and mechanism exist among the three, all of which involve the causal relationship among economic development, energy consumption, and carbon emissions. Therefore, in this trend, the research on social and economic development, energy consumption, and carbon emissions has become the focus of global climate change research.

Our contributions are threefold. First, this paper aims to look at the long-run relationship between CO_2 emissions, energy consumption, and economic growth using the spatial Durbin model. Second, we estimated CO_2 emissions based on the amount of fuel burned and the default emission factors. Third, our empirical results, to some extent, indicate that there is a two-way relationship between energy consumption (ENC) and carbon dioxide emissions (CO_2).

2. A Brief Literature Review

The literature on economic growth, energy consumption, and carbon dioxide emissions can be divided into three categories: the first category studies the relationship between economic growth and energy consumption; the second category studies the relationship between economic growth and carbon dioxide emissions; there are abundant literatures about the relationship between the two kinds of research, while the third kind is a new trend, that is to study the relationship among economic growth, energy consumption, and carbon dioxide emission in a unified framework.

The research on the relationship between economic growth and carbon dioxide emission mainly focuses on testing the existence of "Environmental Kuznets hypothesis (EKC)." Many scholars, such as Schmalensee, Thomas, and Saboori, have verified the inverted U-shaped relationship between economic growth and carbon dioxide emissions, but in the existing literature, "EKC hypothesis" is often regarded as a phenomenon to be tested [1, 2]. Apergis tested for the validity of the Environmental Kuznets Curve (EKC) using both panel-based and time-series-based methodological approaches of cointegration and used data from fifteen countries, spanning the period 1960–2013 [3]. Salahuddin et al. applied the autoregressive distributed lag (ARDL) bounds testing approach and found that cointegration exists among the series. Findings indicate that economic growth, electricity consumption, and FDI stimulate CO_2 emissions in both the short and long run [4]. Coondoo et al. adopted the Environmental Kuznets Curve (EKC) to examine the presence or otherwise of an inverted U-shaped relationship between the level of pollution and the level of income. Customarily, in the diagram of EKC, the level of income is shown on the horizontal axis and that of pollution is shown on the vertical axis [5].

There are some researches on the relationship between energy consumption and output. This relationship shows that economic growth and energy consumption may be jointly determined, because economic growth is closely related to energy consumption, and higher economic growth requires more energy consumption. Since Hussain [6], many scholars have used Granger's causality test and cointegration model to evaluate the relationship between economic growth and energy consumption, such as Narayan, Stamler, and Zeng [7–9]. Based on the data of France from 1960 to 2000, Ang used cointegration analysis and error correction model (ECM) to test the relationship between economic growth, energy consumption, and carbon dioxide emissions. It was found that in the long run, the relationship between economic growth and carbon dioxide emissions was

inverted U-shaped, and economic growth promoted the increase of energy consumption, while the increase of energy consumption led to the increase of carbon dioxide emissions [10]. Apergis and Payne used VECM to study the relationship between economic growth, energy consumption, and carbon dioxide emissions in six Central American countries from 1971 to 2004 [11, 12]. Combining ECM and ARDL model, Halicioglu studied Turkey's per capita income, energy consumption, carbon dioxide emissions, and foreign trade from 1960 to 2005 and found that there was a causal relationship between income and emissions in both the short and long term [13]. Saboori et al. studied the relationship between electricity consumption, economic growth, and carbon dioxide emissions of BRICS countries in 1990–2010 under the framework of panel causal analysis [14]. Feng et al. revealed that the air pollution is affected by not only local environmental regulations, but also regulations implemented in the surrounding cities [15]. Song et al. studied the decoupling relationship between CO_2 emissions and economic development based on two-dimensional decoupling theory; they get some important implications [16–19].

On the whole, scholars at home and abroad have studied the relationship between economic growth and energy consumption, economic growth, and carbon dioxide emissions, which provides a good foundation for this study. As a developing country, China is in a period of high-speed development. It is the second largest energy producer and consumer in the world, and the second largest CO_2 emitter in the world. It is faced with various pressure from the international energy conservation and emission reduction. In this case, this paper chooses the relationship among China's economic growth, energy consumption, and carbon emissions as the research topic, with a view to energy policy and the formulation of energy conservation and emission reduction policies can improve the reference to achieve a positive interaction among economic growth and energy conservation and emission reduction.

3. Methodology

3.1. Spatial Correlation Test. Spatial correlation, also known as spatial dependence, refers to the spatial interdependence, mutual restriction, mutual influence, and interaction between things and phenomena in different regions. It is the inherent spatial economic attribute of things and phenomena and the essential attribute of geospatial phenomena and spatial processes. Based on the complexity of spatial econometric model, before empirical analysis using spatial econometric model, it is necessary to test the spatial correlation of research variables to determine whether they have spatial correlation or not. The commonly used methods of spatial correlation test are Moran's index test, Guillain's index test, cetis index, etc. Among them, Moran's index is a more practical way to measure the spatial correlation of variables. Global Moran's I statistic is a common measure of global spatial autocorrelation. Moran's I index comes from Pearson's correlation coefficient in statistics. The correlation coefficient is extended to autocorrelation coefficient; the autocorrelation coefficient of time series is extended to

autocorrelation coefficient of space series. Finally, the weighted function is used to replace the lag function, and the one-dimensional autocorrelation coefficient is extended to the two-dimensional autocorrelation coefficient to obtain Moran's I index. Moran's I index is actually a standardized spatial autocorrelation [20, 21].

Assuming a vector $x = [x_1, x_2, \ldots, x_n]'$, Moran's I index is expressed in vector form as follows:

$$
\begin{aligned}
I &= \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (x_i - \overline{x})(x_j - \overline{x})}{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \sum_{i}^{n} (x_i - \overline{x})^2} \\
&= \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (x_i - \overline{x})(x_j - \overline{x})}{S^2 \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}},
\end{aligned}
\tag{1}
$$

where n is the total number of regions in the study area and w_{ij} is the spatial weight (set w_{ij} by whether area i and area j are adjacent: when i region is adjacent to region j, $w_{ij} = 1$; when i region is not adjacent to region j, $w_{ij} = 0$); x_i and x_j are attributes of region i and region j, respectively; $\overline{x} = 1/n \sum_{i=1}^{n} x_i$ is the average value of all attributes; and $S^2 = 1/n \sum_{i=1}^{n} (x_i - \overline{x})^2$ is the variance of the attribute. Moran's I index can be regarded as the correlation coefficient between the observation value and its spatial lag. The spatial lag of variable x_i is the average value of x_i in neighborhood j, defined as $x_{i,-1} = \sum_{j=1}^{n} w_{ij} x_{ij} / \sum_{j=1}^{n} w_{ij}$.

Therefore, the value of Moran's I is generally between -1 and 1. If it is greater than 0, it means positive correlation. The closer the value is to 1, it means similar attributes are clustered together (high value is adjacent to high value; low value is adjacent to low value). If it equals 1, it means complete positive correlation. If it is less than 0, it means negative correlation. If it is closer to -1, it means different attributes are clustered together (high value is adjacent to low value. If the Moran I index is close to 0, it indicates that the attribute is randomly distributed, or there is no spatial autocorrelation.

3.2. Setting of Spatial Weight Matrix. The spatial weight matrix reflects the interdependence degree of each variable in different regional space, which is the basis and precondition for spatial correlation analysis. Everything is related to other things to some extent, but according to the law of geography, things with closer space distance are more closely related to each other than things with longer space distance. In view of this, in order to objectively analyze the relationship between energy consumption, carbon dioxide emissions, and economic growth, this paper constructs a 0, 1 adjacency weight matrix w_{ij} for empirical analysis. The spatial weight matrix formula is as follows:

$$
w_{ij} = \begin{cases} 1 \\ 0 \end{cases}
\tag{2}
$$

The 0, 1 adjacency weight matrix w is constructed according to whether provinces and cities are adjacent in geographical location. If provinces and cities i are adjacent to provinces and cities j, w is assigned as 1. If provinces and cities i are adjacent to provinces and cities j, w is assigned as 0.

3.3. Spatial Econometric Model. The Spatial Durbin Model (SDM) is chosen to analyze the relationship among energy consumption, carbon dioxide emission, and economic growth, because the carbon dioxide emission of a province is easily affected by the energy consumption level of neighboring provinces. Spatial Durbin Model (SDM) considers not only the spatial lag term of energy consumption, but also the spatial lag term of carbon dioxide emissions. Considering the spatial correlation of carbon dioxide emissions, it also considers the spatial correlation of energy consumption; that is, the carbon dioxide emissions of the province are not only affected by the energy consumption of the province, but also by the energy consumption of the neighboring provinces. In view of the above content, this paper constructs the following basic form of Spatial Durbin Model (SDM) based on the existing research:

$$
Y = \rho W Y + X\beta + \theta W X + \alpha I_n + \varepsilon.
\tag{3}
$$

Among them, Y stands for the explained variable, X stands for the explained variable, ρ stands for the spatial autocorrelation coefficient, W stands for the constructed spatial weight matrix, WX and WY stand for the explained variable energy consumption and the spatial lag term of carbon dioxide emission of the explained variable, and β and θ stand for the regression coefficient of the model, α for the constant term, I_n for the unit matrix, and ε for the error term.

3.4. Decomposition of Spatial Effect. The spatial lag of energy consumption, carbon dioxide emission, and economic growth is considered in the regression analysis of SDM. Based on the research of Lesage [22], this paper uses the partial differential method to decompose the spatial spillover effect of Spatial Durbin Model into three parts: direct effect, indirect effect, and total effect, so as to reduce or even avoid the biases of the Spatial Durbin Model in testing the spatial spillover effect. Among them, the direct effect refers to the impact of energy consumption and economic growth on carbon dioxide emissions of the province; the indirect effect refers to the impact of energy consumption and economic growth on carbon dioxide emissions of surrounding provinces. The specific calculation formula is as follows:

$$
\begin{aligned}
Y &= (1 - \rho W)^{-1} \alpha I_n + (1 - \rho W)^{-1} (X_t \beta + W X_t \theta) + (1 - \rho W)^{-1} \varepsilon, \\
Y &= \sum_{r=1}^{k} S_r(W) X_r + V(W) I_n \alpha + V(W) \varepsilon.
\end{aligned}
\tag{4}
$$

In which, $S_r(W) = V(W) (I_n \beta + W \theta_r)$, $V(W) = (I_n - \rho W)^{-1}$, I_n is the nth-order unit matrix, by transforming the above formula into matrix form, we can get

$$\begin{bmatrix} y_1 \\ y_2 \\ ... \\ y_n \end{bmatrix} = \sum_{r=1}^{k} \left[\begin{bmatrix} S_r(W)_{11}...S_r(W)_{1n} \\ ... \\ ... \\ S_r(W)_{n1}...S_r(W)_{nn} \end{bmatrix} \right] \begin{bmatrix} x_{1r} \\ x_{2r} \\ ... \\ x_{nr} \end{bmatrix} + V(W)\varepsilon. \tag{5}$$

The total effect ATI is equal to the sum of direct effect ADI and indirect effect AII, and the specific formula is as follows:

$$\begin{aligned} \text{ATI} &= n^{-1}I_n S_r(W)_{I_n}, \\ \text{ADI} &= n^{-1}t_r(S_r(W)), \\ \text{AII} &= \text{ATI} - \text{ADI}. \end{aligned} \tag{6}$$

4. Variables and Data Sources

This paper studies the relationship among carbon dioxide emissions, energy consumption, and economic growth in 30 provinces of China by using the Spatial Durbin Model. The data are from China Statistical Yearbook and China Energy Statistical Yearbook. According to the availability of data, the sample interval of this paper is 2000–2017. Each variable is an annual variable, which adopts the form of logarithmic transformation.

At present, most of these studies use per capita GDP to express the level of economic development. This paper also uses per capita GDP as the representative variable of economic growth. We use energy consumption per capita (kg oil equivalent/person) to represent energy consumption. For carbon dioxide emissions, per capita carbon dioxide emissions (metric tons/person) are selected as the research variable. These three variables are expressed by GDP_{in}, ENC_{in}, and CO_2, respectively, and are shown in Table 1.

5. Estimation of Carbon Dioxide Emissions

In 2006 IPCC guidelines for national green gas inventories, the international panel on climate change of the United Nations Intergovernmental Panel on climate change introduced three methods for estimating the CO_2 emissions from fossil fuel combustion in fixed and mobile sources in detail. Method 1 estimates CO_2 emissions based on the amount of fuel burned and the default emission factors. Although the calculation results of this method are relatively rough, it is relatively simple and easy to operate, and the data requirements are not high (Wang et al.) [23]. In this paper, "method 1" in IPCC (2006) is used to estimate the data of 30 provinces and municipalities (excluding Tibet) in mainland China from 2000 to 2017.

All kinds of energy data consumed by 30 provinces and regions in 2000–2017 are from China Energy Statistical Yearbook in 2000–2017. In order to avoid double calculation, 654 energy balance tables of 30 provinces and regions in 2000–2017 are selected in this paper. Among them, 11 kinds of energy are obtained in 2000–2009: coal, coke, coke oven gas, other gas, crude oil, gasoline, kerosene, diesel oil,

Table 1: Variable description.

Variables	Description
GDP	Level of economic development
ENC	Energy consumption per capita
CO_2	Per capita carbon dioxide emissions

fuel oil, liquefied petroleum gas, and natural gas. In 2010–2017, according to the balance sheet, 14 kinds of energy sources were added, including blast furnace gas, converter gas and liquefied natural gas. The input and loss in the process of energy processing and conversion as well as the part used as raw materials and materials in industrial production are eliminated from various energy sources consumed in each province year by year, so as to obtain the (net) consumption of 30 provinces and regions in 2000–2017. IPCC (2006) not only provides the calculation method of carbon dioxide, but also provides the carbon content and effective carbon dioxide emission factors of various types of fuels. The specific estimation method is as follows:

$$CO_2 = \sum_{i=1}^{14} CO_{2,i} = \sum_{i=1}^{14} E_i \cdot NCV_i \cdot CEF_i. \tag{7}$$

In the formula, CO_2 represents the carbon dioxide emission to be estimated; i represents various energy fuels, including coal, coke, coke oven gas, blast furnace gas, converter gas, other gas, crude oil, gasoline, kerosene, diesel oil, fuel oil, liquefied petroleum gas, natural gas, and liquefied natural gas; E_i represents the combustion consumption of various energy sources; NCV_i is the average low calorific value of various energy sources, which is used to convert various energy consumption into energy units (TJ); and CEF_i represents the carbon dioxide emission factors of various energy sources. The formula is as follows:

$$CEF_i = CC_i \cdot COF_i \cdot \left(\frac{44}{12}\right). \tag{8}$$

In the formula, CC_i is the carbon content of all kinds of energy; COF_i is the carbon oxidation factor of all kinds of energy, usually the value is 1, indicating that the energy is completely oxidized. In this paper, coal and coke are set as 0.99 and the rest as 1 [24]. (44/12) is the molecular weight ratio of carbon dioxide to carbon. The specific values are shown in Table 2.

Source of data: (1) NCV comes from China Energy Statistical Yearbook 2018, in which the average low calorific value of raw coal is selected; CEF is from IPCC (2006).

6. Empirical Results

6.1. Spatial Autocorrelation Test. Firstly, the spatial autocorrelation test of data is used to verify whether there is spatial correlation. The global Moran's I values of carbon dioxide emissions of 30 provinces and cities in China were calculated by Stata software, as shown in Table 3. It can be seen from Table 3 that Moran's I of carbon dioxide emission under the 0, 1 spatial weight is significantly positive, and

TABLE 2: Various indexes and coefficients involved in the calculation of carbon dioxide.

Energy types	NCV (kj/kg)	CEF (kg/TJ)
Coal	20908.0	95977.0
Coke	28435.0	105996.0
Coke oven gas	17981.0	44367.0
Blast furnace gas	3855.0	259600.0
Converter gas	8585.0	181867.0
Other gas	18273.6	44367.0
Crude oil	41816.0	73333.0
Gasoline	43070.0	70033.0
Kerosene	43070.0	71500.0
Diesel oil	42652.0	74067.0
Fuel oil	41816.0	77367.0
Liquefied petroleum gas	50179.0	63067.0
Natural gas	38931.0	56100.0
Liquified natural gas	44200.0	64167.0

TABLE 3: Global Moran's I of carbon dioxide emission in China.

Year	Moran's I	p value*
2000	0.055	0.0052
2001	0.106	0.0023
2002	0.099	0.0536
2003	0.135	0.0061
2004	0.022	0.0911
2005	0.064	0.0793
2006	0.068	0.0761
2007	−0.036	0.0992
2008	0.007	0.0794
2009	0.025	0.0923
2010	0.011	0.0657
2011	0.054	0.0847
2012	0.056	0.0828
2013	0.06	0.0808
2014	0.068	0.0746
2015	0.079	0.0067
2016	0.085	0.0628
2017	0.097	0.0054

both can pass the test under 10% significance level. This shows that there is a significant correlation effect between the provincial carbon dioxide emissions, and the spatial agglomeration characteristics are more obvious. Therefore, in the study of carbon dioxide emissions, we should fully consider the impact of spatial factors and select the spatial econometric model to reduce or even avoid the shortcomings of the general econometric model.

Based on the overall autocorrelation test of the data, the local Moran I of carbon dioxide emission in each province is further obtained. Here, the local Moran scatter plots of 2000, 2005, 2010, 2015, and 2017 are selected for research and analysis, as shown in Figures 1–5.

From the Moran scatter plots of 2000, 2005, 2010, 2015, and 2017, it can be seen that the overall carbon dioxide emissions of provinces in China still present the trend of agglomeration. Most of the provinces and cities in the Moran scatter plots of carbon dioxide emissions are located

in the second and third quadrants, showing the characteristics of "high-high" and "low-low" spatial clusters as a whole.

6.2. Estimation of Spatial Econometric Model

6.2.1. LM, Hausman, and LR Tests. Through the spatial autocorrelation test, it can be determined that the spatial autocorrelation factors should be fully considered in the empirical analysis of carbon dioxide emissions to avoid large differences in research results. On this basis, the paper first determines the spatial error model, spatial lag model, and spatial Durbin model which are more suitable for the spatial measurement model of this study through LM test and then further confirms the specific spatial measurement model according to Hausman test and LR test. The LM, Hausman, and LR tests' results of panel data are shown in Table 4.

It can be seen from Table 4 that in the regression LM Test, the test values of LM error and LM lag reject the original hypothesis at the significance level of 1%; it shows that not only the variables have spatial lag effect, but also the error items have spatial correlation. Therefore, the spatial Durbin model is used as the model of spatial econometric analysis for empirical research. Then, the Hausman test results of panel data show that the test results under 0, 1 adjacency weight matrix are significantly positive, so the fixed effect space Durbin model is selected. The LR test results of this paper show that the LR test value of the spatial fixed effect model is 24.34, which is significant at the level of 1%, indicating that there are significant differences in time and space for each variable. Therefore, the spatial Durbin model is selected as the best model for panel data in our research.

6.2.2. Spatial Econometric Model Test. In order to further confirm the applicability of the selected model, this paper uses LR statistics test to confirm whether the spatial Durbin model (SDM) will degenerate into the other two spatial econometric models. The test results are shown in Table 5.

It can be seen from Table 5 that the LR statistics test results reject the original hypothesis that the spatial Durbin model (SDM) will degenerate into the spatial lag model (SAR) and the spatial error model (SEM) at the significance level of 1%. Therefore, the LR test results still support the use of the spatial Durbin model as the empirical model of this study.

6.2.3. SDM Estimation. In this paper, Stata 13.0 is selected to study the spatial effect between carbon dioxide emissions, energy consumption, and economic growth. The estimation results of spatial Durbin model (SDM) is as shown in Table 6.

It can be seen from the table that the explanatory variable GDP (economic growth) under the 0, 1 adjacency weight matrix is significantly positive at the level of 10%. It can be seen that economic growth can significantly improve carbon dioxide emissions, and China's economic growth level has become a positive driving force for carbon dioxide emissions. The result of Equation Estimation with GDP as the

Moran's scatterplot (Moran's I = −0.059)
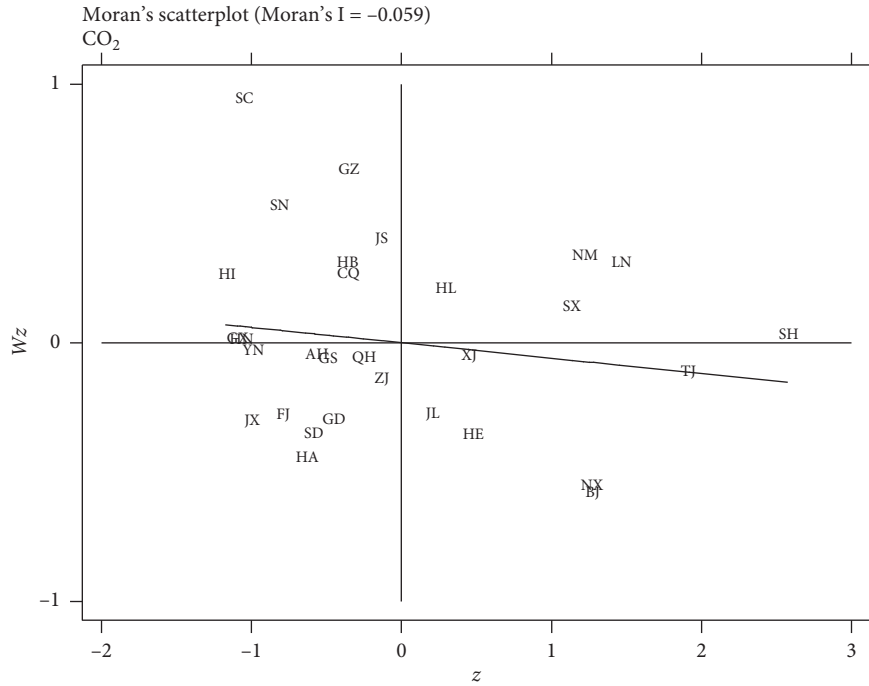
Figure 1: Moran's scatter plot in 2000.

Moran's scatterplot (Moran's I = −0.066)
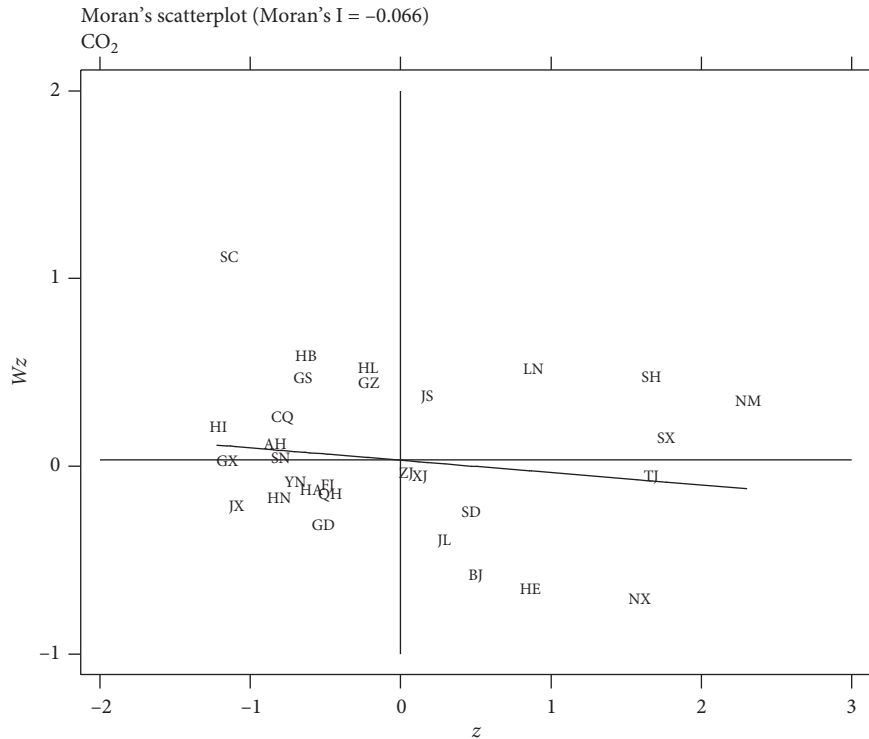
Figure 2: Moran's scatter plot in 2005.

explanatory variable shows that at the level of 10% significance, economic growth will not be significantly affected by the reduction of carbon dioxide emissions. However, if there is a causal relationship between carbon dioxide emissions and economic growth, it means that carbon dioxide emissions contain relevant information of future economic growth, and economic growth will be significantly impacted by "energy conservation and emission reduction" measures.
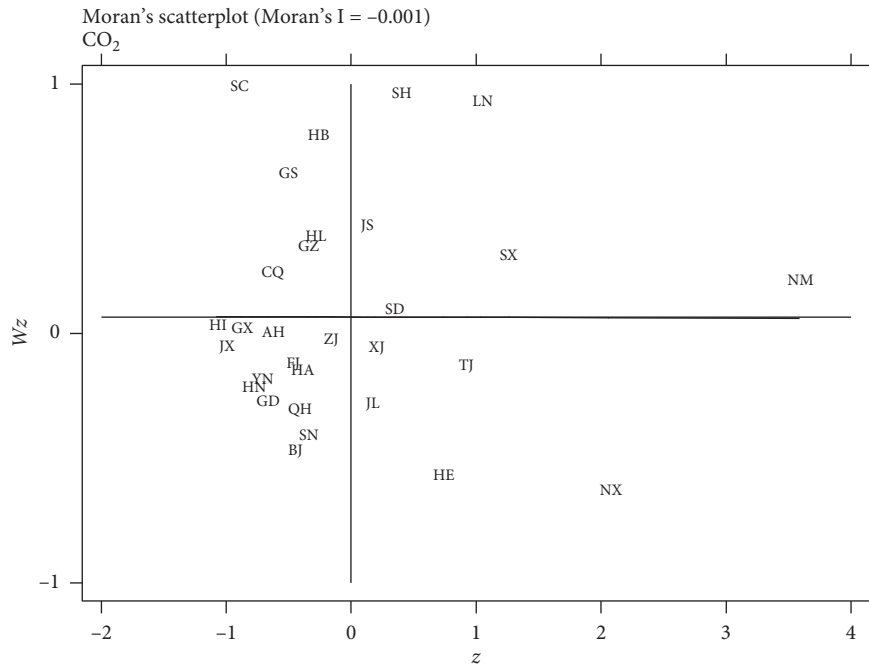
FIGURE 3: Moran's scatter plot in 2010.
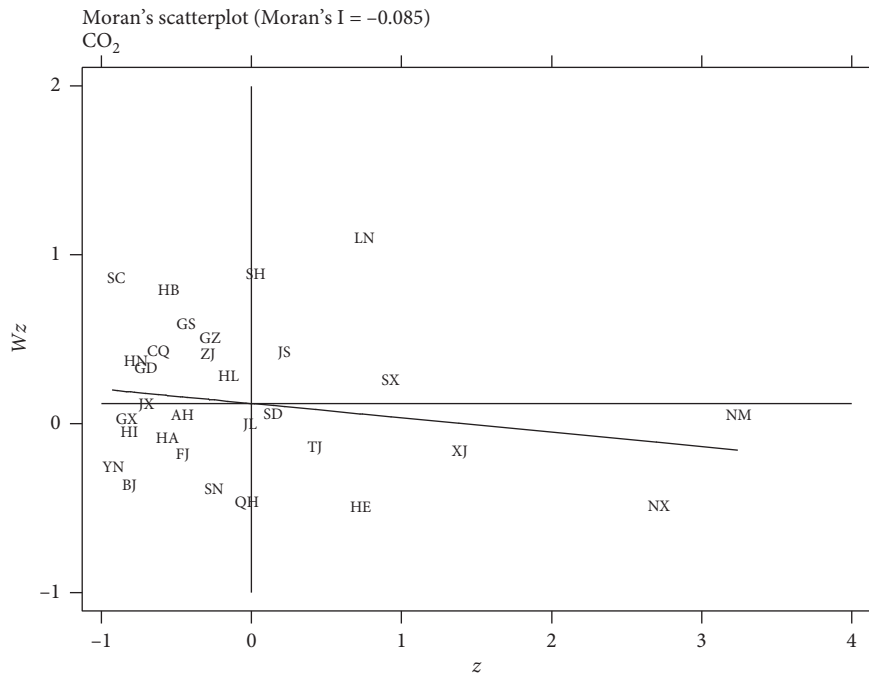
FIGURE 4: Moran's scatter plot in 2015.

In addition, it can be seen from the table that coefficient of WGDP is significantly positive, indicating that the positive spatial spillover effect of economic growth on carbon dioxide emissions is obvious.

There is a two-way relationship between energy consumption (ENC) and carbon dioxide emissions (CO_2). The increase of energy consumption will significantly promote the increase of carbon dioxide emissions, leading to the continuous deterioration of the environment. At the same time, the increase of carbon dioxide emissions will continue to increase energy consumption. For energy-dependent countries such as China, the constraint target of carbon dioxide emission will restrict the continuous growth of energy consumption and form the energy bottleneck constraint of economic development.

Moran's scatterplot (Moran's I = −0.096)
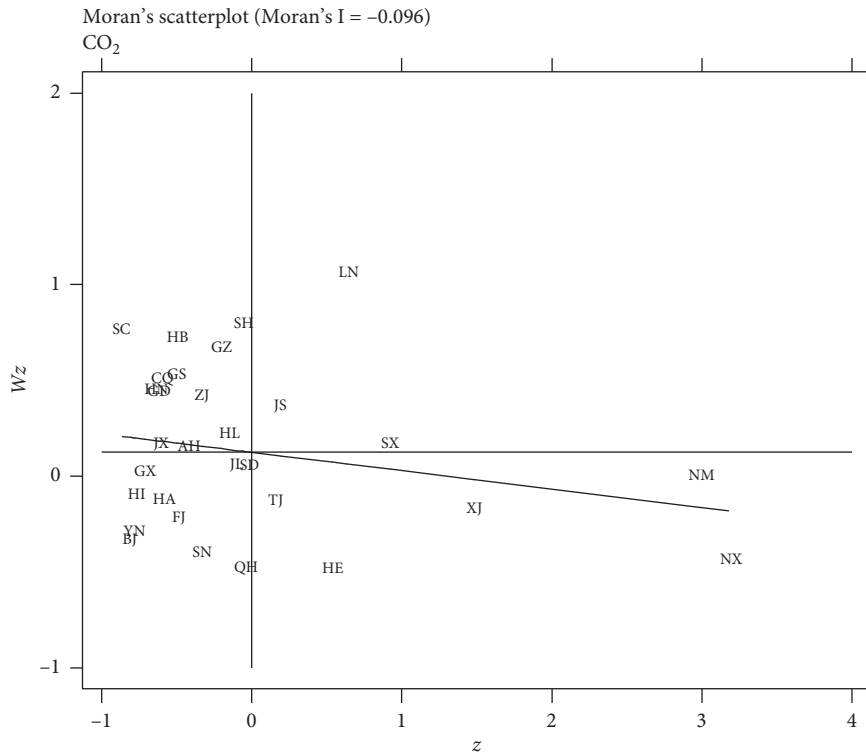
FIGURE 5: Moran's scatter plot in 2017.

TABLE 4: LM, Hausman, and LR tests.

Test type		St	p
LM test	LM (error)	4.508438	0.00
	LM (lag)	8.65328	0.00
Hausman test		0.24	0.06262
LR test		24.34	0.00

TABLE 5: Spatial econometric model test.

Hypothesis	SDM will degenerate into SAR	SDM will degenerate into SEM
LR test value	27.15	32.42
p value	0.0032	0.0014

TABLE 6: Estimation results of SDM.

Variables	CO_2		GDP		ENC	
	Coef.	p value	Coef.	p value	Coef.	p value
CO_2	—	—	0.680361	0.000	0.17246	0.000
GDP	0.197398	0.0353	—	—	0.20499	0.000
ENC	0.41966	0.000	0.845657	0.000	—	—
WCO_2	—	—	0.1235264	0.000	0.5094	0.000
WGDP	0.155793	0.393	—	—	0.19691	0.000
WENC	0.1576883	0.023	0.1436614	0.003	—	—
sigma2_e	0.0110605		0.0004057		0.001725	
Log-likelihood	389.9832		81.5463		528.7699	
R-sq	0.936		0.9103		0.9553	

6.2.4. Direct Effect, Indirect Effect, and Total Effect of Spatial Durbin Model. The coefficient of carbon dioxide emission cannot be used to directly explain its impact on economic growth and energy consumption, which is mainly due to the spatial spillover effect of the spatial Durbin model. For this reason, it is necessary to decompose the total effect of space deeply. According to the direct effect, indirect effect, and total effect of space Durbin model, we can better explain the effect of carbon dioxide emission on carbon dioxide emission in this region and other regions and the spillover effect between regions. The results are shown in Table 7.

The direct effect of spatial Durbin model: under the 0, 1 adjacency weight matrix, the influence coefficient of economic growth (GDP) on carbon dioxide emissions in this

TABLE 7: Direct, indirect, and total effects of SDM.

	Direct effect		Indirect effect		Total effect	
	Coef.	Z-value	Coef.	Z-value	Coef.	Z-value
GDP	0.021308*	0.53	0.152586**	0.71	0.036566*	0.65
ENC	1.035705**	10.55	−0.76671*	−0.95	0.959034**	8.67

region is 0.21308, and it has passed the significance test of 10%, which shows that economic growth can significantly promote carbon dioxide emissions in this region. Economic growth can stimulate people's consumption demand and improve people's consumption capacity. Demand-driven production has become an important driving force for the increase of carbon dioxide emissions. The coefficient of influence of the explanatory variable energy consumption (ENC) on carbon dioxide emissions in this region is 1.035705, and it has passed the 5% significance level test, indicating that energy consumption plays a role in promoting carbon dioxide emissions.

The indirect effect of spatial Durbin model: under the 0, 1 adjacency weight matrix, the influence coefficient of economic growth (GDP) on carbon dioxide emissions of surrounding provinces and cities is 0.152586, which has passed the 5% significance test. The coefficient of influence of the explanatory variable energy consumption (ENC) on the surrounding provinces and cities is −0.76671, which has passed the significance level test of 10%, indicating that the energy consumption has a negative spatial spillover effect on the carbon dioxide emissions of the surrounding provinces and cities.

7. Conclusion

Based on the test of data and the selection of appropriate spatial econometric model, this paper selects the provincial panel data of China from 2000 to 2017, first estimates the carbon dioxide emissions of 30 provinces and cities in China, and then takes the spatial Durbin model (SDM) to conduct in-depth research and discusses on the nexus among carbon dioxide emissions, energy consumption, and economic growth in China. The empirical analysis results show the following:

(1) From the results of the global Moran test, we can see that there is a significant positive spatial correlation effect on China's carbon dioxide emissions, while the local Moran test shows that China's carbon dioxide emissions show the characteristics of "high-high" and "low-low" spatial clusters.

(2) From the SDM test results, we can see that energy consumption has a significant positive effect on carbon dioxide emissions in the region and a significant negative spatial spillover effect on carbon dioxide emissions in the surrounding provinces and cities.

Energy consumption has a decisive impact on carbon dioxide emissions. The energy consumption elasticity of

carbon dioxide emissions is stable at 15%–50%. Energy consumption and carbon emissions are interrelated. Therefore, energy consumption is a crucial factor in carbon dioxide emissions. According to the IPCC Research Report and the world bank's calculation of carbon dioxide emissions, the energy consumption of fossil fuels is the main source of carbon dioxide emissions, among which coal is the fossil energy with the largest emission coefficient. Particularly for China, in the short term, it is difficult to fundamentally change the structure of coal-based high-carbon energy consumption, and the constraint goal of carbon dioxide emissions is difficult to achieve. However, in the long run, the establishment of a positive constraint target on carbon dioxide emissions can serve as a target to control fossil energy consumption and promote the development of new and renewable energy resources, forming a "forced" mechanism."

Energy is an important input factor for economic development. At present, China is still a traditional high-carbon energy system dominated by oil and coal. Therefore, the rigid demand of economic growth for energy consumption leads to the continuous increase of carbon dioxide emissions, forming the relationship chain of "economic growth⟶energy consumption⟶carbon dioxide emission," which makes the global climate change face an increasingly urgent situation. China must change the mode of economic growth and optimize industrial structure and energy structure. The government can release relevant policies to reduce the proportion of fossil energy and improve the efficiency of energy use [13].

Data Availability

This paper studies the relationship among carbon dioxide emissions, energy consumption, and economic growth in 30 provinces of China by using the Spatial Durbin Model. The data are from China Statistical Yearbook and China Energy Statistical Yearbook. According to the availability of data, the sample interval of this paper is 2000–2017. Each variable is an annual variable, which adopts the form of logarithmic transformation. At present, most of these studies use per capita GDP to express the level of economic development. This paper also uses per capita GDP as the representative variable of economic growth. The authors use energy consumption per capita (kg oil equivalent/person) to represent energy consumption. For carbon dioxide emissions, per capita carbon dioxide emissions (metric tons/person) are selected as the research variable. These three variables are expressed by GDP_{in}, ENC_{in}, and CO_2 respectively.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this paper.

Authors' Contributions

All authors contributed equally to this work. All authors read and approved the final manuscript.

Acknowledgments

References

[1] R. Schmalensee, R. A. Stoker, and R. A. Judson, "World carbon dioxide emissions: 1950-2050," *Review of Economics and Statistics*, vol. 80, no. 1, pp. 15–27, 1998.

[2] B. Saboori and J. Sulaiman, "CO2 emissions, energy consumption and economic growth in Association of Southeast Asian Nations (ASEAN) countries: a cointegration approach," *Energy*, vol. 55, pp. 813–822, 2013.

[3] N. Apergis, "Environmental Kuznets curves: new evidence on both panel and country-level CO_2 emissions," *Energy Economics*, vol. 54, pp. 263–271, 2016.

[4] M. Salahuddin, K. Alam, I. Ozturk et al., "The effects of electricity consumption, economic growth, financial development and foreign direct investment on CO2, emissions in Kuwait," *Renewable and Sustainable Energy Reviews*, vol. 81, pp. 2002–2010, 2017.

[5] D. Coondoo and S. Dinda, "Causality between income and emission: a country group-specific econometric analysis," *Ecological Economics*, vol. 40, no. 3, pp. 351–367, 2002.

[6] M. Hussain, M. I. Javaid, and P. R. Drake, "The CO2 Emission, Energy Consumption & Economic Growth of Pakistan," *International Journal of Energy Sector Management*, vol. 6, no. 4, pp. 518–533, 2012.

[7] P. K. Narayan and R. Smyth, "Energy consumption and real GDP in G7 countries: new evidence from panel cointegration with structural breaks," *Energy Economics*, vol. 30, no. 5, pp. 2331–2341, 2008.

[8] J. Stamler, R. Stamler, and R. B. Shekelle, "Energy consumption and real GDP in G7 countries: new evidence from panel co-integration with structural breaks. Energy Economics," *Energy Economics*, vol. 30, no. 5, pp. 2331–2341, 2010.

[9] S.. Zeng, "Research on energy consumption, economic growth and energy demand prediction in China," *Management Review*, vol. 23, no. 2, pp. 38–44, 2011.

[10] J. B. Ang, "CO2 emissions, energy consumption, and output in France," *Energy Policy*, vol. 35, no. 10, pp. 4772–4778, 2007.

[11] N. Apergis and J. E. Payne, "CO2 emissions, energy usage, and output in Central America," *Energy Policy*, vol. 37, no. 8, pp. 3282–3286, 2009.

[12] M. E. H. Arouri, A. Ben Youssef, H.M henni, C. Rault et al., "Energy consumption, economic growth and CO2 emissions in Middle East and North African countries," *Energy Policy*, vol. 45, pp. 342–349, 2012.

[13] F. Halicioglu, "An econometric study of CO2 emissions, energy consumption, income and foreign trade in Turkey," *Energy Policy*, vol. 37, no. 3, pp. 1156–1164.

[14] B. Saboori, J. E. Rasoulinezhad, and J. Sung, "The nexus of oil consumption, CO2 emissions and economic growth in China, Japan and South Korea," *Environmental Science and Pollution Research*, vol. 24, no. 8, pp. 7436–7455, 2017.

[15] T. Feng, H. Du, Z. Lin, and J. Zuo, "Spatial spillover effects of environmental regulations on air pollution: evidence from urban agglomerations in China," *Journal of Environmental Management*, vol. 272, Article ID 110998, 2020.

[16] Y. Song, J. Sun, M. Zhang, and B. Su, "Using the Tapio-Z decoupling model to evaluate the decoupling status of China's CO2 emissions at provincial level and its dynamic trend," *Structural Change and Economic Dynamics*, vol. 52, pp. 120–129, 2020.

[17] S. Yan, Z. Ming, and S. Cheng, "Research on the Decoupling Trend and Mitigation Potential of CO2 Emissions from China's Transport Sector," *Energy*, vol. 183, pp. 837–843, 2019.

[18] Y. Song, M. Zhang, and M. Zhou, "Study on the decoupling relationship between CO2 emissions and economic development based on two-dimensional decoupling theory: a case between China and the United States," *Ecological Indicators*, vol. 102, pp. 230–236, 2019.

[19] Y. Song, S. M. Zhang, and S. Dai, "Study on China's energy-related CO2 emission at provincial level," *Natural Hazards*, vol. 77, no. 1, pp. 89–100, 2015.

[20] E. S. Mills and B. W. Hamilton, *Urban Economics*, pp. 23–66, Harper Collins College Publishers, NY, Newyork, 1994.

[21] K. J. Button, *Urban Economics: Theory and Policy*, pp. 105–121, The Mac Millan Press, London, UK, 1976.

[22] P. Lesage and R. K. Pace, *Introduction to Spatial Econometrics*, Taylor and Francis Group, Oxfordshire, UK, 2009.

[23] F. Wang, F. Genfu, and L. Wu, "Decomposition of provincial and regional contribution to the decline of carbon intensity in China's economic growth," *Economic Research*, no. 8, pp. 143–155, 2013.

[24] S. Chen, "The fluctuating and decreasing pattern of China's carbon emission intensity and its economic explanation," *World Economy*, vol. 4, pp. 124–143, 2011.

Research Article

Optimizing the Procurement of IaaS Reservation Contracts via Workload Predicting and Integer Programming

Huamin Zhu ⓘ, Jun Luo ⓘ, and Hongyao Deng ⓘ

College of Big Data and Intelligent Engineering, Yangtze Normal University, Chongqing, China

Correspondence should be addressed to Huamin Zhu; zhuhuamin2001@163.com

Cloud-based web applications are proliferating fast. Owing to the elastic capacity and diverse pricing schemes, cloud Infrastructure-as-a-Service (IaaS) offers great opportunity for web application providers to optimize resource cost. However, such optimization activities are confronting the challenges posed by the uncertainty of future demand and the increasing reservation contracts. This work investigates the problem of how to minimize IaaS rental cost associated with hosting web applications, while meeting the demand in the future business cycle. First, an integer liner program model is developed to optimize reservation-contract procurement, in which reserved and on-demand resources are planned for multiple provisioning stages as well as a long-term plan, e.g., twelve stages in an annual plan. Then, a Long Short-Term Memory (LSTM) based algorithm is designed to predict the workload in the future business cycle. In addition, the approaches for determining virtual instance capacity and the baseline workload of planning time slot are also presented. Finally, the experimental prediction results show the LSTM-based algorithm gains an advantage over several popular models, such as the Holter–Winters, the Seasonal Autoregressive Integrated Moving Average (SARIMA), and the Support Vector Regression (SVR). The simulations of resource planning show that the provisioning scheme based on our reservation-optimization model obtains significant cost savings than other typical provisioning schemes, while satisfying the demands.

1. Introduction

Cloud computing is a large-scale distributed computing paradigm in which a pool of computing resources is available via the Internet. As the most widely applied service model in cloud computing, the IaaS liberates organizations from the expensive infrastructure investment with the virtually infinite resources and the elasticity. In this model, infrastructure resources such as computing, storages, and networks can be rented to customers in the form of virtual machine instances. Each instance belongs to a specific instance type specifying the hardware configuration (CPU cores and speed, memory, and I/O channels). The consumers can quickly deploy the packaged OS and application images to the leased IaaS instances and start them. Meanwhile, web application workload generally exhibits inherent seasonality, stochastic volatility, and aggregated volatility [1]. Therefore, web applications are well suitable for the deployment on the instances rented from IaaS providers, as it makes easy to quickly scale resources so as to deal with varying workload. For example, the 12306 e-ticket site in China is now very stable [2], but before being deployed to the cloud, it gets stuck or even crashed almost whenever the peak of visits appears.

IaaS providers usually offer customers two types of resource provisioning plans, namely, on-demand and reservation plans with different charging schemes. The on-demand plans charge customers on a pay-as-you-go basis and enable them to start or terminate instances at any moment according to needs without paying any penalty. However, comparing the unit price, the on-demand resources are often more expensive than the reserved ones. With the reservation plans, virtual instances are reserved in the form of long-term contracts. Through the use of reservation plans, customers can get significant price discount compared with on-demand plans and pay once for the

contract duration (e.g., 1 month, 3 months, 6 months, 9 months, or 1 to 5 years at Aliyun [3], 6 to 36 months at Rackspace [4], and 1 year or 3 years at Amazon [5]). Taking an Aliyun's ecs.g5.large instance in the Region Qingdao of North China, for example, compared with the on-demand plan, the discount rates of monthly fees for 1 month, 3 months, 6 months, and 1 year reservation contracts are 60.9%, 64.8%, 66.7%, and 69.0%, respectively.

In fact, for the web applications with time-varying workload, using only reserved resources or on-demand resources is generally not the best choice. Imagine a web application with changing resources demand, as shown in the curve of Figure 1. If only reserved resources are planed, e.g., N_H instances are reserved, then lots of instances will not be efficiently utilized, resulting in significant waste of resources.. On the contrary, if only on-demand resources are used, high unit price of resources will lead to a large total cost. Apparently, the best decision is to reserve N_R (namely, a number between N_H and N_L) instances and then supplement several on-demand instances when needed. As such, an optimal total cost can be obtained, while meeting the workload demand.

Nowadays, more and more web applications are migrated to the cloud. Meanwhile, more and more IaaS reservation contracts are also offered by cloud providers. For web applications providers, it has become very necessary to optimize the provisioning of IaaS resources for saving cost. However, most of the existing approaches have employed the deterministic resources provisioning schemes [6–10]. In these studies, the uncertain nature of the user's demands is neglected by assuming the demand as a deterministic value. To address the demand uncertainties, in [11–14], some dynamic resource provisioning schemes are proposed. These schemes are more flexible and provision resources dynamically to meet fluctuating workload. However, these studies do not exploit the cost benefits of reservation contracts, resulting in failure to achieve economical solutions. Given the disadvantages of two categories of schemes above, several studies have employed the hybrid schemes to provision resources [15–18]. Although the decision making is more complex, the hybrid schemes take advantage of reserved and on-demand resources simultaneously so as to save cost, while better meeting varying demand. The hybrid provisioning scheme is generally carried out in two phases. Prior to the start of the workload cycle, the resource-reservation contract procurement is planned in advance based on an estimated or predicted workload. During the workload cycle, the previous obtained reservation plans are carried out successively and then the reserved resources are utilized, while additional on-demand resources may be provisioned whenever necessary.

For cloud-based web applications, we prefer the hybrid scheme and believe that an excellent provisioning scheme should use as many reserved resources as possible to satisfy long-term stable demands in the future and only use a small amount of on-demand resources to deal with sudden demands so as to minimize the total resources cost. However, as more and more IaaS reservation contracts are offered, for

a long web application workload cycle, how to combine multiple reservation contracts as well as determine the numbers and start times of them so as to optimize the total cost? is the first major challenge for the IaaS resource decision-makers of web applications.

Besides, planning resource for the future business cycle of a web application requires an estimation/prediction of the future workloads. Some studies have employed the simulated workloads [6, 9, 11]; meanwhile, some studies directly take the workloads in the historical cycle as an estimation of the workloads in the future cycle [14, 15, 17]. But, the two approaches generally could not obtain a good accuracy. There are also some studies to develop the stochastic programming models for future workloads based on the historical workloads' summary [18–20] (e.g., the mean and standard deviation). However, such models are only applicable to stochastic workload series and cannot handle the workload series with the trend and seasonality. The most widely employed schemes are workload predictions. One group of prediction approaches for web workloads is statistical models such as Autoregression (AR), Autoregressive Moving Average (ARMA), Autoregressive Integrated Moving Average (ARIMA) [21], Exponential Smoothing (ES) [22], and Linear Regression (LR) [13] models. These statistical models are effective for the short-term prediction of stationary series, while their prediction accuracy for nonstationary series is very poor. This is because erratic fluctuations, which are typical for web workload series, are practically impossible to predict. This problem can be resolved by using machine-learning techniques such as the Support Vector Regression (SVR) [23–25] and the Deep Belief Networks (DBN) [26]. The advantages of these approaches are that they can learn from historical data (search connections among features) and build a prediction model for future workloads [27]. However, due to the lack of long-term memory ability, these models are still difficult to learn long-term inherent patterns of workload series. In view of the facts that web workload is often affected by many factors and the workload cycle is usually long, how to deal with the uncertainty of workload prediction in the future business cycle for web applications? is the second major challenge for the IaaS resource decision-makers of web applications.

The Recurrent Neural Network (RNN) is a novel neural network architecture specially designed for the sequence data and has been proven successful in time series prediction tasks [28]. However, a traditional RNN performs poorly at handling long-term dependencies, mainly due to the exploding and vanishing gradient problem [29, 30]. As a redesigned architecture of RNN, the LSTM network addresses the shortcomings by replacing the RNN cell with an LSTM cell in the hidden layer and thus has the ability of learning long-term dependencies [31]. Owing to the superior long-term memory ability, the LSTM exhibits excellent potential for predicting the long time series. There have been several good attempts on applying the LSTM to carry out the mid- and long-term prediction for time series, such as traffic flow [32], bank business [33], and earthquakes [34]. Under such background, we choose to employ the LSTM for
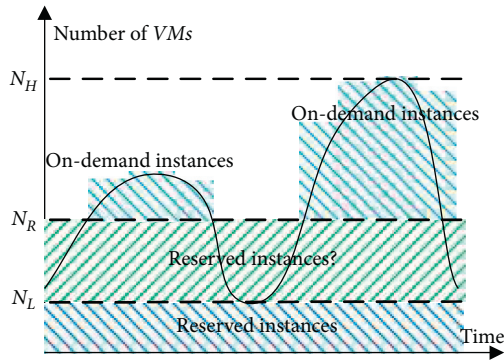
FIGURE 1: A workload example.

predicting the future cycle workload of web applications and evaluate its prediction performance by comparing with other popular approaches.

In addition, planning resource for the future business cycle of a web application depends not only on the future workload but also on the processing capacity of each IaaS instance. However, this value is not fixed and is closely related to the threshold of service response time. At the same time, for a long planning cycle, because it is hard to perform fairly fine granularity of prediction, the duration of planning time slot is not usually short, such as a day. Also, the predicted workload is generally the total or average number of requests in a time slot. If the predicted average values are directly used as the baseline workload for planning slot, it is clear that the underprovisioning will happen frequently. Therefore, how to determine the processing capacity of single instance and the baseline workload of planning time slot so that planned resources can better meet the demand? is the third major challenge for the IaaS resource decision-makers of web applications.

This work focuses on planning resource reservations prior to the beginning of the workload cycle, aimed at achieving the optimal plan of IaaS reservation contract procurement through the use of workload prediction. We studied the above several challenge problems in depth, and the major contributions are threefold.

(1) Based on the divisions of the provisioning cycle and the description of reservation contracts, an integer linear program model is developed to optimize reservation contract procurement.

(2) Given the inherent pattern of web cycle workload series, a Long Short-Term Memory (LSTM) network-based algorithm is designed to predict the cycle workload of web applications.

(3) The approach for determining instance capacity is presented by using an $M/M/n$ queuing system. Time-slot baseline workload is also determined based on the average of historical peak workloads.

The remainder of this paper is organized as follows. Section 2 discusses the related work. Section 3 presents the problem domain and assumptions. Section 4 develops the reservation contract procurement-optimization model. In Section 5, the LSTM-based prediction algorithm is designed. In Section 6, the approaches for determining the instance capacity and time-slot baseline workload are introduced. Experimental settings and results are presented in Section 7, followed by conclusions and future work in Section 8.

2. Related Work

In the past years, the problem of cloud resource planning has attracted many researchers' attention to develop resource provisioning algorithms and techniques [6, 9, 35–37]. A deeply survey can separate the studies into three categories: deterministic resource provisioning, dynamic resource provisioning, and hybrid resource provisioning. In the following sections, the existing studies are discussed in these categories, and finally the prediction approaches for web application workload are also discussed.

2.1. Deterministic Resource Provisioning. Most of the studies model this problem as a single phase optimization algorithm that only considers resources with reserved contracts from IaaS providers. These studies neglect the uncertainty of users' demands and regard the demands as fixed values and then employ deterministic provisioning schemes to deal with future workload [8, 9]. Mireslami et al. [6] planned the number of service instances according to the instance's minimum service rate. Imai et al. [7] used an expensive overprovisioning scheme for the worst-case demand. Jiao et al. [38] designed a cost optimization model for online social network deployment in geo-distributed clouds. The work regarded the demand of each cloud as a deterministic value. Similarly, in [10], a multiobjective algorithm was developed to minimize total deployment cost and maximize service of quality (QoS) performance. Chen et al. [9] constructed a resource cost optimization model for periodical workflow applications based on fixed workload.

Deterministic resource provisioning is better suited for constant workload scenarios (e.g., batch processing tasks) rather than web applications with varying workload.

2.2. Dynamic Resource Provisioning. In order to deal with the uncertainty of users' demands, some studies employ elastic mechanisms to provision cloud resources. Zhao et al. [11] constructed a resource cost optimization model for computational and data intensive applications, which is performed periodically at hourly intervals. Antonescu et al. [13] dynamically adjusted resources to meet predicted short-term workload so as to minimize the cost, while avoiding the service level agreement (SLA) violations. Sniezynski et al. [14] used linear regression, neural networks, etc., to learn resource usage patterns from the historical records so as to predict and update resource capacity periodically.

Although these dynamic provisioning schemes better meet the varying demands, they result in considerable cost because of using only expensive on-demand resources.

2.3. Hybrid Resource Provisioning. The hybrid resource provisioning uses deterministic reserved resources to deal with long-term stable workload and uses dynamic on-demand resources to deal with short-term sudden workload. Stijven et al. [39] proposed a scheme to plan reserved resources based on short-term workload prediction but only one kind of contract could be used. Candeia et al. [15] designed the algorithms to select IaaS reservation markets and determine the numbers of instances as well as their lifespans, without considering multiple kinds of contracts simultaneously. Similarly, Chen et al. [17] also presented a hybrid short-term provisioning scheme that could only include one contract type. Mireslami et al. [18] proposed two-stage provisioning scheme for web applications. In the first stage, they decide which contract to purchase based on the minimum workload, and in the second stage, additional on-demand resources was provisioned dynamically. Their scheme is similar to this work but only one type of contracts is considered.

For all above studies, only one reservation-contract type can be included, and the reserved resources are constant. In this work, the problem of reservation-resource planning for the entire workload cycle is investigated. A workload cycle is divided into multiple provisioning stages uniformly so that multiple reservation contract types with different durations can be combined to provision resources so as to obtain a minimum total cost.

2.4. Prediction of Web Application Workload. Calheiros et al. [21] presented the realization of a workload prediction module for cloud-based applications based on the ARIMA. However, the ARIMA cannot deal with the seasonal variations of workload series. Dhib et al. [40] employed the SARIMA to fit the workload of the Massively Multiplayers Online Gaming and allocated resources according to predicted workload. Although the experimental results show that the quality of experience is improved, the SARIMA still cannot fit the nonlinear variations of the workload well. Ma et al. [26] designed a workload-prediction algorithm for web applications based on the Deep Belief Networks but only verified its short-term prediction effect. Zhao et al. [23] employed the SVR to predict the workload of web application, and the prediction accuracy reached 89% but only verified the short-term prediction for future three steps. Singh et al. [41] proposed an adaptive prediction model for web application workload using Linear Regression, ARIMA, and SVR models. Similarly, they only verified short-term prediction effect. Given the sufficient long-term memory ability of the LSTM, some scholars attempted to employ it for predicting the long time series. Tian et al. [32], Liu et al. [33], and Wang et al. [34] designed the mid- and long-term prediction models for the traffic flow, the reserve requirements of bank outlets, and the earthquakes based on the LSTM, respectively. As a result, they all obtained good prediction accuracy. However, so far the LSTM was still seldom applied in the mid- and long-term prediction for web-application workloads. Kumar et al. [28] employed the LSTM to carry out the long-term prediction for HTTP requests to web servers in cloud datacenter and claimed to have obtained ideal results, but they did not present the details of the design. Tran et al. [42] designed a LSTM-based algorithm for predicting cloud resource consumption with multivariate time series, but only verified the short-term prediction effect. By contrast, we specially designed a LSTM-based long-term prediction algorithm for the future cycle workload of web application and presented the design details. From experimental results, our LSTM-based prediction algorithm outperforms existing common models and achieves a good accuracy.

3. Problem Domain and Assumptions

First, this work targets at interactive web applications deployed in IaaS cloud. There are various applications deployed in IaaS cloud, such as interactive applications [2, 43], scientific computing [44, 45], and batch processing tasks [46, 47]. Among them, interactive applications usually have a certain business cycle (e.g., one year), the workloads of which generally show similar patterns in the long run, while being stochastic in the short run. Due to the complexity of enterprise-level application architecture, it is difficult to conduct general research on the resource planning of whole application. But, the large application is generally orchestrated by multiple web services or subapplications. Especially in the rise of microservices architecture today, more and more web subapplications run independently as services. This work focuses on planning reserved resources for such web services or subapplications. In addition, such a subapplication is generally composed of several components such as web server, database server, and hard disk. Among them, nonservice components can be statically configured, and service-oriented components need to be scalable. According to the experience knowledge of web development and operation, as long as the numbers of instances of several service-oriented components satisfy a certain ratio with each other, the system can be in a stable state. This ratio can be obtained through application-specific benchmarking.

Next, this work does not involve the IaaS discovery and the selection of cloud providers. These problems belong to another research domain. We assume that the matching instance type of each service component has been found, and the provider has also been selected.

Additionally, only the horizontal scaling scheme is considered in this work. Horizontal scaling adjusts service capacity through dynamically changing the number of instances, while vertical scaling does this through dynamically changing the instance's configuration. However, most providers have not yet opened the services to support dynamic vertical scaling.

Finally, in view of the fact that most providers have sufficient resource capacity nowadays, it is assumed that the provisioning of on-demand instances is not restricted by the quantity. Also, we assume that all reservation contracts are paid completely in advance so as to obtain a larger discount and simplify the problem although several providers also support partial payment.

4. Problem Description and Model Construction

4.1. Provisioning Phases. As illustrated in Figure 2, over the provisioning time horizon, there are three provisioning phases: reservation, utilization, and on-demand phases. The corresponding actions of these phases are performed in different points of time (or events). In the short reservation phase, the decision maker develops a resource-reservation plan and conducts it. In the following utilization phase, the reserved instances are used to deal with incoming workload. During the ongoing utilization phase, once the workload exceeds the processing capacity of reserved instances, an on-demand phase starts, during which additional on-demand instances are provisioned. The reservation and utilization phases always appear in pairs in a sequential order. A utilization phase may contain several on-demand phases. Over the provisioning horizon, there may be multiple pairs of reservation and utilization phases, and the reservation durations may be contained or overlapped by each other.

4.2. Divisions of Resource Provisioning Cycle. As illustrated in Figure 3, we regard a web-application's business cycle as its resource provisioning cycle, namely, resource planning cycle, which consists of several equal-duration provisioning stages.

4.2.1. Resource Planning Cycle. Let T denote a resource planning cycle, which is a relatively long workload-processing cycle defined by the web-application provider. The cycle has a definite beginning and a definite end. During the cycle, although the workload seems to fluctuate randomly in the short term, there is usually a certain pattern implied in workloads from long-term observations. This makes it possible and meaningful to plan resources for a business cycle. Since such a cycle is generally long (e.g., one year), multiple reservation contracts with equal duration or unequal duration can be included in the plan so as to obtain a lower total cost.

4.2.2. Provisioning Stage. As shown in Figure 3, a resource planning cycle T can be divided into several provisioning stages uniformly. Let T_i be the i-th provisioning stage. The duration of a provisioning stage is generally equal to the greatest common divisor of the durations of all reservation contracts so as to ensure that each contract can cover an integer number of stages. For example, an annual planning cycle can be divided into 12 monthly provisioning stages T_1, $T_2, \ldots, T_{12}$. Each provisioning stage can contain one reservation phase ΔT and the whole or part of utilization phases (namely, a utilization phase may cover one or more provisioning stages), as well as one or more on-demand provisioning phases. In particular, as seen in Figure 3, the optimal procurement plan of reservation contracts for the entire cycle T is decided in the phase ΔT_1 of the first stage T_1, and the subplan of procurements corresponding to T_1 is also carried out in ΔT_1. In each subsequent ΔT, its corresponding contract procurements are carried out according to the optimal plan developed in ΔT_1.

4.2.3. Provisioning Time Slot. Due to the workload is usually fluctuating during a provisioning stage, it is not appropriate to provide fixed resources during a provisioning stage. Therefore, as presented in Figure 3, we divide each provisioning stage into several provisioning time slots (e.g., T_{11}, T_{12}, and T_{24}) uniformly for planning resources. Due to the duration of any reserved contract is not shorter than the one of any provisioning stages, the available reserved resources are exactly same for all slots in the same stage. Given it is difficult to obtain the fairly fine granularity of predicted workloads, the duration of each slot is usually set as one day.

4.3. Reservation Contracts. An IaaS provider usually offers multiple reservation-contract types with different durations for consumers. Let K be the set of reservation-contract types, and any contract type $k \in K$ can be expressed as $\langle v, l, p^r \rangle$, where v, l, and p^r denote the offered instance type, contract duration, and unit price, respectively. To describe the conditionality of procurement and utilization of reservation contracts, an annual plan case with 12 months (K_1), 6 months (K_2), 3 months (K_3), and 1 month (K_4) reservation contracts is illustrated in Figure 4. The boxes over the time horizon represent the time coverage of these contracts.

We take the planning cycle $T = \{T_1, T_2, \ldots, T_{12}\}$. Let $|k|$ denote the duration (in unit of provisioning stages) of any k-type contract. Due to only the contracts with a duration of not longer than T are considered, $|k| \leq |T|$. Let T_k denotes the set of stages at which IaaS providers can start provisioning resources with a k-type contract, and then T_k can be expressed as formula (1). According to Section 4.2.2, T_k is also the set of stages at which a k-type contract can be purchased. This is because only when a k-type contract is purchased at the stages from T_k can this contract be properly terminated during T.

$$\mathbf{T}k = \{1, \ldots, |T| - |k| + 1\}. \tag{1}$$

Let F_{ki} denote the set of stages at which some resources reserved by a k-type contract can be utilized in the stage T_i. It means that only when the k-type contracts are purchased at stages belonging to F_{ki}, the resources reserved by these contracts can be utilized during the i-th stage. F_{ki} can be expressed as formula (2). In Figure 4, any set F_{ki} can be obtained. For example, $F_{K_1 T_3} = \{T_1\}$, $F_{K_2 T_3} = \{T_1, T_2, , T_3\}$, $F_{K_3 T_{11}} = \{T_9, T_{10}\}$, and $F_{K_4 T_3} = \{T_3\}$. Let n^r_{vki} be the number the k-type contracts with instance type v available in the stage T_i. Let r_{vkj} be the number of the k-type contracts with instance type v purchased at the stage T_j. Based on F_{ki} and r_{vkj}, n^r_{vki} is calculated by using the following formula:

$$F_{ki} = \{\max(1, i - |k| + 1) \& \min(i, |T| - |k| + 1)\}, \tag{2}$$

$$n^r_{vki} = \sum_{j \in F_{ik}} r_{vkj}. \tag{3}$$
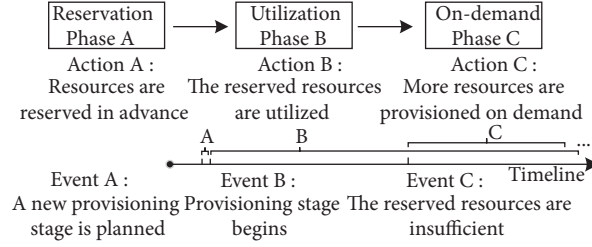
Figure 2: Transition of provisioning phases.
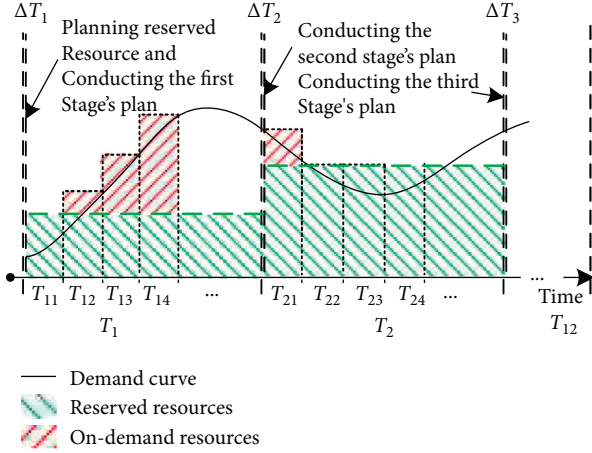
Figure 3: Divisions of a resource planning cycle.
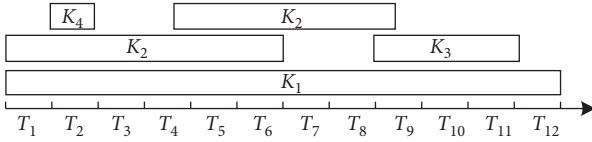
Figure 4: An annual plan case with 5 reservation contracts.

4.4. Model Construction. We choose to perform resources planning based on each time slot rather than each provisioning stage. Owing to the fine granularity of a time slot, the planned resources based on time slots are more adaptable to fluctuating demand, and the amount of overprovisioning and underprovisioning can be reduced greatly. Thus, for a web application service, the resource planning goal is to minimize the total cost of reserved and on-demand resources, while meeting any time-slot's demand in the entire business cycle.

For a specific web application, according to the assumptions in Section 3, several instance types have been selected for its service components. The processing capacity of single web-server instance as well as the optimal ratios between the other server instances and the web-server instances have been determined (the method for determining

the former is presented in Section 6, while the latter can be obtained by benchmarking). Besides, it is also assumed that the workload of each time slot has been predicted. Based on these assumptions, the numbers of various service instances required in each time slot are determined. Finally, we have defined some necessary parameters as presented in Table 1 so as to construct the optimization model.

In particular, in the context of horizontal scaling, a service component is deployed on a cluster of instances with the same type; therefore, the instance types correspond to the component types one by one. In addition, let S_i represent the number of time slots in the i-th provisioning stage for $S_i \in S$, while let s represent any one in S_i. Based on defined parameters and variables above, the model is constructed as follows.

First, the cost of all reservation contracts charged to v-type instances in T_i, namely, c_{vi}^r, is expressed as formula (4). Note that r_{vki} is equal to 0 when $i \notin \mathbb{F}_k$.

$$c_{vi}^r = \sum_{k \in K} p_{vk}^r r_{vki}. \tag{4}$$

Next, because the number of v-type instances reserved by k-type contracts available in T_i, namely, n_{vki}^r, is obtained by formula (3), the number of v-type instances available in T_i, namely, n_{vi}^r, can be expressed as follows:

$$n_{vi}^r = \sum_{k \in K} \sum_{t \in F_{ki}} r_{vkt}. \tag{5}$$

In addition, the cost of all v-type instances provisioned on demand in T_i, namely, c_{vi}^o, is expressed as follows:

$$c_{vi}^o = \sum_{j \in S_i} p_v^o \cdot n_{vij}^o \cdot |s|, \tag{6}$$

where the number n_{vij}^o is equal to the maximum of 0 and $n_{vij}^d - n_{vi}^r$, n_{vij}^d denotes the number of v-type instances required in the j-th time slot of the stage T_i, and $|s|$ denotes the number of hours in time slot s.

As a result, the total cost of v-type instances provisioned in T_i, namely, c_{vi}, is equal to the sum of reservation cost c_{vi}^r and on-demand cost c_{vi}^o in T_i. Therefore, the total cost of v-type instances provisioned in the entire planning cycle, namely, c_v, can be expressed as follows:

TABLE 1: Notation box.

Symbol	Definition
Input parameter	
K	Set of reservation contract types
I	Set of service instance types
T	Set of provisioning stages
P^r	Set of unit prices of reservation contracts while $p^r_{vk} \in P^r$ denotes the price of the k-type reservation-contract charged to a v-type instance
P^o	Set of unit prices of on-demand instances while $p^o_v \in P^o$ denotes the unit price of on-demand v-type instance
S	Set of numbers of time slots in each stage
D	Set of numbers of various instances required in each time slot while $n^d_{vij} \in D$ denotes the number of v-type instances required in the j-th time slot of the stage T_i
Decision variable	
r_{vki}	Number of k-type reservation-contracts with v-type instance purchased in T_i
Other parameters	
F_k	Set of stages at which k-type reservation-contracts can be purchased for planning cycle T (obtained by formula (1))
F_{ki}	Set of stages at which some instances reserved by k-type contracts can be utilized during T_i (obtained by formula (2))
n^r_{vki}	Number of v-type instances reserved by k-type contracts available in T_i (obtained by formula (3))
n^r_{vi}	Number of v-type reserved instances available in T_i
n^o_{vij}	Number of v-type instances provisioned on demand for the j-th time slot in T_i
c_v	Total cost of v-type instances provisioned in T
c_{vi}	Cost paid to v-type instances provisioned in T_i
c^r_{vi}	Cost of all reserved contracts charged to v-type instances in T_i
c^o_{vi}	Cost of all v-type instances provisioned on demand in T_i
c^o_{vij}	Cost of all v-type instances provisioned on demand in the j-th time slot in T_i

$$c_v = \sum_{i \in T} \left(\sum_{k \in K} p^r_{vk} r_{vki} + \sum_{j \in S_i} p^o_v \cdot n^o_{vij} \cdot |s| \right). \qquad (7)$$

Finally, for the entire planning cycle T, the optimization model of reservation contract procurement for various required instances is constructed as follows:

$$
\begin{aligned}
\min \sum_{v \in I} \sum_{i \in T} & \left(\sum_{k \in K} p^r_{vk} r_{vki} + \sum_{j \in S_i} p^o_v \cdot n^o_{vij} \cdot |s| \right) \\
\text{s.t.} \quad & r_{vki} \le \max\left(n^d_{vij}\right), v \in I, k \in K, i \in T, j \in S_i, n^d_{vij} \in D \\
& r_{vki} = 0, v \in I, k \in K, i \in T, _{\mathsf{T}k} \\
n^o_{vij} = \max & \left(0, n^d_{vij} - \sum_{k \in K} \sum_{t \in F_{ki}} r_{vkt} \right) \begin{array}{l} v \in I, i \in T, j \in S_i, k \in K, n^d_{vij} \in D \\ r_{vki} \in \mathrm{N}, p^r_{vk} \in P^r, p^o_v \in P^o, s \in S, S_i \subset S, \end{array}
\end{aligned}
\qquad (8)
$$

where only r_{vki} is the decision variables, and the objective function is the linear function of r_{vki}; therefore, this is a Pure Integer Linear Programming (PILP) problem, which can be solved by using the classical Branch and Bound method.

5. Workload Prediction

Considering that the LSTM is designed to combine the short-term and long-term temporal information and exhibits superior long time-series prediction performance, we attempt to design a LSTM-based algorithm for predicting the future cycle workload of a web application.

5.1. Prediction Algorithm Based on the LSTM

5.1.1. Typical LSTM Architecture and Principles. The key to the LSTM is the cell state. Figure 5 illustrates the typical architecture of the LSTM memory cell and the cell's state transition at time $t-1$, t, and $t+1$, and in practice the transition flow usually contains more moments. It can be seen that the cell state runs straight down the entire chain with only some linear interactions, which makes it easy for information to be propagated over time. For the memory cell at time t, there are three inputs: the current input $\mathbf{x}_t$, the previous output $\mathbf{h}_{t-1}$, and the previous state $\mathbf{c}_{t-1}$, and two outputs: the current output $\mathbf{h}_t$ and the current state $\mathbf{c}_t$. The
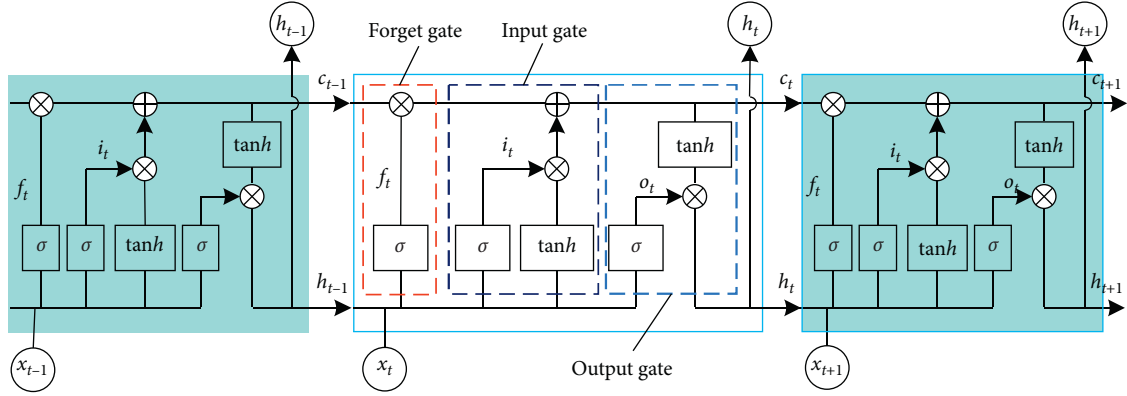
FIGURE 5: The LSTM memory cell and its state transition.

LSTM uses three gates to control the cell state transition. The forget gate determines how much information of the previous state $\mathbf{c}_{t-1}$ is retained to the current state $\mathbf{c}_t$, while the input gate determines how much information of the current input $\mathbf{x}_t$ is saved to the state $\mathbf{c}_t$. The output gate determines how much information of the current state $\mathbf{c}_t$ is output to $\mathbf{h}_t$, which controls the influence of long-term memory on the current output. The forward calculation of the LSTM is expressed as follows:

$$\mathbf{f}_t = \sigma\left(\mathbf{W}_f \mathbf{h}_{t-1} + \mathbf{U}_f \mathbf{x}_t + \mathbf{b}_f\right), \tag{9}$$

$$\widetilde{c}_t = \tan h\left(\mathbf{W}_c \mathbf{h}_{t-1} + \mathbf{U}_c \cdot \mathbf{x}_t + \mathbf{b}_c\right), \tag{10}$$

$$\mathbf{i}_t = \sigma\left(\mathbf{W}_i \cdot \mathbf{h}_{t-1} + \mathbf{U}_i \cdot \mathbf{x}_t + \mathbf{b}_i\right), \tag{11}$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \widetilde{c}_t, \tag{12}$$

$$\mathbf{o}_t = \sigma\left(\mathbf{W}_o \cdot \mathbf{h}_{t-1} + \mathbf{U}_o \cdot \mathbf{x}_t + \mathbf{b}_o\right), \tag{13}$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tan h\left(\mathbf{c}_t\right), \tag{14}$$

where $\mathbf{f}$, $\mathbf{i}$, and $\mathbf{o}$ denote the forget gate, the input gate, and the output gate, respectively, $\mathbf{W}$ and $\mathbf{U}$ matrices are the network parameters, $\mathbf{b}$ denotes the bias, σ is a sigmoid function, and $\odot$ denotes the product operation.

The LSTM is trained with the Back Propagation Through Time (BPTT) algorithm, which is similar to the Back Propagation (BP) algorithm in principle. The main process is as follows: (1) obtain the outputs by the forward calculation (formulas (9)–(14)); (2) calculate the loss function of each LSTM cell from two backward propagation directions of time and network; and (3) select a gradient optimization algorithm to minimize the loss function and hence optimize system parameters. There are several commonly used gradient optimization algorithms such as the SGD, AdaGrad, RMSProp, and Adam optimizers. Among them, the Adam is a stochastic gradient descent algorithm that combines the advantages of the AdaGrad and RMSProp and can adaptively adjust the learning rate of parameters. By comparison, the Adam performs better in practice.

5.1.2. Prediction Framework Based on the LSTM. As web workload is usually influenced by many factors, such as date, time, and business events, we express web workload as a multivariate time series for training and predicting. The designed workload prediction framework based on the LSTM is illustrated in Figure 6, which contains four functional parts, namely, the data, the LSTM network, the training, and the prediction parts. The data part performs preprocessing on raw historical workload data, such as missing data processing, abnormal data processing, feature extraction, workload series generation, supervised data generation, normalization, and division of training and test sets. The designed LSTM network contains an input layer, a hidden layer, and an output layer. The number of nodes in the input layer and the number of LSTM cells in the hidden layer are both equal to the number of time steps of a workload sequence sample. In Figure 6, $\mathbf{c}$ and $\mathbf{h}$ are the state and output of each cell, respectively. The output layer contains an output node p^y, which saves the output of an input sample. In the training part, the process is as follows: (1) the samples are continuously fed into the network, and then the errors are calculated based on formula (15) (where *num* is the number of samples); (2) the network parameters are updated by the Adam optimizer based on the errors; (3) the two steps above are performed iteratively a specified number of times, and finally the network parameters are saved. In the prediction part, the trained network is used to iteratively predict future workload.

$$\text{loss} = \sum_{i=1}^{\text{num}} \frac{\left(p_i^y - y_i\right)^2}{\text{num}}. \tag{15}$$

5.1.3. Supervised Data Generation. The workload at each moment is not only related to its previous values but also related to the date, time, holiday, and other information. Therefore, we express a workload series with length n as $F' = \{X_1', X_2', \ldots, X_n'\}$, where $X_i' = \{x_{i1}', x_{i2}', \ldots, x_{ik}', r_i'\}$ represents the observations at time i, r_i' denotes the workload value, and x_{i1}' to x_{ik}' denote the measurements of the k variables related to r_i'. To avoid the influence of inconsistent
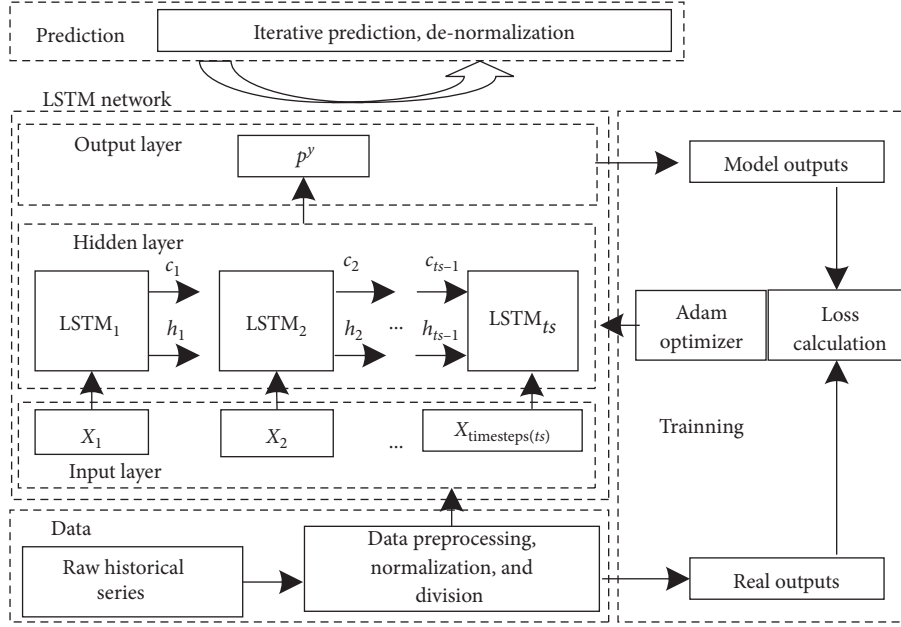
FIGURE 6: The LSTM-based prediction framework for workload series.

dimensions on the learning, all feature values in the series F' are normalized to unified dimension $[0, 1]$. As such, a new normalized workload series is obtained as $\mathbf{F} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\}$, $\mathbf{X}_i = \{x_{i1}, x_{i2}, \ldots, x_{ik}, r_i\}$, where r_i and x_{i1} to x_{ik} denote normalized r_i' and x_{i1}' to x_{ik}', respectively. Let ts be the number of time steps of a workload sequence that is used as an input of the LSTM network, while let $\mathbf{S}$ be the set of inputs. Based on ts and $\mathbf{F}$, $\mathbf{S}$ is expressed as follows:

$$\begin{aligned} \mathbf{S} &= \{\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_{n-ts}\}, \\ \mathbf{S}_i &= \{\mathbf{X}_i, \mathbf{X}_{i+1}, \ldots, \mathbf{X}_{i+ts-1}\}, \end{aligned} \tag{16}$$

where $i \le n - ts$ ($\mathbf{S}_{n-ts+1}$ has been removed from $\mathbf{S}$ so as to make each input correspond to an output) and $\mathbf{S}_i$ denotes the i-th input sequence. Let $\mathbf{Y}$ be the corresponding output set, which is expressed as follows:

$$\mathbf{Y} = \{y_1, y_2, \ldots, y_{n-ts}\}. \tag{17}$$

Here, y_i is equal to r_{i+ts}. As a result, n-ts samples are obtained based on $\mathbf{S}$ and $\mathbf{Y}$, and then the training and test sets are also be easily obtained after a simple division.

5.1.4. Network Training and Data Predicting. To obtain the best result, we use grid search to optimize three key hyperparameters of the LSTM network: ts (the number of time steps of a input sequence), $units$ (the number of neurons in the hidden layer), and η (the Adam optimizer's initial learning rate). Other hyperparameters are set according to general experience. The designed training and predicting process is presented in Algorithm 1. There are several inputs, where ts_l, ts_u, $step_{ts}$, $units_l$, $units_u$, and $step_{units}$ denote the lower bounds, upper bounds, and growth step sizes of ts as well as $units$, respectively. η_Array, m, $seed$, and $epochs$ denote the value range of learning rate η, sample division ratio, random-number seed, and iteration times,

respectively. The outputs include possible combinations of hyperparameters in the grid and their corresponding test errors, as well as the optimal predicted result and its error.

The algorithm traverses the hyperparameters space, looping the training, and predicting process as shown in lines 7 to 22. In particular, based on the $\mathbf{S}$, $\mathbf{Y}$, and m, line 4 obtains the input set $\mathbf{S}_{tr}$ and the output set $\mathbf{Y}_{tr}$ for the training, as well as the corresponding sets $\mathbf{S}_{te}$ and $\mathbf{Y}_{te}$ for the testing. Line 7 and line 8, respectively, create and initialize the pLSTM model, lines 9 to 12 perform training, and lines 13 to 17 iteratively predict test data. Line 17 uses the current predicted result to update the workload value of the last observations in the next input sequence. $\mathbf{P}_{tr}^y$ and $\mathbf{P}_{te}^y$ are the output sets of the training and the testing, respectively.

6. Determination of Instance Capacity and Time-Slot Baseline Workload

6.1. Determination of Service Instance Capacity. For interactive web applications, service response time is the most important QoS index, and the system designer usually specifies an upper bound for it so as to ensure a good user experience. In fact, there is an inherent relationship among service response time, request arrival rate, and system service capacity. Let C be the service capacity of a virtual instance, which refers to the maximum request arrival rate supported by this instance while the response time index is met. Due to the web request arrival process is a Poisson process and the service time complies with negative exponential distribution, therefore, a service instance with n vCPUs can be modeled as an $M/M/n$ queuing system. Let μ and λ be the average service rate and request arrival rate, respectively, then service intensity ρ is equal to $(\lambda/(n\mu))$. Let ρ_1 be (λ/μ) and p_k be the probability of the state that there are k requests in the system. According to K's algebraic equation, when $k < n$,

$p_k = ((n^n \rho^k p_0)/k!)$ and when $k \geq n$, $p_k = ((n^n \rho^k p_0)/n!)$ [48]. Obviously, $(\sum_{k=0}^{\infty} p_k = 1)$. After deducing the formulas, p_0 is expressed as follows (where $(\rho < 1)$):

$$p_0 = \left(\sum_{k=0}^{n-1} \frac{\rho_1^k}{k!} + \frac{\rho_1^n}{n!} \frac{1}{1-\rho} \right)^{-1}. \tag{18}$$

Additionally, let L_s, L_q, and L_{busy} be the average number of requests, the number of queued requests, and the number of busy vCPUs in the system, respectively. Apparently, $L_s = L_q + L_{busy}$, $L_q = \sum_{k=n}^{+\infty} (k-n) p_k$, and $L_{busy} = \rho_1$. After some derivations, L_s is expressed as follows:

$$L_s = \frac{\rho \rho_1^n p_0}{n! (1-\rho)^2} + \rho_1. \tag{19}$$

According to Little's formula, the service response time, namely, the average staying time $\bar{t}_s$ of a request in the system is calculated as follows:

$$\bar{t}_s = \frac{L_s}{\lambda} = \frac{\rho_1^n p_0}{\mu n \cdot n! (1-\rho)^2} + \frac{1}{\mu}. \tag{20}$$

If $\bar{t}_{max}$ is the upper bound of acceptable response time, that is, $\bar{t}_s \leq \bar{t}_{max}$, then the allowable maximum request arrival rate is determined based on $\bar{t}_{max}$ by the formulas (18) and (20), which is exactly the service capacity C of the instance.

6.2. Determination of Time-Slot Baseline Workload.

We consider that the baseline workload for planning slot should be set this way as far as possible to meet all workload demands after excluding few abnormal values. As the workload distribution of the adjacent planning cycle is similar, the workload statistics of the last cycle can be used to transform current predicted workload so as to obtain the reasonable baseline workload. It is assumed that the last historical cycle contains m time-slots (e.g., a year contains 365 days), any one of which contains ΔT (e.g., 10 minutes), and the number of requests during each ΔT has been counted. First, the request numbers of all ΔTs are sorted in the descending order, and then a two-dimensional array d is obtained, where d_{ij} represents the j-th largest workload in time slot i. Next, we specify a workloads-ratio threshold fr (e.g., $fr = 0.1$) and then calculate the average of time-ratios of m slots, namely, tr, as follows:

$$tr = \frac{\left(\sum_{i=1}^{m} (l_i/n) \right)}{m}, \tag{21}$$

where l_i is calculated based on the constraints: $((\sum_{j=1}^{l_i-1} d_{ij})/(\sum_{j=1}^{n} d_{ij})) \leq fr \leq ((\sum_{j=1}^{l_i} d_{ij})/(\sum_{j=1}^{n} d_{ij}))$.

In particular, tr denotes the average cumulative-time-ratio of several sequenced peak workloads with a cumulative-workloads-ratio fr for time slots in a stage. For example, $fr = 0.2$ and $tr = 0.1$ means that, on average, 20 percent of peak workloads in a slot only takes up 10 percent of time. Then, the average of a certain percentage of sequenced peak workloads can be used as the time-slot's baseline workload. Let D_s denotes the number of requests in slot s, while $|s|$

denotes the number of seconds in s, then the baseline workload λ_s of time slot s is calculated as follows:

$$(\lambda_s = (\text{fr} \cdot D_s)/(\text{tr} \cdot |s|)). \tag{22}$$

7. Experimental Evaluation

7.1. Experimental Environment, Datasets, and Evaluation Criteria

7.1.1. Experimental Environment. The experiments were run on an OS Win10 machine with 16 GB of memory and 3.0 GHz Intel Core i7 processor. By using Python 3.7 under the PyCharm 2019.1.3, the LSTM-based prediction algorithm was developed through the use of the Tensorflow 1.13.1 framework, and the experimental SARIMA and Holter–Winters models were developed based on Statsmodels 0.10.1 package, while the experimental SVR model was developed based on the machine-learning toolkit Sklearn 0.21. To solve the optimization problem, LINGO 15 [49] solver was used.

7.1.2. Datasets. The LAcity.org website traffic dataset from Kaggle [50] was chosen to evaluate the workload prediction approaches. This is a dataset hosted by the city of Los Angeles, which contains detailed daily traffic data from January 1, 2014, to July 12, 2019, for lacity.org, the main website for the city of Los Angeles. We obtained the numbers of daily requests after preprocessing and then intercepted the data from January 1, 2014, to December 31, 2018, for evaluations. The distribution of workloads is shown in Figure 7. To obtain fine-grained web-traffic data to simulate the determination of time-slot baseline workload, the YOOCHOOSE dataset was also downloaded from Kaggle [51], in which all clicks of users over a retailer's website had been recorded. After preprocessing, we obtained the numbers of requests per 10 minutes over the website from June 1, 2014 to August 31, 2014. The distribution of workloads is shown in Figure 8.

7.1.3. Evaluation Criteria. We mainly used the Mean Absolute Error (MAE), the Mean Absolute Percentage Error (MAPE), and the Root Mean Square Error (RMSE) as the evaluation criteria to gauge the prediction accuracy, which were calculated as the formulas (27)–(29), respectively, where the parameter n denotes the number of observations, y_i is the actual workload, and $\hat{y}_i$ represents the predicted workload.

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i|, \tag{23}$$

$$\text{MAPE} = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{\hat{y}_i - y_i}{y_i} \right|, \tag{24}$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2}. \tag{25}$$
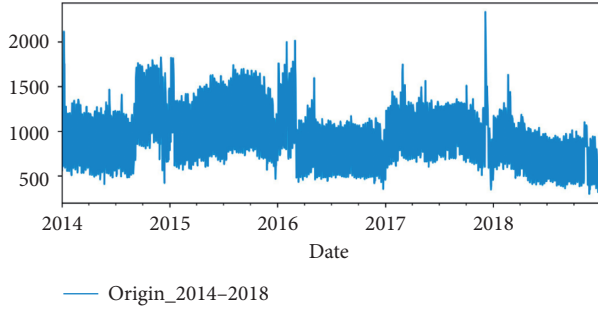
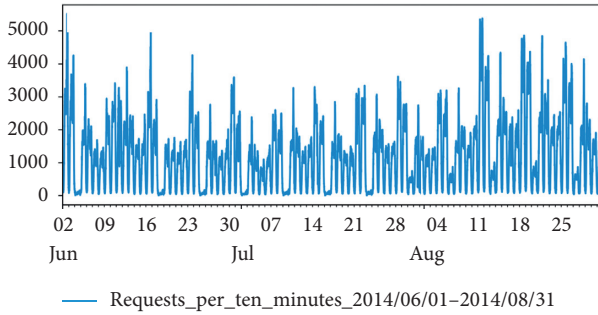FIGURE 7: Distribution of daily request rate for lacity.org.

FIGURE 8: Distribution of requests per 10 minutes for a retailer's website.

7.2. Evaluation for Workload Estimation/Prediction.

In this section, we first introduce several typical estimation or prediction approaches for future cycle workloads and then present the result of the LSTM-based prediction algorithm.

7.2.1. Historical-Workload-Based Estimations.

Some studies directly used the historical cycle's workloads as an estimation of the current cycle's workloads [14, 15, 17]. Similarly, we used the workloads from 2014 to 2017 as the estimations of the workloads in the following years, namely, 2015 to 2018, respectively. The results are shown in Figure 9, where only dark red overlapping areas are accurately estimated areas. Their Mean Absolute Percentage Errors (MAPEs) are 27.0%, 69.1%, 29.4%, and 51.7%, respectively. Apparently, the accuracies are poor.

7.2.2. Holter–Winters Seasonal Models.

Several typical exponential smoothing models are often used in workload predicting, which include single exponential smoothing, double exponential smoothing, and multiple-parameter exponential smoothing (namely, Holter–Winters seasonal models). Among them, cubic exponential smoothing-based Holter–Winters seasonal models can deal with seasonality and trends, which are classified as additive model and multiplicative model. In the additive model, several components such as level value, seasonal trend, and linear trend are considered to be independent of each other, and so they are directly added. In the multiplicative model, these components are considered to be influenced by each other, and so they are directly multiplied. Given the workload

fluctuations is relatively gentle, we selected the Holter–Winters additive model for predicting. The prediction formula is shown as follows:

$$\widehat{x}_{t+k} = a_t + b_t k + S_{t-s+k}, \quad k \in N^+, \tag{26}$$

where a_t and b_t are the intercepts and S_{t-s+k} and s denote seasonal component and period length, respectively. The first three parameters are calculated as follows:

$$
\begin{aligned}
a_t &= \alpha(x_t - S_{t-s}) + (1 - \alpha)(a_{t-1} + b_{t-1}), \\
b_t &= \beta(a_t - a_{t-1}) + (1 - \beta)b_{t-1}, \\
S_t &= \gamma(x_t - a_t) + (1 - \gamma)S_{t-s}.
\end{aligned}
\tag{27}
$$

Here, α, β, and γ are three damping factors, and $\alpha, \beta, \gamma \in (0, 1)$. For the purpose of comparison, we used both additive and multiplicative models for predicting, and the results are shown in Figure 10. It can be seen that the trends predicted by the multiplicative model decay dramatically from the beginning so that the prediction cannot continue after a while. The additive model can basically predict the trends and periods, but the MAPE reached 33.8%, and obviously the overall accuracy is still low.

7.2.3. SARIMA Model.

AR, MA, ARMA, ARIMA, and SARIMA are several typical time series models. The first three models are only suitable for stationary series, while the ARIMA can make some nonstationary series become stationary through differencing. Given the SARIMA can further deal with the seasonal trends compared with the ARIMA, we employed the SARIMA to perform the prediction and the comparison. The SARIMA model is generally expressed as follows:

$$\phi(B)\Phi(^{BS})^{\nabla d}\nabla^D_{S\,xt} = c + \theta(B)\Theta(BS)\varepsilon_t. \tag{28}$$

Here,

$$
\begin{cases}
\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots \phi_p B^p, \\
\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots \theta_q B^q, \\
\Phi(B^S) = 1 - \Phi_1 B^S - \Phi_2 B^{2S} - \cdots - \Phi_P B^{PS}, \\
\Theta(B^S) = 1 - \Theta_1 B^S - \Theta_2 B^{2S} - \cdots - \Theta_Q B^{QS}, \\
\nabla^d = (1 - B)^d, \\
\nabla^D_S = (1 - BS)^D,
\end{cases}
\tag{29}
$$

where S, D, d, ε_t, and c denote the length of seasonal period, the times of seasonal difference and ordinary difference, the white Gaussian noise, and the constant term, respectively. $\phi(B)$ is an autoregressive polynomial, $(\phi_1, \phi_2, \ldots, \phi_p)$ are the autoregressive coefficients, $\Phi(B^S)$ is a seasonal autoregressive polynomial, and $(\Phi_1, \Phi_2, \ldots, \Phi_P)$ are the seasonal autoregressive coefficients. Meanwhile, $\theta(B)$ is a moving average polynomial, $(\theta_1, \theta_2, \ldots, \theta_q)$ are the moving average coefficients, $\Theta(B^S)$ is a seasonal moving average polynomial, and $(\Theta_1, \Theta_2, \ldots, \Theta_Q)$ are the seasonal moving average coefficients. Here, p, P, q, and Q are the orders of $\phi(B)$, $\Phi(B^S)$, $\theta(B)$, and $\Theta(B^S)$, respectively. In addition, B is the ordinary
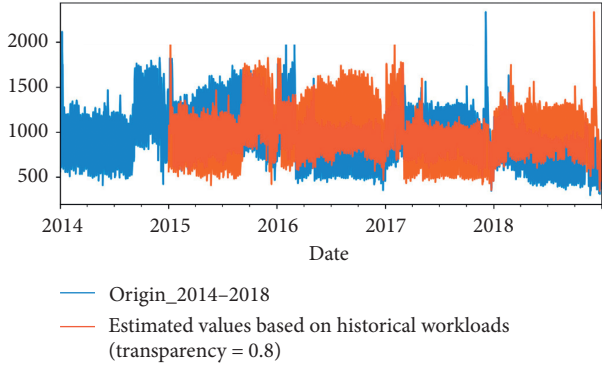
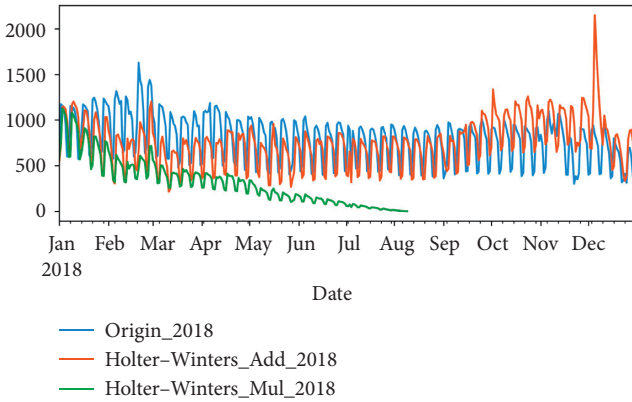FIGURE 9: Estimated results based on the historical cycle's workloads.

FIGURE 10: Comparison of predicted results with origin data.

lag operator, B^S is the seasonal lag operator, and ∇^d is the ordinary difference operator, while ∇_S^D is the seasonal difference operator. $\nabla^d \nabla_S^D x_t$ represents a stationary time series. The model expressed as equation (28) can be abbreviated as SARIMA $(p, d, q) (P, D, Q)_S$, which is constructed based on p, d, q, P, D, Q, and S. The process of determining these parameters is as follows.

First, the S-steps (S is equal to period length) periodic difference is performed D times to eliminate seasonal trends, and then the ordinary difference is performed d times based on the results of stationarity checking so that the series become stationary. In this process, S is determined by observing the time-series diagram, while D is equal to the times of periodic difference and d is equal to the times of ordinary difference. In general, D and d do not exceed 3. Second, the order p can be determined based on the tailing or truncation of partial autocorrelation coefficients in the partial autocorrelogram. Meanwhile, the order q can be determined based on the tailing or truncation of autocorrelation coefficients in the autocorrelogram. Similarly, the orders P and Q can also be determined based on the tailing or truncation of the autocorrelation and partial autocorrelation coefficients over the time-lag points with several times of period length. Finally, the SARIMA model is created based on the determined parameters above and then is fitted based on the

samples. The results of iterative prediction are shown in Figure 11. The MAPE is 22.3%, the MAE is 190.5, and the RMSE is 253.3, respectively. It can be seen that the overall prediction for seasonal and linear trends is relatively accurate, but the detailed prediction is poor.

7.2.4. SVR Model. The Support Vector Machine (SVM) is an innovative statistical learning model proposed by Cortes and Vapnik based on the principle of structural risk minimization [52]. It has excellent generalization capabilities and can deal with small sample, nonlinear, high-dimensional learning problems. The SVR is the application of the SVM in the data regression and prediction. The applications of SVR in workload forecasting have been also widely studied [23, 53–57]. In the process of the SVR nonlinear regression and prediction, the original data is mapped to the high-dimensional space through the use of a nonlinear mapping, where a linear function can be found to fit the input and output values of samples, and then the prediction is done based on this function. Given the workload of the LAcity.org exhibits obvious nonlinear characteristics, we choose to use the nonlinear $\varepsilon - \mathrm{SVR}$ model for the prediction and the comparison.

Suppose there is a sample set: $\{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \ldots, (\mathbf{x}_n, \mathbf{y}_n)\}$, $\mathbf{x}_i \in \mathbf{R}^d$, and $y_i \in \mathbf{R}$, where d denotes the feature dimension and n denotes the number of samples. After the set is mapped to the high-dimensional space, its linear fitted function can be expressed as follows:

$$f(\mathbf{x}) = \boldsymbol{\omega} \cdot \varphi(\mathbf{x}) + b, \tag{30}$$

where $\varphi(\mathbf{x})$ is the nonlinear mapping function from the original data to the high-dimensional space, ω is the coefficient vector, and b is the offset. According to the principle of the $\varepsilon - \mathrm{SVR}$ model, the goal of learning is to make $f(\mathbf{x})$ and y as close as possible but tolerate a deviation with the maximum value ε between $f(\mathbf{x})$ and y; that is, the loss is calculated only when the deviation is greater than ε. Also, considering a few samples is still unable to be fitted under the accuracy ε, the slack variables (ξ_i) and (ξ_i^*) are introduced. Thus, based on the principle of structural risk minimization, the function estimation problem is transformed into the following optimization problem:

$$\min \frac{1}{2}\|\boldsymbol{\omega}\|^2 + C\frac{1}{n}\sum_{i=1}^{n}\left(\xi_i + \xi_i^*\right)$$

$$\text{s.t.} \begin{cases} y_i - \boldsymbol{\omega} \cdot \varphi(\mathbf{x}_i) - b \le \varepsilon + \xi_i \\ \boldsymbol{\omega} \cdot \varphi(\mathbf{x}_i) + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0, \quad i = 1, 2, \ldots, n, \end{cases} \tag{31}$$

where C is the penalty coefficient, which determines how well the regression function fits the data. In order to facilitate solving the problem, the Lagrange multipliers α and α^* are introduced, and the above problem is transformed into the dual problem:
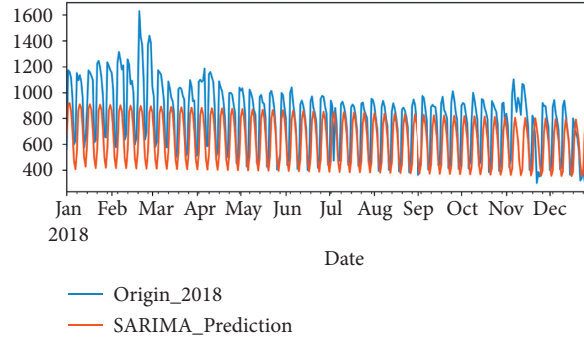
FIGURE 11: Distribution of predicted values from the SARIMA model.

$$\min \frac{1}{2} \sum_{i,j=1}^{n} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) K\langle \mathbf{x}_i, \mathbf{x}_j \rangle + \varepsilon \sum_{i=1}^{n} (\alpha_i + \alpha_i^*) + \sum_{i=1}^{n} y_i (\alpha_i - \alpha_i^*) \text{s.t.} \quad \left\{ \sum_{i}^{n} (\alpha_i - \alpha_i^*) = 0 \, 0 \leq \alpha_i, a_i^* \leq C, \quad i = 1, 2, \dots, n, \right.$$

(32)

where through the use of kernel function $K\langle \mathbf{x}_i, \mathbf{x}_j \rangle = \langle \varphi(\mathbf{x}_i), \varphi(\mathbf{x}_j) \rangle$, the calculation of vector inner product in the high-dimensional space is converted to the corresponding calculation in the original low-dimensional space, avoiding the problem of dimension explosion. By solving this problem, α_i and a_i^* are obtained, then multiple samples satisfying $0 < \alpha_i, a_i^* < C$ can be chosen to solve for b and the average value of b is used (namely, $\bar{b}$), so the regression function is obtained as follows:

$$f(\mathbf{x}) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K\langle \mathbf{x}_i, \mathbf{x} \rangle + \bar{b}.$$

(33)

After data preprocessing, we obtained the normalized daily-workload series of the LAcity.org from 2014 to 2018: $R = \{r_1, r_2, \dots, r_n\}$, the corresponding month-feature series: $M = \{m_1, m_2, \dots, m_n\}$, and the corresponding workday-feature series: $D = \{d_1, d_2, \dots, d_n\}$, where n is the length of these series, r_i denotes the value of the i-th workload (namely, the number of daily requests), m_i denotes the month-number feature corresponding to the i-th workload, and d_i denotes the corresponding workday-number feature. The reason for choosing the feature m and d is that the analysis found that the daily workload is closely related to its date attributes. Then, we designed the input set as $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{n-lag}\}$ and the output set as $\mathbf{Y} = \{y_1, y_2, \dots, y_{n-lag}\}$ for the SVR model, where $\mathbf{x}_i = \{r_i, r_{i+1}, \dots, r_{i+lag-1}, m_{i+lag}, d_{i+lag}\}$, $y_i = r_{i+lag}$, and the adjustable parameter lag denotes the length of time lag. The value of lag implies that the current workload is most relevant to the recent lag historical workloads. Finally, we developed the SVR-based prediction algorithm through the use of the toolkit Sklearn. Considering the good adaptability of the radial basis function, we chose it as the kernel function (namely, $K\langle x_i, x \rangle = \exp(-\gamma\|x_i - x\|^2)$). Also, we applied grid search and cross-validation to determine the values of lag, C (penalty coefficient) and γ (width coefficient of the

radial basis function) and sorted the predicted results according to the MAPE. The top five optimal combinations of hyperparameters, corresponding errors, and time cost are illustrated in Table 2. Meanwhile, the distributions of predicted results and original workloads are shown in Figure 12.

From the results, the prediction accuracy reaches 86% and the computational overhead is low, which is mainly due to the use of kernel function. As can be seen from Figure 12, the predicted results fit well with the original series in terms of the level values, the seasonality, and the trends, and the prediction of the details is also good. The disadvantage is that the predicted values of valley workloads are generally higher than the actual values.

7.2.5. Prediction Algorithm Based on the LSTM. We implemented the training and prediction process of the LSTM network according to Algorithm 1. First, several general parameters were set empirically, where the random-number seed was set as 1 and the number of iterations was set as 200. Then, the value range of three key hyperparameters was set. We let the number of time steps of a sequence sample, namely, ts, belong to $\{2, 3, \dots, 60\}$, let the neuron number of the hidden layer, namely, $units$, belong to $\{2, 3, \dots, 60\}$, and let the learning rate, namely, η, belong to $\{0.001, 0.003, 0.005, 0.007, 0.01, 0.02, 0.03, 0.04, 0.05, 0.07, 0.1\}$. The step sizes of ts and $units$ were all set to be 1, and the loss function was set as the Mean Square Error (MSE) according to formula (15). Finally, we ran this program to traverse all combinations of hyperparameters. According to the MAPE values, the top five optimal combinations, corresponding errors, and time cost are illustrated in Table 3. Meanwhile, the distributions of predicted workloads and original workloads are shown together in Figure 13.

From the results, the LSTM model makes a better prediction for the annual workloads than previous several

Input: $(ts_l, ts_u, \text{step}_{ts})$, $(units_l, units_u, \text{step}_{units})$, η_Array, m, $seed$, $epochs$, and $min\_error = +\infty$
Output: $pra\_results$, $best\_pred$, and min_error
(1) $\mathbf{F}$ = normalize (F');
(2) **for** each ts in ts_l: ts_u by step_{ts}
(3)   get $\mathbf{S}$, $\mathbf{Y}$ from $\mathbf{F}$ by ts;
(4)   get $\mathbf{S}_{tr}$, $\mathbf{Y}_{tr}$, $\mathbf{S}_{te}$, $\mathbf{Y}_{te}$ from $\mathbf{S}$, $\mathbf{Y}$ by m;
(5)   **for** each η in η_Array
(6)   **for** each q in $units_l$: $units_u$ by step_{units}
(7)   create pLSTM by ts, q;
(8)   initialize pLSTM by $seed$;
(9)   **for** each step in 1: $epochs$
(10) $\mathbf{P}_{tr}^{y}$ = pLSTM$_{\text{forward}}$ $(\mathbf{S}_{tr})$;
(11)   get $loss$ from $\mathbf{P}_{tr}^{y}$, $\mathbf{Y}_{tr}$;
(12)   update pLSTM by Adam with $loss$ and η;
(13)   **for** each i in 0: length $(\mathbf{S}_{te}) - 1$
(14) p_i^{y} = pLSTM $(\mathbf{S}_{te}[i])$;
(15)   append p_i^{y} to $\mathbf{P}_{te}^{y}$;
(16)   **if** $i <$ length $(\mathbf{S}_{te}) - 1$
(17) $\mathbf{S}_{te}[i+1][ts-1][k-1] = p_i^{y}$;
(18) $\mathbf{P}_{te}^{y}$ = denormalize $(\mathbf{P}_{te}^{y})$;
(19)   get $error$ by $\mathbf{P}_{te}^{y}$, $\mathbf{Y}_{te}$;
(20)   append $[ts, \eta, q, error]$ to $pra\_results$;
(21)   **if** $error < min\_error$
(22) $best\_pred = \mathbf{P}_{te}^{y}$; $min\_error = error$;
(23) **return** $pra\_results$, $best\_pred$, min_error;

ALGORITHM 1: Network training and data predicting.
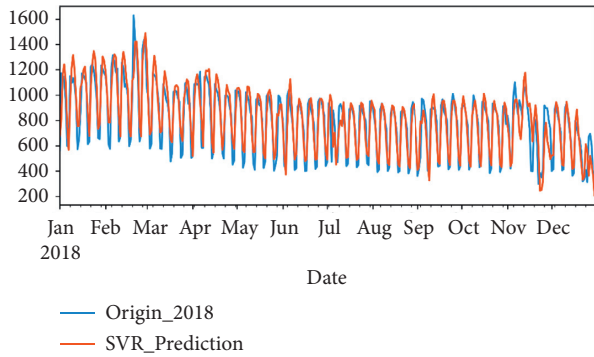
FIGURE 12: Distribution of predicted values from the SVR model.
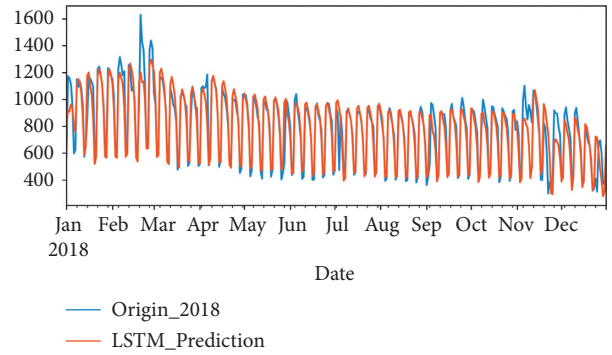
FIGURE 13: Distribution of predicted values from the LSTM model.

TABLE 2: The best five hyperparameter combinations of the SVR model.

Rank	lag	C	γ	Mape	Rmse	mae	time_cost
1	9	16	0.25	13.2%	117	90	4.61
2	45	32	0.004	13.3%	114	85	6.19
3	21	8	0.125	13.7%	120	95	5.17
4	25	2	0.125	14.4%	119	93	5.73
5	12	8	0.25	14.7%	123	97	4.38

approaches. For the several optimal hyperparameter combinations, the average accuracy is more than 90%, and the computational overhead is also low. As can be seen from Figure 13, the predicted results fit well with the original series in terms of the level values, the seasonality, the trends, and the details. Compared with the SVR model, the LSTM model is superior in terms of overall accuracy and detailed prediction. The LSTM model exhibits excellent prediction performance for long time series, which is mainly attributed to its strong ability of learning the long-term and short-term temporal information simultaneously.

7.3. Evaluation for the Optimization Model

7.3.1. Simulations of Resource Provisioning Scenarios. Ideally, the evaluation for the reservation-contract-procurement-optimization model should be based on a real web application and its workloads. However, there are only some public web-traffic datasets are available. Given the simulation of the running of web applications does not affect the

evaluation, we used the predicted workloads of LAcity.org in 2018 from the LSTM model to simulate the running of a real web application. We assumed that the application contained two elastic components, namely, the web server and the database server. Meanwhile, we also assumed that Aliyun's ecs.g5.large and mysql.n4.medium instances in North China [3] had been selected for the two servers, respectively. For the ecs.g5.large and mysql.n4.medium instances, the on-demand unit prices (namely, hourly rate), the reservation-contract unit prices, and their corresponding discount rates of monthly cost compared with the on-demand plan are listed in Table 4 in turn, where the unit of cost is RMB Yuan. Apparently, these reservation contracts offer considerable cost discounts, and the longer the contract duration, the higher the discount rate.

7.3.2. Determination of Instance Capacity. In real scenarios, the instance's average service rate can be obtained through benchmarking, combined with the specified response time index; the instance processing capacity can be calculated according to formulas (18) and (20). However, the capacity is difficult to be determined without the related application suite or benchmarking abilities. As this number only has an influence on the absolute cost figure and does not affect the evaluation of optimization model. Therefore, we assumed that the capacity of single web-server instance in the simulated application was 50 requests per second. Moreover, it was also assumed that the system performance was optimal when the instance numbers of web server and database server meet the ratio of 1 : 1.

7.3.3. Determination of Time-Slot Baseline Workload. We used the YOOCHOOSE dataset to simulate the determination of baseline workload for time slot. After pre-processing, the numbers of requests per 10 minutes from June 1, 2014, to August 31, 2014, were obtained. According to Section 6.2, we treated a day as a time slot and then calculated the baseline workloads from August 1 to August 31 based on the statistics of workloads from June 1 to July 31. First, we let *fr* belong to [0.001, 0.4], namely, the range of cumulative-peak-workloads ratio and traversed this range in a step size equaling to 0.002 to calculate the corresponding *tr*, namely, the average cumulative-time ratio from June 1 to July 31. Second, the daily baseline workloads in August were calculated, respectively, based on each pair of *fr* and *tr* above, and then the average probability that time-slot workloads were met (let *avg_fullfilled_rate* represents this probability) was obtained for each pair of *fr* and *tr*. Finally, as presented in Table 5, the best five results are listed after ranking *avg_fullfilled_rate*. In fact, in this way, the baseline workload is equal to *fr/tr* times of the average original workload. As can be seen from Table 5, the best five *avg_fullfilled_rates* are above 96%. To obtain a higher *avg_fullfilled_rate*, the *fr/tr* coefficient can be finely increased manually, also resulting in more resource costs. In short, determining the time-slot's baseline workload in this way greatly alleviates the adverse impact of coarse-grained predicted workloads on resource planning.

7.3.4. Model Evaluation. Based on the above simulated scenarios and related data, combined with the predicted daily workloads of LAcity.org in 2018, the reservation-contract-procurement-optimization model for the simulated application was constructed and solved in LINGO15. To evaluate the effect, several resource provisioning schemes were compared, and the results were presented in Table 6, where the Reservation Contract Procurement Optimization Based on Predicted Workload (RCPOBPW) scheme means to first determine the reservation contract procurement plan based on the approaches presented in this paper and then carry out the plan as well as supplement necessary on-demand resources during the future business cycle. The Reservation Contract Procurement Based on Average Predicted Workloads (RCPBAPW) scheme means to first purchase the fixed number of reservation contracts once based on average predicted workload and then supplement necessary on-demand resources during the future cycle. The Using only Reserved Resources Provisioned by One Kind of Contracts (URRPOC) scheme means to purchase the fixed number of reservation contracts once based on the maximum predicted workload and do not use any on-demand resources. Additionally, the Using only Reserved Resources (URR) scheme and the Using only On-demand Resources (UOR) scheme mean to use only reserved resources and use only on-demand resources during the future cycle, respectively. The Reservation Contract Procurement Optimization Based on Real Workloads (RCPOBRW) scheme is theoretically optimal, which differs from RCPOBPW only in that it determines the reservation plan based on real workloads. In Table 6, $WN_{C1} \sim WN_{C4}$ are, respectively, the numbers of web server instances with 1 month, 3 months, 6 months, and 1 year reservation contracts, while $DN_{C1} \sim DN_{C4}$ are, respectively, the corresponding numbers of database server instances. C, C_0/C, and R_S denote total resource cost, the ratio of on-demand resource costs to total costs, and the cost ratio of each scheme to the RCPOBPW scheme, respectively, while R_{SLA} denotes the SLA satisfaction rate of each scheme.

As can be seen from the *WN* and *DN* columns, except for the UOR, all schemes use reserved resources, and the reservation contracts with the longest duration are purchased the most. Except that the RCPBAPW and URRPOC scheme only purchase the longest-duration contracts based on fixed workloads, other schemes using reserved resources (e.g., RCPOBPW, URR, and RCPOBRW) have purchased various contracts. From the total cost, the UOR scheme using only on-demand resources is the highest, the URRPOC scheme using only reserved resources provided by one kind of contracts is the second, and followed by the RCPBAPW and URR. Obviously, our RCPOBPW is the least costly practical scheme, and its cost is only 0.4% more than the theoretical optimal scheme. From the SLA satisfaction rate, all schemes can fully meet the demands except for the URRPOC and URR schemes, which do not use on-demand resources. Overall, our RCPOBPW scheme is the best among the five practical schemes. Finally, several conclusions can be drawn as follows: (1) it is not appropriate to use completely on-demand resources, which will result in huge expenditures; (2) it is also not appropriate to use completely reserved

TABLE 3: The best five hyperparameter combinations of the LSTM model.

Rank	ts	lr	$units$	$Mape$ (%)	$Rmse$	mae	$time\_cost$ (s)
1	42	0.05	36	9.12	104	65	5.87
2	42	0.05	33	9.31	106	66	5.21
3	45	0.007	45	9.34	106	68	7.23
4	45	0.005	33	9.37	107	68	6.95
5	36	0.05	30	9.40	107	67	4.65

TABLE 4: The unit prices and discount rates of on-demand plans and reservation contracts for the cs.g5.large and mysql.n4.medium instances.

Duration	1 hour	1 month	3 months	6 months	12 months
Price1	0.91	256.25	692.25	1308.00	2437.80
Discount rate1	0	60.9%	64.8%	66.7%	69.0%
Price2	1.03	346	975.63	1821.72	3387.84
Discount rate2	0	53.3%	56.1%	59.1%	61.9%

TABLE 5: The best five pairs of fr and tr as well as corresponding results.

fr	tr	fr/tr	$avg\_fullfilled\_rate$ (%)
0.111	0.061	1.828	96.5
0.125	0.068	1.827	96.5
0.141	0.077	1.825	96.5
0.137	0.075	1.821	96.4
0.113	0.062	1.820	96.4

TABLE 6: Comparison of several typical resource provisioning schemes.

Scheme	WN_{C1}	WN_{C2}	WN_{C3}	WN_{C4}	DN_{C1}	DN_{C2}	DN_{C3}	DN_{C4}	C	C_O/C (%)	R_S (%)	R_{SLA} (%)
RCPOBPW	3	2	2	18	4	3	2	17	123611	6.85	100	100
RCPBAPW	0	0	0	16	0	0	0	16	134183	33.4	108.6	100
URRPOC	0	0	0	27	0	0	0	27	157292	0	127.2	98.33
URR	14	5	1	19	14	5	1	19	130588	0	106	96.67
UOR	0	0	0	0	0	0	0	0	277079	100	224.2	100
RCPOBRW	7	1	2	18	8	8	8	18	123174	6.2	99.6	100

resources, as it is likely that some unexpected workloads cannot be handled; and (3) it is advisable to use a combination of on-demand and reserve resources and take as many contracts as possible into account to maximize the share of reserved resources so as to achieve the greatest cost discounts while meeting the demand.

8. Conclusions and Future Work

In this paper, we investigated the resource-reservation-planning problems for cloud-based web applications. First, we developed an integer linear program model for optimizing the reservation-contracts procurement. Then, we designed the LSTM-based algorithm for predicting the business cycle's workloads of web applications. Thereafter, the approaches for determining the instance capacity and the baseline workload of time slot were also presented. Finally, experimental evaluations were carried out based on several real datasets. From the comparison of predicted results, our LSTM-based algorithm achieves better effect than the Holter–Winters, SARIMA, and SVR models, with an accuracy of about 90%. This result is attributed to the LSTM network's good memory and learning ability for long time series and also related to its learning of workload-related information such as date and time. Meanwhile, from the comparative results of several typical practical provisioning schemes, the scheme based on the optimization model presented in this paper achieves the least resource cost while entirely satisfying future demands.

However, for a cloud-based web application, although the optimal resource-reservation plan can be obtained based on the proposed solution in this paper, the problem of how to dynamically provision on-demand resources during the business cycle remains to be solved, which is worth in-depth study.

Data Availability

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 they have no conflicts of interest regarding the publication of this paper.

Acknowledgments

References

[1] P. A. Dinda and D. R. O'hallaron, "Host load prediction using linear models," *Cluster Computing*, vol. 3, no. 4, pp. 265–280, 2000.

[2] 12306, https://www.12306.cn/index/, 2019.

[3] Aliyun, https://www.aliyun.com/, 2019.

[4] Rackspace, https://www.rackspace.com/, 2019.

[5] Aws, https://aws.amazon.com/cn/ec2/pricing/reserved-instances/, 2019.

[6] S. Mireslami, L. Rakai, M. Wang et al., "Minimizing deployment cost of cloud-based web application with guaranteed QoS," in *Proceedings of the IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, San Diego, CA, USA, December 2014.

[7] S. Imai, S. Patterson, and C. A. Varela, "Cost-Efficient elastic stream processing using application-agnostic performance prediction," in *Proceedings of the IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid)*, pp. 604–607, Cartagena, CO, USA, May 2016.

[8] M. Anastasopoulos, A. Tzanakaki, and D. Simeonidou, "Stochastic energy efficient cloud service provisioning deploying renewable energy sources," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 12, pp. 3927–3940, 2016.

[9] L. Chen, X. Li, and R. Ruiz, "Resource renting for periodical cloud workflow applications," *IEEE Transactions on Services Computing*, vol. 13, no. 1, pp. 130–143, 2020.

[10] S. Mireslami, L. Rakai, B. H. Far, and M. Wang, "Simultaneous cost and QoS optimization for cloud resource allocation," *IEEE Transactions on Network and Service Management*, vol. 14, no. 3, pp. 676–689, 2017.

[11] H. Zhao, M. Pan, X. Liu, X. Li, and Y. Fang, "Exploring fine-grained resource rental planning in cloud computing," *IEEE Transactions on Cloud Computing*, vol. 3, no. 3, pp. 304–317, 2015.

[12] Y. Ran, B. Yang, W. Cai et al., "Cost-Efficient provisioning strategy for multiple services in distributed clouds," in *Proceedings of the International Conference on Cloud Computing Research & Innovations (ICCCRI)*, Singapore, May 2016.

[13] A.-F. Antonescu and T. Braun, "Simulation of SLA-based VM-scaling algorithms for cloud-distributed applications," *Future Generation Computer Systems*, vol. 54, pp. 260–273, 2016.

[14] B. Sniezynski, P. Nawrocki, M. Wilk, M. Jarzab, and K. Zielinski, "VM reservation plan adaptation using machine learning in cloud computing," *Journal of Grid Computing*, vol. 17, no. 4, pp. 797–812, 2019.

[15] D. Candeia, R. A. Santos, and R. Lopes, "Business-driven long-term capacity planning for SaaS applications," *IEEE Transactions on Cloud Computing*, vol. 3, no. 3, pp. 290–303, 2015.

[16] A. Wolke, M. Bichler, and T. Setzer, "Planning vs. Dynamic control: resource allocation in corporate clouds," *IEEE Transactions on Cloud Computing*, vol. 4, no. 3, pp. 322–335, 2016.

[17] J. Chen, H. Zhou, and W. Wang, "A provision algorithm for cloud resources with cost optimization," *Journal of Xi'an Jiaotong University*, vol. 51, no. 10, pp. 135–141, 2017.

[18] S. Mireslami, L. Rakai, M. Wang, and B. H. Far, "Dynamic cloud resource allocation considering demand uncertainty," *IEEE Transactions on Cloud Computing*, vol. 51, p. 1, 2019.

[19] Y. Lei and H. Shen, "Towards bandwidth guarantee for virtual clusters under demand uncertainty in multi-tenant clouds," *IEEE Transactions on Parallel and Distributed Systems*, vol. 29, no. 2, pp. 450–465, 2018.

[20] Z. Cao, J. Lin, C. Wan et al., "Optimal cloud computing resource allocation for demand side management in smart grid," *IEEE Transactions on Smart Grid*, vol. 8, no. 4, pp. 1943–1955, 2017.

[21] R. N. Calheiros, E. Masoumi, R. Ranjan, and R. Buyya, "Workload prediction using ARIMA model and its impact on cloud applications' QoS," *IEEE Transactions on Cloud Computing*, vol. 3, no. 4, pp. 449–458, 2015.

[22] R. V. D. Bossche, K. Vanmechelen, and J. Broeckhove, "IaaS reserved contract procurement optimisation with load prediction," *Future Generation Computer Systems*, vol. 53, pp. 13–24, 2015.

[23] L. Zhao, "Load forecasting model of cloud computing resources based on support vector machine," *Journal of Nanjing University of Science and Technology*, vol. 42, no. 6, pp. 687–692, 2018.

[24] Z. Zhang, S. Ding, and Y. Sun, "A support vector regression model hybridized with chaotic krill herd algorithm and empirical mode decomposition for regression task," *Neurocomputing*, vol. 410, no. 14, pp. 184–201, 2020.

[25] W.-C. Hong, Y. Dong, C.-Y. Lai, L.-Y. Chen, and S.-Y. Wei, "SVR with hybrid chaotic immune algorithm for seasonal load demand forecasting," *Energies*, vol. 4, no. 6, pp. 960–977, 2011.

[26] A. Ma, C. Zhang, B. Zhang et al., "Load prediction approach for cloud application based on Deep Belief networks," *Journal of Northeastern University(Natural Science)*, vol. 38, no. 2, pp. 209–213, 2017.

[27] K. Cetinski and M. B. Juric, "AME-WPC: advanced model for efficient workload prediction in the cloud," *Journal of Network and Computer Applications*, vol. 55, pp. 191–201, 2015.

[28] J. Kumar, R. Goomer, and A. K. Singh, "Long short term memory recurrent neural network (lstm-rnn) based workload forecasting model for cloud datacenters," *Procedia Computer Science*, vol. 125, pp. 676–682, 2018.

[29] X. Ma, Z. Tao, Y. Wang, H. Yu, and Y. Wang, "Long short-term memory neural network for traffic speed prediction using remote microwave sensor data," *Transportation Research Part C: Emerging Technologies*, vol. 54, pp. 187–197, 2015.

[30] X. Wang, J. Wu, C. Liu et al., "Fault time series prediction based on LSTM recurrent neural network," *Journal of Beijing University of Aeronautics and Astronautics*, vol. 44, no. 4, pp. 772–784, 2018.

[31] A. Graves, "Long short-term memory," *Studies in Computational Intelligence*, vol. 385, 2012.

[32] Y. Tian, K. Zhang, J. Li, X. Lin, and B. Yang, "LSTM-based traffic flow prediction with missing data," *Neurocomputing*, vol. 318, no. 27, pp. 297–305, 2018.

[33] Y. Liu, S. Dong, M. Lu et al., "LSTM based reserve prediction for bank outlets," *Tsinghua Science and Technology*, vol. 24, no. 1, pp. 77–85, 2018.

[34] Q. Wang, Y. Guo, L. Yu et al., "Earthquake prediction based on spatio-temporal data mining: an LSTM network approach," *IEEE Transactions on Emerging Topics in Computing*, vol. 8, no. 1, pp. 148–158, 2017.

[35] N. Sfika, A. Korfiati, C. Alexakos et al., "Dynamic cloud resources allocation on multidomain/multiphysics problems," in *Proceedings of the International Conference on Future Internet of Things and Cloud*, pp. 31–37, Rome, Italy, August 2015.

[36] S. Khatua, P. K. Sur, R. K. Das, and N. Mukherjee, "Heuristic-based resource reservation strategies for public cloud," *IEEE Transactions on Cloud Computing*, vol. 4, no. 4, pp. 392–401, 2016.

[37] R. I. Meneguette, A. Boukerche, A. H. M. Pimenta et al., "A resource allocation scheme based on Semi-Markov Decision Process for dynamic vehicular clouds," in *Proceedings of the IEEE International Conference on Communications (ICC)*, pp. 1–6, Paris, France, May 2017.

[38] L. Jiao, J. Li, T. Xu, W. Du, and X. Fu, "Optimizing cost for online social networks on geo-distributed clouds," *IEEE/ACM Transactions on Networking*, vol. 24, no. 1, pp. 99–112, 2016.

[39] S. Stijven, R. Van den Bossche, E. Vladislavleva et al., "Optimizing a cloud contract portfolio using genetic programming-based load models," *Genetic Programming Theory and Practice XI*, vol. 22, pp. 47–63, 2014.

[40] E. Dhib, N. Zangar, N. Tabbane et al., "Impact of seasonal ARIMA workload prediction model on QoE for massively multiplayers online gaming," in *Proceedings of the 5th International Conference on Multimedia Computing and Systems (ICMCS)*, pp. 737–741, Marrakech, Morocco, September 2016.

[41] P. Singh, P. Gupta, and K. Jyoti, "Tasm: technocrat arima and svr model for workload prediction of web applications in cloud," *Cluster Computing*, vol. 22, no. 2, pp. 619–633, 2019.

[42] N. Tran, T. Nguyen, B. M. Nguyen, and G. Nguyen, "A multivariate fuzzy time series resource forecast model for clouds using LSTM and data correlation analysis," *Procedia Computer Science*, vol. 126, pp. 636–645, 2018.

[43] Alibaba, https://www.taobao.com, 2020.

[44] Huawei, https://lab.huaweicloud.com, 2020.

[45] Chinese Academy of Sciences Innovative Center in Quantum Information and Quantum Physics, http://quantumcomputer.ac.cn, 2020.

[46] Alibaba, https://www.aliyun.com/product/rsimganalys, 2020.

[47] Tableau, https://www.tableau.com/about/mission, 2020.

[48] C. Lu, *Queueing Theory*, Beijing University of Posts and Telecommunications Press, Beijing, China, Second edition, 2009.

[49] Lindo, https://www.lindo.com, 2020.

[50] Lacity.org, https://www.kaggle.com/cityofLA/lacity.org-websitetraffic, 2019.

[51] Yoochoose, https://www.kaggle.com/jacobgreen4477/weblog, 2020.

[52] C. Cortes and V. Vapnik, "Support vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[53] G.-F. Fan, S. Qing, H. Wang, W.-C. Hong, and H.-J. Li, "Support vector regression model based on empirical mode decomposition and auto regression for electric load forecasting," *Energies*, vol. 6, no. 4, pp. 1887–1901, 2013.

[54] Y. H. Chen, W. C. Hong, W. Shen et al., "Electric load forecasting based on a least squares support vector machine with fuzzy time series and global harmony search algorithm," *Energies*, vol. 9, no. 2, 2016.

[55] Y. Dong, Z. Zhang, and W. C. Hong, "A hybrid seasonal mechanism with a chaotic cuckoo search algorithm with a support vector regression model for electric load forecasting," *Energies*, vol. 11, no. 4, 2018.

[56] Z. Zhang and W.-C. Hong, "Electric load forecasting by complete ensemble empirical mode decomposition adaptive noise and support vector regression with quantum-based dragonfly algorithm," *Nonlinear Dynamics*, vol. 98, no. 2, pp. 1107–1136, 2019.

[57] Z. Zhang, W.-C. Hong, and J. Li, "Electric load forecasting by hybrid self-recurrent support vector regression model with variational mode decomposition and improved cuckoo search algorithm," *IEEE Access*, vol. 8, pp. 14642–14658, 2020.

Research Article

Optimal Control and Simulation for Enterprise Financial Risk in Industry Environment

Yanjun Liang [ID],[1] **Wei-hua Zhang** [ID],[2] **Youjun Lu**,[1] **and Zhong-Sheng Wang**[3]

[1]*School of Data Science and Information Engineering, Guizhou Minzu University, Guiyang 550025, China*
[2]*College of Economics and Management, Shanghai Ocean University, Shanghai 201306, China*
[3]*Department of Automation, Guangdong Polytechnic Normal University, Guangzhou 510000, China*

Correspondence should be addressed to Wei-hua Zhang; whzhang@shou.edu.cn

A technique for enterprise financial risk optimal control with exponential decay rate and simulation is developed in industry environment. The factors of industry environment risks to enterprise financial activities are considered, based on the evaluation method taking into both subjectivity and objectivity, seven kinds of industry environment risks influencing enterprise financial activities are chosen as state variables, and the enterprise financial risk dynamical system model is established for the first time. In order to reduce the risk of enterprise financial activity subjected to industry environment, an average performance index with exponential decay rate is chosen for the systems. Using the optimal control approach, an optimal vibration controller with exponential decay rate is designed. Numerical simulation results illustrate the effectiveness of the proposed technique.

1. Introduction

In the period of postfinancial crisis, enterprises are facing constantly changing external conditions, many kinds of crises are emerging, and financial risks are more serious than before. Therefore, the construction of enterprise financial risk control or early warning system is imperative.

Enterprise financial risk control or early warning system has drawn much more attention recently, and there are about three kinds of work completed now. Firstly, the necessity of the construction of enterprise financial risk control or early warning system has been discussed [1–4]. For example, the relation of managing methods, financial risks, and financial systems is discussed in [3], and the authors think that sophisticated managing methods and financial systems may not do good to reduce financial risks, and in [4], the importance and standard process of financial risk engineering for electric power enterprises are emphasized, and the risks which are most likely to occur in business activities are identified. Secondly, the models of enterprise financial risk control or early warning system have been constructed and analyzed by many scholars [5–8]. In [5], a new financial

early warning logit model is developed and improved the accuracy of prediction and stability. To forecast the bankruptcy risk of enterprises in Latin America and Central Europe in [6], the author has used statistical and soft computing methods to program the prediction models. To predict financial crises, based on a multinomial logit model, in [7], a new early warning system model is developed. In [8], an early warning model of China's energy price is analyzed from the aspects of price fluctuation and price ratio structure through fitting the risk distributions of indices and applying the computable general equilibrium model. Furthermore, some methods are employed to control and to early warn financial risks [9–14]. Early warning system is treated as a pattern recognition problem in [9], and using a pattern classifier, based on distinctive features of economics, crisis critical and normal economical situations are distinguished. Systemic banking crises and early warning systems in low-income countries are studied, and a multinomial logit approach is proposed in [10]. To reduce abortions in dairy cattle in Denmark, a modified two-stage method for detecting an unusual increase in the abortion incidence is applied [11]. To reduce financial risks and early warn the

risks, an adaptive fuzzy measure by using the dynamic information in the single classifier pattern recognition results is proposed in [12]. For China's burgeoning real estate enterprises, in [13], the z-score model is used to reduce the financial risk in early warning models. By using data mining, a financial early warning system is developed, and 15 risk indicators that affected financial distress are detected [14].

With the development of computer and internet technology, many scholars begin to study financial risk problem in the new environment. Based on data science and computer technology, financial risk is studied [15–20]. Recently, more and more scholars consider the dynamic property of financial risk control systems [21]. In this paper, the enterprise financial risk dynamical system model is established, and using the optimal control method, the optimal controller is designed to reduce the financial risk in the industry environment.

The remainder of this paper is structured as follows. In Section 2, enterprise financial risk dynamical model and optimal control problem in the industry environment are presented. In Section 3, optimal controller with α exponential decay rate and algorithm are designed for the financial risk control system. In Section 4, numerical experiments are presented. Finally, in Section 5, some conclusions are drawn.

2. Dynamical System Model and Optimal Control Problem

Enterprise financial activity is comprised by six subsystems, namely, purchase subsystem, production subsystem, sale subsystem, investment subsystem, financing subsystem, and profit subsystem, and they work together to complete the financial circulation of capital collection, investment, consumption, income, and distribution, and seven kinds of industry environment risks may emerge: industry resource risk, industry competition risk, industry life cycle risk, industry technological change risk, industry credit risk,

industry tax rate risk, and industry interest rate risk. Enterprise finance system is a highly open system, and each subsystem is connected widely with the industry environment. In this paper, in order to establish an enterprise financial risk system, seven kinds of industry environment risks influencing enterprise financial activities are considered, and the system model is as follows:

$$
\begin{aligned}
\dot{z}_1 &= f_1(z_1, z_2, \cdots, z_7), \\
\dot{z}_2 &= f_2(z_1, z_2, \cdots, z_7), \\
\dot{z}_3 &= f_3(z_1, z_2, \cdots, z_7), \\
\dot{z}_4 &= f_4(z_1, z_2, \cdots, z_7), \\
\dot{z}_5 &= f_5(z_1, z_2, \cdots, z_7), \\
\dot{z}_6 &= f_6(z_1, z_2, \cdots, z_7), \\
\dot{z}_7 &= f_7(z_1, z_2, \cdots, z_7),
\end{aligned}
\tag{1}
$$

where z_1 is industry resource risk, z_2 is industry competition risk, z_3 is industry life cycle risk, z_4 is industry technological change risk, z_5 is industry credit risk, z_6 is industry tax rate risk, and z_7 is industry interest rate risk.

In (1), industry resource risk z_1 is determined by purchase price index of raw material, fuel, and power z_{11} and supplier concentration ratio z_{12}; industry competition risk z_2 is indicated by concentration ratio z_{21} and enterprise yearly increment rate z_{22} of the industry; industry life cycle risk z_3 is represented by sales growth rate z_{31} and investment in the fixed asset growth rate z_{32} of the industry; industry technological change risk z_4 is influenced by industry technological investment rate and is controlled; industry credit risk z_5 is described by cash flow current ratio z_{51} and bad debt rate z_{52} of the industry; and industry tax rate risk z_6 and industry interest rate risk z_7 are determined, respectively, by tax rate level and interest rate level of the industry and are constants in general. According to economic principles, the enterprise financial risk system (1) is redescribed in the following form:

$$
\begin{aligned}
\dot{z}_{11} &= -k_{1111}z_{11} + k_{1112}z_{12} - k_{1121}z_{21} + k_{1122}z_{22} + k_{1131}z_{31} + k_{1132}z_{32} + k_{1151}z_{51}, \\
\dot{z}_{12} &= k_{1211}z_{11} - k_{1212}z_{12}, \\
\dot{z}_{21} &= -k_{2121}z_{21} + k_{2122}z_{22} - k_{2131}z_{31} + k_{2132}z_{32}, \\
\dot{z}_{22} &= -k_{2221}z_{21} - k_{2222}z_{22} + k_{2231}z_{31} - k_{2232}z_{32}, \\
\dot{z}_{31} &= -k_{3111}z_{11} - k_{3112}z_{12} - k_{3121}z_{21} + k_{3122}z_{22} - k_{3131}z_{31} + k_{3132}z_{32} + k_{314}z_4, \\
\dot{z}_{32} &= k_{3221}z_{21} - k_{3222}z_{22} + k_{3231}z_{31} - k_{3232}z_{32} - k_{324}z_4 - k_{3251}z_{51} - k_{3252}z_{52}, \\
\dot{z}_4 &= -k_{421}z_{21} + k_{422}z_{22} - k_{431}z_{31} - k_{432}z_{32} - k_{44}z_4 - k_{451}z_{51} - k_{452}z_{52} + bu, \\
\dot{z}_{51} &= -k_{1111}z_{11} - k_{5112}z_{12} + k_{5121}z_{21} - k_{5122}z_{22} + k_{5131}z_{31} + k_{5132}z_{32} - k_{5151}z_{51} - k_{5152}z_{52}, \\
\dot{z}_{52} &= k_{5221}z_{21} - k_{5222}z_{22} + k_{5231}z_{31} + k_{5232}z_{32} - k_{5251}z_{51} - k_{5252}z_{52},
\end{aligned}
\tag{2}
$$

where $k_{imjn} > 0$ is coefficient, $i = 1, 2, \cdots, 5$, $m = 1, 2$, $j = 1, 2$, and $n = 1, 2$, u is control input, and b is its coefficient.

The dynamic system model of enterprise financial risk control system in industry environment is rewritten in the state-space representation:

$$\dot{x}(t) = Ax(t) + Bu(t),$$
$$x(0) = x_0, \tag{3}$$

where $x = [x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9,]^T = [z_{11}, z_{12}, z_{21}, z_{22}, z_{31}, z_{32}, z_4, z_{51}, z_{52}]^T$, $B = [0, 0, 0, 0, 0, 0, b, 0, 0,]^T$, and

$$A = \begin{bmatrix} -k_{1111} & k_{1112} & -k_{1121} & k_{1122} & k_{1131} & k_{1132} & 0 & k_{1151} & 0 \\ k_{1211} & -k_{1212} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -k_{2121} & k_{2122} & -k_{2131} & k_{2132} & 0 & 0 & 0 \\ 0 & 0 & -k_{2221} & -k_{2222} & k_{2231} & -k_{2232} & 0 & 0 & 0 \\ -k_{3111} & -k_{3112} & -k_{3121} & k_{3122} & -k_{3131} & k_{3132} & k_{314} & 0 & 0 \\ 0 & 0 & k_{3221} & -k_{3222} & k_{3231} & -k_{3232} & -k_{324} & -k_{3251} & -k_{3252} \\ 0 & 0 & -k_{421} & k_{422} & -k_{431} & -k_{432} & -k_{44} & -k_{451} & -k_{452} \\ -k_{1111} & -k_{5112} & k_{5121} & -k_{5122} & k_{5131} & k_{5132} & 0 & -k_{5151} & -k_{5152} \\ 0 & 0 & k_{5221} & -k_{5222} & k_{5231} & k_{5232} & 0 & -k_{5251} & -k_{5252} \end{bmatrix}. \tag{4}$$

Optimal control theory is a subject to study and solve the problem of finding the optimal solution from all possible control schemes. In order to study the enterprise financial risk optimal control problem, choose an average performance index for system (3) as follows:

$$J = \lim_{T \to \infty} \frac{1}{T} \int_0^T e^{2\alpha t} \left[x^T(t) Q x(t) + R u^2(t) \right] dt, \tag{5}$$

where $Q \in \mathbb{R}^{9 \times 9}$ are positive semidefinite matrices, $R \in \mathbb{R}$ is a positive definite matrix, and $\alpha \geq 0$ is the exponential decay rate. We can balance the control effect and control energy by changing the parameter of the performance index because the control effect is influenced by the weighting matrix Q and the control energy is altered by the weighting coefficient R.

The objective of this paper is to find a control law $u^*(t)$ for system (3) and make the value of performance index (5) a minimum.

3. Optimal Controller and Algorithm Design

Then, we design a controller for financial risk control system (3). The optimal control law with α exponential decay rate can be presented in the following theorem.

Theorem 1. *Consider the optimal control problem described by system (3) with performance index (5). The optimal control law $u^*(t)$ exists and is unique. Its form is as follows:*

$$u^*(t) = -R^{-1} B^T P_1 x(t), \tag{6}$$

where P_1 is the unique positive definite solution of the following Riccati matrix equation:

$$(A + \alpha I)^T P_1 + P_1 (A + \alpha I) - P_1 S P_1 + Q = 0, \tag{7}$$

where $S = R^{-1} B^T$.

Proof. Introduce model transform for system (3) with performance index (5):

$$\tilde{x}(t) = x(t) e^{\alpha t},$$
$$\tilde{u}(t) = u(t) e^{\alpha t}, \tag{8}$$

$$\tilde{A} = A + \alpha I,$$
$$\tilde{B} = B, \tag{9}$$
$$\tilde{D} = D,$$

then, we have

$$\dot{\tilde{x}}(t) = \dot{x}(t) e^{\alpha t} + \alpha x(t) e^{\alpha t}$$
$$= Ax(t) e^{\alpha t} + \alpha x(t) e^{\alpha t} + Bu(t) e^{\alpha t} \tag{10}$$
$$= \tilde{A} \tilde{x}(t) + \tilde{B} \tilde{u}(t),$$

$$J = \lim_{T \to \infty} \frac{1}{T} \int_0^T e^{2\alpha t} \left[x^T(t) Q x(t) + R u^2(t) \right] dt$$
$$= \lim_{T \to \infty} \frac{1}{T} \int_0^T \left[\left(x(t) e^{\alpha t} \right)^T Q \left(x(t) e^{\alpha t} \right) + R \left(u(t) e^{\alpha t} \right)^2 \right] dt$$
$$= \lim_{T \to \infty} \frac{1}{T} \int_0^T \left[\tilde{x}^T(t) Q \tilde{x}(t) + R \tilde{u}^2(t) \right] dt. \tag{11}$$

The maximum principle in optimal control is the necessary condition to obtain the optimal control in the problem of maximizing the objective functional, which is named after making the Hamiltonian function reach the maximum value. Applying the maximum principle to the optimal control problem in (10) and (11), the optimal control law can be written as follows:

$$\tilde{u}^*(t) = -R^{-1} \tilde{B}^T \tilde{\lambda}(t), \tag{12}$$

where $\tilde{\lambda}(t)$ is the solution to the following two-point boundary value problem:

Step 1: give the exponential decay rate α;
Step 2: calculate the matrix P_1 from the Riccati matrix equation (7);
Step 3: substituting the given exponential decay rate α and obtained matrices P_1 into (6), the optimal vibration controller is obtained.

ALGORITHM 1: The algorithm to solve the enterprise financial risk optimal control law with exponential decay rate.
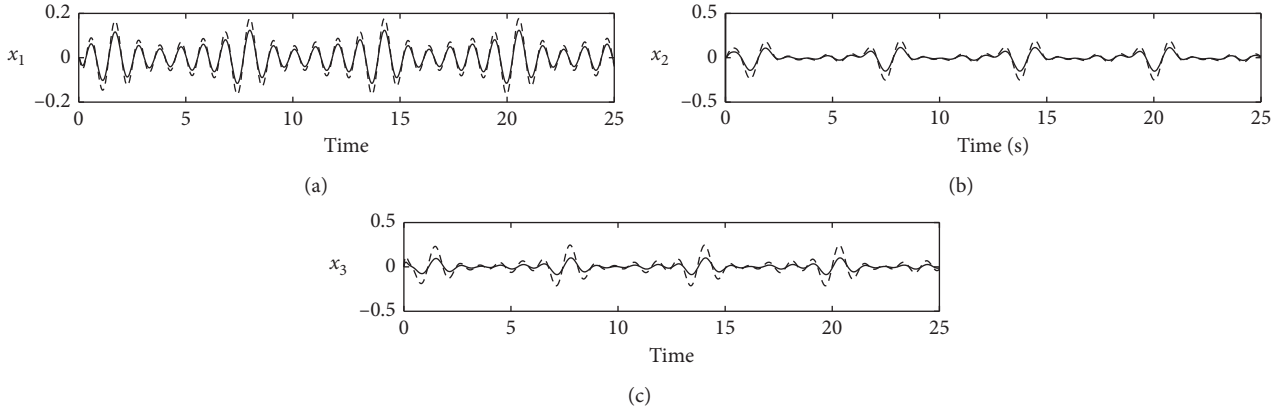

(a)

(b)

(c)

FIGURE 1: x_1, x_2, and x_3 comparison curves. (a) Purchase price index of raw material, fuel, and power. (b) Supplier concentration ratio. (c) Concentration ratio.

$$-\dot{\tilde{\lambda}}(t) = Q\dot{x}(t) + A^T\tilde{\lambda}(t),$$
$$\dot{\tilde{x}}(t) = A\tilde{x}(t) - R^{-1}\tilde{B}^T\tilde{\lambda}(t),$$
$$\tilde{x}(0) = x_0 e^{\alpha t}, \tag{13}$$
$$\tilde{\lambda}(\infty) = 0.$$

To solve (13), let

$$\tilde{\lambda}(t) = P_1\tilde{x}(t). \tag{14}$$

Substituting the equations of (10) and 12) into the first derivatives of (14), we get

$$\dot{\tilde{\lambda}}(t) = P_1\dot{\tilde{x}}(t) = \left(P_1\tilde{A} - P_1SP_1\right)\tilde{x}(t). \tag{15}$$

From equations 13 and (14), we obtain

$$\dot{\tilde{\lambda}}(t) = -\left(Q + \tilde{A}^T P_1\right)\tilde{x}(t). \tag{16}$$

Comparing the coefficients of (15) and (16), we obtain the matrix equation:

$$\tilde{A}^T P_1 + P_1\tilde{A} - P_1SP_1 + Q = 0. \tag{17}$$

Then, adding the model transform (9), we can obtain

$$(A + \alpha I)^T P_1 + P_1(A + \alpha I) - P_1SP_1 + Q = 0. \tag{18}$$

In order to implement the control law described in Theorem 1, we design an algorithm as follows (Algorithm 1). □

4. Numerical Experiment

Considering the data of Turkey in 2007, the data of firms were obtained from the Turkish Central Bank with permission and are shown in Table 1 [14]. Financial data gained from balance sheets and income statements are used to calculate the matrices of the system.

Employing Matlab software, a numerical experiment is carried out for the proposed optimal vibration controller.

The main purpose of risk control is to reduce enterprise financial risks which indicate the limit of the enterprise financial activities. So, to evaluate the effectiveness of the proposed control strategy, the risks of the enterprise finance system are considered. Then, the corresponding curves of open loop and controlled by the proposed optimal controller are compared and shown in Figures 1–3.

The curves of the enterprise financial risks are shown in Figures 1–3, in which solid lines represent the open loop results of the enterprise financial risk system in industry environment and dotted lines describe the results of the enterprise financial risk system controlled by the proposed control strategy. It can be seen from these numerical results that the proposed optimal controller with α exponential decay rate is efficient, real-time, and robust in reducing the enterprise financial risks in industry environment. Enterprises, financial managers, and researchers can balance the control effect and control energy by adjusting the parameters of the performance index, namely, changing the value of the weighting matrix Q and the weighting coefficient R, thereby improving further safety of the enterprise financial activities.

TABLE 1: Financial indicators that influenced financial risks.

Financial indicators	Value (10^{-4})
Profit before tax to own funds	0.700
Return on equity	0.300
Cumulative profitability ratio	1.000
Short-term liabilities to total loans	1.000
Total loans to total assets	230
Interest expenses to net sales	11.0
Fixed assets to long-term loans and own funds	27.0
Long-term liabilities to total liabilities	0.800
Gross profit to net sales	332
Bank loans to total assets	12.0
Inventory dependency ratio	1.00
Own funds turnover	432
Short-term receivables to total assets	121
Operating expenses to net sales	149
Receivables turnover	0.400

(a)

(b)

(c)
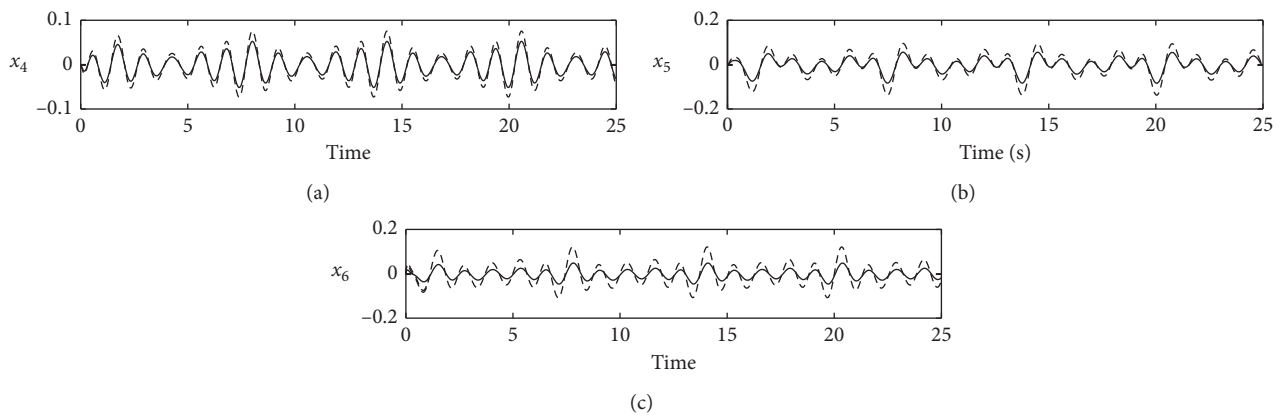
FIGURE 2: x_4, x_5, and x_6 comparison curves. (a) Enterprise yearly increment rate. (b) Sales growth rate. (c) Investment in the fixed asset growth rate.

(a)

(b)

(c)
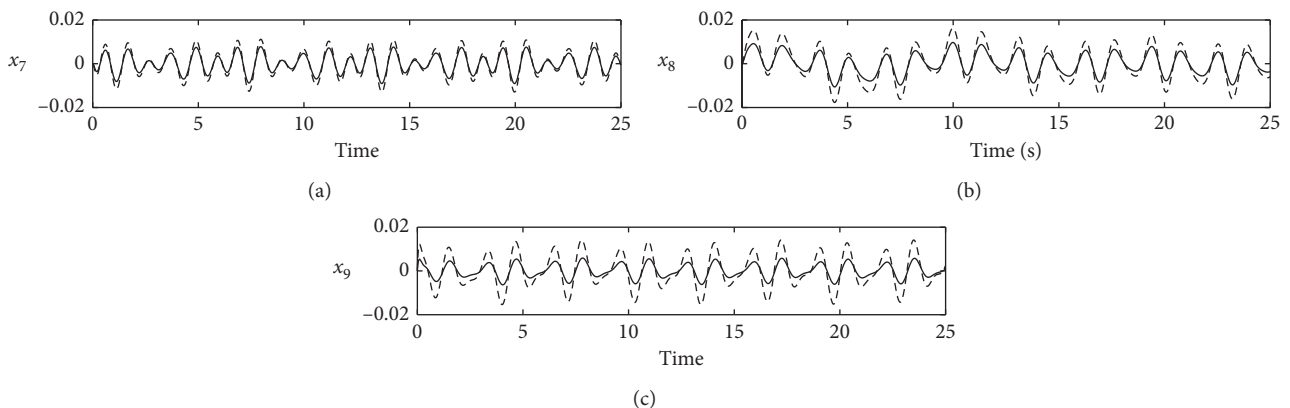
FIGURE 3: x_7, x_8, and x_9 comparison curves. (a) Industry technological change risk. (b) Cash flow current ratio. (c) Bad debt rate.

5. Conclusions

The industry environment is an external environment, the most basic of enterprises, and the influence of the macro-environment for the enterprises generally spreads to specific enterprises through the industry environment; therefore, the adjustment of the enterprise internal environment is based on the requirements of the industry environment. In order to study the situation of enterprise finance, its industry environment and the relationship between them have to be

considered. In this paper, seven kinds of industry environment risks influencing enterprise financial activities are chosen as state variables and the enterprise financial risk dynamical system model is established. The optimal controller with exponential decay rate is designed for the enterprise financial risk system. Numerical simulation results demonstrated that the proposed strategy is efficient, real-time, and robust. In the next studies, we will focus on the nonlinear systems and time-delay systems of the enterprise financial risk control systems, and fuzzy control strategy [22–25] can also be applied to the subject.

Data Availability

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 they have no conflicts of interest.

Acknowledgments

References

[1] J. J. Ahn, K. J. Oh, T. Y. Kim, and D. H. Kim, "Usefulness of support vector machine to develop an early warning system for financial crisis," *Expert Systems with Applications*, vol. 38, no. 4, pp. 2966–2973, 2011.

[2] J. Sun and H. Li, "Financial distress early warning based on group decision making," *Computers & Operations Research*, vol. 36, no. 3, pp. 885–906, 2009.

[3] S. Richard Sylla, "Financial systems, risk management, and entrepreneurship: Historical perspectives," *Japan and the World Economy*, vol. 15, no. 4, pp. 447–458, 2003.

[4] Z. Li, K. Liu, K. Wang, and X. Shen, "Research on financial risk management for electric power enterprises," *Systems Engineering Procedia*, vol. 4, pp. 54–60, 2012.

[5] S. Li and S. Wang, "A financial early warning logit model and its efficiency verification approach," *Knowledge-Based Systems*, vol. 70, pp. 78–87, 2014.

[6] T. Korol, "Early warning models against bankruptcy risk for Central European and Latin American enterprises," *Economic Modelling*, vol. 31, pp. 22–30, 2013.

[7] M. Bussiere and M. Fratzscher, "Towards a new early warning system of financial crises," *Journal of International Money and Finance*, vol. 25, no. 6, pp. 953–973, 2006.

[8] Y.-X. He, Y. Zhou, B. Wang, W. Xiong, and H.-Y. He, "Early warning model for risks of energy prices and energy price ratios in China's energy engineering," *Systems Engineering Procedia*, vol. 3, pp. 22–29, 2012.

[9] W. J. Yoon and K. S. Park, "A study on the market instability index and risk warning levels in early warning system for economic crisis," *Digital Signal Processing*, vol. 29, pp. 35–44, 2014.

[10] G. Caggiano, P. Calice, and L. Leonida, "Early warning systems and systemic banking crises in low income countries: a multinomial logit approach," *Journal of Banking & Finance*, vol. 47, pp. 258–269, 2014.

[11] T. E. Carpenter, M. Chrièl, and M. Greiner, "An analysis of an early-warning system to reduce abortions in dairy cattle in Denmark incorporating both financial and epidemiologic aspects," *Preventive Veterinary Medicine*, vol. 78, no. 1, pp. 1–11, 2007.

[12] Y. Cao, "Aggregating multiple classification results using Choquet integral for financial distress early warning," *Expert Systems with Applications*, vol. 39, no. 2, pp. 1830–1836, 2012.

[13] Y. Wang, "Z-score model on financial crisis early-warning of listed real estate companies in China: A financial engineering perspective," *Systems Engineering Procedia*, vol. 3, pp. 153–157, 2012.

[14] A. S. Koyuncugil and N. Ozgulbas, "Financial early warning system model and data mining application for risk detection," *Expert Systems with Applications*, vol. 39, no. 6, pp. 6238–6253, 2012.

[15] Q. Yang, Y. Wang, and Y. Ren, "Research on financial risk management model of internet supply chain based on data science," *Cognitive Systems Research*, vol. 56, pp. 50–55, 2019.

[16] M. Jin, Y. Wang, and Y. Zeng, "Application of data mining technology in financial risk analysis," *Wireless Personal Communications*, vol. 102, no. 4, pp. 3699–3713, 2018.

[17] S. Srinivasan and T. Kamalakannan, "Multi criteria decision making in financial risk management with a multi-objective genetic algorithm," *Computational Economics*, vol. 52, no. 2, pp. 443–457, 2018.

[18] K. Valaskova, T. Kliestik, L. Svabova, and P. Adamko, "Financial risk measurement and prediction modelling for sustainable development of business entities using regression analysis," *Sustainability*, vol. 10, no. 7, p. 2144, 2018.

[19] X. X. Zhang, "Financial risk early warning system for small and medium-sized enterprises," *Agro Food Industry Hi-Tech*, vol. 28, pp. 2798–2801, 2017.

[20] Y. Tao, "Corporate financial risks and risk prevention in internet era," *Agro Food Industry Hi-Tech*, vol. 28, pp. 3633–3636, 2017.

[21] B. Candelon, E.-I. Dumitrescu, and C. Hurlin, "Currency crisis early warning systems: Why they should be dynamic," *International Journal of Forecasting*, vol. 30, no. 4, pp. 1016–1029, 2014.

[22] W. Sun, Y. Wu, and Z. Sun, "Command Filter-Based Finite-Time Adaptive Fuzzy Control for Uncertain Nonlinear Systems with Prescribed Performance," *IEEE Transactions on Fuzzy Systems*, p. 1, 2020.

[23] Y. Chang, Y. Q. Wang, F. E. Alsaadi, and G. D. Zong, "Adaptive fuzzy output-feedback tracking control for switched stochastic pure-feedback nonlinear systems," *International Journal of Adaptive Control and Signal Processing*, vol. 33, 2019.

[24] L. Ma, X. Huo, X. Zhao, and G. D. Zong, "Observer-based adaptive neural tracking control for output-constrained switched MIMO nonstrict-feedback nonlinear systems with unknown dead zone," *Nonlinear Dynamics*, vol. 99, no. 2, pp. 1019–1036, 2020.

[25] Z.-M. Li, X.-H. Chang, and J. H. Park, "Quantized static output feedback fuzzy tracking control for discrete-time nonlinear networked systems with asynchronous event-triggered constraints," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, pp. 1–12, 2019.

Research Article

A Comparison of Hour-Ahead Solar Irradiance Forecasting Models Based on LSTM Network

Xiaoqiao Huang (ID),[1,2,3] **Chao Zhang,**[2,3] **Qiong Li,**[1] **Yonghang Tai,**[2,3] **Bixuan Gao,**[2] **and Junsheng Shi** (ID)[2,3]

[1]*Solar Energy Research Institute, Yunnan Normal University, Kunming, Yunnan 650500, China*
[2]*School of Physics and Electronic Information, Yunnan Normal University, Kunming, Yunnan 650500, China*
[3]*Yunnan Key Lab of Optoelectronic Information Technology, Kunming, Yunnan 650500, China*

Correspondence should be addressed to Junsheng Shi; shi-js@263.net

The intermittence and fluctuation character of solar irradiance places severe limitations on most of its applications. The precise forecast of solar irradiance is the critical factor in predicting the output power of a photovoltaic power generation system. In the present study, Model I-A and Model II-B based on traditional long short-term memory (LSTM) are discussed, and the effects of different parameters are investigated; meanwhile, Model II-AC, Model II-AD, Model II-BC, and Model II-BD based on a novel LSTM-MLP structure with two-branch input are proposed for hour-ahead solar irradiance prediction. Different lagging time parameters and different main input and auxiliary input parameters have been discussed and analyzed. The proposed method is verified on real data over 5 years. The experimental results demonstrate that Model II-BD shows the best performance because it considers the weather information of the next moment, the root mean square error (RMSE) is 62.1618 W/m^2, the normalized root mean square error (nRMSE) is 32.2702%, and the forecast skill (FS) is 0.4477. The proposed algorithm is 19.19% more accurate than the backpropagation neural network (BPNN) in terms of RMSE.

1. Introduction

Along with the rapid increase of solar power generation, more and more solar power is connected to the grid, which has already shown its substantial economic impact. Based on the statistics of the International Renewable Energy Agency (IRENA), the total installed capacity for PV has reached 205.493 GW in China at the end of 2019 [1]. However, power generation from photovoltaic systems is highly variable due to its dependence on meteorological conditions. There is a severe challenge to the security of the power grid because of the fluctuation of solar power. Therefore, an effective method of solar irradiance forecasting can mitigate intermittency as it gives information about future trends and allows users to make decisions beforehand.

Solar forecasting is a timely topic, and several short-term solar irradiance forecasting approaches have been presented recently. Broadly, prediction can be divided into five categories based on forecast methods as follows [2]: (1) time series; (2) regression; (3) numerical weather prediction; (4) image-based forecasting; and (5) machine learning. A time series is a sequence of observations taken sequentially in time. That is divided into stationary and nonstationary time series forecasting models. Autoregressive (AR), moving average (MA), and autoregressive moving average (ARMA) are commonly used to forecast stationary trends; integrated moving average (IMA), autoregressive integrated moving average (ARI-MA), seasonal autoregressive integrated moving average (SARIMA), and other models are used to forecast nonstationary trends [3–6]. Regression is a statistical process for estimating the relationships among variables; it is a handy tool to describe the relationship between solar irradiance and exogenous variables [7, 8]. Numerical weather prediction (NWP) models directly simulate the

irradiance fluxes at multiple levels in the atmosphere, separately considering the shortwave and longwave parts of the solar spectrum [9, 10]. Image-based forecasting method is using satellite cloud images and all-sky images as main or auxiliary data sources to forecast irradiance. This can effectively increase forecasting skills, as it provides warning of approaching clouds at a lead time of several minutes to hours [11–13]. The machine learning method, as a branch of artificial intelligence, can learn from datasets and construct a nonlinear mapping between input and output data. Nowadays, machine learning (ML) is perhaps the most popular approach in solar forecasting and load forecasting [2]. Although artificial neural networks (ANNs) and support vector machines (SVMS) are still the basis of machine learning methods in solar irradiance prediction, many other approaches have been used recently, such as k-nearest neighbors (kNN), random forest (RF), gradient boosted regression (GBR), hidden Markov models (HMMs), fuzzy logic (FL), wavelet networks (WNN), and long short-term memory networks (LSTM) [14–22]. Meanwhile, some hybrid algorithms are used to improve the prediction accuracy. For example, the metaheuristic algorithms, such as cuckoo search (CS) algorithm, krill herd (KH) algorithm, and chaotic immune algorithm, are combined with a support vector regression (SVR) model to predict electric load [19, 23–26]. Some signal preprocessing methods, such as variational mode decomposition (VMD) method and empirical mode decomposition (EMD), are also used in the hybrid model [24, 25]. Obviously, the abovementioned methods are not detailed lists. Many other applications of machine learning algorithms in solar radiation prediction can be found in recent literature [27].

As a novel machine learning tool, LSTM has successful applications in solar irradiance forecasting [28–30]. Due to its special maintaining a memory cell structure, it can preserve the important features which should be remembered during the learning process and improve performance. Therefore, using LSTM to predict irradiance can not only obtain the correlation during continuous hours but also extract its long-term (e.g., seasonal) behavior trends [30]. Yu et al. [29] proposed an LSTM-based approach for short-term global horizontal irradiance (GHI) prediction under complicated weather conditions, the result indicated that LSTM outperforms ARIMA, SVR, and NN models, especially on cloudy days and mixed days. Qing and Niu [30] proposed a novel hourly day-ahead solar irradiance predicted method using weather forecasts based on LSTM networks. The proposed algorithm uses the hourly weather forecasts of the same day and data information at the predicted time as the input variables, and the hourly irradiance values of the same anticipated day are taken as the output variable. Experimental results show that the proposed learning algorithm is more accurate than persistence, linear least squares regression method (LR), and BPNN due to the consideration of time dependence. Srivastava and Lessmann [28] studied the ability of LSTM in predicting solar irradiance, demonstrated the robustness of LSTM, and showed that the LSTM model with optimally configured outperforms GBR and FFNN for day-ahead GHI forecasting. Abdel-Nasser

and Mahmoud [31] proposed a method based on LSTM to forecast the output power of PV systems accurately. Liu et al. [32–34] proposed a new hybrid approach for the wind speed high-accuracy predictions based on some decomposition algorithm (such as secondary decomposition algorithm (SDA), empirical wavelet transform (EWT), and VMD) and the LSTM networks.

However, the LSTM methods mentioned above do not deeply study the effects of different parameters and structures on experimental results, but these factors will affect the prediction accuracy. In this paper, two different models based on traditional LSTM network are applied, and the effects of various parameters are investigated; meanwhile, four models based on a novel LSTM-MLP structure with two-branch input is proposed. For the new LSTM-MLP model, we use historical irradiance (or historical irradiance and meteorological parameters) as the main input and the meteorological parameters at the current time or the next time as the auxiliary input to predict the irradiance at the next time through the multilayer LSTM-MLP network. Experimental results show that the proposed model can achieve better prediction results.

The main innovations of this study are as follows: (1) An LSTM-MLP structure with two branches, including main input and auxiliary input, is proposed, which can provide a reference for similar models. (2) It is confirmed that the lagging time plays an important role when the input variables of the LSTM model are small. Still, for more input information, it is not that the more the lagging parameters, the higher the accuracy. (3) The meteorological parameters at the next moment play a vital role in the prediction accuracy, which can be gained by the weather forecast.

The organization of this paper is as follows: The methodology is described in detail in Section 2. Section 3 provides information about the dataset. Experimental results and discussion are presented in Section 4. Finally, conclusions are given in Section 5.

2. Method

2.1. Long Short-Term Memory Network.
In the learning phase, the traditional neural network cannot use the information learned in the previous time step to model the data of the current step. This is the main shortcoming of conventional neural networks. RNNs attempt to solve this problem by using loops that pass information from one step of the network to the next, ensuring the persistence of the information. In other words, the RNNs connect the previous information to the current task. Using previous sequence samples may help to understand the current sample.

The LSTM network, which has the time-varying inputs and targets, is a special RNN and was initially introduced by Hochreiter and Schmidhuber [35]. Due to the excellent ability to solve the long-term and short-tern dependency problem, the LSTM network often has satisfactory performance in processing time series. A general architecture is composed of a cell (the memory part of the LSTM unit) and three "regulators" (usually called gates), of the flow of information inside the LSTM unit: an input gate, an output

gate, and a forget gate. The memory unit is an essential parameter of the LSTM network, which can store information over an arbitrary time. The input gate, forget gate, and output gate can control the actual input signal by adding or deleting information to the signal state.

A schematic of the LSTM block can be seen in Figure 1. Every time a new input comes, its information will be accumulated to the cell if the input gate is activated. The prior cell status could be forgotten in this process if the forget gate is activated. Whether the latest cell output will be propagated to the final state is further controlled by the output gate.

The model input is denoted as $x = (x_1, x_2, \ldots, x_T)$, and the output sequence is denoted as $y = (y_1, y_2, \ldots, y_T)$, where T is the prediction period. In the context of solar irradiance forecasting, x can be considered as historical input data (e.g., irradiance and meteorological parameters), and y is the forecasting data. The predicted irradiance will be iteratively calculated by the following equations [36]:

$$i_t = \sigma\left(W_{ix}x_t + W_{im}m_{t-1} + W_{ic}c_{t-1} + b_i\right), \tag{1}$$

$$f_t = \sigma\left(W_{fx}x_t + W_{fm}m_{t-1} + W_{fc}c_{t-1} + b_f\right), \tag{2}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g\left(W_{cx}x_t + W_{cm}m_{t-1} + b_c\right), \tag{3}$$

$$o_t = \sigma\left(W_{ox}x_t + W_{om}m_{t-1} + W_{oc}c_{t-1} + b_o\right), \tag{4}$$

$$m_t = o_t \odot h\left(c_t\right), \tag{5}$$

$$y_t = W_{ym}m_t + b_y, \tag{6}$$

where i_t denotes the input gate, f_t is the forget gate, c_t is the activation vectors for each cell, o_t is the output gate, m_t is the activation vectors for each memory block, w is the weight matrices, b is the bias vectors, and "$\odot$" represents the scalar product of two vectors, and $\sigma(\cdot)$ denotes the standard logistics sigmoid function defined as follows:

$$\sigma(x) = \frac{1}{1 + e^{-x}}. \tag{7}$$

$g(\cdot)$ is a centered logistic sigmoid function defined as follows:

$$g(x) = \frac{4}{1 + e^{-x}} - 2, \quad x \in [-2, 2]. \tag{8}$$

$h(\cdot)$ is a centered logistic sigmoid function defined as follows:

$$h(x) = \frac{2}{1 + e^{-x}} - 1, \quad x \in [-1, 1]. \tag{9}$$

2.2. Model Development. As previously mentioned, the primary objective of this study is to examine the feasibility of the LSTM network for short-term solar irradiance forecasting and find the optimal structure of the LSTM for the forecast. In this section, firstly, the standard LSTM solar irradiance forecasting pipeline is introduced. Then, a
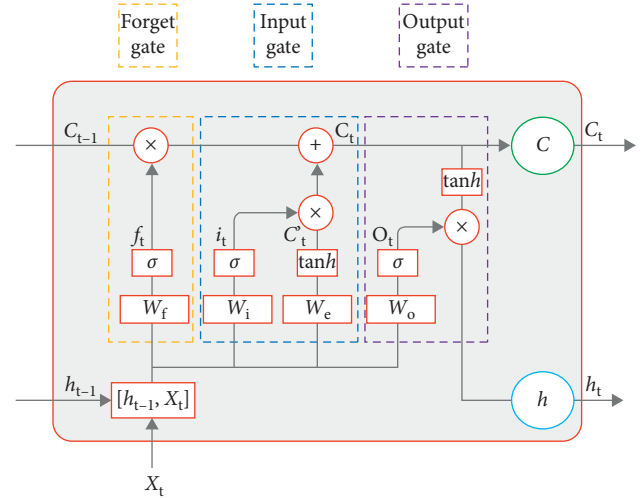
FIGURE 1: Detailed schematic of the long short-term memory block.

classical LSTM model with two input structures and a novel model with four different input structures were conducted to discuss the performance of the LSTM network.

Figure 2 presents a standard pipeline for solar irradiance forecasting through LSTM. The data is divided into training, validation, and test. The feed-forward and feed-backward are the two types of LSTM models that are used to process the data and train network further. The error calculation is carried out when the models are developed, which can be used to describe the training accuracy and decide the feed-backward. At the final stage, the selection of a successful model for prediction is established.

The structure of the conventional LSTM model (we call it Model I) for solar irradiance forecasting can be seen in Figure 3. The network structure contains 1 input layer, 2 LSTM layers (or 1 LSTM layer), and 1 output layer. The input layer includes two different structures in which the input A is the data of historical irradiance, and input B is the data of historical irradiance and meteorological parameters. These structures can be expressed as I-A and I-B. For input A (I-A), the historical irradiance at $t-1, t-2, \ldots, t-m$ time is feed LSTM layer 1; for input B (I-B), the historical irradiance and meteorological parameters at $t-1, t-2, \ldots, t-m$ time is feed LSTM layer 1, and m is the length of the lagging window in time.

Meanwhile, the novel LSTM-MLP structure is proposed in Figure 4 (named Model II). A two-branch structure is designed, including one main input, one auxiliary input, one main output, and one auxiliary output. The data of history irradiance (or irradiance and meteorological parameters) is as main input, which is feed to LSTM layers. When the data is output from the LSTM layer, one part is output as auxiliary output, and the other part is previously combined with the meteorological parameters (auxiliary input) at the current or next time and sent to a new MLP structure. After several hidden layers of MLP, the final output is the main output, which is the irradiance prediction value at the next time.

The simplified expression of the above operation is as follows:

FIGURE 2: The standard LSTM solar irradiance forecasting pipeline.

FIGURE 3: The framework of the traditional LSTM Model I.

FIGURE 4: The framework of the proposed LSTM Model II.

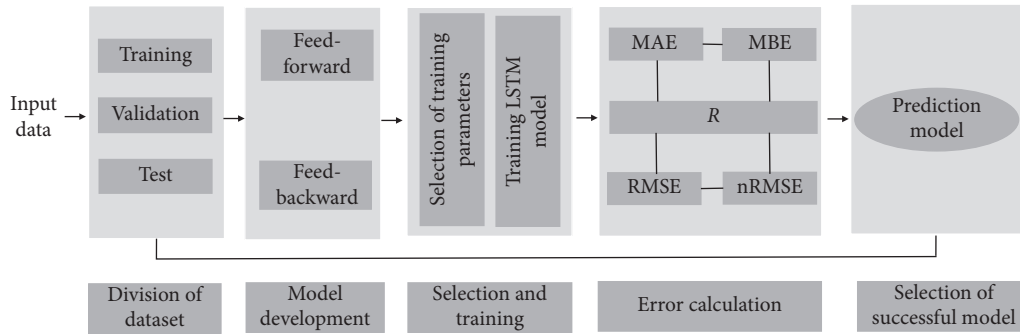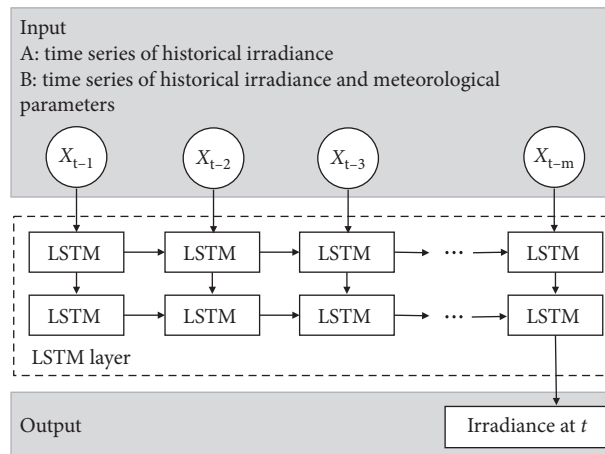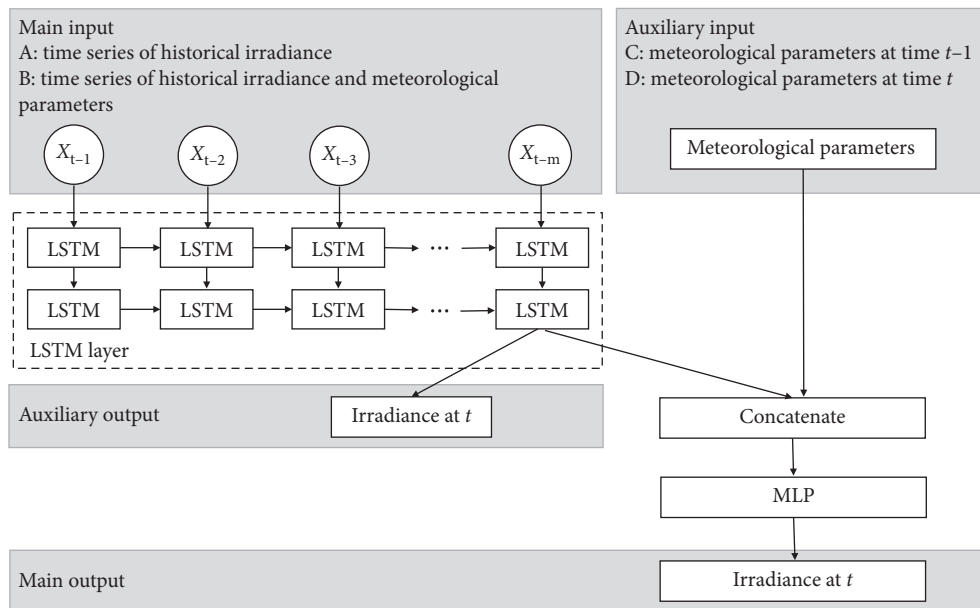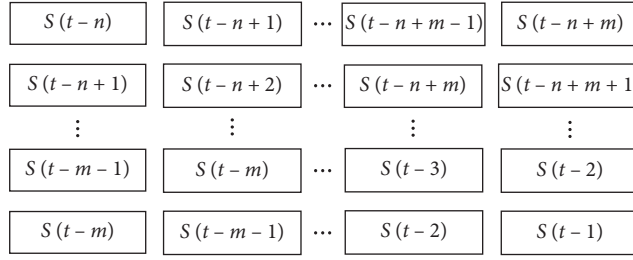| $S(t-n)$ | $S(t-n+1)$ | $\cdots$ | $S(t-n+m-1)$ | $S(t-n+m)$ |
$S(t-n+1)$	$S(t-n+2)$	$\cdots$	$S(t-n+m)$	$S(t-n+m+1)$
$\vdots$	$\vdots$		$\vdots$	$\vdots$
$S(t-m-1)$	$S(t-m)$	$\cdots$	$S(t-3)$	$S(t-2)$
$S(t-m)$	$S(t-m-1)$	$\cdots$	$S(t-2)$	$S(t-1)$

Figure 5: Input (or main input) time series structure of the train samples.

$$
\begin{aligned}
y_{\text{aux}} &= \mathscr{F}_{\text{LSTM}}\left(x_{\text{input}_{\text{main input}}}\right), \\
h_{\text{LSTM}} &= \mathscr{F}_{\text{LSTM}}\left(x_{\text{input}_{\text{main input}}}\right), \\
h_{\text{dense}} &= \text{dense}\left(h_{\text{LSTM}} \oplus x_{\text{aux input}}\right), \\
y_{\text{main}} &= \mathscr{F}_{\text{MLP}}\left(h_{\text{dense}}\right),
\end{aligned}
\tag{10}
$$

where $x_{\text{input}_{\text{main input}}}$ represents the main input, which is the time series of historical irradiance (or together with historical meteorological), $\mathscr{F}_{\text{LSTM}}$ denotes the LSTM layer, h_{LSTM} represent the output through the LSTM layer, $x_{\text{aux input}}$ denotes the auxiliary input described in Figure 4, $\oplus$ means the concatenate operator, dense means the fully connected layer, $\mathscr{F}_{\text{MLP}}$ denotes the MLP layer, and y_{aux} and y_{main} denote the auxiliary output and main output, respectively.

As can be seen in Figure 4, there are two input methods for the main input and auxiliary input, respectively. According to the different combinations of main inputs A and B and auxiliary inputs C and D in Figure 4, the model can be expressed as II-AC, II-AD, II-BC, and II-BD. In order to find better network parameters, six experiments are designed with two models mentioned above. There are (1) Model I-A; (2) Model I-B; (3) Model II-AC; (4) Model II-AD; (5) Model II-BC; and (6) Model II-BD, where the influence of different lagging time parameters (e.g., from Lagging 1 and Lagging 12) is discussed.

Figure 5 shows the input (or main input) time series structure of the train samples. $S(t)$ is the current data, n is the number of train samples, and m is the number of input data in each group, which is the number of lagging time and the length of the lagging window in time. For example, we used $S(t-m)$, $S(t-m-1)$, $S(t-2)$, and $S(t-1)$ as training input and $S(t)$ as training output. Then the data are shifted; the input has become $S(t-m-1)$ to $S(t-2)$, the output is $S(t-1)$, and so on.

2.3. Forecasting Accuracy Evaluation.

To assess the prediction performance of the involved models, four error measures, which include the root mean square error (RMSE), the normalized root mean square error (nRMSE), the mean absolute error (MAE), the mean bias error (MBE), and R (Pearson's correlation coefficients) are utilized in the forecasting experiments.

These indexes can be defined as follows:

$$
\begin{aligned}
\text{RMSE} &= \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(Y_i' - Y_i\right)^2}, \\
\text{nRMSE} &= \frac{\text{RMSE}}{\overline{Y}}, \\
\text{MAE} &= \frac{1}{N}\sum_{i=1}^{N}\left|Y_i - Y_i'\right|, \\
\text{MBE} &= \frac{1}{N}\sum_{i=1}^{N}\left(Y_i' - Y_i\right), \\
R &= \frac{\sum_{i=1}^{N}\left(Y_i - \overline{Y}\right)\left(Y_i' - \overline{Y'}\right)}{\sqrt{\sum_{i=1}^{N}\left(Y_i - \overline{Y}\right)^2}\sqrt{\sum_{i=1}^{N}\left(Y_i' - \overline{Y'}\right)^2}},
\end{aligned}
\tag{11}
$$

where N denotes the number of testing instances, Y_i' denotes the prediction value of the models, $\overline{Y'}$ denotes the mean value of Y_i', Y_i denotes the measured value, and $\overline{Y}$ denotes the mean value of Y_i.

Besides, forecast skill (FS) is an indicator that compares a selected model with a reference model (usually with the persistence model), regardless of the prediction horizon and location [37,38], which is a fair-minded approach to evaluating performance in solar irradiance prediction, as described by the following equation [2]:

$$
\text{FS} = 1 - \frac{\text{nRMSE}}{\text{nRMSE}_{\text{persistence}}}.
\tag{12}
$$

The persistence model is one of the most basic prediction models, which is often applied to compare the performance of other prediction models. The definitions of this model are varied; this paper adopts the most basic definition, which is to assume that the predicted value at the next time is the same as the present value [39,40]:

$$
\text{GHI}_{t+\Delta t} = \text{GHI}_t.
\tag{13}
$$

To further evaluate the performance of the adopted model compared with the benchmark model, the promoting percentage of RMSE (P) is employed to make a further comparison. The formulas are as follows:

$$P = \frac{\text{RMSE}_{\text{benchmark}} - \text{RMSE}_{\text{comparison}}}{\text{RMSE}_{\text{benchmark}}} \times 100\%, \qquad (14)$$

where P is promoting percentage of RMSE, and $\text{RMSE}_{\text{benchmark}}$ and $\text{RMSE}_{\text{comparison}}$ are the root mean square error computed from the benchmark model and comparison model, respectively.

3. Data and Analysis

The data used in this study came from a solar power plant in Denver, Colorado, USA. Average global horizontal irradiance (GHI; in this paper, solar irradiance represents GHI) and meteorological data (such as ambient temperature, relative humidity, wind velocity, atmospheric pressure, precipitation, and so on) have been collected in a one-hour resolution during January 1, 2012, to December 31, 2016, from NREL Solar Radiation Research Laboratory [41]. The data from 2012 to 2015 is used for training and validation; the data from 2016 is used for testing. The main statistical characteristics of solar irradiance in this dataset are shown in Table 1.

Pearson's correlation coefficient is the test statistics that measures the statistical relationship or association between two continuous variables. The relationship between irradiation and wind speed, atmospheric pressure, air temperature, and relative air humidity was analyzed to determine whether these variables should be included as inputs and which parameters to choose as inputs in this network. Table 2 shows the Pearson correlation coefficient between the five weather variables and the solar irradiance on the dataset. It can be observed that only temperature and humidity have a high correlation. However, the irradiance is not correlated with wind speed, precipitation, and pressure, so these three meteorological parameters are excluded. Figure 6 shows the average hourly irradiance distribution for different months in 2016. It can be noticed that there is a strong correlation between hours for each day and solar irradiance. Obviously, the irradiance value is low at the beginning of the day and increases to the peak value at noon and then gradually decreases in the afternoon. Meanwhile, it can be noticed that the peak of irradiance is different every month. The highest peak is between June and July, and the lowest peak value is between December and January. Consequently, the time must be used as an input variable.

Autocorrelation function (ACF) refers to the degree of similarity between time series and their own lag series in a continuous-time interval. However, irradiance is a time-series data, which can be characterized by ACF. Let $\{X_t\}$ be a time series with length T. Denote $\{X_{t-h}\}$ the lagged time series by h periods. The autocorrelation of $\{X_t\}$ at lag h is given by

$$\rho_X(h) = \text{Cov}(X_t - X_{t-h}) = \frac{\gamma_X(h)}{\gamma_X(0)}$$

$$= \frac{E[(X_t - \mu_X)(X_{t-h} - \mu_X)]}{E(X_t - \mu_X)^2}, \qquad (15)$$

Table 1: Main statistical features of solar irradiance in the dataset.

Samples	Statistical indicator (GHI (W/m^2))			
	Number	Max	Mean	Std.
All samples	43824	1090	188.8933	270.6560
Training samples	35040	1090	187.9399	270.3011
Testing samples	8784	1050	192.6963	272.0491

Table 2: Pearson's R correlation coefficients between meteorological parameters and solar irradiation.

Meteorological parameters	Pearson's R correlation coefficients
Wind speed	0.0053
Atmospheric pressure	0.0429
Precipitation	−0.0328
Relative air humidity	−0.3044
Air temperature	0.3745
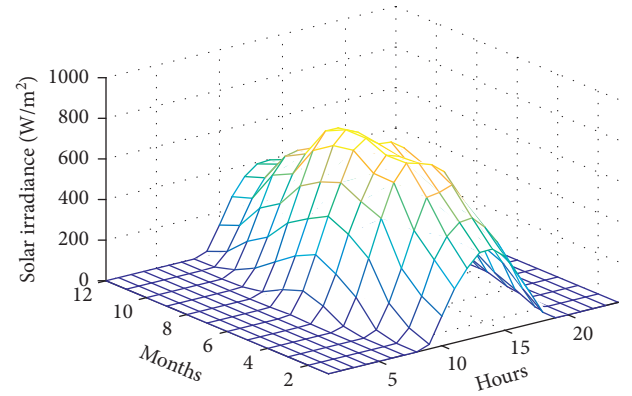
Figure 6: Hourly average solar irradiance data for different months in 2016.

where $\gamma_X(h)$ is the autocovariance of $\{X_t\}$ at lag h, $\gamma_X(0)$ is the autocovariance of $\{X_t\}$ at lag 0, and μ_X is the expected value of $\{X_t\}$.

From the ACF plot above, we can see that our daily period consists of 24 timesteps (where the ACF has the second-largest positive peak). While it was easily apparent from the natural law, it can also be seen from Figure 7 that the time interval of the maximum positive and negative correlation is 12 hours. At the same time, in the actual model calculation, when the lagging time is between 12 and 24, the performance is very similar. Therefore, in this paper, we choose a 12-hour lagging time.

The training dataset is optimized by Adam algorithm, and the sigmoid function is used in the output layer for all models. The program code of this paper is performed on an Intel® Core™ I7-8600 CPU using Python 3.7.5 and Keras 2.3.1 with TensorFlow 2.0.0 backend.

4. Results and Discussion

In this section, the above six models were simulated and calculated to verify the performance of the proposed method. We discuss the effect of the input length
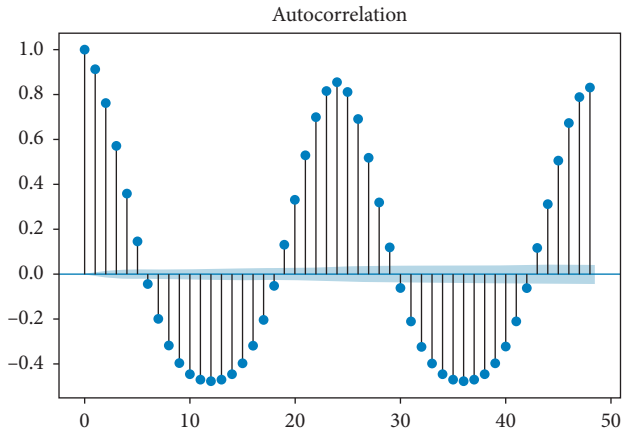
FIGURE 7: ACF plot of hourly solar irradiance. The abscissa represents lag time, and the ordinate represents the autocorrelation coefficient.

(determined by the lagging time). The forecasting results under different lengths of the input sequence with different models are shown in Tables 3–6. The details of forecasting results and analysis are given as follows:

(1) For Model I, since it has only one single-branch input, the number of input variables directly affects the prediction accuracy. As can be seen in Table 3, it is clear that with the increase of lagging time parameter, the RMSE and nRMSE decrease continuously. This fact implies that, for this case study, data from previous points in time is vital for forecasting, especially when only historical irradiance is used for prediction.

(2) However, when the historical irradiance and meteorological parameters are input to the LSTM network at the same time, the influence of the lagging time parameters on forecasting accuracy has a significant downward trend. When the lagging time is only one hour, the RMSE of Model I-A is 110.64 W/m^2, and the RMSE of Model I-B is 75.4654 W/m^2, which shows that when the lagging time is fixed, the information of meteorological parameters helps the prediction of irradiance very well.

(3) As can be seen in Tables 3 and 4, in general, the prediction accuracy will increase with the increase of lagging time in 1–12 hours. However, the expansion of lagging time will lead to a rise in input variables, increasing in operation time. Considering these factors, we need to choose a more reasonable lagging time. In this case, although the best lagging time is 10 hours and 11 hours for Model I-A and Model I-B, respectively, we think the 8 hours lagging time is reasonable. Without a doubt, the perfect lagging time may be different for different datasets.

(4) For Model II-AC and Model II-AD, compared with Model I, we add an independent branch with meteorological parameters (C: meteorological parameters at the current time; D: meteorological parameters at the next time) as input, which plays an important role. Comparing the results with Models I-A and II-AC in Tables 3 and 5, with the same lagging time, the prediction accuracy has a noticeable difference; especially when the lagging time is small, the difference is more prominent. For instance, when the lagging time is 1 hour, the RMSE is 110.64 W/m^2 in Model I-A, but the RMSE is 73.2477 W/m^2 in Model II-AC. The best prediction accuracy of the two models is 75.22 W/m^2 and 71.0791 W/m^2 by RMSE, respectively, which shows that the proposed new branch can improve the prediction accuracy. Meanwhile, it can also be seen from Table 5 that historical irradiance is used as the main input, and whether the auxiliary input is the meteorological parameter at the current time or the next time, the prediction accuracy is the same.

(5) Comparing Tables 5 and 6, we find that using the meteorological parameters of the next moment can better take advantage of the proposed new branch structure. As shown in Model II-BD in Table 6, when historical irradiance and meteorological parameters are the main input and the meteorological parameters at the next moment are the auxiliary input, the prediction effect is the best; the RMSE and nRMSE are 62.1618 W/m^2 and 32.2702, respectively. For Model II-BC, because the current meteorological parameters in the auxiliary input already exist in the main input, the accuracy improvement effect is not apparent.

The best parameters and architecture of the LSTM network for 1-hour-ahead forecasting with the proposed six models are shown in Table 7. In Model I, two LSTM layers within 100 and 40 neurons (100–40) are used with lagging time 10 and 11, respectively, but in Model II, a 64–32 MLP hidden layer is added, and most of them used only one LSTM layer.

The performance of the six models with the optimal parameters and structure can be seen in Table 8 and Figure 8. Compared with the persistence model, the performance of the forecast skill (FS) and the promoting percentage of RMSE (P) of each model is significantly improved. Compared with BPNN, the P of each model has also improved, and the improvement advantage of Model II-BD is more visible, reaching 19.19%.

The RMSE and time cost curve of the different models with different lagging time are shown in Figure 9. It can be seen from the figures that with the increase of lagging time (the dimension of the input variable increases), the time cost increases approximately linearly (especially, in Figure 9(f), there is a sudden change in time cost because the number of LSTM layers increased from 1 to 2). This is because the increase in input variables leads to an increase in the amount of calculation. Meanwhile, except for Model I-A, the RMSE of other models does not decrease linearly with the rise of lagging time, but only shows a certain downward trend, and the whole curve is fluctuant. This indicates that the optimal lagging time is not the maximum lagging time; we need to

Table 3: The performance of Model I-A with 1–12-hour historical irradiance (from Lagging 1 to Lagging 12).

	RMSE (W/m²)	nRMSE (%)	MAE (W/m²)	MBE (W/m²)	R
Lagging 1 (1 h)	110.64	55.67	66.32	−1.57	0.9216
Lagging 2 (2 h)	84.77	43.02	45.52	2.65	0.9524
Lagging 3 (3 h)	80.37	40.75	40.88	1.57	0.9574
Lagging 4 (4 h)	79.56	40.35	43.15	−0.07	0.9583
Lagging 5 (5 h)	78.45	39.78	42.71	0.63	0.9596
Lagging 6 (6 h)	76.95	39.00	40.38	1.07	0.9610
Lagging 7 (7 h)	77.53	39.21	42.17	0.22	0.9607
Lagging 8 (8 h)	75.91	38.47	38.26	−0.94	0.9621
Lagging 9 (9 h)	75.81	38.43	37.43	0.85	0.9622
Lagging 10 (10 h)	**75.22**	**38.12**	**36.90**	**−2.41**	**0.9628**
Lagging 11 (11 h)	75.26	38.15	37.36	2.02	0.9628
Lagging 12 (12 h)	75.91	38.25	41.15	2.25	0.9624

Table 4: The performance of Model I-B with 1–12 hour historical irradiance (from Lagging 1 to Lagging 12).

	RMSE (W/m²)	nRMSE (%)	MAE (W/m²)	MBE (W/m²)	R
Lagging 1 (1 h)	75.4654	39.2068	34.179	−7.2412	0.9615
Lagging 2 (2 h)	73.0181	37.9311	32.2728	−5.3405	0.9637
Lagging 3 (3 h)	72.0597	37.429	33.022	−8.7159	0.9648
Lagging 4 (4 h)	73.1294	37.9803	35.0303	−13.1121	0.9649
Lagging 5 (5 h)	72.5747	37.6879	32.0244	−5.851	0.9641
Lagging 6 (6 h)	73.3373	38.0796	34.9869	−10.9972	0.9638
Lagging 7 (7 h)	72.0033	37.3826	30.9602	−3.4511	0.9648
Lagging 8 (8 h)	71.9089	37.3294	30.8381	−3.7222	0.9647
Lagging 9 (9 h)	71.4177	37.0707	31.2055	−2.3624	0.9656
Lagging10 (10 h)	72.6381	37.7026	30.7169	−0.6251	0.9642
Lagging 11 (11 h)	**71.0791**	**36.8959**	**32.8573**	**−7.3356**	**0.9656**
Lagging 12 (12 h)	71.8434	37.2962	30.6385	−1.9523	0.9649

Table 5: The performance of Model II-AC and II-AD with 1–12-hour historical irradiance (from Lagging 1 to Lagging 12).

	Model II-AC		Model II-AD	
	RMSE (W/m²)	nRMSE (%)	RMSE (W/m²)	nRMSE (%)
Lagging 1 (1 h)	73.2477	38.0547	71.843	37.3249
Lagging 2 (2 h)	71.9784	37.391	71.2248	36.9995
Lagging 3 (3 h)	71.4207	37.0971	71.2623	37.0148
Lagging 4 (4 h)	71.3815	37.0725	71.2267	36.9921
Lagging 5 (5 h)	71.1067	36.9256	71.8271	37.2996
Lagging 6 (6 h)	71.3472	37.0462	70.0358	36.3653
Lagging 7 (7 h)	71.2828	37.0086	70.3305	36.5142
Lagging 8 (8 h)	71.6838	37.2126	**69.7479**	**36.2076**
Lagging 9 (9 h)	71.1645	36.9392	71.8127	37.2757
Lagging 10 (10 h)	**70.2837**	**36.4806**	71.2756	36.9954
Lagging 11 (11 h)	72.045	37.3973	71.3583	37.0408
Lagging 12 (12 h)	70.4213	36.558	70.1675	36.4262

choose the appropriate lagging time according to the actual dataset and the required accuracy.

The one-hour-ahead irradiance forecasted results for the proposed Model II-BD with the best parameters and architecture are shown in Figure 10. As can be seen in Figure 10(a), the blue circle (O) in the figure represents the measured value, the red asterisk (*) denotes the forecasted value, and the predicted value and the actual value can remain the same for most of the time. It can be shown more clearly from the local enlarged drawing that the difference between measured and forecasted values is small. It is clear from Figure 10(b) that the predicted values are strongly

correlated with the measured solar irradiance data, and the linear regression coefficient reaches 0.9642. So, in summary, the forecasted values of the solar irradiance have good agreement with the measured values.

Through the above experimental results, we found that the Model II-BD structure of the LSTM-MLP model has the best prediction accuracy. The following LSTM-MLP model specifically represents the LSTM-MLP model with a Model II-BD structure.

Six experimental simulations were performed to verify the performance of the proposed LSTM-MLP model, including BP network, general RNN network, random forest

TABLE 6: The performance of Model II-BC and II-BD with 1–12-hour historical irradiance (from Lagging 1 to Lagging 12).

	Model II-BC		Model II-BD	
	RMSE (W/m^2)	nRMSE (%)	RMSE (W/m^2)	nRMSE (%)
Lagging 1 (1 h)	72.3486	37.5875	67.076	34.8482
Lagging 2 (2 h)	72.7865	37.8108	66.3851	34.4854
Lagging 3 (3 h)	71.5573	37.168	66.7471	34.6695
Lagging 4 (4 h)	72.5258	37.6668	66.0398	34.2982
Lagging 5 (5 h)	72.8937	37.8535	64.2217	33.3502
Lagging 6 (6 h)	70.9322	36.8307	63.4328	32.9368
Lagging 7 (7 h)	**70.4761**	**36.5898**	64.4025	33.4365
Lagging 8 (8 h)	71.8012	37.2735	63.4285	32.927
Lagging 9 (9 h)	71.4277	37.0758	62.955	32.678
Lagging 10 (10 h)	71.2123	36.9626	63.0588	32.7305
Lagging 11 (11 h)	70.8344	36.7688	64.965	33.7222
Lagging 12 (12 h)	71.6534	37.1976	**62.1618**	**32.2702**

TABLE 7: The best parameters and architecture of the six LSTM models for 1 hour ahead forecasting.

Model	Input shape		Structure (hidden layers)	Optimizer/epoch
	Main input	Auxiliary input		
Model I-A	$(k \times 10 \times 1)$	—	100-40 (LSTM)	Adam/200
Model I-B	$(k \times 11 \times 5)$	—	100-40 (LSTM)	Adam/200
Model II-AC	$(k \times 10 \times 1)$	$(k \times 3 \times 1)$	32 (LSTM)-64-32 (MLP)	Adam/200
Model II-AD	$(k \times 8 \times 1)$	$(k \times 3 \times 1)$	32 (LSTM)-64-32 (MLP)	Adam/200
Model II-BC	$(k \times 7 \times 5)$	$(k \times 3 \times 1)$	32 (LSTM)-64-32 (MLP)	Adam/200
Model II-BD	$(k \times 12 \times 5)$	$(k \times 3 \times 1)$	30-10 (LSTM)-64-32 (MLP)	Adam/200

k: the sample size of the minibatch.

TABLE 8: The performance of the six models with the optimal parameters and structure.

Model	Model I-A	Model I-B	Model II-AC	Model II-AD	Model II-BC	Model II-BD	Persistence	BPNN
RMSE (W/m^2)	75.22	71.0791	70.2837	69.7479	70.4761	62.1618	112.5854	76.9272
nRMSE (%)	38.12	36.895	36.4806	36.2076	36.5898	32.2702	58.4263	39.9215
FS	0.3476	0.36856	0.37566	0.3803	0.3737	0.4477	0	0.3167
P1 (%)	33.19	36.87	37.57	38.05	37.40	44.79	0	0.3167
P2 (%)	2.22	7.60	8.64	9.33	8.39	19.19	−46.35	0

P1: the benchmark model is persistence model; P2: the benchmark model is BPNN model.
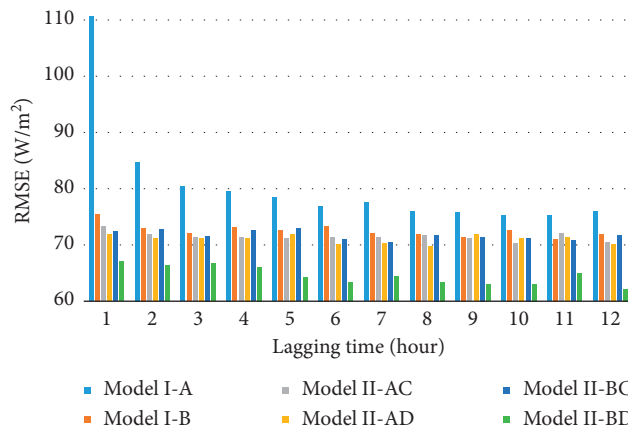
FIGURE 8: The RMSE of the six models with different lagging time (from 1 hour to 12 hours) based on the optimal parameters and structure.

network, SVM network, general LSTM, and LSTM-MLP model. The forecasting results are shown in Table 9. As can be seen from the table, the RMSE, nRMSE, MAE, MBE, and R criteria of the proposed LSTM-MLP model outperformed the other five general machine learning models. Compared with BPNN, RNN, random forest, SVM, and LSTM, the promoting percentage of RMSE (P) was improved by 19.19%, 20.15%, 11.68%, 19.31%, and 13.48%, respectively. Obviously, the
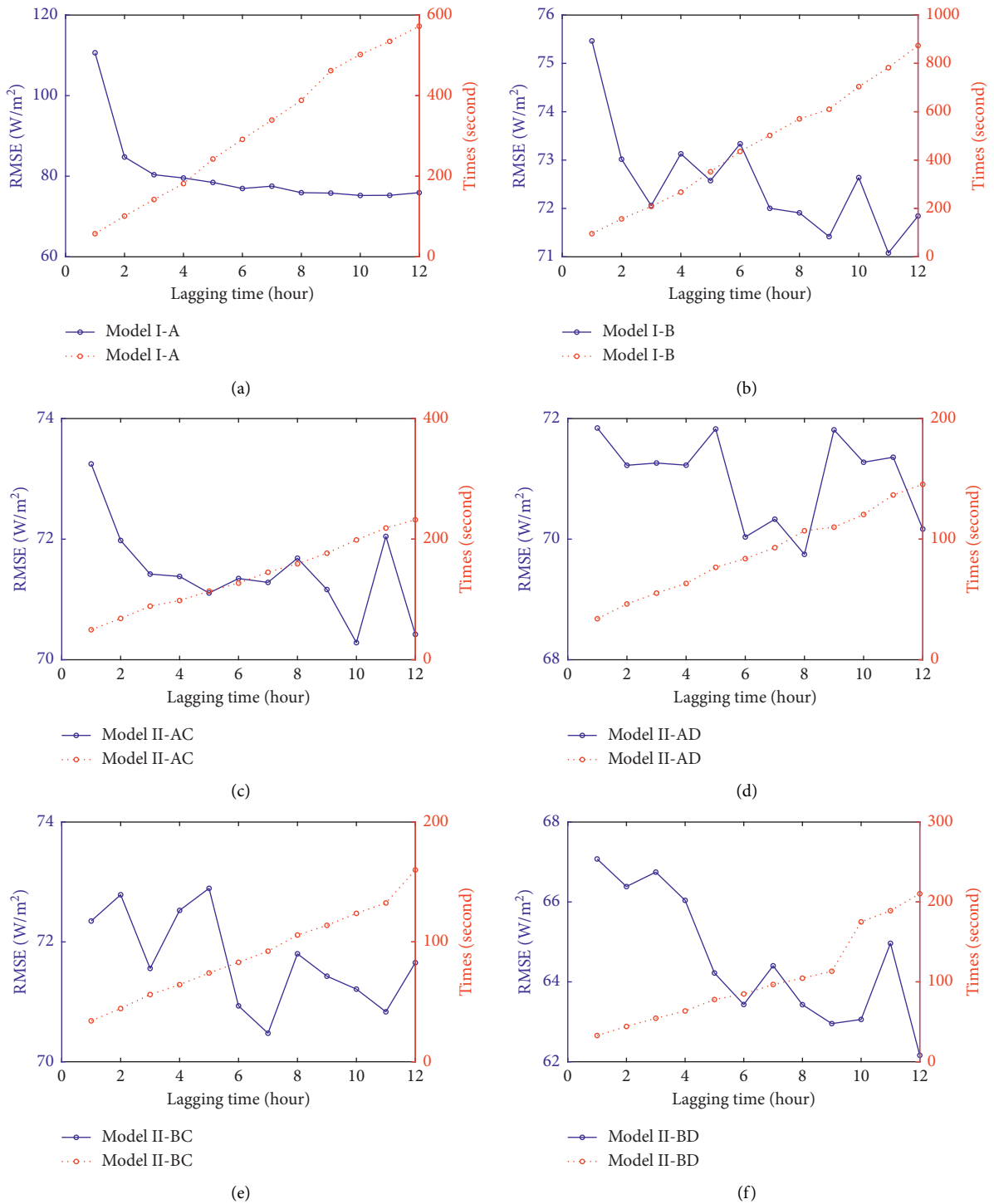
Figure 9: The RMSE and time cost curve of the different models with different lagging time. (a) Model I-A, (b) Model I-B, (c) Model II-AC, (d) Model II-AD, (e) Model II-BC, and (f) Model II-BD.

LSTM-MLP model's predicted results are better than those of the other five models. This is because, for the LSTM-MLP model, it is a mixture model of LSTM and MLP models, in which the MLP model adds a new input containing the future hidden information, so it can improve the prediction accuracy.

Furthermore, the data of three weather conditions are randomly selected from the test set, and the results are shown in Figure 11 and Table 10. On a sunny day (June 27, 2016), all the prediction curves and the measurement curve are in good agreement, LSTM model has the best prediction result, but LSTM-MLP also has high precision, and the nRMSE is only 4.73%. On a cloudy day (June 8, 2016), the measured values show significant volatility, and the predicted values of different models have similar trend curves,

(a)

(b)
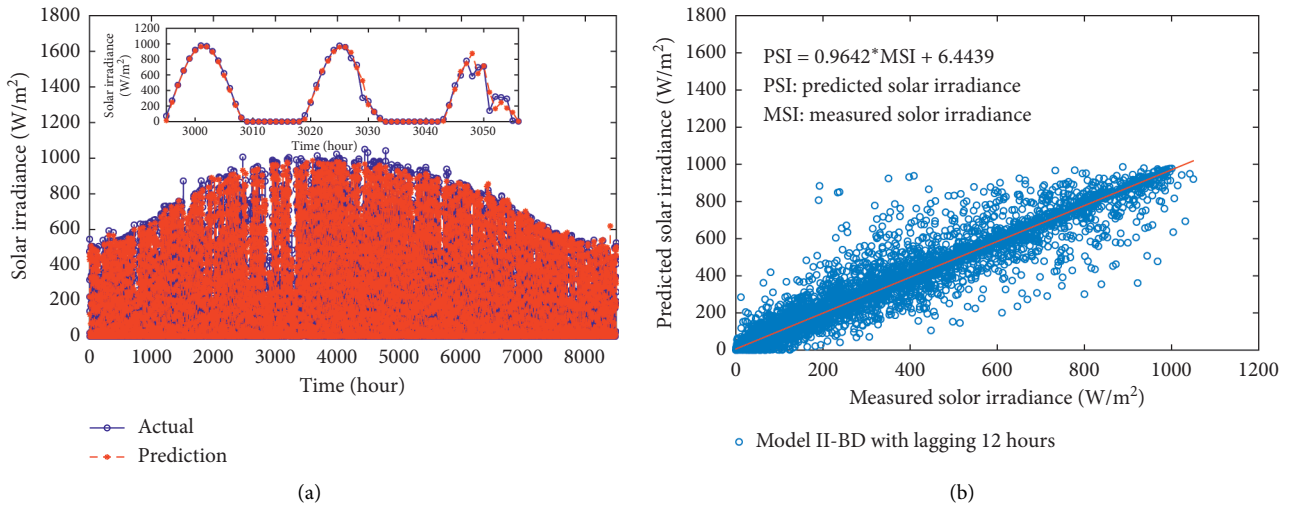
FIGURE 10: Scatter plots of actual vs. predicted values for the proposed Model II-BD with the best parameters and architecture.

TABLE 9: Performance comparison between the proposed LSTM-MLP model and the general machine learning method.

Model	RMSE (W/m²)	nRMSE (%)	MAE	MBE	R
BPNN	76.9272	39.9215	33.2872	−4.7798	0.9601
RNN	77.8444	40.4115	40.0704	−16.2864	0.9602
Random forest	70.3808	36.5369	30.5618	−0.9593	0.9659
SVM	77.0384	39.9931	50.9564	−1.6776	0.9593
General LSTM	71.8434	37.2962	30.6385	−1.9523	0.9649
The proposed LSTM-MLP	62.1618	32.2702	26.6538	−0.4547	0.9737

(a)

(b)

FIGURE 11: Continued.

(c)

Figure 11: Comparison between measured and forecasted hourly solar irradiance for three types of weather with different methods. (a) Sunny (June 27, 2016), (b) cloudy (June 8, 2016), and (c) rainy (May 17, 2016).

Table 10: Root mean square error (RMSE, W/m^2) and normalized root mean square error (nRMSE, %) of different models in sunny, cloudy, and rainy weather.

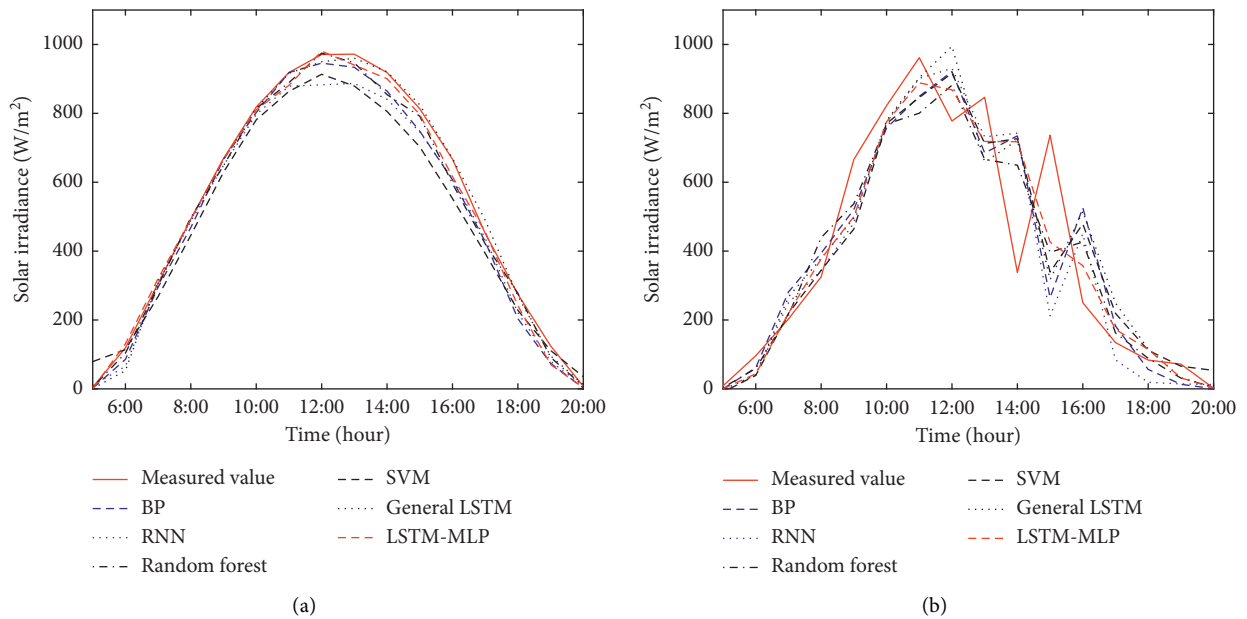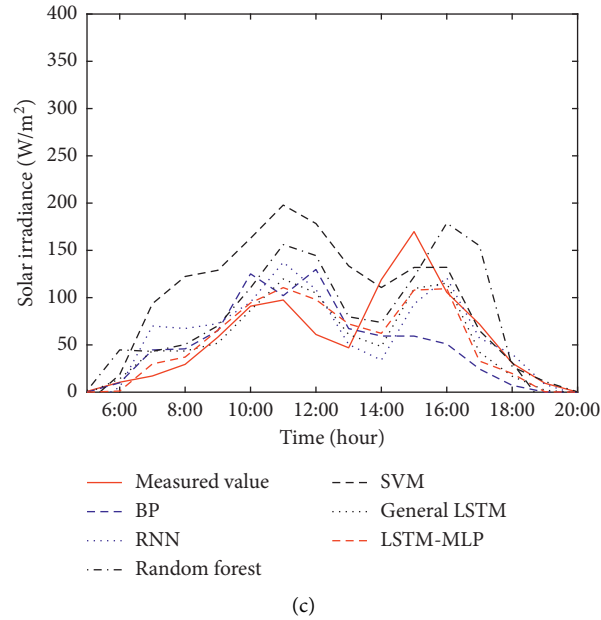| Model	Sunny (June 27, 2016)		Cloudy (June 8, 2016)		Rainy (May 17, 2016)	
	RMSE	nRMSE	RMSE	nRMSE	RMSE	nRMSE
BPNN	38.0148	7.1308	186.9453	47.3051	42.6140	74.164
RNN	49.4941	9.2841	188.9789	47.8197	37.2885	64.8956
Random forest	29.4791	5.5297	147.7844	37.3957	43.9405	76.4725
SVM	66.8935	12.5478	171.4332	43.3799	61.1678	106.4545
General LSTM	**21.8439**	**4.0975**	183.1270	46.3389	29.0062	50.4815
The proposed LSTM-MLP	25.2291	4.7325	**141.8265**	**35.8881**	**26.6292**	**46.3446**

Table 11: Comparison between the best result obtained in this study and conventional methods.

Reference	Model type	Location of data used	Performance
Lee et al. [42]	Angstrom-type equations	South Korea (Seoul)	nRMSE (%) = 43.09 nMBE (%) = −12.24
Brabec et al. [43]	Heteroscedastic model (HR)	Romania (Timisoara)	nRMSE (%) = 45.80
Voyant et al. [44]	MLP	France (Ajaccio, Corsica)	nRMSE (%) = 40.55
Voyant et al. [44]	ARMA	France (Ajaccio, Corsica)	nRMSE (%) = 40.32
Voyant et al. [44]	Hybrid MLP-ARMA	France (Ajaccio, Corsica)	nRMSE (%) = 36.59
Trapero et al. [45]	ARIMA	Spain (Castilla-La Mancha)	nRMSE (%) = 37.34 nMBE (%) = 1.89
Akarslan and Hocaoglu [46]	Adaptive approach	Turkey (Çanakkale)	nRMSE (%) = 34.86 RMSE = 68.4132 W/m^2
Zhao et al. [47]	3D-CNN with AR	NREL database	nRMSE (%) = 34.97 nMBE (%) = 0.65 FS = 5.89
Bae [48]	ANN	South Korea (Yuseong-gu, Daejeon)	RMSE = 71.41 W/m^2 R = 0.9264
Bae [48]	NAR	South Korea (Yuseong-gu, Daejeon)	RMSE = 111.41 W/m^2 R = 0.7912
Bae [48]	SVM	South Korea (Yuseong-gu, Daejeon)	RMSE = 58.72 W/m^2 R = 0.9562

TABLE 11: Continued.

Reference	Model type	Location of data used	Performance
Qing and Niu [30]	LSTM	Cape Verde (Santiago)	RMSE = 76.245 W/m^2
Yu [29]	LSTM	USA (Hawaii)	RMSE = 66.69 W/m^2 MAE = 46.04 W/m^2 R = 0.95
Yu [29]	SVR	USA (Hawaii)	RMSE = 144.43 W/m^2 MAE = 96.25 W/m^2 R = 0.75
This study	Model II-BD based on LSTM-MLP	USA (Denver, Colorado)	nRMSE (%) = 32.27 RMSE = 62.16 W/m^2 MAE = 26.65 W/m^2 MBE = −0.4547 W/m^2 R = 0.9737

but the error is more prominent. The rapid change of the cloud layer in the cloudy day brings enormous difficulties with irradiance prediction. In contrast, the proposed model shows a good prediction effect; the nRMSE is 35.89%. On a rainy day (May 17, 2016), the measured value of irradiance is low, which can be seen from the solid red line in Figure 11, but the predicted value of the red dotted line can better follow the change of measured value. This indicates that the proposed LSTM-MLP model shows better performance on rainy days. All related results are reported in Table 10.

In order to place the work with other published works, the results with the proposed approach and results from different studies of others are compared in Table 11. The results are similar.

5. Conclusions

In this work, a new novel LSTM-MLP structure with two-branch input is proposed. The proposed LSTM-MLP includes one main input, one auxiliary input, one main output, and one auxiliary output. The data of historical irradiance (or irradiance and meteorological parameters) is as main input, which is feed to LSTM layers. One part from the LSTM layer is output as auxiliary output, and the other part is previously combined with the meteorological parameters (auxiliary input) and sent to a new MLP structure. The output from several hidden layers of MLP is the main output, which is the final irradiance prediction value. Four network structures based on LSTM-MLP and two network structures based on traditional LSTM are designed and developed. A real-world test case in Denver, which consists of 5 years of data, is used to verify and discuss the potential of each model.

The experimental results demonstrate that the proposed Model II-BD, which with historical irradiance and meteorological parameters as main input and the next moment meteorological parameters as an auxiliary input, significantly outperforms other models in terms of three widely used evaluation criteria. The RMSE is 62.1618 W/m^2, the nRMSE is 32.2702%, and FS is 0.4477. Compared with BPNN, the promoting percentage of RMSE (P) of Model II-BD is 19.19%. The meteorological parameters at the next moment play a vital role in the prediction accuracy, which can be gained by the weather forecast. The lagging time is a

significant variable for the input of LSTM, especially when only historical irradiance is used as input (e.g., Model I-A).

Data Availability

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 article.

Acknowledgments

References

[1] IRENA, *Renewable Capacity Statistics 2020 International Renewable Energy Agency (IRENA)*, Abu Dhabi, UAE, 2020, https://www.irena.org/-/media/Files/IRENA/Agency/Publication/2020/Mar/IRENA_RE_Capacity_Statistics_2020.pdf.

[2] D. Yang, J. Kleissl, C. A. Gueymard, H. T. C. Pedro, and C. F. M. Coimbra, "History and trends in solar irradiance and PV power forecasting: a preliminary assessment and review using text mining," *Solar Energy*, vol. 168, pp. 60–101, 2018.

[3] W. Ji and K. C. Chee, "Prediction of hourly solar radiation using a novel hybrid model of ARMA and TDNN," *Solar Energy*, vol. 85, no. 5, pp. 808–817, 2011.

[4] C. Voyant, M. Muselli, C. Paoli, and M.-L. Nivet, "Numerical weather prediction (NWP) and hybrid ARMA/ANN model to predict global radiation," *Energy*, vol. 39, no. 1, pp. 341–355, 2012.

[5] Y. Wang, C. Wang, C. Shi, and B. Xiao, "Short-term cloud coverage prediction using the ARIMA time series model," *Remote Sensing Letters*, vol. 9, no. 3, pp. 274–283, 2017.

[6] J. Hassan, "ARIMA and regression models for prediction of daily and monthly clearness index," *Renewable Energy*, vol. 68, pp. 421–427, 2014.

[7] Z. Dong, D. Yang, T. Reindl, and W. M. Walsh, "A novel hybrid approach based on self-organizing maps, support vector regression and particle swarm optimization to forecast solar irradiance," *Energy*, vol. 82, pp. 570–577, 2015.

[8] H. Jiang, Y. Dong, and L. Xiao, "A multi-stage intelligent approach based on an ensemble of two-way interaction model for forecasting the global horizontal radiation of India," *Energy Conversion and Management*, vol. 137, pp. 142–154, 2017.

[9] H. Verbois, R. Huva, A. Rusydi, and W. Walsh, "Solar irradiance forecasting in the tropics using numerical weather prediction and statistical learning," *Solar Energy*, vol. 162, pp. 265–277, 2018.

[10] R. Perez, E. Lorenz, S. Pelland et al., "Comparison of numerical weather prediction solar irradiance forecasts in the US, Canada and Europe," *Solar Energy*, vol. 94, pp. 305–326, 2013.

[11] J. O. Kamadinata, T. L. Ken, and T. Suwa, "Sky image-based solar irradiance prediction methodologies using artificial neural networks," *Renewable Energy*, vol. 134, pp. 837–845, 2019.

[12] Y. Chu, H. T. C. Pedro, M. Li, and C. F. M. Coimbra, "Real-time forecasting of solar irradiance ramps with smart image processing," *Solar Energy*, vol. 114, pp. 91–104, 2015.

[13] F. Wang, Z. Zhen, C. Liu et al., "Image phase shift invariance based cloud motion displacement vector calculation method for ultra-short-term solar PV power forecasting," *Energy Conversion and Management*, vol. 157, pp. 123–135, 2018.

[14] G. M. Yagli, D. Yang, and D. Srinivasan, "Automatic hourly solar forecasting using machine learning models," *Renewable and Sustainable Energy Reviews*, vol. 105, pp. 487–498, 2019.

[15] C. Fan, J. Wang, W. Gang, and S. Li, "Assessment of deep recurrent neural network-based strategies for short-term building energy predictions," *Applied Energy*, vol. 236, pp. 700–710, 2019.

[16] F. Olaiya and A. B. Adeyemo, "Application of data mining techniques in weather prediction and climate change studies," *International Journal of Information Engineering and Electronic Business*, vol. 4, no. 1, pp. 51–59, 2012.

[17] E. Akarslan, F. O. Hocaoğlu, and R. Edizkan, "A novel M-D (multi-dimensional) linear prediction filter approach for hourly solar radiation forecasting," *Energy*, vol. 73, pp. 978–986, 2014.

[18] F. Wang, Z. Xuan, Z. Zhen et al., "A minutely solar irradiance forecasting method based on real-time sky image-irradiance mapping model," *Energy Conversion and Management*, vol. 220, Article ID 113075, 2020.

[19] Y. Dong, Z. Zhang, and W.-C. Hong, "A hybrid seasonal mechanism with a chaotic cuckoo search algorithm with a support vector regression model for electric load forecasting," *Energies*, vol. 11, no. 4, p. 1009, 2018.

[20] Z. Peng, D. Yu, D. Huang, J. Heiser, and P. Kalb, "A hybrid approach to estimate the complex motions of clouds in sky images," *Solar Energy*, vol. 138, pp. 10–25, 2016.

[21] L. Benali, G. Notton, A. Fouilloy, C. Voyant, and R. Dizene, "Solar radiation forecasting using artificial neural network and random forest methods: application to normal beam, horizontal diffuse and global components," *Renewable Energy*, vol. 132, pp. 871–884, 2019.

[22] S. Monjoly, M. André, R. Calif, and T. Soubdhan, "Hourly forecasting of global solar radiation based on multiscale decomposition methods: a hybrid approach," *Energy*, vol. 119, pp. 288–298, 2017.

[23] W.-C. Hong, Y. Dong, C.-Y. Lai, L.-Y. Chen, and S.-Y. Wei, "SVR with hybrid chaotic immune algorithm for seasonal load demand forecasting," *Energies*, vol. 4, no. 6, pp. 960–977, 2011.

[24] Z. Zhang, W.-C. Hong, and J. Li, "Electric load forecasting by hybrid self-recurrent support vector regression model with

[25] Z. Zhang, S. Ding, and Y. Sun, "A support vector regression model hybridized with chaotic krill herd algorithm and empirical mode decomposition for regression task," *Neurocomputing*, vol. 410, pp. 185–201, 2020.

[26] G.-F. Fan, S. Qing, H. Wang, W.-C. Hong, and H.-J. Li, "Support vector regression model based on empirical mode decomposition and auto regression for electric load forecasting," *Energies*, vol. 6, no. 4, pp. 1887–1901, 2013.

[27] C. Voyant, G. Notton, S. Kalogirou et al., "Machine learning methods for solar radiation forecasting: a review," *Renewable Energy*, vol. 105, pp. 569–582, 2017.

[28] S. Srivastava and S. Lessmann, "A comparative study of LSTM neural networks in forecasting day-ahead global horizontal irradiance with satellite data," *Solar Energy*, vol. 162, pp. 232–247, 2018.

[29] Y. Yu, J. Cao, and J. Zhu, "An LSTM short-term solar irradiance forecasting under complicated weather conditions," *IEEE Access*, vol. 7, pp. 145651–145666, 2019.

[30] X. Qing and Y. Niu, "Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM," *Energy*, vol. 148, pp. 461–468, 2018.

[31] M. Abdel-Nasser and K. Mahmoud, "Accurate photovoltaic power forecasting models using deep LSTM-RNN," *Neural Computing and Applications*, vol. 31, no. 7, pp. 2727–2740, 2017.

[32] Y. Li, H. Wu, and H. Liu, "Multi-step wind speed forecasting using EWT decomposition, LSTM principal computing, RELM subordinate computing and IEWT reconstruction," *Energy Conversion and Management*, vol. 167, pp. 203–219, 2018.

[33] H. Liu, X. Mi, and Y. Li, "Smart multi-step deep learning model for wind speed forecasting based on variational mode decomposition, singular spectrum analysis, LSTM network and ELM," *Energy Conversion and Management*, vol. 159, pp. 54–64, 2018.

[34] H. Liu, X.-w. Mi, and Y.-f. Li, "Wind speed forecasting method based on deep learning strategy using empirical wavelet transform, long short term memory neural network and Elman neural network," *Energy Conversion and Management*, vol. 156, pp. 498–514, 2018.

[35] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997, http://www7.informatik.tu-muenchen.de/~hochreithttp://www.idsia.ch/~juergen.

[36] X. Ma, Z. Tao, Y. Wang, H. Yu, and Y. Wang, "Long short-term memory neural network for traffic speed prediction using remote microwave sensor data," *Transportation Research Part C: Emerging Technologies*, vol. 54, pp. 187–197, 2015.

[37] B. Joshi, M. Kay, J. K. Copper, and A. B. Sproul, "Evaluation of solar irradiance forecasting skills of the Australian bureau of meteorology's ACCESS models," *Solar Energy*, vol. 188, pp. 386–402, 2019.

[38] X. Huang, J. Shi, B. Gao, Y. Tai, Z. Chen, and J. Zhang, "Forecasting hourly solar irradiance using hybrid wavelet transformation and elman model in smart grid," *IEEE Access*, vol. 7, pp. 139909–139923, 2019.

[39] L. Martín, L. F. Zarzalejo, J. Polo, A. Navarro, R. Marchante, and M. Cony, "Prediction of global solar irradiance based on time series analysis: application to solar thermal power plants energy production planning," *Solar Energy*, vol. 84, no. 10, pp. 1772–1781, 2010.

[40] R. Blaga, A. Sabadus, N. Stefu, C. Dughir, M. Paulescu, and V. Badescu, "A current perspective on the accuracy of incoming solar energy forecasting," *Progress in Energy and Combustion Science*, vol. 70, pp. 119–144, 2019.

[41] A. Andreas and T. Stoffel, "NREL Solar Radiation Research Laboratory (SRRL): Baseline Measurement System (BMS); Golden, Colorado (Data)," "NREL Report No. DA-56488, National Renewable Energy Laboratory, Golden, CO, USA, 1981.

[42] K. Lee, H. Yoo, and G. J. Levermore, "Quality control and estimation hourly solar irradiation on inclined surfaces in South Korea," *Renewable Energy*, vol. 57, pp. 190–199, 2013.

[43] M. Brabec, M. Paulescu, and V. Badescu, "Tailored vs black-box models for forecasting hourly average solar irradiance," *Solar Energy*, vol. 111, pp. 320–331, 2015.

[44] C. Voyant, C. Darras, M. Muselli, C. Paoli, M.-L. Nivet, and P. Poggi, "Bayesian rules and stochastic models for high accuracy prediction of solar radiation," *Applied Energy*, vol. 114, pp. 218–226, 2014.

[45] J. R. Trapero, N. Kourentzes, and A. Martin, "Short-term solar irradiation forecasting based on dynamic harmonic regression," *Energy*, vol. 84, pp. 289–295, 2015.

[46] E. Akarslan and F. O. Hocaoglu, "A novel adaptive approach for hourly solar radiation forecasting," *Renewable Energy*, vol. 87, pp. 628–633, 2016.

[47] X. Zhao, H. Wei, H. Wang, T. Zhu, and K. Zhang, "3D-CNN-based feature extraction of ground-based cloud images for direct normal irradiance prediction," *Solar Energy*, vol. 181, pp. 510–518, 2019.

[48] K. Y. Bae, H. S. Jang, and D. K. Sung, "Hourly solar irradiance prediction based on support vector machine and its error analysis," *IEEE Transactions on Power Systems*, vol. 32, no. 2, pp. 935–945, 2017.

Research Article

Time Series Prediction of Electricity Demand Using Adaptive Neuro-Fuzzy Inference Systems

Amevi Acakpovi [ID],[1] **Alfred Tettey Ternor,**[2] **Nana Yaw Asabere,**[3] **Patrick Adjei,**[4] **and Abdul-Shakud Iddrisu**[5]

[1]*Department of Electrical/Electronic Engineering, Accra Technical University, Accra, Ghana*
[2]*Department of Computer Engineering, University of Ghana, Accra, Ghana*
[3]*Department of Computer Science, Accra Technical University, Accra, Ghana*
[4]*Department of Electrical Engineering, Regional Maritime University, Accra, Ghana*
[5]*Department of Telecommunications Engineering, Ghana Technology University College, Accra, Ghana*

Correspondence should be addressed to Amevi Acakpovi; acakpovia@gmail.com

This paper is concerned with the reliable prediction of electricity demands using the Adaptive Neuro-Fuzzy Inference System (ANFIS). The need for electricity demand prediction is fundamental and vital for power resource planning and monitoring. A dataset of electricity demands covering the period of 2003 to 2018 was collected from the Electricity Distribution Company of Ghana, covering three urban areas namely Mallam, Achimota, and Ga East, all in Ghana. The dataset was divided into two parts: one part covering a period of 0 to 500 hours was used for training of the ANFIS algorithm while the second part was used for validation. Three scenarios were considered for the simulation exercise that was done with the MATLAB software. Scenario one considered four inputs sampled data, scenario two considered an additional input making it 5, and scenario 3 was similar to scenario 1 with the exception of the number of membership functions that increased from 2 to 3. The performance of the ANFIS algorithm was assessed by comparing its predictions with other three forecast models namely Support Vector Regression (SVR), Least Square Support Vector Machine (LS-SVM), and Auto-Regressive Integrated Moving Average (ARIMA). Findings revealed that the ANFIS algorithm can perform the prediction accurately, the ANFIS algorithm converges faster with an increase in the data used for training, and increasing the membership function resulted in overfitting of data which adversely affected the RMSE values. Comparison of the ANFIS results to other previously used methods of predicting electricity demands including SVR, LS-SVM, and ARIMA revealed that there is merit to the potentials of the ANFIS algorithm for improved predictive accuracy while relying on a quality data for training and reliable setting of tuning parameters.

1. Introduction

Forecasting electricity demand is vital for power generation and planning. Accurate forecast of electricity demands presents a better understanding of the electricity network expansion and generation to sustainably cater for future demands [1–3].

Similarly, a forecast of electricity demand informs network operators about the growth of the grid and provides optimal development and expansion of the Grid system. Load forecasting is important to satisfy the power demand and meet consumers growing electrical needs [4]. Load forecasting is a major field in electrical engineering especially power systems. The power industry requires a forecast to predict production and financial outlay. It is important to predict energy demand as well as peak power demand. Accurate information on electricity demand is a requirement in the reliable operation of power systems. Since electricity is expensive to store efficiently in large quantities, the amount generated at any given time must meet all the demand from consumers as well as grid losses.

Load forecasting can, therefore, be defined as the estimation of future load conditions and their effect on power

generation based on previous data [4]. Load forecasting is very crucial to decision-making in electrical energy production and distribution. To carry out a load forecast, past data records are necessary input decision variables. A load forecast can be more specific and may cover a range of possible outcomes.

Load forecasting is used to decide whether extra generation must be injected into the grid by either increasing the output of the online generator system, by commissioning one or more extra generating units or by interchanging power with a less loaded generating system. In addition, load forecasts are used to decide whether the output of existing online generator must be reduced or switched off, which is determined by the generation control strategy used by the power system company like scheduling, unit-commitment, interchange evaluation, reserve management, economic load dispatch, and co-ordination [5, 6]. Consequently, the accuracy of forecasted loads directly impacts the financial outlay and reliability of system operations.

Load forecasting is divided into three types according to the time period in which the forecast is undertaken. The three types of load forecasts are short-term, medium-term, and long-term load forecasting [7]. In short-term load forecasting [8], the aim is to estimate the load for the next half hour up to the next two weeks. Short-term load forecasting helps power system operators with various decision-making in the power system, including supply planning, generation reserve, system security, dispatching scheduling, demand-side management, and financial planning. Medium-Term Load Forecasting (MTLF) is a category of electric load forecasting that covers a time span of up to one year. It suits outages and maintenance planning, as well as load switching operation. Long-Term Load Forecasting (LTLF) is load forecasting that usually covers forecasting horizons of one to ten years and sometimes extends to several decades [9]. It provides weekly/monthly forecasts for peak and valley loads which are important to expand generation, transmission, and distribution systems. In addition, long-term forecasts are used for investment planning and the expansion of power system infrastructure.

Load forecasting must be carefully done to avoid wrong planning or causing financial burden. An overestimation of load demand causes economic waste because huge financial commitment will be required for the construction of additional power capacity, while underestimation will result in unreliable supply or shortage of supply to customers.

Load forecasting can also be classified into quantitative and qualitative methods. Quantitative methods are based on established mathematical analysis like the regression analysis, exponential smoothing, Box-Jenkins methods, and decomposition methods. Qualitative load forecasting methods sample views from experts in order to forecast future load intuitively. The latter method is used when previous or past data are not available for the analysis. These qualitative methods include the Delphi method, Technological comparisons, and subjective curving.

This paper is concerned with a medium-term load forecasting Electricity demand for a selected town in the Greater Accra Region, Mallam town, which is densely populated. This study is useful in the generation capacity planning for future network upgrades due to the increasing load demand of the community. The rest of the paper is structured to cover the methodology, the result, discussion, and conclusion sections, respectively.

2. Literature Review

Electricity load forecasting is an activity that has been carried regularly in the past with numerous methods. The accuracy and reliability of this prediction varies based on the methodology used; some of the prominent methods used in the past to conduct time series electricity forecasting include the following: simple moving average (SMA), weighted moving average (WMA), simple exponential smoothing (SES), Holt linear trend (HL), Holt-Winters (HW), ARIMA models, vector autoregressive (VAR) forecasting models, and artificial neural networks and support vector machines (SVM). This section presents the strengths and weaknesses of the above-mentioned methods and makes a case for the consideration of the Adaptive Neuro-Fuzzy Inference System (ANFIS) in predicting electricity load forecast.

Moving average is based on the calculation of average price over a given period, an average that is termed as moving. In other terms, it is the average of a selected range of prices by the period in that range. SMA is popularly used to determine price direction either upward or downward. SMA has been used for electricity demand prediction by a number of available studies [6, 10] even though its original strength has to do with cost prediction of goods that vary often like fuel. The exponential moving average (EMA) is an enhancement of the SMA that weights recent price action better. It has been proved that the longer the period of the SMA, the smoother the result. However, the same longer period introduces lags between the SMA prediction and its source.

Weighted Moving Average (WMA) is a similar method to the Simple Moving Average but puts more weight on recent previous data than past ones. In WMA, a heavier weighting is assigned to more current data than past data. This approach enhances the prediction, and the method seems to be generally more accurate than the SMA. WMA also helps determine trend direction.

Like the WMA, the exponential moving average also considers more weighting to recent data than past data; however, the difference between previous and current data follows an exponential model rather than being simply additive or incremental. The main difference is that the EMA reacts more significantly to recent price changes.

Moving average (MA) techniques, however, have several limitations. Just like any other time series prediction tool, MA does not take into account the fundamental factors that affect electricity consumption and pattern, but it is only based on a history of recorded data. MA techniques have the flexibility to be spread out over any time period and this poses some challenges in terms of accuracy of prediction. In this regard, a prediction that seems to be upward for a 50-day prediction may turn to become downward for 200-days prediction. The emphasis on recent data seems to be

challenging because the real factors affecting the pattern of electricity demand are not necessarily linked to recent data but to seasonal behavior of people, rain and dry seasons, and many more factors; the prediction may be more accurate depending on vast data.

On the other hand, Holt-Winters (HW) methodologies are used to cater for seasonal changes in predicting electricity demand. There is great merit to this technique since electricity consumption for typical communities follows regular patterns based on seasons. In actual fact seasons determine the type and form of activities that communities carry and at the same time their capacity to generate electricity locally with renewables which subsequently reduce their reliance on grids. The Holt-Winter's method is one of several exponential smoothing methods that have the power to directly analyse time series data. Several academic papers used the Holt-Winters method in the past to predict electricity demands [6, 11, 12]. Holt-Winters method falls under the category of exponential smoothening problems which are affected by some few factors. The first factor known as lag implies that the prediction often lags behind the actual trends [12, 13]. Also, exponential smoothening may be best suited for forecasts that are short-term rather than seasonal or cyclic.

Moreover, ARIMA, known for Auto-Regressive Integrated Moving Average, is actually a robust method of forecast that determines its future values based on the known past value of the series only. The strength of ARIMA compared to exponential smoothening methods resides in the fact that ARIMA uses the autocorrelation between series of values while the latter targets trends and seasonality. According to [14], ARIMA models are mainly "backward looking" meaning their prediction lack accuracy at a turning point. ARIMA has been used intensively to predict electricity demand by the authors of [6, 11, 15, 16].

Furthermore, Artificial Neural Network (ANN) has recently emerged as a reliable model for electricity demand prediction. According to [17], an Adaptive Neural Network is a computational model that is derived from a biological neural network. It is made up of highly interconnected processing elements called neurons that work together to resolve specific problems. ANNs are generally composed of three main elements, namely, neurons, interconnections, and learning rules. Neural networks have the capability to learn from real-life scenarios which may be very complex. ANN can easily handle nonlinear relationships between dependent and independent variables and are best suited for complex information processing. The main disadvantage has to do with the black-box nature of ANN. The black box implies that while approximating a function, there is less information on the internal structure of the ANN and how it approximates the function. ANN is however proving reliable and has gained widespread use in forecasting and optimisation theories. The following papers [4, 18–21] adopted ANN in forecasting electricity demand in various location and for varying durations.

Likewise, support vector regression is another family of a neural network used for classification and forecasting of time series data with considerable accuracy. They are provided with learning algorithms, capable of handling linear and nonlinear problems especially with the recent introduction of the loss function of Vapbik's ε-intensity [22]. The authors of [23–27] used SVR models to predict several relevant variables using time series forecasting. Similarly, Dong et al. [28] combined SVR with a chaotic cuckoo search model. The Chaotic Cuckoo Search model is expected to improve on the capabilities of the original Cuckoo Search algorithm by diversifying the population and avoiding local optima. The objective of SVR is to find a function which may be defined for instance as

$$f(x) = w^T \varphi(x) + b, \tag{1}$$

where $f(x)$ denotes the forecasting values; w and b are adjustable coefficients and $\varphi(x)$ represents the mapping function that maps the training data into a high-dimensional feature space [22]. The coefficient is estimated by using any appropriate optimization method like the Lagrange multipliers or by minimizing some empirical risk functions.

Hybridising SVR with other metaheuristic optimization techniques may lead to improved accuracy and results in predicting electricity demand. In this regard, Fan et al. [29] presented a hybridised model of SVR with an empirical mode decomposition (EMD) method and autoregression (AR) which was applied to data collected from New South Wales in Australia. The result of this hybridisation was mainly improved accuracy and interpretability. Chen et al. [30] performed a similar hybridisation using least square SVR, Fuzzy time series, and global harmony search algorithm. The resulting accuracy shows superiority in terms of predictive accuracy over existing methods such as ARIMA, Genetic Algorithm, and simple least square SVR. A similar hybrid system was developed by Hong et al. [31] who combined SVR with a Chaotic Immune Algorithm and demonstrated that the new system yielded better accuracy than the ARIMA and other forms of a hybrid system using SVR. Zhang et al. [32] further combined SVR with empirical mode decomposition (EMD) and the krill herd (KH) algorithm. His hybrid system also demonstrated superiority over the existing methods in terms of accuracy of prediction.

SVR has many advantages that make it attractive especially when combined with other heuristic optimization techniques like simple Fuzzy logic, GA, EMD, and KH. Some of these advantages include effectiveness with high-dimensional spaces and memory efficiency [33]. On the other hand, SVR disadvantages include its computational complexity irrespective of the dimensionality of its input space [34]. SVR algorithms are also not suitable for large datasets especially in the presence of more noise.

Recently there has been the development of an advanced form of ANN among which the combination of fuzzy logic and artificial neural network known as Adaptive Neuro-Fuzzy Inference Systems (ANFIS). While Fuzzy logic on its own is considered as a reliable method, its combination with ANN gives some advantages that are highly enviable in forecasting. ANFIS is a type of ANN that is derived from the Takagi–Sugeno fuzzy system [35–37]. It combines the benefits of fuzzy systems and ANN in a single framework, and it is perceived for this reason as the universal estimator.

While it is believed that the prediction capabilities of ANFIS surpass all the previously listed methods in the literature, it is also remarkable to highlight the fact that there is no prevalence of previous studies on electricity load forecasting that adopted ANFIS. This is a claim of novelty in this study. ANFIS is therefore used in this study to predict electricity demands with more accuracy and reliability and also minimise considerably the lag problems encountered with the exponential and moving average methods.

3. Methodology

This study makes use of data on electricity demands collected from Mallam, Achimota, and Ga East, all towns in the Greater Accra Region of Ghana to predict future electricity need using ANFIS. The data was collected from the Electricity Company of Ghana that is the sole supplier of electricity to residents and companies in the Mallam area. The dataset used to train the model and to test for prediction comprises monthly electricity demand from the region for the period of June 2013 to February 2018. In total, there were 1200 samples in the dataset.

3.1. ANFIS Architecture. Consider a fuzzy inference system having two inputs x and y for simplicity and one output z. Based on a first-order Sugeno fuzzy model, the rule set using the "if-then rules" is depicted below:

Rule 1: If x is A_1 and y is B_1, then $f_1 = p_1 x + q_1 y + r_1$,

Rule 2: If x is A_2 and y is B_2, then $f_2 = p_2 x + q_2 y + r_2$.

$$(2)$$

The reasoning of the Sugeno model and its corresponding ANFIS architecture is illustrated in Figures 1(a) and 1(b). In the diagram, the output of the ith node on layer l is denoted $O_{l,i}$.

Layer 1: Every node belonging to this layer is designated as an adaptive node with an expression below:

$$O_{1,i} = \mu_{A_i}(x), \quad \text{for } i = 1, 2,$$
$$\text{or } O_{1,i} = \mu_{B_{i-2}}(y), \quad \text{for } i = 3, 4,$$

$$(3)$$

where x (or y) is the input to node i and A_i, B_{i-2} is a label associated with the node. In other terms, $O_{1,i}$ is the membership grade of a fuzzy set $A (A_1, A_2, B_1, B_2)$ and it specifies the degree to which the input x (or y) satisfies the quantifier A. The membership function A can be any parametrized membership function like the Bell function defined as follows:

$$\mu_A(x) = \frac{1}{1 + \left| x - c_i/a_i \right|^{2b_i}},$$

$$(4)$$

where a_i, b_i, c_i are the parameter sets. Variation in the values of these parameters affects the shape of the Bell function, thus exhibiting varying forms of membership functions.

Layer 2: The output of nodes in this layer is the product of all the incoming signals:

$$O_{2,i} = w_i = \mu_{A_i}(x)\mu_{B_{i-2}}(y), \quad i = 1, 2.$$

$$(5)$$

Each node output represents the firing strength of a rule.

Layer 3: At this level, the ith node calculates the ratio of the ith rule's firing strength to the sum of all rules' firing strengths:

$$O_{2,i} = \overline{w_i} = \frac{w_i}{w_1 + w_2}, \quad i = 1, 2.$$

$$(6)$$

Output of layer 3 are called normalized firing strengths.

Layer 4: All nodes in this layer are called adaptive node with a node function defined as follows:

$$O_{4,i} = \overline{w_i} f_i = \overline{w_i} (p_i x + q_i y + r_i),$$

$$(7)$$

where $\overline{w_i}$ represents a normalized firing strength from layer 3 and p_i, q_i, r_i, the parameter set of this node. Parameters in this layer are called consequent parameters.

Layer 5: The single node in this layer computes the overall output as the summation of all incoming signals:

$$\text{overall output} = O_{5,1} = \sum_i \overline{w_i} f_i = \frac{\sum_i w_i f_i}{\sum_i w_i}.$$

$$(8)$$

The five layers described above constitutes an adaptive network that is functionally equivalent to a Sugeno fuzzy model. The structure can be altered by combining for instance layers 3 and 4 to obtain an equivalent network with only four layers. The Sugeno model can further be extended to Tsukamoto ANFIS model. Furthermore, the Mamdani model can be derived from the first two models by using some discrete approximations and is finally far more complicated than the first two.

3.2. Application of ANFIS to the Data with Matlab. To use the ANFIS algorithm to train and then consequently use it for prediction, there are four main important steps to complete as illustrated in Figure 2. These steps are discussed in the following points.

3.2.1. Constructing the Learning Contextual Elements. This is the point where the various contextual elements are formulated. The contextual elements are basically the factors that influence electricity demand. For electricity demand generally, the following factors are considered: economic growth, housing and population growth, income growth, and weather. Particularly for this paper, we considered the history of electricity demand that covers all the other factors listed. Historical data on these factors will help construct our feature set for a good training of the model.
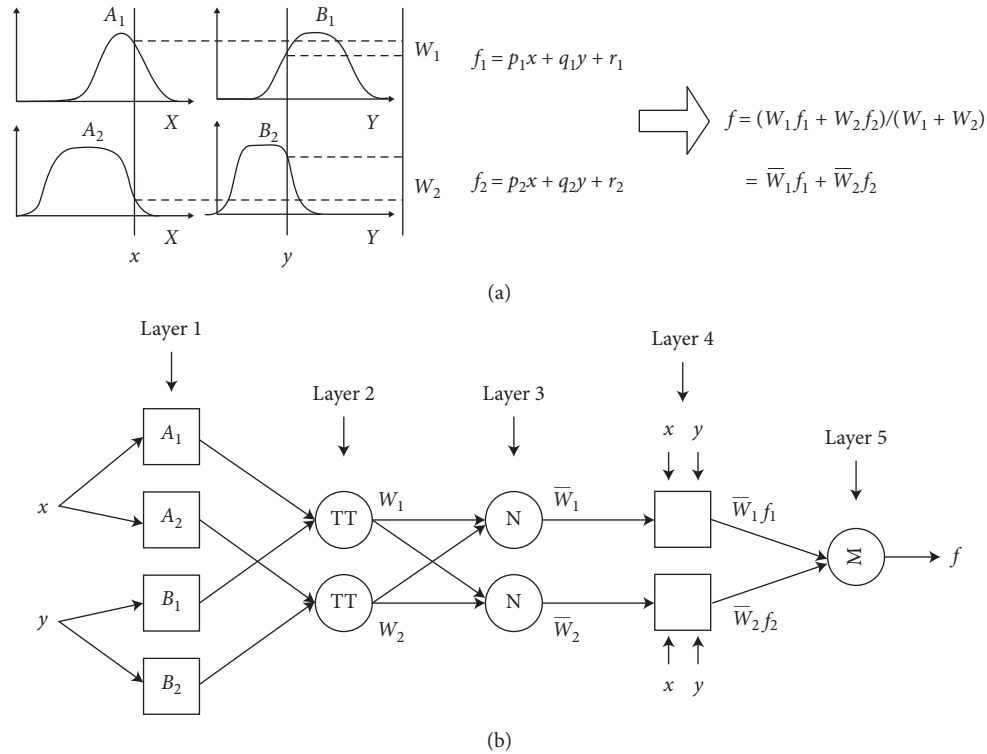
(a)

(b)

FIGURE 1: (a) A two-input first-order Sugeno Fuzzy model with two rules. (b) Equivalent ANFIS structure.
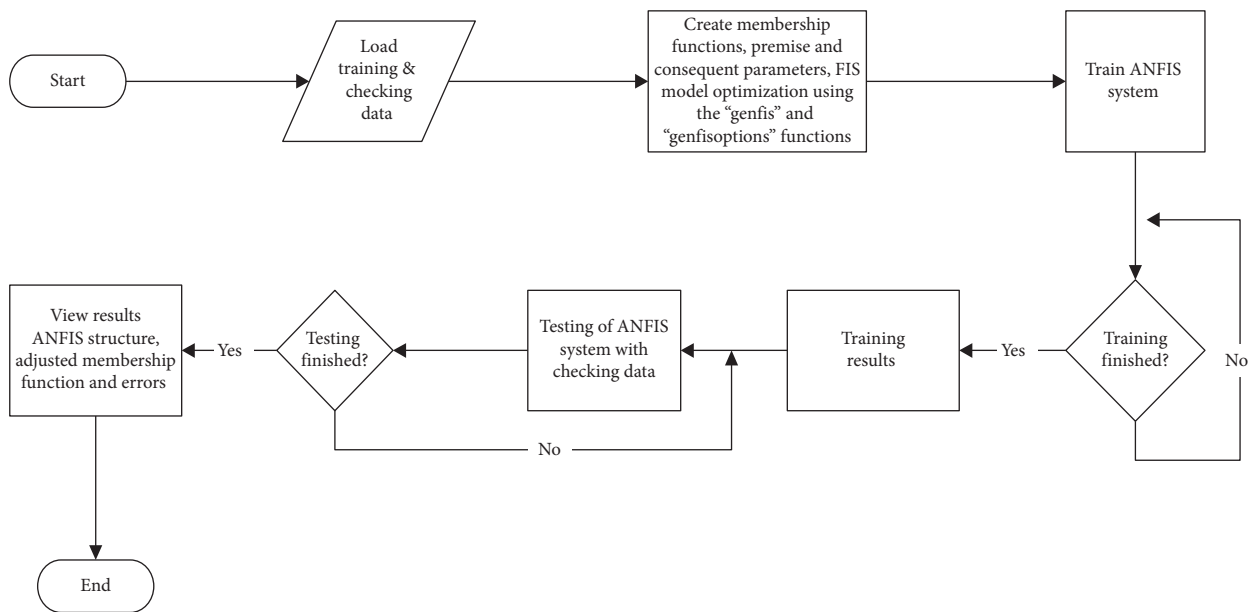

FIGURE 2: Flow diagram of training and testing of ANFIS system.

3.2.2. Input Selection for the System. After constructing the learning contextual elements, there is the need for the selection of the input for the system. The objective in selecting inputs is to eliminate irrelevant inputs or input that does not have a significant influence on the output and consequently reduce the time required to construct the model.

3.2.3. ANFIS System Training Process. After data gathering and selection of the input parameters and contextual parameters, the ANFIS training system process begins by loading the training and checking dataset, which forms the two vectors used in the training of the ANFIS system. The data is loaded into the FIS model using the "genfis" function, to generate the initial membership functions of the premise

parameters from the training data. After the initial membership functions are created, the "anfis" function is used to train the ANFIS system. The training data and checking data are fed to the "anfis" function as a matrix. The membership function created is also fed to the function to be used in the training of the data with the use of the "anfisOptions" function. The same membership function is used to supply the checking data for validation and checking. The "anfisOptions" is also used to provide the number of training epochs for the ANFIS system. After training is complete, the "evalfis" function is used to evaluate the performance of the system and provide the final output of the ANFIS system.

3.3. Prediction. At this point, the trained model is then used to make predictions for the future.

The training of the model and execution of the prediction in MATLAB software for the simulation for this paper was done with the following procedures:

(1) The dataset was loaded and then plotted to obtain a graphical representation of the data. The dataset was then divided into two parts: the training data and the checking data which contained each 500 data values. The first 500 data values were used for training and the last 500 data values after the training data were used for testing and validation. The data division in MATLAB was performed with the two instructions below:

$$
\begin{aligned}
\text{trnData} &= \text{Data}(1{:}500, {:}); \\
\text{chkData} &= \text{Data}(501{:}\text{end}, {:}).
\end{aligned}
\tag{9}
$$

(2) For time series prediction, a mapping from D sample data points sampled every Δ units in time $x(t - (D - 1)\Delta, \ldots, x(t - \Delta), x(t))$ is created to predict a future value $x(t + P)$. This was how the various input vectors and output vectors deduced as will be seen in the scenarios below.

Subsequently, a number of scenarios have been considered to cover many reasonable assumptions under which the simulations were conducted.

Scenario 1. For the first scenario, four input sampled data units were used as input training data for the ANFIS algorithm. It was a four-column vector of the form $w_1(t) = [x(t - 30), x(t - 20), x(t - 10), x(t)]$ and the output training data for this scenario was $x(t+10)$ and this is the future value from the above past values that we wish to predict in this scenario. The GENFIS was used to generate the initial FIS structure to train our ANFIS algorithm. The default values of the GENFIS function were used; therefore, a FIS structure with 2 member functions and two generalized bell membership functions on each of the four inputs was generated to train our ANFIS algorithm. This structure is supplied to the ANFIS algorithm through the "anfisOption" function. The check data was supplied to the ANFIS algorithm to perform validation. The ANFIS algorithm was trained with 10 epochs after which there was convergence.

Scenario 2. Five column vector of the form $w_2(t) = [x(t - 40), x(t - 30), x(t - 20), x(t - 10), x(t)]$ and output $x(t + 10)$ was used. All other parameters were the same as in Scenario 1. The objective was to increase the history of past data which may subsequently affect the accuracy of the prediction.

Scenario 3. For this scenario, Scenario 1 was revisited, but this time around, the number membership of function for the FIS structure to be created was increased from 2 to 3. But all other parameters remained the same. The aim was to see the effect of the number of membership functions on the prediction.

4. Results and Interpretations

Figure 3 shows a graphical representation of the entire dataset. This gives a pictorial representation of the electricity demand from 2013 to 2018.

4.1. Scenario 1. Membership functions are used to graphically represent a fuzzy set. Usually, each fuzzy set is mapped to a value between 0 and 1. Figure 4 shows the plot of the initial membership function generated by the "genfis" function for each of the four inputs.

The height of the graph indicates the maximum value of the membership function. Membership function gives a pictorial view of the value of each fuzzy set. A fuzzy set of height less than 1 is considered a subnormal fuzzy set and that with a height of 1 is a normal fuzzy set. This membership function is the fuzzy rule set that forms part of the initial FIS structure used to start training the ANFIS algorithm.

Figure 5 shows the plot of the final membership function that is derived at the end of the training. It contains the final fuzzy set rules for the ANFIS algorithm. The final membership function exhibited the least minimum checking error because the validation data option was properly set with enough data for training the ANFIS algorithm.

Figure 6 shows the prediction of data done by the ANFIS algorithm in comparison with the actual data. It also shows the plot of the error in the prediction done by the ANFIS algorithm. The actual data is plotted in blue while the predicted data is in orange as indicated by the legend. The errors are really small as illustrated by the second graph on errors. Some peak values were however obtained towards a time limit of 1000 s. These may be due to some aberrant data that probably ought to be corrected with some more cleaning of data.

4.2. Scenario 2. After increasing the record of past data from 4 to 5 thereby creating a fifth input vector as a feature of the fuzzy set, the following results illustrated in Figure 6 were obtained.

Increasing the input vector by 1 doubled the number of fuzzy rules generated by the model which also had an impact on the speed and accuracy of the prediction. There is a minor difference between the predictions in Figure 6 (Scenario 1) and Figure 7 (Scenario 2).
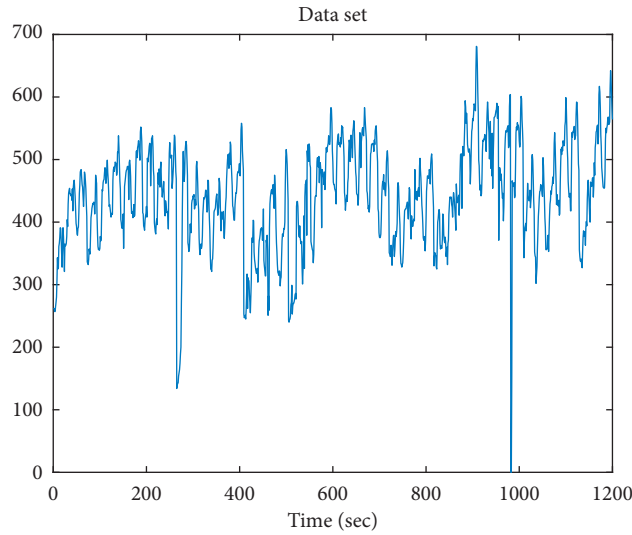
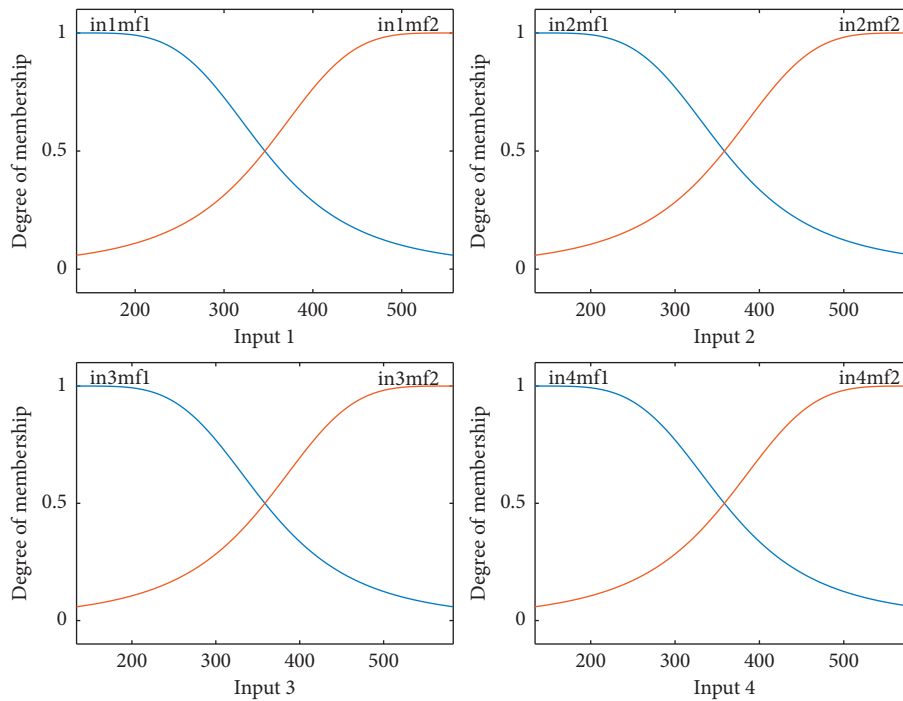FIGURE 3: Graphical representation of the dataset.

FIGURE 4: Initial membership function (Scenario 1).

4.3. *Scenario 3.* Figures 8–10 show the results obtained after increasing the number of membership functions for the same parameters of Scenario 1. The default number of the membership function is usually 2 which was applied for Scenario 1. For this test, the number of membership functions was increased to 3.

Increasing the number of member function by 1 has considerably affected the number of fuzzy rules generated to be used in the prediction, an increase of about triple the initial number of rules. The number of rules subsequently affects the speed and accuracy of the prediction as it can be

observed that there is a slight difference between the predictions in Figure 6 (Scenario 1) and Figure 10 (Scenario 3).

4.4. *Comparative Analysis with Other Methods of Electricity Demand Prediction.* The prediction is repeated in Matlab software with the same input data using the SVR approach, the Least Square SVM, and the ARIMA model. The SVR model was applied in Matlab with the help of some special functions included in the LIBSVM library which provides a combination of functions that were used to apply the State
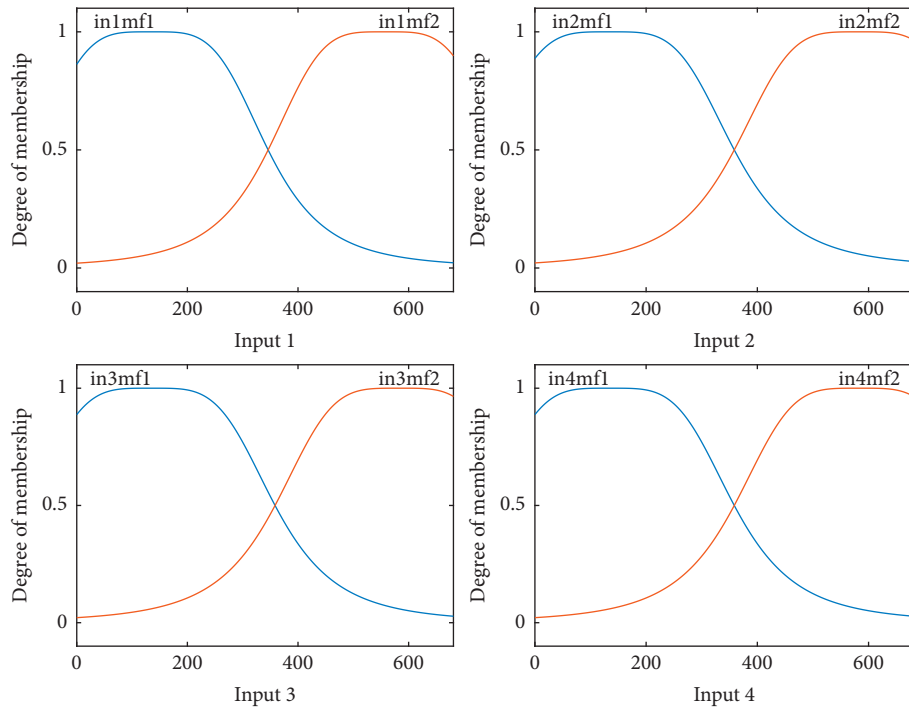
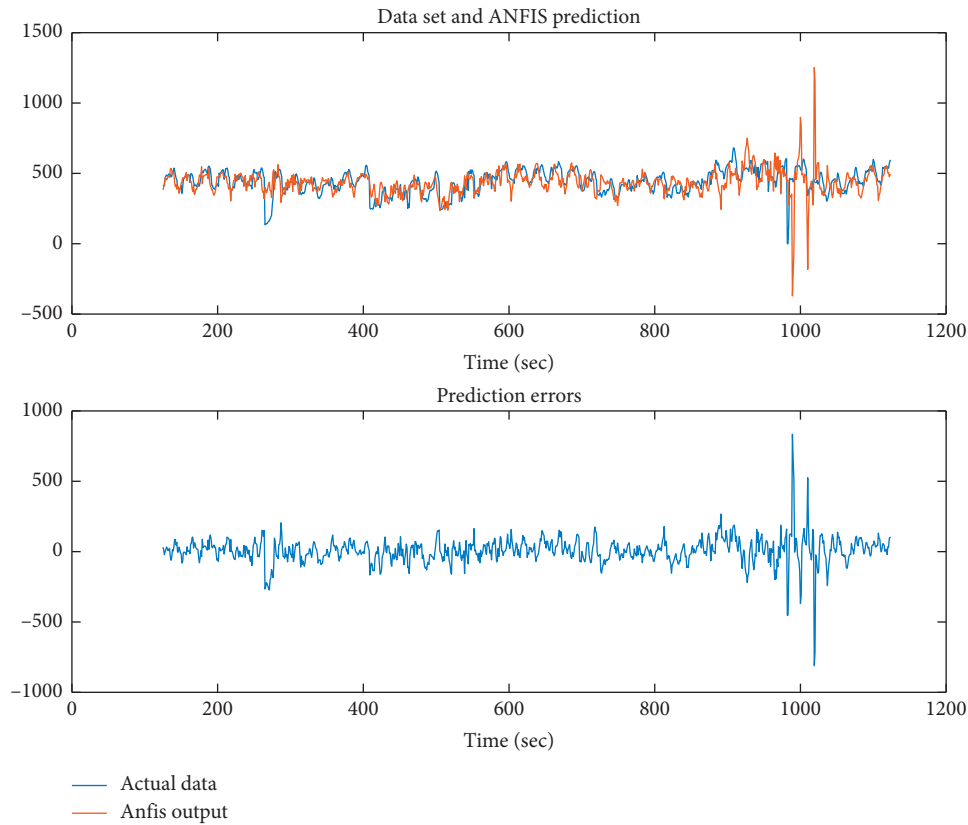FIGURE 5: Final membership function (Scenario 1).

FIGURE 6: Dataset vs. ANFIS output and prediction error (Scenario 1).

Vector Regression approach on the same dataset. Some of the useful functions used in developing the code included the FITCSVM function previously known as SVMTRAIN which fits a classification Support Vector Machine and the SVMPREDICT function which classifies observations using support vector machine (SVM) classifier.
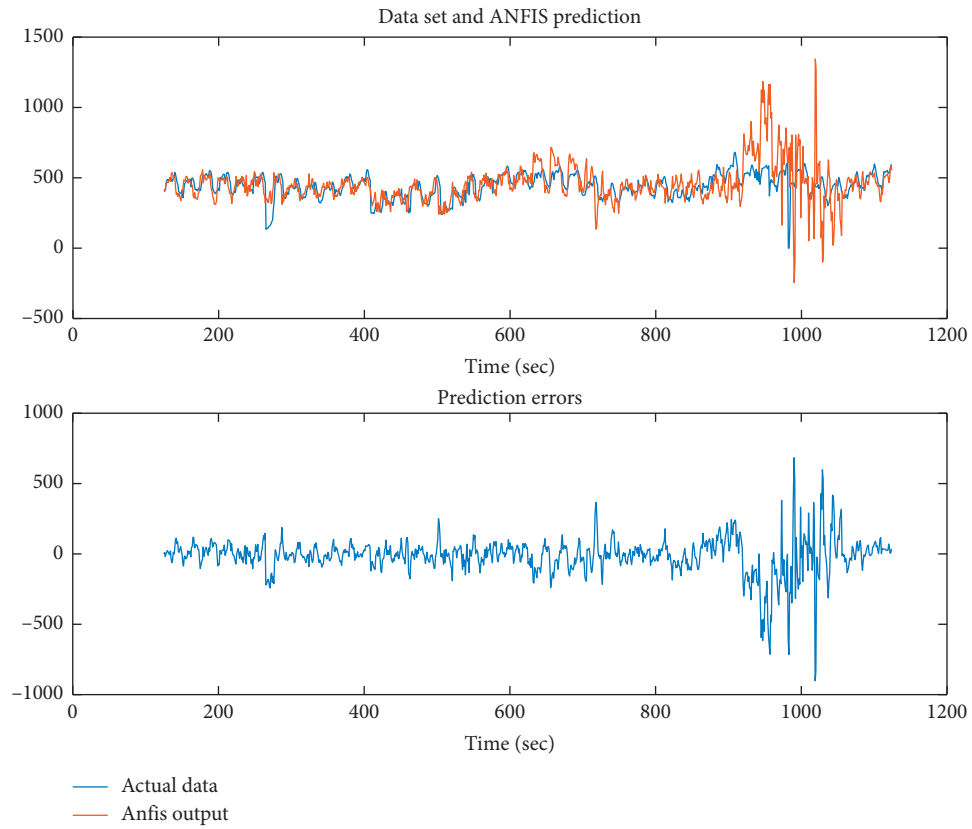
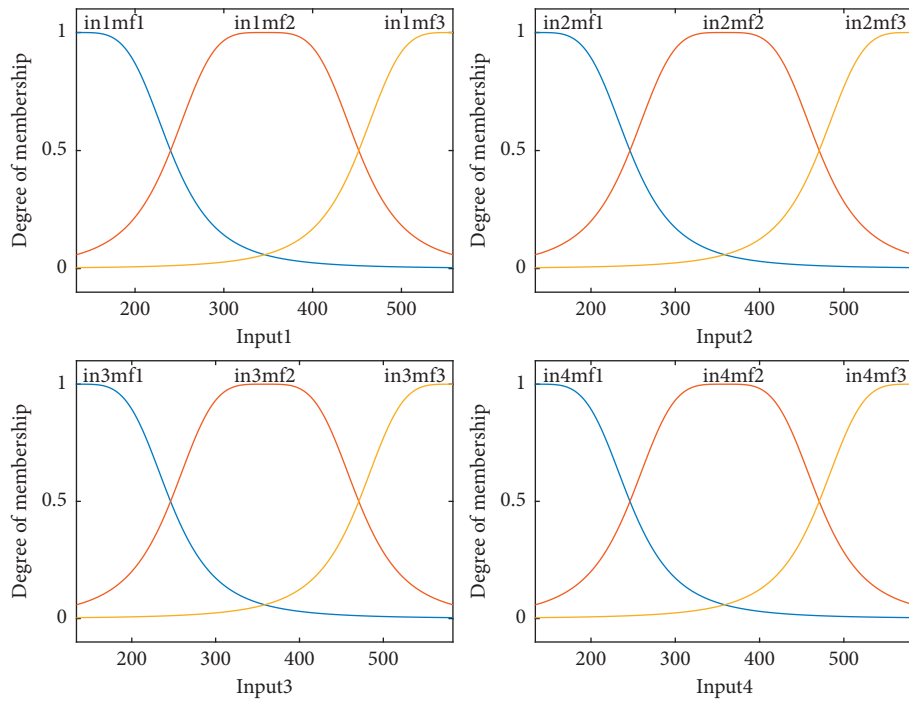FIGURE 7: Actual data vs. ANFIS prediction and prediction error (Scenario 2).

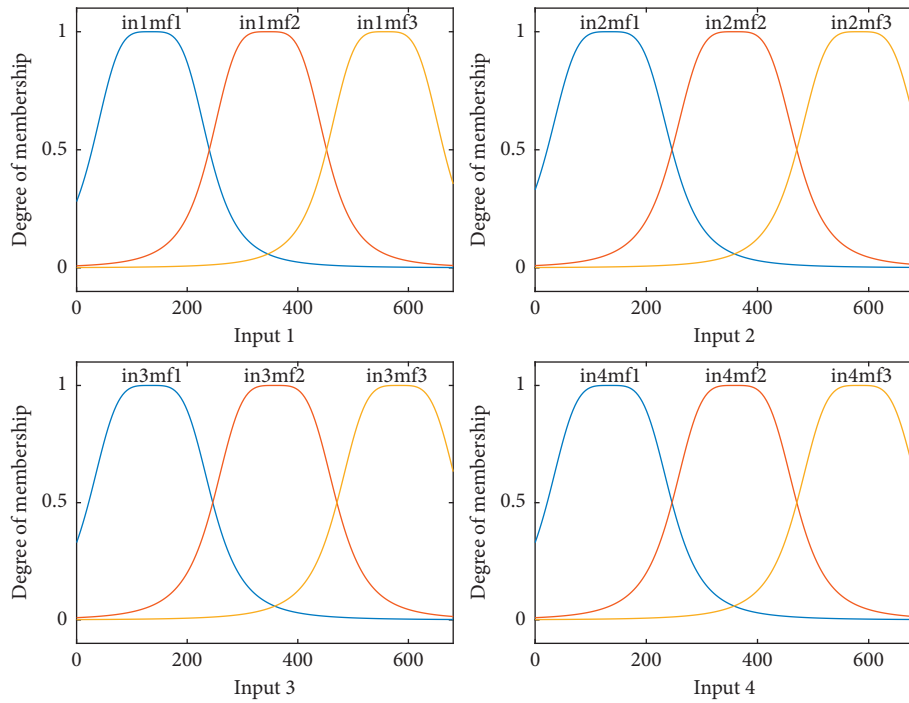FIGURE 8: Initial membership function (Scenario 3).

FIGURE 9: Final membership function (Scenario 3).
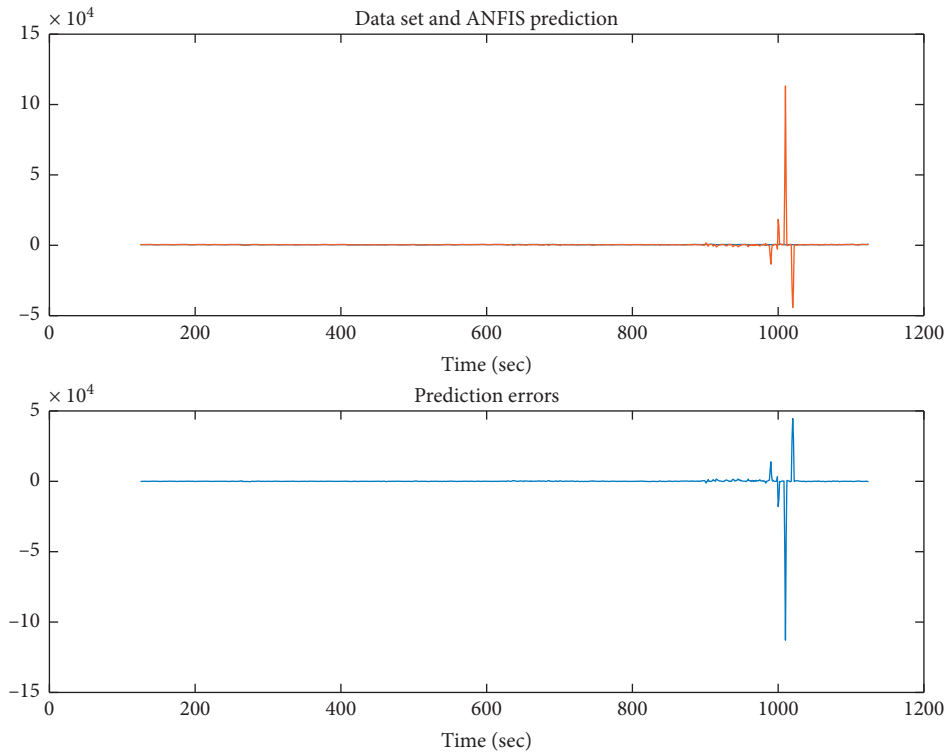
FIGURE 10: Actual data vrs. ANFIS prediction and prediction error (Scenario 3).

Furthermore, the Least Square Support Vector Machine toolbox in Matlab was used to model the same data and perform the prediction with Matlab. Additionally, the ARIMA model was established and tested in Matlab using the estimate function which helped to evaluate the parameters of the ARIMA model. For each of the model tested, the Root Mean Square Error at the end of each epoch were evaluated and are summarized in Table 1. Figures 11 and 12 give a histogram plot of the RMSE values of the different scenarios.

TABLE 1: Root mean square values for each epoch considering all the scenarios.

| Epoch number	ANFIS Scenario 1		ANFIS Scenario 2		ANFIS Scenario 3		SVR		LS-SVR		ARIMA	
	RMSE values for training	RMSE values for checking	RMSE values for training	RMSE values for checking	RMSE values for training	RMSE values for checking	RMSE values for training	RMSE values for checking	RMSE values for training	RMSE values for checking	RMSE values for training	RMSE values for checking
1	57.2167	112.558	46.4971	103.663	34.3291	6537.14	66.4006	170.887	65.6463	102.701	70.3049	148.376
2	57.2117	112.631	46.4912	102.59	34.3163	6478.66	65.3564	170.896	65.4131	109.136	70.3043	148.409
3	57.2067	112.705	46.4853	101.52	34.3409	6517.89	64.8084	170.909	65.2659	116.989	70.3038	148.432
4	57.2017	112.779	46.4795	105.151	34.3347	6467.86	64.6973	170.922	65.098	117.018	70.3032	148.489
5	57.1968	112.853	46.4736	105.868	34.3302	6480.19	64.7074	170.934	64.6657	120.457	70.3027	148.513
6	57.1919	113.928	46.4677	111.795	34.3323	6557.93	64.5821	170.948	64.2683	121.738	70.3022	148.493
7	57.1864	113.01	46.4613	117.977	34.3842	6594.07	64.6992	170.96	63.6042	119.821	70.3016	148.513
8	57.181	113.093	46.4549	119.558	34.3316	6551.48	64.5605	170.977	63.1597	122.85	70.301	148.559
9	57.1751	113.176	46.4485	131.342	34.3406	6524.04	64.5173	170.992	63.0961	123.919	70.3004	148.596
10	57.1703	113.259	46.4421	132.514	34.3338	6552.78	64.4919	171.008	63.2443	84.4919	70.2998	148.62
Average	57.19383	112.9992	46.47012	113.1978	34.33737	6526.204	64.88211	170.9433	64.34616	113.91209	70.30239	148.5
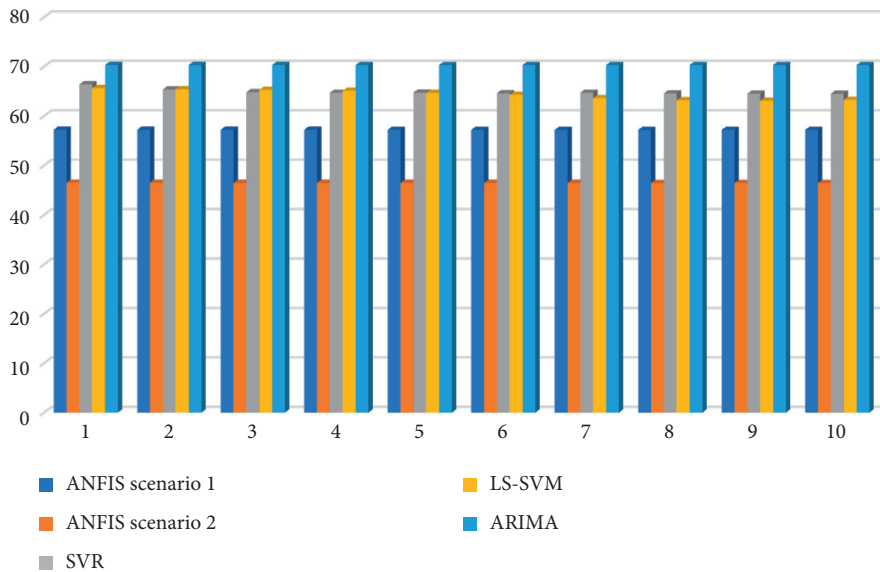
FIGURE 11: Comparative result of RMSE for training for different methods.

5. Discussion

Figures 6, 7, and 10 illustrate the forecast done by the ANFIS algorithm after training and predicting the electricity demand. After training, the ANFIS algorithm makes a prediction for the entire 1000 hours period which was compared with the actual data that is plotted in blue. The algorithm is only trained with 500 hours set of data but was able to make predictions beyond the 500-hour mark, and this is cross-checked with the checking dataset for the validity of the prediction. Additionally, the data on the prediction error gives a better understanding of the accuracy of the proposed ANFIS algorithm in the prediction of electricity demand. From the results, it can be observed that the prediction errors are usually not huge, and depending on the dataset used, increasing the membership function to 3 resulted in overfitting. The fuzzy rules define the characteristics of the clustering type that was used to define the membership function. For each of the scenarios, the default clustering type which is the "GridPartition" was used. For this type of clustering, the system generated input membership functions by uniformly partitioning the input variable range and then creating a single-output Sugeno fuzzy system. This is the reason why the membership function plots illustrated in Figures 3, 4, 7, and 8 were all uniformly spaced and covered the entire input space. Scenario 1–3 after training, resulted in 16, 32, and 81 fuzzy rules respectively. All these rules have been stored and used in coming up with the predictions. Increasing the input vector by 1 resulted in doubling the number of fuzzy rules generated by the model. Increasing the number of member function by 1 also nearly tripled the number of fuzzy rules generated to be used in the prediction. Consequently, the number of rules affects the speed and accuracy of the
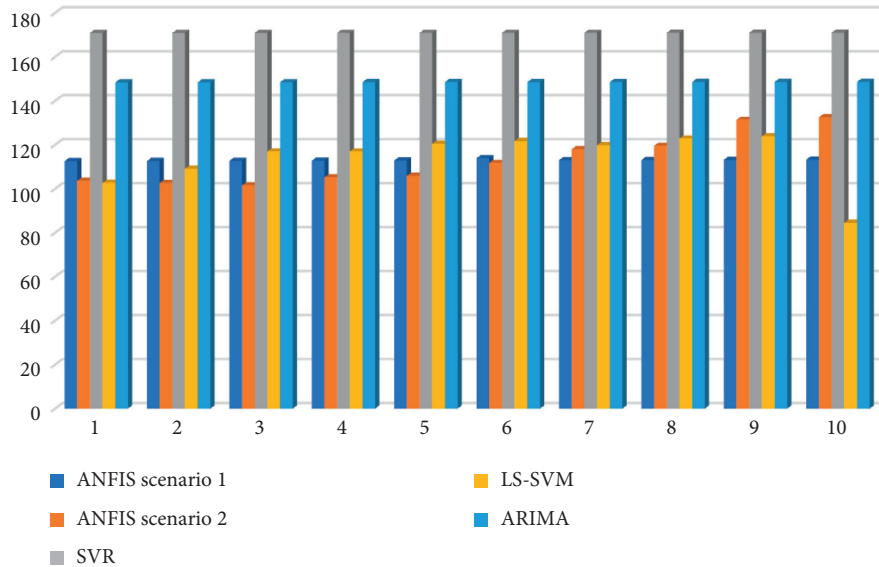
FIGURE 12: Comparative result of RMSE for checking for different methods.

prediction. The ANFIS uses a combination of optimization techniques such as the gradient descent back propagation and mean least squares algorithms to fine-tune the model.

The rules implemented by the ANFIS model helped to improve upon the prediction by reducing the error at the end of each epoch. Table 1 shows the Root Mean Square Error at the end of each epoch for each of the scenarios. From Table 1 above, it can be noticed that, increasing the amount of past data used for learning actually improved upon the speed at which the algorithm converges to an optimal value. Scenario 1 had a minimum training and checking RMSE of 57.17 and 112.55, respectively. Scenario 2 had a minimum training and checking RMSE values of 46.44 and 101.52, respectively.

Moreover, RMSE values keep decreasing and increasing around a specific range of values as illustrated in Table 1 for the first three scenarios relating to the ANFIS model. This implies that, for the current dataset, increasing the number of membership functions is not a good decision to improve upon the performance of the model and this led in actual fact to very high values of errors in the prediction. Subsequently, the smoothness of the data used for training is very important in ANFIS modeling. The data determines what input membership type, number of input member functions, and other decisions to consider when producing an efficient and effective system for prediction.

It can be observed that with the exception of the 3-membership function which resulted in data overfitting and subsequently led to some high RMSE values, the SVR could not perform better than the ANFIS in the first two scenarios. The scenario with three (3) membership functions was worse for this dataset hence, its poor performance in the prediction. Comparing the RMSE of the additional two methods, namely, LS-SVM and ARIMA, as illustrated in Figure 11 for the training data, it can be observed that the ANFIS scenarios one and two overperformed all the other methods by indicating lower values of RMSE and therefore led to improved accuracy.

Comparing the average RMSE over the ten epochs sampled in Table 1, it can be observed that the ANFIS Scenario 1 and 2 have training RMSE values of 57.19 and 46.47 compared to the RMSE values of SVR, LS-SVM, and ARIMA that are respectively 64.88, 64.34, and 70.30. It is then obvious that the well-set two scenarios based on ANFIS are more accurate than the counterpart methods listed above.

Similarly, a comparison of the checking RMSE shows a similar trend with Scenarios 1 and 2 having the least averages as compared to the other methods. However, a critical look at Figure 12 shows that the LS-SVM is very close in performance to the ANFIS method and has outperformed them at the 10th epoch. In average, Scenarios 1 and 2 have recorded checking RMSE values of 112.99 and 113.20. SVR, LS-SVM, and ARIMA recorded the following averages for checking RMSE, respectively, 170.94, 113.91, and 148.5. Therefore, the ANFIS methods used in Scenarios 1 and 2 have the minimum RMSE values for checking followed by the LS-SVM.

This shows that the ANFIS algorithms were mostly more accurate in the prediction of the checking data when compared with the SVR, LS-SVM, and ARIMA. However, Scenario 3 of the ANFIS algorithm due to its high overfitting problem did not perform quite well in the prediction of error. In conclusion, the ANFIS algorithm given the right parameters performs with high accuracy. Unlike the SVR, ANFIS prediction is even more accurate as the volume of the time series data increases.

6. Conclusion

This paper considered the prediction of electricity demand using the ANFIS method. The study made the following fundamental findings: the speed of convergence of the ANFIS algorithm can be significantly improved with an increase in the initial data used for learning; the prediction

errors are considerably small and depending on the dataset used; increasing the membership function to three resulted in overfitting; and increase in the membership function causes RMSE values to vary uncontrollably. Moreover, the ANFIS result was compared to the other three methods including SVR, LS-SVM, and ARIMA models using the same input series data. The analysis of the outcome shows that the similarities in prediction were high; meanwhile, the ANFIS scenario performed better than all other predictions and therefore predicted the data with more accuracy. It was therefore inferred that the efficiency of an ANFIS prediction model depends considerably on the data quality and the tuning parameters.

In view of these findings, ANFIS can be considered as a reliable and promising method that could challenge already known electricity demand prediction models including moving average, ARIMA, SVR, LS-SVM methods provided the training data and essential parameters are properly set. We, therefore, envisage in the future study a framework to accurately predict optimal parameters for ANFIS algorithm and compare results with other emerging techniques for time series data prediction.

Data Availability

The data were collected from Electricity Company of Ghana and are available upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] Z. Dilaver Zafer and L. C. Hunt, "Industrial electricity demand for Turkey: a structural time series analysis," *Energy Economics,* vol. 33, 2011.

[2] E. Erdogdu, "Electricity demand analysis using cointegration and ARIMA modelling: a case study of Turkey," *Energy Policy*, vol. 35, no. 2, pp. 1129–1146, 2007.

[3] J. F. M. Pessanha and N. Leon, "Forecasting long-term electricity demand in the residential sector," *Procedia Computer Science*, vol. 55, pp. 529–538, 2015.

[4] D. Keles, J. Scelle, F. Paraschiv, and W. Fichtner, "Extended forecast methods for day-ahead electricity spot prices applying artificial neural networks," *Applied. Energy*, vol. 162, pp. 218–230, 2016.

[5] Z. Hu, J. Ma, L. Yang, L. Yao, and M. Pang, "Monthly electricity demand forecasting using empirical mode decomposition-based state space model," *Energy & Environment*, vol. 30, no. 7, pp. 1236–1254, 2019.

[6] Y. W. Lee, K. G. Tay, and Y. Y. Choy, "Forecasting electricity consumption using time series model," *International Journal Engineering & Technology*, vol. 7, no. 4.30, p. 218, 2018.

[7] I. Ghalehkhondabi, E. Ardjmand, G. R. Weckman, and W. A. Young, "An overview of energy demand forecasting methods published in 2005-2015," *Energy Systems*, vol. 8, no. 2, pp. 411–447, 2017.

[8] J. W. Taylor and P. E. McSharry, "Short-term load forecasting methods: an evaluation based on European data," *IEEE Transactions on Power Systems*, vol. 22, no. 4, pp. 2213–2219, 2007.

[9] A. Eshragh, B. Ganim, and T. Perkins, *The Importance of Environmental Factors in Forecasting Australian Power Demand*, pp. 1–15, Cornell University, Ithaca, NY, USA, 2019.

[10] S. A. A. Karim and S. A. Alwi, "Electricity load forecasting in UTP using moving averages and exponential smoothing techniques," *Applied Mathematical Sciences*, vol. 7, no. 77–80, pp. 4003–4014, 2013.

[11] I. Shah, H. Iftikhar, S. Ali, and D. Wang, "Short-term electricity demand forecasting using components estimation technique," *Energies*, vol. 12, no. 13, pp. 1–17, 2019.

[12] W. Jiang, X. Wu, Y. Gong, W. Yu, and X. Zhong, "Holt-Winters smoothing enhanced by fruit fly optimization algorithm to forecast monthly electricity consumption," *Energy*, vol. 193, p. 116779, 2020.

[13] N. Regoli, "Advantages and disadvantages of exponential smoothing," 2020, https://connectusfund.org/5-advantages-and-disadvantages-of-exponential-smoothing.

[14] J. Dai and S. Chen, "The application of ARIMA model in forecasting population data," *Journals of Physics:Conference Series*, vol. 1324, no. 1, 2019.

[15] S. Katara, A. Faisal, and G. M. Engmann, "A time series analysis of electricity demand in tamale, Ghana," *International Statistics and Applied Mathematics,* vol. 4, no. 6, pp. 269–275, 2014.

[16] B. Nepal, M. Yamaha, A. Yokoe, and T. Yamaji, "Electricity load forecasting using clustering and ARIMA model for energy management in buildings," *Japan Architectural Review*, vol. 3, no. 1, pp. 62–76, 2020.

[17] S. Bhadouria and A. Jayant, "Development of ANN models for demand forecasting American," *Journal of Engineering Research,* vol. 6, no. 12, p. 2012, 2017.

[18] P. Mandal, T. Senjyu, N. Urasaki, T. Funabashi, and A. K. Srivastava, "A novel approach to forecast electricity price for PJM using neural network and similar days method," *IEEE Transactions on Power Systems,* vol. 22, no. 4, pp. 2058–2065, 2007.

[19] A. Azadeh, S. F. Ghaderi, and S. Sohrabkhani, "Annual electricity consumption forecasting by neural network in high energy consuming industrial sectors," *Energy Conversion and Management*, vol. 49, no. 8, pp. 2272–2278, 2008.

[20] A. Azadeh, S. F. Ghaderi, S. Tarverdian, and M. Saberi, "Integration of artificial neural networks and genetic algorithm to predict electrical energy consumption," *Applied Mathematics and Computation*, vol. 186, no. 2, pp. 1731–1741, 2007.

[21] B. Dong, Z. Li, S. M. M. Rahman, and R. Vega, "A hybrid model approach for forecasting future residential electricity consumption," *Energy Building*, vol. 117, pp. 341–351, 2016.

[22] Z. Zhang, W. C. Hong, and J. Li, "Electric load forecasting by hybrid self-recurrent support vector regression model with variational mode decomposition and improved cuckoo search algorithm," *IEEE Access*, vol. 8, pp. 14642–14658, 2020.

[23] S. Lahmiri, "Minute-ahead stock price forecasting based on singular spectrum analysis and support vector regression," *Applied Mathematics and Computation*, vol. 320, pp. 444–451, 2018.

[24] J. Wang, R. Hou, C. Wang, and L. Shen, "Improved v -Support vector regression model based on variable selection and brain storm optimization for stock price forecasting," *Applied Soft Computing*, vol. 49, pp. 164–178, 2016.

[25] H. Jiang and Z. Wang, "GMRVVm-SVR model for financial time series forecasting," *Expert Systems with Applications*, vol. 37, no. 12, pp. 7813–7818, 2010.

[26] G. Sermpinis, C. Stasinakis, K. Theofilatos, and A. Karathanasopoulos, "Modeling, forecasting and trading the EUR exchange rates with hybrid rolling genetic algorithms-Support vector regression forecast combinations," *European Journal of Operational Research*, vol. 247, no. 3, pp. 831–846, 2015.

[27] W.-C. Hong, "Electric load forecasting by support vector model," *Applied Mathematical Modelling*, vol. 33, no. 5, pp. 2444–2454, 2009.

[28] Y. Dong, Z. Zhang, and W. C. Hong, "A hybrid seasonal mechanism with a chaotic cuckoo search algorithm with a support vector regression model for electric load forecasting," *Energies*, vol. 11, no. 4, 2018.

[29] G.-F. Fan, S. Qing, H. Wang, W.-C. Hong, and H.-J. Li, "Support vector regression model based on empirical mode decomposition and auto regression for electric load forecasting," *Energies*, vol. 6, no. 4, pp. 1887–1901, 2013.

[30] Y. H. Chen, W. C. Hong, W. Shen, and N. N. Huang, "Electric load forecasting based on a least squares support vector machine with fuzzy time series and global harmony search algorithm," *Energies*, vol. 9, no. 2, pp. 1–13, 2016.

[31] W.-C. Hong, Y. Dong, C.-Y. Lai, L.-Y. Chen, and S.-Y. Wei, "SVR with hybrid chaotic immune algorithm for seasonal load demand forecasting," *Energies*, vol. 4, no. 6, pp. 960–977, 2011.

[32] Z. Zhang, S. Ding, and Y. Sun, "A support vector regression model hybridized with chaotic krill herd algorithm and empirical mode decomposition for regression task," *Neurocomputing*, vol. 410, pp. 185–201, 2020.

[33] Z. Zhang and W.-C. Hong, "Electric load forecasting by complete ensemble empirical mode decomposition adaptive noise and support vector regression with quantum-based dragonfly algorithm," *Nonlinear Dynamics*, vol. 98, no. 2, pp. 1107–1136, 2019.

[34] M. Awad, R. Khanna, M. Awad, and R. Khanna, *Support Vector Regression Efficient Learning Machines*, Apress, New York, NY, USA, 2015.

[35] D. Karaboga and E. Kaya, "Adaptive network based fuzzy inference system (ANFIS) training approaches: a comprehensive survey," *Artificial Intelligence Review*, vol. 52, pp. 2263–2293, 2019.

[36] J.-S. Roger Jang, "01_NeuroFuzzyApproach.pdf." p. 614, 2000.

[37] B. Mrinal, "Adaptive network based fuzzy inference system (ANFIS) as a tool for system identification with special emphasis on training data minimization," vol. 141, Department of Electronics and Communication Engineering, Indian Institute of Technology Guwahati, India, 2008.