ADVANCED MODELS OF ENERGY FORECASTING

EDITED BY: Xun Zhang, Jian Chai, Bo Meng and Lean Yu

PUBLISHED IN: Frontiers in Energy Research

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-83250-681-3
DOI 10.3389/978-2-83250-681-3

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

ADVANCED MODELS OF ENERGY FORECASTING

Topic Editors:
Xun Zhang, Academy of Mathematics and Systems Science, Chinese Academy of Sciences (CAS), China
Jian Chai, Xidian University, China
Bo Meng, Japan External Trade Organization (JETRO), Japan
Lean Yu, Beijing University of Chemical Technology, China

Citation: Zhang, X., Chai, J., Meng, B., Yu, L., eds. (2022). Advanced Models of Energy Forecasting. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-83250-681-3

Table of Contents

ORIGINAL RESEARCH
published: 19 August 2021
doi: 10.3389/fenrg.2021.707937

A New Two-Stage Approach with Boosting and Model Averaging for Interval-Valued Crude Oil Prices Forecasting in Uncertainty Environments

Bai Huang[1], *Yuying Sun*[2,3,4]* and *Shouyang Wang*[2,3,4]

[1]School of Statistics and Mathematics, Central University of Finance & Economics, Beijing, China, [2]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China, [3]Center for Forecasting Science, Chinese Academy of Sciences, Beijing, China, [4]School of Economics and Management, University of Chinese Academy of Sciences, Beijing, China

OPEN ACCESS

Edited by:
Farhad Taghizadeh-Hesary,
Tokai University, Japan

Reviewed by:
Ehsan Rasoulinezhad,
University of Tehran, Iran
Robina Iram,
Jiangsu University, China

Correspondence:
Yuying Sun
sunyuying@amss.ac.cn

Specialty section:
This article was submitted to Sustainable Energy Systems and Policies,
a section of the journal Frontiers in Energy Research

Received: *12 May 2021*
Accepted: *16 July 2021*
Published: *19 August 2021*

Citation:
Huang B, Sun Y and Wang S (2021) A New Two-Stage Approach with Boosting and Model Averaging for Interval-Valued Crude Oil Prices Forecasting in Uncertainty Environments. Front. Energy Res. 9:707937. doi: 10.3389/fenrg.2021.707937

In view of the intrinsic complexity of the oil market, crude oil prices are influenced by numerous factors that make forecasting very difficult. Recognizing this challenge, numerous approaches have been introduced, but little work has been done concerning the interval-valued prices. To capture the underlying characteristics of crude oil price movements, this paper proposes a two-stage forecasting procedure to forecast interval-valued time series, which generalizes point-valued forecasts to incorporate uncertainty and variability. The empirical results show that our proposed approach significantly outperforms all the benchmark models in terms of both forecasting accuracy and robustness analysis. These results can provide references for decision-makers to understand the trends of crude oil prices and improve the efficiency of economic activities.

Keywords: crude oil prices forecasting, forecast combination, interval-valued time series, model averaging, vector L2-boosting

1 INTRODUCTION

As one of the most important commodities, crude oil plays a vital role in various fields. In the past decades, crude oil prices have been extremely volatile (see **Figure 1**). The oil-related industries are highly sensitive to oil price changes (Ebrahim et al., 2014; Taghizadeh-Hesary et al., 2016). Accurate prediction of crude oil prices and the market volatility is valuable for market participants to make risk management plans and investment decisions (Zaabouti et al., 2016; Zhang et al., 2020). The crude oil prices are volatile, and are dependent on many factors such as market trends, sentiments and stock markets. The aforementioned factors make the crude oil prices unstable and makes its prediction complicated and challenging. Thus, we aim to develop a reliable model for crude oil price forecasting.

In recent literatures, most of the existing methods focus on the point-valued crude oil closing prices (Abramson and Finizza, 1995; Zhang et al., 2008; Kilian, 2009; Zhang et al., 2009; Shin et al., 2013; Zhao et al., 2017; Binder et al., 2018; Álvarez-Díaz, 2019). However, the use of closing prices has the disadvantage that it does not take into account the oil price variation information within a given period time, e.g., the midpoint and range of crude oil prices in October 2008 are about \$76.61/bbl and \$36.31/bbl respectively. While the midpoint and range of crude oil prices in November 2009 are around \$77.99/bbl and \$5.42/bbl respectively.

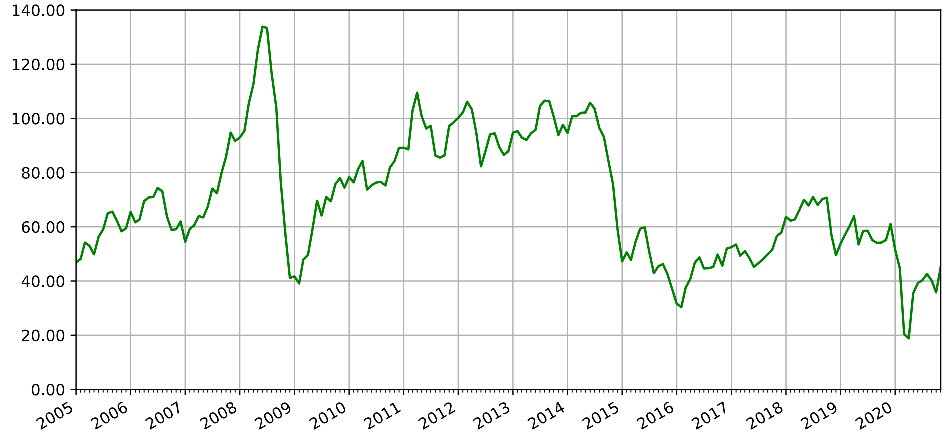


FIGURE 1 | Crude oil West Texas Intermediate (WTI), January 2005–December 2020.

Such forecasts with point-valued crude oil price data have not been particularly successful when compared with the interval-valued time series forecasts (see Sun et al., 2019). What is more, recent studies also provide empirical evidence suggesting that ITS models have achieved great success on improving the forecast accuracy in a wide range of fields such as stock price forecasting (Maia and de Carvalho, 2011; Xiong et al., 2017) and forecasting in energy markets, such as electric power demand (García-Ascanio and Maté, 2010; Hu et al., 2015), and crude oil prices (Yang et al., 2016). By accessing more information (e.g., highs, lows, midpoints, and range), an interval-based method is expected to be superior to the point-based method (Sun et al., 2018). Here, highs and lows are points of inflection for prices. The price range is the difference between two boundaries, which gives the interval length. It can be regarded as a measure of volatility to reflect the price fluctuation. For example, instead of traditional point-based method, Yang et al. (2012) introduce interval dummy variables in the autoregressive conditional interval models. Sun et al. (2019) apply a threshold autoregressive interval-valued model. Qiao et al. (2019) develop an interval-valued factor pricing model. Conclusions from prior studies suggest that interval-valued time series (ITS) models may produce more accurate forecasts.

Therefore, the desirable characteristics of the interval modeling make them ideal candidates for the prediction of crude oil prices. In addition, it is well known that a large set of factors are responsible for changes in the crude oil price, including overall economic conditions, demand and supply, monetary policy, as well as speculative trading (Hamilton, 2008; Yoshino and Taghizadeh-Hesary, 2014). Thus, the number of potential predictors can be very large. In such cases, interval-valued variable selection is considered necessary and becomes the critical step in achieving promising forecasting performances in data-rich environments. On the other hand, in practice, when only some of the variables are selected to include as the predictors in a model, model misspecification is unavoidable, which can worsen the model forecast performance of the model. Therefore, model averaging is considered to take a weighted average of possible combinations of selected interval-valued predictors.

For these reasons, this paper proposes a new two-stage procedure for interval valued crude oil price forecasting based on boosting and model averaging. First, we extend the L_2 boosting method by Buhlmann (2006) to achieve variable selection for the interval model. Several penalized methods have been proposed to achieve variable selection. Examples include the class of Bridge estimators (Frank and Friedman, 1993), where the Lasso-type estimators are included a special case (Knight and Fu, 2000), or the smoothly clipped absolute deviation (SCAD) estimator (Fan and Li, 2001). Instead of these regularized (penalized) methods, Donald et al. (2009) apply information criteria for moment selection, Ng and Bai (2009) develop boosting for variable selection, where variable selection and shrinkage are performed simultaneously to increase prediction accuracy. The proposed vector boosting algorithm can achieve significant dimension reduction when a long list of interval-valued variables is available.

Next, we extend the LsoMA method developed by Liao et al. (2019) to average predictions from interval models with interval-valued exogenous variables to reduce model uncertainty. The idea of model averaging (MA) is first introduced to combine predictions from many forecasting models by Bates and Granger (1969) and has received great interest in econometrics and statistics. Model averaging is an extension of model selection which can substantially reduce the selection bias induced by selecting only one candidate model. Hoeting et al. (1999) provide a comprehensive summary of previous research on Bayesian model averaging (BMA) where models are weighted by the posterior model probabilities. Unlike BMA, frequentist model averaging (FMA) usually select the optimal weighting with the smallest information criteria scores (Buckland et al., 1997; Hjort and Claeskens, 2003; Hjort and Claeskens, 2006; Zhang and Liang, 2011; Zhang et al., 2012; Xu et al., 2014), Mallows model averaging (MMA) by Hansen (2007), jackknife model averaging

(JMA) by Hansen and Racine (2012). Liao and Tsay (2016) extend MMA to the situation of the VAR models.

Univariate and bivariate methods are broadly the two main approaches in the interval modeling literature. In the univariate method, models are presented separately for a pair of attributes of interval variables (e.g., midpoint and range). The two attributes are estimated separately (De Carvalho et al., 2004; Maia et al., 2008), thus only information of one attribute is used in estimating model parameters at a time. Unlike the univariate method, the bivariate method estimates the two attributes simultaneously (e.g., Cheung et al., 2009; He et al., 2010; Lima Neto and De Carvalho, 2010; Arroyo et al., 2011; González-Rivera and Lin, 2013), which is more desirable in ITS forecasting. Therefore, in this paper, in order to consider possible interdependence between midpoint and range, the LsoMA methods are constructed following the bivariate modeling approach to efficiently use the contained information.

This paper proposes a two-stage vector boosting model averaging (2SVBMA) forecasting framework: Stage 1 uses vector L_2 Boosting to select interval-valued variables; Stage 2 uses the leave-subject-out cross-validation model averaging method with exogenous interval-valued variables to average interval-valued predictions. Our procedure combines the merits of these two techniques and can be easily adapted to any new situation. We compare our 2SVBMA method with other competing methods including model selection methods by Akaike information criterion (AIC), Bayesian information criterion (BIC), Hannan-Quinn (HQ), and model averaging methods by smoothed AIC, smoothed BIC (Buckland et al., 1997), smoothed HQ, and MMA in interval model. The empirical results indicate that the 2SVBMA method has better forecasting performance than the commonly used model selection and averaging methods.

Our proposed 2SVBMA forecasting procedure has a few appealing features. First, this approach extends the forecasting success of point-valued data models of crude oil price to interval-valued data models, which is capable of assessing and forecasting the changes in both the trend and volatility of crude oil prices simultaneously due to the informational gain from interval-valued data. Second, our vector boosting method provides a parsimony and feasible solution to the interval-valued variable selection problem for interval models. Third, the extended interval-valued LsoMA model with interval-valued exogenous variables demonstrates the gains in forecast accuracy through forecast combination. By doing so, our approach improves crude oil price forecasting performances significantly.

The remainder of this paper is organized as follows. **Section 2** first proposes 2SVBMA methodology, starts with extended L_2 boosting to interval-valued variable selection and develops the LsoMA with interval-valued model with interval-valued exogenous variables. **Section 3** provides the empirical implementations. **Section 4** discusses the empirical results. **Section 5** concludes.

2 METHODOLOGY

2.1 Model Framework

Let $(\Omega, \mathcal{F}, P)$ be a probability space, where Ω is the set of elementary events, $\mathcal{F}$ is the σ-field of events, and $P\colon \mathcal{F} \to [0,1]$ is the σ-additive probability measure. An interval random variable is defined as a measurable mapping $X\colon \mathcal{F} \to [x_L, x_U] \in \mathbb{R}$, such that for all $x \in [x_L, x_U]$ there is a set $A_X(x) \in \mathcal{F}$, where $A_X(x) = \{w \in \Omega | X(w) = x\}$ with $x \in [x_L, x_U]$ (Arroyo et al., 2011; González-Rivera and Lin, 2013). A stochastic ITS $\{y_t = [y_{L,t}, y_{U,t}]\}_{t=1}^{T}$ can be represented by its midpoint and range, i.e., $y_t = <y_{c,t}, y_{r,t}>$, where $y_{c,t} = \frac{1}{2}(y_{L,t} + y_{U,t})$ and $y_{r,t} = y_{U,t} - y_{L,t}$. Assume that $\{y_t\}$ is stationary and follows a vector autoregressive models with interval-valued exogenous variables:

$$\begin{aligned} \mathbf{y}_t &= \sum_{i=1}^{p} \boldsymbol{\alpha}_i \mathbf{y}_{t-i} + \sum_{j=1}^{q} \boldsymbol{\beta}_j \mathbf{x}_{t-j} + \boldsymbol{\varepsilon}_t \\ &\equiv \boldsymbol{\Pi}' \mathbf{z}_t + \boldsymbol{\varepsilon}_t, t = 1, \ldots, T, \end{aligned} \tag{1}$$

where $\mathbf{y}_t \triangleq (y_{c,t}, y_{r,t})'$, $\mathbf{x}_{t-j} \triangleq (x_{c,t-j}, x_{r,t-j})'$, and $\boldsymbol{\varepsilon}_t = (\varepsilon_{c,t}, \varepsilon_{r,t})'$ is an interval-valued sequence with mean zero and covariance matrix $\mathbb{E}\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t' \equiv \Sigma$, and $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_j$ are the coefficient matrix that satisfies $\sum_{i=1}^{p} \|\boldsymbol{\alpha}_i\| < \infty$ and $\sum_{j=1}^{q} \|\boldsymbol{\beta}_j\| < \infty$, $\mathbf{z}_t = (\mathbf{y}_{t-1}', \ldots, \mathbf{y}_{t-p}', \mathbf{x}_{t-1}', \ldots, \mathbf{x}_{t-q}')'$ is a $2(p+q) \times 1$ vector, $\boldsymbol{\Pi} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q)'$ is a $2(p+q) \times 2$ vector, and the assumed initial data are $\{\mathbf{y}_t\}_{t=-p+1}^{0}$. This data generating process guarantees the natural order of the intervals, i.e., the lower bound is smaller than or equal to the upper bound.

In matrix form, (1) is represented by

$$\mathbf{Y}_c = \mathbf{Z}\boldsymbol{\Pi}_c + \boldsymbol{\varepsilon}_c, \tag{2}$$

and

$$\mathbf{Y}_r = \mathbf{Z}\boldsymbol{\Pi}_r + \boldsymbol{\varepsilon}_r, \tag{3}$$

where $\mathbf{Y}_c = (y_{c,1}, \ldots, y_{c,T})'$, $\mathbf{Y}_r = (y_{r,1}, \ldots, y_{r,T})'$, $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_T)'$, $\boldsymbol{\Pi} \equiv (\boldsymbol{\Pi}_c, \boldsymbol{\Pi}_r)$, $\boldsymbol{\varepsilon}_c = (\varepsilon_{c,1}, \ldots, \varepsilon_{c,T})'$, and $\boldsymbol{\varepsilon}_r = (\varepsilon_{r,1}, \ldots, \varepsilon_{r,T})'$.

The least squares estimators of $\boldsymbol{\Pi}_c$ and $\boldsymbol{\Pi}_r$ are given by

$$\hat{\Pi}_c = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}_c, \tag{4}$$

and

$$\hat{\Pi}_r = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}_r. \tag{5}$$

2.2 First Stage: Vector Boosting

We first extend L_2Boosting regularization method to interval model to select a subset of interval-valued variables. $\mathbf{Z}_k$ is the k^{th} row in $\mathbf{Z}$. They are the potential interval-valued variables that will be selected by vector boosting. $\mathbf{Z}_{k,t}$ is the t^{th} element in $\mathbf{Z}_k$ and $\boldsymbol{\Pi}_k$ is the corresponding k^{th} interval-valued coefficient of $\boldsymbol{\Pi}$, where $k = 1, \ldots, p+q$. Let m denote the m^{th} iteration in the vector boosting procedure, and $\bar{M}$ denote the maximum number of iteration. At each step m, the interval-valued variable $\underset{km}{\hat{\Pi}}$ that is most relevant to the "current interval-valued residual" is selected. Denote $\mathbf{F}_{m,t}$ as the strong learner and $\mathbf{f}_{m,t}$ as the weak learner for $k = 1, \ldots, p+q$. Let $\hat{\varepsilon}_m = (\hat{\varepsilon}_{m,1}, \ldots, \hat{\varepsilon}_{m,T})'$, $\mathbf{f}_m = (\mathbf{f}_{m,1}, \ldots, \mathbf{f}_{m,T})'$ and $\mathbf{F}_m = (\mathbf{F}_{m,1}, \ldots, \mathbf{F}_{m,T})'$.

Vector L_2 Boosting performs an interval-valued variable selection for $\mathbf{Y}$ using the following procedure:

1. When $m = 0$, the initial weak learner for $\mathbf{y}_t$ is

$$\mathbf{F}_{0,t} = \mathbf{f}_{0,t} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{y}_t. \tag{6}$$

2. For each step. $m = 1, \ldots, \bar{M}$
   1) Compute the "current interval-valued residual," $\hat{\varepsilon}_{m,t} = \mathbf{y}_t - \mathbf{F}_{m-1,t}$.
   2) Regress the current interval-valued residual $\hat{\varepsilon}_{m,t} = (\hat{\varepsilon}_{c,m,t}, \hat{\varepsilon}_{r,m,t})'$ on each $\mathbf{Z}_{k,t}$. The estimator $\mathbf{\Pi}_k$ is obtained as

$$\hat{\Pi}_{c,k} = \min_{\mathbf{\Pi}_{c,k}}\sum_{t=1}^{T}\left(\hat{\varepsilon}_{c,m,t} - \mathbf{Z}_{k,t}\mathbf{\Pi}_{c,k}\right)^2, \tag{7}$$

$$\hat{\Pi}_{r,k} = \min_{\mathbf{\Pi}_{r,k}}\sum_{t=1}^{T}\left(\hat{\varepsilon}_{r,m,t} - \mathbf{Z}_{k,t}\mathbf{\Pi}_{r,k}\right)^2. \tag{8}$$

The interval-valued variables that has the minimum sum of squared residuals is picked up, such that

$$k_m = \text{argmin}_{k\in\{1,\ldots,p+q\}}\sum_{t=1}^{T}\left(\hat{\varepsilon}_{m,t} - \mathbf{Z}_{k,t}\hat{\Pi}_k\right)^2. \tag{9}$$

3) The weak learner is

$$\mathbf{f}_{m,t} = \mathbf{Z}_{k_m,t}\hat{\Pi}_{k_m}, \tag{10}$$

where $\mathbf{Z}_{k_m,t}$ is the interval-valued variable that is selected.

4) The strong learner $\mathbf{F}_{m,t}$ is updated as

$$\mathbf{F}_{m,t} = \mathbf{F}_{m-1,t} + c_m\mathbf{f}_{m,t}, \tag{11}$$

with $c_m > 0$, where c_m is a learning rate, which can be seen as a small step size when updating $\mathbf{F}_{m,t}$.

To avoid overfitting, a version of AIC is used to choose the optimal number of iteration M . Define $\mathbf{P}_m = \mathbf{Z}_{k_m}(\mathbf{Z}_{k_m}{}'\mathbf{Z}_{k_m})^{-1}\mathbf{Z}_{k_m}{}'$ to be an $T \times T$ matrix. From Equation **(10)**,

$$\mathbf{Z}_{k_m}\hat{\Pi}_{k_m} = \mathbf{P}_m\hat{\varepsilon}_m\mathbf{f}_m = \mathbf{P}_m(\mathbf{Y} - \mathbf{F}_{m-1}). \tag{12}$$

The strong learner at each step m is

$$\begin{aligned}\mathbf{F}_m &= \mathbf{F}_{m-1} + c_m\mathbf{P}_m(\mathbf{Y} - \mathbf{F}_{m-1})\\ &= \left[\mathbf{I}_{T\times T} - \prod_{a=0}^{m}\left(\mathbf{I}_{T\times T} - c_{k_a}\mathbf{P}_{k_a}\right)\right]\mathbf{Y} =: \mathbf{B}_m\mathbf{Y}.\end{aligned}$$

AIC is given as

$$AIC(m) = \log(\hat{\sigma}_m^2) + \frac{1 + \text{trace}(\mathbf{B}_m)/T}{1 - (\text{trace}(\mathbf{B}_m) + 2)/T}. \tag{13}$$

where $\log(\hat{\sigma}_m^2) = \frac{1}{T}\sum_{t=1}^{T}(\hat{\varepsilon}_m - c_m\mathbf{f}_{m,t})^2$. Then $\hat{M} = \arg\min_{m=1,\ldots,\bar{M}}AIC(m)$.

2.3 Second Stage: LsoMA

After selecting these important exogenous interval-valued variables, LsoMA technique is extended to interval candidate models with interval-valued exogenous variables, which is adopted to reduce model uncertainty and increase forecast accuracy.

Consider S candidate models used to approximate the DGP in **Eq. (1)** with S to be infinite if the sample size is going to infinity. The sth $(1 \le s \le S)$ candidate model is given by

$$\begin{aligned}\mathbf{y}_t &= \sum_{i=1}^{i_s}\boldsymbol{\alpha}_i\mathbf{y}_{t-i} + \sum_{j=1}^{j_s}\boldsymbol{\beta}_j\mathbf{x}_{t,j} + \boldsymbol{\varepsilon}_t,\\ &\equiv \mathbf{z}_t^{(s)'}\mathbf{\Pi}^{(s)} + \boldsymbol{\varepsilon}_t, \quad t = S+1, \ldots, T,\end{aligned}$$

where $\mathbf{z}_t^{(s)} = (\mathbf{y}_{t-1}', \ldots, \mathbf{y}_{t-i_s}', \mathbf{x}_{t,1}', \ldots, \mathbf{x}_{t,j_s}')'$, $\mathbf{\Pi}^{(s)} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{i_s}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{j_s})'$, and $1 \le i_s, j_s \le S$. Then in matrix form, we have

$$\mathbf{Y} = \mathbf{Z}^{(s)}\mathbf{\Pi}^{(s)} + \boldsymbol{\varepsilon},$$

where $\mathbf{Y} = (\mathbf{y}_{S+1}, \ldots, \mathbf{y}_T)'$, $\mathbf{Z}^{(s)} = (\mathbf{z}_{S+1}^{(s)}, \ldots, \mathbf{z}_T^{(s)})'$, and $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_{S+1}, \ldots, \boldsymbol{\varepsilon}_T)'$. For each candidate model, we use multivariate least squares (LS) method to estimate parameters and thus the LS estimator of $\mathbf{\Pi}^{(s)}$ is $\hat{\Pi}^{(s)} = (\mathbf{Z}^{(s)'}\mathbf{Z}^{(s)})^{-1}\mathbf{Z}^{(s)'}\mathbf{Y}$, and the corresponding estimator of conditional mean $\boldsymbol{\mu}$ is $\hat{\mu}^{(s)} = \mathbf{Z}^{(s)}\hat{\Pi}^{(s)}$ in sth candidate model.

Let the weight vector $\mathbf{w} = (w_1, \ldots, w_S)' \in \mathcal{W} = \{w \in [0,1]^S\colon \sum_{s=1}^{S}w_s = 1\}$. Then the model averaging estimator of conditional mean $\boldsymbol{\mu}$ is $\hat{\mu}(\mathbf{w}) = \sum_{s=1}^{S}w_s\hat{\mu}^{(s)}$. To obtain the optimal weights, it is common to minimize the following squared loss function:

$$L(\mathbf{w}) = \|\boldsymbol{\mu} - \boldsymbol{\mu}(\mathbf{w})\|^2. \tag{14}$$

However, this loss is infeasible because of the unknown conditional mean $\boldsymbol{\mu}$. We follow the spirit of Liao et al. (2019) to use the following feasible leave-subject-out cross-validation criterion of choosing weights

$$LsoMA(\mathbf{w}) = \text{trace}\{(\mathbf{Y} - \tilde{\mu}(\mathbf{w}))\mathbf{\Sigma}^{-1}(\mathbf{Y} - \tilde{\mu}(\mathbf{w}))'\}, \tag{15}$$

where $\tilde{\mu}^{(s)} = (\tilde{\mu}_{S+1}^{(s)'}, \ldots, \tilde{\mu}_T^{(s)'})'$, $\tilde{\mu}_{S+t}^{(s)} = \boldsymbol{\psi}_t^{(s)}\tilde{\mu}_{[t]}^{(s)}$, $\boldsymbol{\psi}_t^{(s)}$ is the selected matrix to select observations at time point $S + t$, $\tilde{\mu}_{[t]}^{(s)}$ is the leave-subject-out cross-validation estimator after deleting some observations around $S + t$, and $\tilde{\mu}(\mathbf{w}) = \sum_{s=1}^{S}w_s\tilde{\mu}^{(s)}$; see more discussions in Liao et al. (2019). Minimizing this criterion, we have

$$\hat{\mathbf{w}} = arg\,min_{\mathbf{w}\in\mathcal{W}}LsoMA(\mathbf{w}), \tag{16}$$

and thus the model averaging estimator is $\hat{\mu}(\hat{\mathbf{w}})$. As Liao et al. (2019) proved, the weight obtained by minimizing the feasible cross-validation criterion $LsoMA(\mathbf{w})$ is asymptotically optimal in the sense of achieving the lowest possible quadratic errors, i.e.,

$$\frac{L(\hat{\mathbf{w}})}{\inf_{\mathbf{w}\in H_T}L(\mathbf{w})} = 1 + o_p(1).$$

This shows that the squared error loss obtained from the selected weight vector $\hat{\mathbf{w}}$ is asymptotically equivalent to the infeasible optimal averaging estimator.

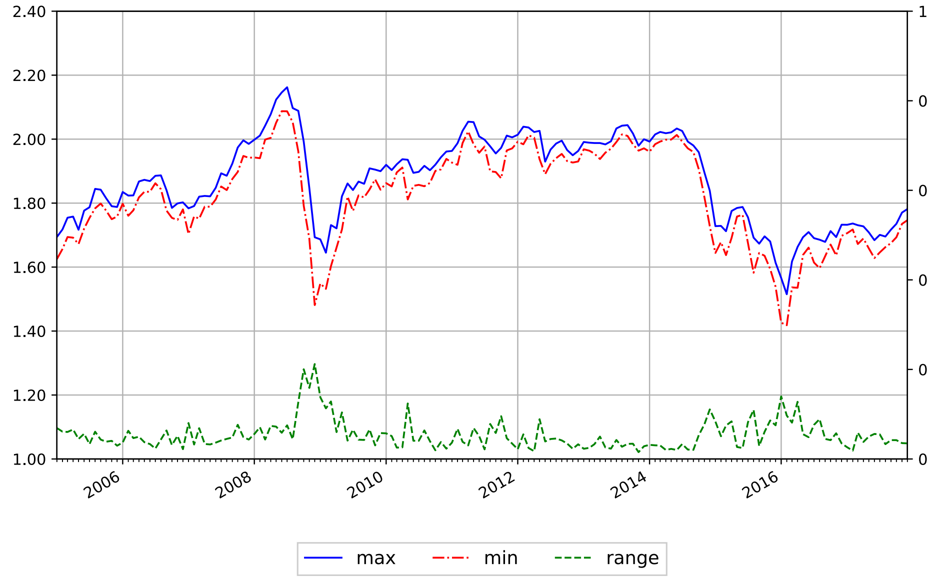


FIGURE 2 | Interval valued crude oil prices from EIA, January 2005–December 2017.

TABLE 1 | Basic statistical analysis on monthly interval-valued crude oil prices.

	Mean	Median	Maximum	Minimum	Std. dev	Skewness	Kurtosis
$y_{U,t}$	75.63	73.19	145.31	32.74	24.32	0.35	–0.71
$y_{L,t}$	67.31	65.26	122.30	26.19	22.78	0.23	–1.01
$y_{avg,t}$	71.41	69.54	133.88	30.32	23.54	0.30	–0.86
$Dy_{U,t}$	0.06	0.06	0.32	–0.15	0.08	0.32	0.93
$Dy_{L,t}$	–0.06	–0.04	0.13	–0.64	0.11	–1.95	6.30
$y_{r,t}$	0.12	0.10	0.49	0.04	0.07	2.11	8.33
$y_{c,t}$	0.00	0.01	0.22	–0.39	0.09	–1.03	2.83

3 EMPIRICAL IMPLEMENTATIONS

This section applies the proposed 2SVBMA procedure to forecast the real price of crude oil. Data and preliminary analysis are introduced in **Section 3.1**. Then the selected interval-valued factors are introduced in **Section 3.2**. **Section 3.3** introduces the candidate models. **Section 3.4** provides competing methods.

3.1 Data and Preliminary Analysis

Following Wang et al. (2017), Chai et al. (2018) and Yu et al. (2019), the daily point-valued WTI crude oil prices are used to construct the interval-valued monthly prices. $y_{U,t}$ and $y_{L,t}$ denote the daily maximum and minimum prices within tth month. $y_{c,t} = (y_{U,t} + y_{L,t})/2$ and $y_{r,t} = y_{U,t} - y_{L,t}$ are the midpoint and range from an interval-valued price observation $y_t = \langle y_{c,t}, y_{r,t} \rangle$. The data period used in the research is from January 2005 to December 2017. Data on crude oil prices are collected from the US Energy Information Administration (EIA). **Figure 2** presents the interval-valued crude oil prices: the range ($y_{r,t}$, right y-Axis), the maximum ($y_{U,t}$, left y-Axis), and minimum ($y_{L,t}$, left y-Axis) prices within 1 month, where we can see that the boundaries and ranges are interlinked, e.g., a strong increase in volatility ($y_{r,t}$) is accompanied by a significant decrease in crude oil prices during the second half of 2008.

Table 1 presents the summary of statistical characteristics. First, it is shown that the spread of ranges is slightly smaller than the volatility in the boundaries ($Dy_{U,t} = y_{U,t} - y_{avg,t-1}$ and $Dy_{L,t} = y_{L,t} - y_{avg,t-1}$), where $y_{avg,t}$ is the monthly prices from EIA. In addition, the skewness and leptokurtic kurtosis are different among $y_{r,t}$, $Dy_{L,t}$ and $Dy_{U,t}$. Compared with $Dy_{L,t}$ and $Dy_{U,t}$, $y_{r,t}$ is with greater skewness and higher leptokurtic. We can see from **Table 1** that the interval-valued data can capture more information than the point-valued data.

3.2 Interval-Valued Control Variables in the First Stage

The potential choices of monthly interval-valued explanatory variables from various aspects are considered in this section, including the stock market, commodity market, technology factor, search query data, speculation, monetary market and currency market (Pan et al., 2014; Wang et al., 2016; Wang et al., 2017; Chai et al., 2018; Yu et al., 2019); see **Table 2** for more discussions. First, the Augmented Dickey-Fuller tests suggest that

TABLE 2 | Monthly interval-valued exogenous variables.

Variables	Description	Transformation	Explanation
$SP_t = [SP_{c,t}, SP_{r,t}]$	S&P 500 index	Δ ln	Affect expected cash flows and/or discount rates,
$DJ_t = [DJ_{c,t}, DJ_{r,t}]$	Dow Jones industrial index	Δ ln	be affected through the expected rate of inflation and the expected real interest rate
$GF_t = [GF_{c,t}, GF_{r,t}]$	COMEX gold future closing prices	Δ ln	Safe haven against oil price movements
$CF_t = [CF_{c,t}, CF_{r,t}]$	LME copper future closing prices	Δ ln	
$WB_t = [WB_{c,t}, WB_{r,t}]$	WTI-Brent spot price spread	Level	Measure of the technology influence
$FD_t = [FD_{c,t}, FD_{r,t}]$	Federal funds rate	Level	As oil prices increased, so did concerns about increasing inflation
$RD_t = [RD_{c,t}, RD_{r,t}]$	Generalized real US dollar index		Oil price is dollar-denominated
$GT_t = [GT_{c,t}, GT_{r,t}]$	The key word of oil price in the Google trend search engine	Level	Reflect psychological behaviors of investors
$NL_t = [NL_{c,t}, NL_{r,t}]$	Non-commercial net long ratio	Level	Provide liquidity to offset risks

Note: (1) These interval-valued variables after transformations are used in candidate models. Transformations are (i) level: $X_t = S_t$; (2) Δ ln: $X_t = \ln S_t - \ln S_{t-1}$; (iii) Δ: $X_t = S_t - S_{t-1}$, where S_t is the original series obtained from EIA or Wind database.

TABLE 3 | Basic statistical analysis on monthly interval-valued explanatory variables.

	Mean	Median	Maximum	Minimum	Std. dev	Skewness	Kurtosis
$\Delta SP_{r,t}$	0.05	0.04	0.31	0.01	0.04	3.63	17.41
$\Delta SP_{c,t}$	0.00	0.00	0.06	−0.16	0.03	−1.82	7.59
$\Delta DJ_{r,t}$	0.05	0.04	0.28	0.01	0.04	3.47	16.36
$\Delta DJ_{c,t}$	0.00	0.00	0.05	−0.14	0.03	−1.59	5.90
$\Delta GF_{r,t}$	0.07	0.06	0.24	0.02	0.03	1.77	4.24
$\Delta GF_{c,t}$	−1.81	−1.73	−1.22	−2.58	0.31	−0.55	−0.35
$\Delta CF_{r,t}$	0.09	0.08	0.51	0.02	0.06	3.06	17.02
$\Delta CF_{c,t}$	1.81	1.73	2.52	1.26	0.32	0.54	−0.49
$WB_{r,t}$	2.29	1.69	15.36	0.01	2.15	2.39	9.12
$WB_{c,t}$	1.12	0.86	12.24	−2.87	1.77	2.03	9.76
$GT_{r,t}$	0.28	0.21	0.97	0.04	0.19	1.14	0.80
$GT_{c,t}$	3.88	4.01	4.54	2.60	0.47	−0.93	0.30
$NL_{r,t}$	0.03	0.03	0.12	0.01	0.02	1.48	2.97
$NL_{c,t}$	0.11	0.12	0.25	−0.09	0.07	−0.23	−0.70
$FD_{r,t}$	0.03	0.02	0.10	0.00	0.02	0.90	0.84
$FD_{c,t}$	−0.15	−0.18	0.01	−0.21	0.06	1.92	1.88
$RD_{r,t}$	0.19	0.09	2.75	0.01	0.32	4.61	28.40
$RD_{c,t}$	1.34	0.28	5.32	0.06	1.77	1.26	0.08

the null hypothesis for the original control variables is hardly rejected at the 5% significance level, except for non-commercial net long ratio (NL_t) and the Federal funds rate (FD_t). For stationarity, we use the Hukuhara's difference of interval-valued exogenous variables. The Hukuhara's difference between a pair of intervals is essentially equal to the regular difference between points in intervals. As Yang et al. (2016) mentioned, the concept of interval with Hukuara's difference is useful and suitable for econometric analysis of interval data. Take S&P 500 index (SP_t) as an example. It is defined as $\Delta SP_t = SP_t - SP_{t-1} = [\tilde{\Delta} SP_{c,t}, \tilde{\Delta} SP_{r,t}]$, where $\tilde{\Delta}$ is the Hukuhara's difference between intervals, and Δ is the regular difference between intervals. This implies that the midpoints and centers of these interval-valued exogenous variables are stationary after Hukuhara's difference. Similarly, we have ΔDJ_t, ΔGF_t, ΔCF_t and ΔRD_t; see specific definitions in **Table 2**.

Second, **Table 3** provides a summary of statistical characteristics. It is shown that no matter whether the time series is transferred by Hukuhara's difference, the midpoints and ranges for interval-valued control variables appear to have different skewness and leptokurtic kurtosis properties. This suggests that using one attribute of ITS contains partial information only. Thus, it is highly desirable to utilize the information contained in interval-valued data.

Third, we use the extended L_2 Boosting regularization method to select interval-valued control variables. Specifically, we set the lag length $L = 12$ for every control variable and thus the number of the potential explanatory interval-valued variables equals $12 \times 9 = 108$. For vector boosting, we start with the learning rate $c = 0.01$, iteration = 100 times. These parameters are adjusted during training. After using various training sets, ΔSP_{t+h-1}, ΔGF_{t+h-1}, ΔGF_{t+h-2}, ΔGF_{t+h-3}, WB_{t+h-1}, and GT_{t+h-4} are selected with duplicates removed and used to do h-step-ahead out-of-sample forecasts of interval-valued crude oil prices.

Furthermore, these selected interval-valued control variables have important economic interpretation for crude oil prices as follows:

ΔSP_{t+h-1}: It provides information of fundamentals and volatility contained in S&P 500. The movement of S&P 500 Index may closely mirror that of the crude oil prices (e.g., Kilian, 2009; Miller and Ratti, 2009; Balcilar et al., 2015; Ding et al., 2016). As discussed in Kilian (2009) and Miller and Ratti (2009), the oil price shocks influence stock prices by affecting expected cash flows and discount rates, since crude oil is an important input in production and its price can influence the costs for the manufacturing and transport sectors.

ΔGF_{t+h-j} (j = 1,2,3): It is the logarithmic difference between Comex gold future prices at $t+h-j$ and $t+h-j-1$, which provides information in Comex gold future market (e.g., Baur and Lucey, 2010; Reboredo, 2013; Souček, 2013; Kang et al., 2017). Gold serves as store of value especially during periods of economic uncertainties. Oil prices can affect levels of inflation (Zhao et al., 2016). Gold investment can be used as a hedge against inflation and currency depreciation. It can also be viewed as a safe haven against the stock market turbulence for investors.

WB_{t+h-1}: It is WTI-Brent spot price spread, which is the price difference between crude oil and the byproducts refined from it. The crack spread gives the profit margin that a refinery can expect. Thus, a tight spread can be seen as a indicator that refiners may slow production to tighten supply.

GT_{t+h-4}: It is the search query data collected from Internet, which has been widely applied as indicator when analyzing the crude oil prices and has been demonstrated to be effective in improving forecasts performance (Fantazzini and Fomichev, 2014; Li et al., 2015a; Wu et al., 2021; Yang et al., 2021). The keyword "oil price" is searched in the Google Trend search engine. Search query data is expected to reflect the psychological aspects of investors when they making strategic investment decisions in the crude oil market (Li et al., 2015b).

3.3 Model Averaging in the Second Stage

3.3.1 Candidate Models

We consider 6 lagged dependent variables $\mathbf{y}_{t-1}, \ldots, \mathbf{y}_{t-6}$ and 6 exogenous variables selected from vector boosting. As we use monthly interval-valued crude oil prices, the maximum lag is set to 6, including the past half year information. Exogenous variables are sorted by relevance to $\mathbf{y}_t$ during the estimation period. Then, 12 nested interval predictive candidate models are considered as:

Model 1. $\mathbf{y}_{t+h} = \boldsymbol{\alpha}_1 \mathbf{y}_{t+h-1} + \boldsymbol{\varepsilon}_{t+h}$.

Model 2. $\mathbf{y}_{t+h} = \sum_{i=1}^{2} \boldsymbol{\alpha}_i \mathbf{y}_{t+h-i} + \boldsymbol{\varepsilon}_{t+h}$.

Model 3. $\mathbf{y}_{t+h} = \sum_{i=1}^{3} \boldsymbol{\alpha}_i \mathbf{y}_{t+h-i} + \boldsymbol{\varepsilon}_{t+h}$.

Model 4. $\mathbf{y}_{t+h} = \sum_{i=1}^{4} \boldsymbol{\alpha}_i \mathbf{y}_{t+h-i} + \boldsymbol{\varepsilon}_{t+h}$.

Model 5. $\mathbf{y}_{t+h} = \sum_{i=1}^{5} \boldsymbol{\alpha}_i \mathbf{y}_{t+h-i} + \boldsymbol{\varepsilon}_{t+h}$.

Model 6. $\mathbf{y}_{t+h} = \sum_{i=1}^{6} \boldsymbol{\alpha}_i \mathbf{y}_{t+h-i} + \boldsymbol{\varepsilon}_{t+h}$.

Next, 6 exogenous variables are added to Model 6 to construct Models 7–12, sorted by relevance to $\mathbf{Y}$:

Model 7. $\mathbf{y}_{t+h} = \sum_{i=1}^{6} \boldsymbol{\alpha}_i \mathbf{y}_{t+h-i} + \boldsymbol{\beta}_1 \Delta SP_{t+h-1} + \boldsymbol{\varepsilon}_{t+h}$.

Model 8. $\mathbf{y}_{t+h} = \sum_{i=1}^{6} \boldsymbol{\alpha}_i \mathbf{y}_{t+h-i} + \boldsymbol{\beta}_1 \Delta SP_{t+h-1} + \boldsymbol{\beta}_2 \Delta GF_{t+h-1} + \boldsymbol{\varepsilon}_{t+h}$.

Model 9. $\mathbf{y}_{t+h} = \sum_{i=1}^{6} \boldsymbol{\alpha}_i \mathbf{y}_{t+h-i} + \boldsymbol{\beta}_1 \Delta SP_{t+h-1} + \boldsymbol{\beta}_2 \Delta GF_{t+h-1} + \boldsymbol{\beta}_3 \Delta GF_{t+h-2} + \boldsymbol{\varepsilon}_{t+h}$.

Model 10. $\mathbf{y}_{t+h} = \sum_{i=1}^{6} \boldsymbol{\alpha}_i \mathbf{y}_{t+h-i} + \boldsymbol{\beta}_1 \Delta SP_{t+h-1} + \boldsymbol{\beta}_2 \Delta GF_{t+h-1} + \boldsymbol{\beta}_3 \Delta GF_{t+h-2} + \boldsymbol{\beta}_4 \Delta GF_{t+h-3} + \boldsymbol{\varepsilon}_{t+h}$.

Model 11. $\mathbf{y}_{t+h} = \sum_{i=1}^{6} \boldsymbol{\alpha}_i \mathbf{y}_{t+h-i} + \boldsymbol{\beta}_1 \Delta SP_{t+h-1} + \boldsymbol{\beta}_2 \Delta GF_{t+h-1} + \boldsymbol{\beta}_3 \Delta GF_{t+h-2} + \boldsymbol{\beta}_4 \Delta GF_{t+h-3} + \boldsymbol{\beta}_5 WB_{t+h-1} + \boldsymbol{\varepsilon}_{t+h}$.

Model 12. $\mathbf{y}_{t+h} = \sum_{i=1}^{6} \boldsymbol{\alpha}_i \mathbf{y}_{t+h-i} + \boldsymbol{\beta}_1 \Delta SP_{t+h-1} + \boldsymbol{\beta}_2 \Delta GF_{t+h-1} + \boldsymbol{\beta}_3 \Delta GF_{t+h-2} + \boldsymbol{\beta}_4 \Delta GF_{t+h-3} + \boldsymbol{\beta}_5 WB_{t+h-1} + \boldsymbol{\beta}_6 GT_{t+h-4} + \boldsymbol{\varepsilon}_{t+h}$.

These candidate models are used for LsoMA in the second stage. We do h-step-ahead prediction with $h \in \{1, 4, 8, 12\}$.

3.4 Competing Methods

In this paper, we compare 2SVBMA forecasts with various competing methods, including AIC, BIC, HQ, Mallows model averaging (MMA; Liao et al., 2019), smoothed AIC (SAIC), smoothed BIC (SBIC) and smoothed Hannan-Quinn (SHQ) based on the same set of candidate models (model 1 - model 12).

The AIC criterion for the sth candidate model ($1 \leq s \leq 15$) is $AIC^{(s)} = \ln|\hat{\Sigma}^{(s)}| + 2s2^2/T$, where $\hat{s}$ minimizes $AIC^{(s)}$ and $\hat{\Sigma}^{(s)} = (T-S)^{-1}(\mathbf{Y} - \tilde{\mu}^{(s)})'(\mathbf{Y} - \tilde{\mu}^{(s)})$ as the residual covariance matrix from the sth candidate model. Similarly, BIC and HQ are model selection methods, minimizing the corresponding criteria $BIC^{(s)} = \ln|\hat{\Sigma}^{(s)}| + (\ln\, T)s2^2/T$, $HQ^{(s)} = \ln|\hat{\Sigma}^{(s)}| + 2(\ln\, \ln T)s2^2/T$, respectively. These three selected candidate models ares used as benchmark models.

Four model averaging (or forecast combination) methods are considered here. MMA proposed by Liao and Tsay (2016) is an extension of Mallows criterion to vector regression models. Specifically, the multivariate Mallow criterion for model averaging takes the following form:

$$C_T(\mathbf{w}) = (T-S)\ \mathrm{trace}\Big(\tilde{\Sigma}(S)^{-1}\hat{\Sigma}(\mathbf{w})) + 2 \cdot 2^2 \mathbf{s}'\mathbf{w}$$

where $\tilde{\Sigma}(S) = \frac{1}{T-S-2S}\sum_{t=S+1}^{T} \hat{\varepsilon}_t(S)\hat{\varepsilon}_t(S)'$, $\hat{\Sigma}(\mathbf{w}) = \frac{1}{T-S}\sum_{t=S+1}^{T} \hat{\varepsilon}\, t(\mathbf{w})\hat{\boldsymbol{\varepsilon}}_t(\mathbf{w})'$, and $\mathbf{s}'\mathbf{w} = \sum_{s=1}^{S} w(s)s$. The Mallows weight vector is defined by:

$$\hat{\mathbf{w}} = \arg\min_{w \in W} C_T(\mathbf{w}).$$

SAIC, SBIC and SHQ are simple model averaging methods with the weights

$$w_{AIC,s} = \exp\Big(-AIC^{(s)}/2\Big) / \sum_{s=1}^{S}\Big(-AIC^{(s)}/2\Big),$$

and

$$w_{BIC,s} = \exp\Big(-BIC^{(s)}/2\Big) / \sum_{s=1}^{S}\Big(-BIC^{(s)}/2\Big),$$

and

$$w_{HQ,s} = \exp\Big(-HQ^{(s)}/2\Big) / \sum_{s=1}^{S}\Big(-HQ^{(s)}/2\Big),$$

respectively.

4 EMPIRICAL RESULTS

This section compares the forecasting performance of the proposed 2SVBMA approach with various competing methods presented in previous studies by using interval-valued crude oil prices. The whole sample from 2005 January to 2017 December are divided into two parts: one is used for parameter estimation, and the other is used for out-of-sample forecasting. Various

TABLE 4 | MSPE (10^{-2}) of the recursive prediction for interval-valued crude oil prices (I).

Estimation: 2005–2010; Forecast:2011–2013

h		2SVBMA	MMA	SAIC	SBIC	SHQ	AIC	BIC	HQ
1	midpoints	**0.75**	2.09	1.58	1.10	1.39	5.59	1.10	5.59
	ranges	**0.70**	2.62	1.79	1.33	1.60	4.99	3.92	5.15
4	midpoints	**0.32**	2.17	1.20	0.90	1.09	3.50	3.29	3.60
	ranges	**1.20**	4.71	2.79	2.11	2.52	5.42	6.86	5.41
8	midpoints	**0.38**	1.55	1.06	0.85	0.98	1.71	1.93	1.78
	ranges	**1.05**	2.99	1.98	1.65	1.85	3.61	3.68	3.64
12	midpoints	**0.39**	2.10	1.20	0.91	1.09	3.67	3.17	3.67
	ranges	**0.65**	3.20	1.64	1.24	1.48	6.89	4.90	6.89

Estimation: 2006–2011; Forecast:2012–2014

h		2SVBMA	MMA	SAIC	SBIC	SHQ	AIC	BIC	HQ
1	midpoints	**0.22**	0.66	0.47	0.36	0.43	1.21	0.91	1.17
	ranges	**0.33**	1.03	0.73	0.55	0.66	2.07	1.61	1.84
4	midpoints	**0.26**	0.97	0.64	0.49	0.58	1.43	1.50	1.34
	ranges	**0.41**	0.97	0.70	0.56	0.64	1.39	1.27	1.35
8	midpoints	**0.22**	0.75	0.47	0.33	0.41	1.29	1.20	1.23
	ranges	**0.40**	0.80	0.50	0.42	0.47	1.57	1.22	1.50
12	midpoints	**0.09**	0.45	0.22	0.14	0.18	1.36	0.61	1.20
	ranges	**0.25**	0.42	0.29	0.27	0.28	0.65	0.56	0.66

Estimation: 2007–2012; Forecast:2013–2015

h		2SVBMA	MMA	SAIC	SBIC	SHQ	AIC	BIC	HQ
1	midpoints	**0.07**	0.16	0.14	0.11	0.13	0.38	0.18	0.34
	ranges	**0.13**	0.22	0.19	0.16	0.18	0.60	0.26	0.35
4	midpoints	**0.09**	0.39	0.24	0.17	0.21	0.58	0.74	0.58
	ranges	**0.18**	0.20	0.21	0.18	0.20	0.48	0.25	0.29
8	midpoints	**0.17**	0.24	0.24	0.23	0.23	0.30	0.35	0.36
	ranges	**0.37**	0.39	0.46	0.42	0.44	1.36	0.31	0.79
12	midpoints	**0.22**	0.27	0.25	0.24	0.24	0.37	0.26	0.35
	ranges	**0.49**	0.73	0.56	0.54	0.55	0.83	0.67	0.97

Note: "Estimation" denotes the sample during this period used to estimate parameters, and "Forecast" denotes the sample during this period used to do out-of-sample forecasts. The best forecasts are marked by boldface, and the second best forecasts are marked by underline.

subsamples for estimation and forecast are used to test prediction accuracy; see **Tables 4**, **5**.

Tables 4, **5** report the MSPEs of h-step-ahead (1,4,8,12) forecasts for the interval-valued crude oil prices using various estimation and forecast samples. First, it is worth noticing that for the horizons of 1, 4, 8 and 12 months, the 2SVBMA method outperforms other competing methods in most cases; out of the 48 cases considered, with respect to RMSFE of midpoints and ranges, it yields the best outcomes 42 times and the second best outcomes 6 times. Intuitively, the proposed 2SVBMA method selects the important factors at the first stage and then give the optimal weights averaging across the 12 nested regression forecasts. Second, 2SVBMA based on LsoMA outperforms various model averaging and model selection methods, including MMA. One possible explanation is that leave-subject-out cross-validation is more suitable for vector autoregressive situations with heteroscedastic and auto-correlated errors. Additionally, as shown in Liao et al. (2019), the approximate unbiasedness of LsoMA and its asymptotic optimality in terms of obtaining the lowest quadratic errors are established. This is why LsoMA outperforms other model averaging methods (i.e., SAIC, SBIC, and SHQ) in the second stage.

Second, the SBIC estimators always produce the second-best forecasts after the 2SVBMA estimator among all model averaging methods, while SAIC achieves higher forecast criteria than other model averaging methods. Similarly, BIC always yields best forecasts among all model selection methods, while the AIC estimator achieves higher MSFE in most cases. This happens because AIC prefers selecting the relatively complicated model, which is inappropriate for out-of-sample forecasting even though it has good in-sample fitting. A simple model may be better for out-of-sample forecasting.

Furthermore, it is shown that at the second stage, model averaging forecasts outperform model selection forecasts in almost 90% of all cases. The significant advantages of model averaging support the argument of Rapach et al. (2010) that

TABLE 5 | MSPE (10^{-2}) of the recursive prediction for interval-valued crude oil prices (II).

Estimation: 2008–2013; Forecast:2014–2016

h		2SVBMA	MMA	SAIC	SBIC	SHQ	AIC	BIC	HQ
1	midpoints	**0.15**	0.30	0.29	0.23	0.27	0.61	0.34	0.57
	ranges	**0.81**	1.14	1.03	0.96	1.00	1.50	1.03	1.38
4	midpoints	**0.39**	0.47	0.52	0.46	0.50	0.69	0.65	0.70
	ranges	**1.11**	1.71	1.47	1.31	1.41	2.15	2.01	2.12
8	midpoints	0.47	1.38	0.93	0.77	0.87	2.83	**0**.43	3.29
	ranges	1.07	2.44	1.98	1.61	1.82	5.63	**1**.03	5.34
12	midpoints	**0.54**	2.45	1.18	0.88	1.05	5.84	0.61	5.85
	ranges	**0.90**	2.52	1.67	1.32	1.52	4.40	1.96	4.54

Estimation: 2009–2014; Forecast:2015–2017

h		2SVBMA	MMA	SAIC	SBIC	SHQ	AIC	BIC	HQ
1	midpoints	**0.22**	0.55	0.52	0.40	0.47	1.08	0.33	1.03
	ranges	**1.14**	2.37	2.12	1.71	1.96	4.45	1.21	4.37
4	midpoints	0.82	1.61	1.47	1.14	1.34	2.46	**0.66**	2.45
	ranges	**2.03**	3.46	2.69	2.17	2.48	5.94	2.75	6.01
8	midpoints	0.47	1.67	1.17	0.87	1.05	6.87	**0.35**	4.15
	ranges	**1.52**	2.92	2.51	1.91	2.25	10.06	1.58	9.45
12	midpoints	0.49	5.15	1.69	1.06	1.42	13.58	**0.40**	12.92
	ranges	**0.75**	3.93	2.08	1.44	1.81	8.94	1.51	8.32

Note: "Estimation" denotes the sample during this period used to estimate parameters, and "Forecast" denotes the sample during this period used to do out-of-sample forecasts. The best forecasts are marked by boldface, and the second best forecasts are marked by underline.

"model uncertainty and instability seriously impair the forecasting ability of individual predictive regression models."

Overall, the proposed approach using interval-valued data is capable of assessing and forecasting the changes in both level and volatility. We can see from the results that forecasting with model averaging is generally better than obtaining the predictions from just one model (model selection). Since we may choose a very different model when there are small changes in the original data set, which may lead to a big change in the final conclusions, resulting in non-effective decision-making due to the unstable forecasting process. The proposed method is able to help obtain more stable decision-making when a long list of interval-valued predictors is available in a wide range of fields, for example, the daily trading strategy in the finance field.

5 CONCLUSION

We propose a novel 2SVBMA forecasting procedure to capture the relevant information available in the interval format and the underlying characteristics of crude oil price movements. Vector L_2Boosting in the first stage and LsoMA in the second stage are extended to interval models with interval-valued exogenous variables. Empirical results show that our proposed approach outperforms other competing model averaging and model selection methods in terms of MSFE of midpoints and ranges.

There are some limitations and potential extensions of our study. First, more advanced optimization algorithms for interval-valued variable selection can be proposed in future work. Second, the candidate models with different structures in model averaging methods can further be developed to enhance forecasting. It would also be interesting to develop interval-based machine learning methods to improve forecast accuracy. Furthermore, the proposed methodology in this paper can be extended to the vector autoregressive (VAR) model, which can cover more applications in economics and finance.

In general, 2SVBMA provides a methodological framework for interval-valued data forecasting when there are a large number of potential predictors. For example, this methodology can be used to quantify the impact of COVID-19 pandemic on oil and gas industry. 2SVBMA can also provide implications for the post-COVID recovery management. The accurate prediction of crude oil prices will assist policy makers in understanding issues affecting different oil industry segments, and help governments be better prepared for the recovery.

6 COMPLIANCE WITH ETHICAL STANDARDS

The authors thank a number of the participants at Symposium on Interval Data Modelling: Theory and Applications (SIDM 2019) in Beijing for their valuable comments and suggestions. This work was partially supported by National Natural Science Foundation of China (Nos. 71973116, 71988101, 72073126, 72091212), and the disciplinary funding of Central University of Finance and Economics. The authors declare no competing interests. This article does not contain any studies with human participants performed by any of the authors.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

REFERENCES

Abramson, B., and Finizza, A. (1995). Probabilistic Forecasts from Probabilistic Models: a Case Study in the Oil Market. *Int. J. Forecast.* 11, 63–72. doi:10.1016/0169-2070(94)02004-9

Álvarez-Díaz, M. (2019). Is it Possible to Accurately Forecast the Evolution of Brent Crude Oil Prices? an Answer Based on Parametric and Nonparametric Forecasting Methods. *Empirical Econ.* 59, 1285–1305. doi:10.1007/s00181-019-01665-w

Arroyo, J., González-Rivera, G., and Maté, C. (2011). "Forecasting with Interval and Histogram Data: Some Financial Applications," in *Handbook of Empirical Economics and Finance*. Editors A. Ullah and D. E. A. Giles (New York: Chapman & Hall), 247–279.

Balcilar, M., Gupta, R., and Miller, S. M. (2015). Regime Switching Model of Us Crude Oil and Stock Market Prices: 1859 to 2013. *Energ. Econ.* 49, 317–327. doi:10.1016/j.eneco.2015.01.026

Bates, J. M., and Granger, C. W. J. (1969). The Combination of Forecasts. *Or* 20, 451–468. doi:10.2307/3008764

Baur, D. G., and Lucey, B. M. (2010). Is Gold a Hedge or a Safe haven? an Analysis of Stocks, Bonds and Gold. *Financial Rev.* 45, 217–229. doi:10.1111/j.1540-6288.2010.00244.x

Binder, K. E., Pourahmadi, M., and Mjelde, J. W. (2018). The Role of Temporal Dependence in Factor Selection and Forecasting Oil Prices. *Empirical Econ.* 58, 1–39. doi:10.1007/s00181-018-1574-9

Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997). Model Selection: An Integral Part of Inference. *Biometrics* 53, 603–618. doi:10.2307/2533961

Buhlmann, P. (2006). Boosting for High-Dimensional Linear Models. *Ann. Stat.* 34, 559–583. doi:10.1214/009053606000000092

Chai, J., Xing, L.-M., Zhou, X.-Y., Zhang, Z. G., and Li, J.-X. (2018). Forecasting the Wti Crude Oil price by a Hybrid-Refined Method. *Energ. Econ.* 71, 114–127. doi:10.1016/j.eneco.2018.02.004

Cheung, Y.-L., Cheung, Y.-W., and Wan, A. T. K. (2009). A High-Low Model of Daily Stock price Ranges. *J. Forecast.* 28, 103–119. doi:10.1002/for.1087

De Carvalho, F. A. T., Lima Neto, E. A., and Tenorio, C. P. (2004). "A New Method to Fit a Linear Regression Model for Interval-Valued Data," in *Lecture Notes in Computer Science, K12004 Advances in Artificial Intelligence* (Berlin: Springer-Verlag).

Ding, H., Kim, H.-G., and Park, S. Y. (2016). Crude Oil and Stock Markets: Causal Relationships in Tails? *Energ. Econ.* 59, 58–69. doi:10.1016/j.eneco.2016.07.013

Donald, S. G., Imbens, G. W., and Newey, W. K. (2009). Choosing Instrumental Variables in Conditional Moment Restriction Models. *J. Econom.* 152, 28–36. doi:10.1016/j.jeconom.2008.10.013

Ebrahim, Z., Inderwildi, O. R., and King, D. A. (2014). Macroeconomic Impacts of Oil price Volatility: Mitigation and Resilience. *Front. Energ.* 8, 9–24. doi:10.1007/s11708-014-0303-0

Fan, J., and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its oracle Properties. *J. Am. Stat. Assoc.* 96, 1348–1360. doi:10.1198/016214501753382273

AUTHOR CONTRIBUTIONS

All three authors contributed equally to this work and the order of authorship has nothing other than alphabetical significance.

FUNDING

This work was partially supported by National Natural Science Foundation of China (Nos. 71973116, 71988101, 72073126, 72091212), the funding of Forecasting and Monitoring of COVID-19 in countries along "Belt and Road" and Related Economic Impacts (ANSO-SBA-2020-12), and the disciplinary funding of Central University of Finance and Economics.

Fantazzini, D., and Fomichev, N. (2014). Forecasting the Real price of Oil Using Online Search Data. *Ijcee* 4, 4–31. doi:10.1504/ijcee.2014.060284

Frank, L. E., and Friedman, J. H. (1993). A Statistical View of Some Chemometrics Regression Tools. *Technometrics* 35, 109–135. doi:10.1080/00401706.1993.10485033

García-Ascanio, C., and Maté, C. (2010). Electric Power Demand Forecasting Using Interval Time Series: A Comparison between Var and Imlp. *Energy Policy* 38, 715–725. doi:10.1016/j.enpol.2009.10.007

González-Rivera, G., and Lin, W. (2013). Constrained Regression for Interval-Valued Data. *J. Business Econ. Stat.* 31, 473–490. doi:10.1080/07350015.2013.818004

Hamilton, J. D. (2008). "Understanding Crude Oil Prices,". (no. w14492).

Hansen, B. E. (2007). Least Squares Model Averaging. *Econometrica* 75, 1175–1189. doi:10.1111/j.1468-0262.2007.00785.x

Hansen, B. E., and Racine, J. S. (2012). Jackknife Model Averaging. *J. Econom.* 167, 38–46. doi:10.1016/j.jeconom.2011.06.019

He, A. W. W., Kwok, J. T. K., and Wan, A. T. K. (2010). An Empirical Model of Daily Highs and Lows of West texas Intermediate Crude Oil Prices. *Energ. Econ.* 32, 1499–1506. doi:10.1016/j.eneco.2010.07.012

Hjort, N. L., and Claeskens, G. (2006). Focused Information Criteria and Model Averaging for the Cox hazard Regression Model. *J. Am. Stat. Assoc.* 101, 1449–1464. doi:10.1198/016214506000000069

Hjort, N. L., and Claeskens, G. (2003). Frequentist Model Average Estimators. *J. Am. Stat. Assoc.* 98, 879–899. doi:10.1198/016214503000000828

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian Model Averaging: a Tutorial. *Stat. Sci.* 14, 382–417. doi:10.1214/ss/1009212519

Hu, Z., Bao, Y., Chiong, R., and Xiong, T. (2015). Mid-term Interval Load Forecasting Using Multi-Output Support Vector Regression with a Memetic Algorithm for Feature Selection. *Energy* 84, 419–431. doi:10.1016/j.energy.2015.03.054

Kang, S. H., McIver, R., and Yoon, S.-M. (2017). Dynamic Spillover Effects Among Crude Oil, Precious Metal, and Agricultural Commodity Futures Markets. *Energ. Econ.* 62, 19–32. doi:10.1016/j.eneco.2016.12.011

Kilian, L. (2009). Not all Oil price Shocks Are Alike: Disentangling Demand and Supply Shocks in the Crude Oil Market. *Am. Econ. Rev.* 99, 1053–1069. doi:10.1257/aer.99.3.1053

Knight, K., and Fu, W. (2000). Asymptotics for Lasso-type Estimators. *Ann. Stat.* 28, 1356–1378. doi:10.1214/aos/1015957397

Li, D., Linton, O., and Lu, Z. (2015a). A Flexible Semiparametric Forecasting Model for Time Series. *J. Econom.* 187, 345–357. doi:10.1016/j.jeconom.2015.02.025

Li, X., Ma, J., Wang, S., and Zhang, X. (2015b). How Does Google Search Affect Trader Positions and Crude Oil Prices? *Econ. Model.* 49, 162–171. doi:10.1016/j.econmod.2015.04.005

Liao, J.-C., and Tsay, W.-J. (2016). Multivariate Least Squares Forecasting Averaging by Vector Autoregressive Models. Available at SSRN 2827416.

Liao, J., Zong, X., Zhang, X., and Zou, G. (2019). Model Averaging Based on Leave-Subject-Out Cross-Validation for Vector Autoregressions. *J. Econom.* 209, 35–60. doi:10.1016/j.jeconom.2018.10.007

Lima Neto, E. d. A., and De Carvalho, F. d. A. T. (2010). Constrained Linear Regression Models for Symbolic Interval-Valued Variables. *Comput. Stat. Data Anal.* 54, 333–347. doi:10.1016/j.csda.2009.08.010

Maia, A. L. S., and de Carvalho, F. d. A. T. (2011). Holt's Exponential Smoothing and Neural Network Models for Forecasting Interval-Valued Time Series. *Int. J. Forecast.* 27, 740–759. doi:10.1016/j.ijforecast.2010.02.012

Maia, A. L. S., De Carvalho, F. d. A. T., and Ludermir, T. B. (2008). Forecasting Models for Interval-Valued Time Series. *Neurocomputing* 71, 3344–3352. doi:10.1016/j.neucom.2008.02.022

Miller, J. I., and Ratti, R. A. (2009). Crude Oil and Stock Markets: Stability, Instability, and Bubbles. *Energ. Econ.* 31, 559–568. doi:10.1016/j.eneco.2009.01.009

Ng, S., and Bai, J. (2009). Selecting Instrumental Variables in a Data Rich Environment. *J. Time Ser. Econom.* 1, 4. doi:10.2202/1941-1928.1014

Pan, Z., Wang, Y., and Yang, L. (2014). Hedging Crude Oil Using Refined Product: A Regime Switching Asymmetric Dcc Approach. *Energ. Econ.* 46, 472–484. doi:10.1016/j.eneco.2014.05.014

Qiao, K., Sun, Y., and Wang, S. (2019). Market Inefficiencies Associated with Pricing Oil Stocks during Shocks. *Energ. Econ.* 81, 661–671. doi:10.1016/j.eneco.2019.04.016

Rapach, D. E., Strauss, J. K., and Zhou, G. (2010). Out-of-sample Equity Premium Prediction: Combination Forecasts and Links to the Real Economy. *Rev. Financ. Stud.* 23, 821–862. doi:10.1093/rfs/hhp063

Reboredo, J. C. (2013). Is Gold a Hedge or Safe haven against Oil price Movements? *Resour. Pol.* 38, 130–137. doi:10.1016/j.resourpol.2013.02.003

Shin, H., Hou, T., Park, K., Park, C.-K., and Choi, S. (2013). Prediction of Movement Direction in Crude Oil Prices Based on Semi-supervised Learning. *Decis. Support Syst.* 55, 348–358. doi:10.1016/j.dss.2012.11.009

Souček, M. (2013). Crude Oil, Equity and Gold Futures Open Interest Co-movements. *Energ. Econ.* 40, 306–315. doi:10.1016/j.eneco.2013.07.010

Sun, Y., Han, A., Hong, Y., and Wang, S. (2018). Threshold Autoregressive Models for Interval-Valued Time Series Data. *J. Econom.* 206, 414–446. doi:10.1016/j.jeconom.2018.06.009

Sun, Y., Zhang, X., Hong, Y., and Wang, S. (2019). Asymmetric Pass-Through of Oil Prices to Gasoline Prices with Interval Time Series Modelling. *Energ. Econ.* 78, 165–173. doi:10.1016/j.eneco.2018.10.027

Taghizadeh-Hesary, F., Rasoulinezhad, E., and Kobayashi, Y. (2016). Oil price Fluctuations and Oil Consuming Sectors: An Empirical Analysis of Japan. *Econom. Pol. Ener. Environ.* (2), 33–51. doi:10.3280/EFE2016-002003

Wang, X., Zhang, Z., and Li, S. (2016). Set-valued and Interval-Valued Stationary Time Series. *J. Multivariate Anal.* 145, 208–223. doi:10.1016/j.jmva.2015.12.010

Wang, Y., Liu, L., and Wu, C. (2017). Forecasting the Real Prices of Crude Oil Using Forecast Combinations over Time-Varying Parameter Models. *Energ. Econ.* 66, 337–348. doi:10.1016/j.eneco.2017.07.007

Wu, B., Wang, L., Lv, S.-X., and Zeng, Y.-R. (2021). Effective Crude Oil price Forecasting Using New Text-Based and Big-Data-Driven Model. *Measurement* 168, 108468. doi:10.1016/j.measurement.2020.108468

Xiong, T., Li, C., and Bao, Y. (2017). Interval-valued Time Series Forecasting Using a Novel Hybrid Holti and Msvr Model. *Econ. Model.* 60, 11–23. doi:10.1016/j.econmod.2016.08.019

Xu, G., Wang, S., and Huang, J. Z. (2014). Focused Information Criterion and Model Averaging Based on Weighted Composite Quantile Regression. *Scand. J. Statist* 41, 365–381. doi:10.1111/sjos.12034

Yang, W., Han, A., Cai, K., and Wang, S. (2012). Acix Model with Interval Dummy Variables and its Application in Forecasting Interval-Valued Crude Oil Prices. *Proced. Comp. Sci.* 9, 1273–1282. doi:10.1016/j.procs.2012.04.139

Yang, W., Han, A., Hong, Y., and Wang, S. (2016). Analysis of Crisis Impact on Crude Oil Prices: a New Approach with Interval Time Series Modelling. *Quantitative Finance* 16, 1917–1928. doi:10.1080/14697688.2016.1211795

Yang, Y., Guo, J. e., Sun, S., and Li, Y. (2021). Forecasting Crude Oil price with a New Hybrid Approach and Multi-Source Data. *Eng. Appl. Artif. Intelligence* 101, 104217. doi:10.1016/j.engappai.2021.104217

Yoshino, N., and Hesary, F. T. (2014). Monetary Policy and Oil price Fluctuations Following the Subprime Mortgage Crisis. *Ijmef* 7, 157–174. doi:10.1504/ijmef.2014.066482

Yu, L., Zhao, Y., Tang, L., and Yang, Z. (2019). Online Big Data-Driven Oil Consumption Forecasting with Google Trends. *Int. J. Forecast.* 35, 213–223. doi:10.1016/j.ijforecast.2017.11.005

Zaabouti, K., Ben Mohamed, E., and Bouri, A. (2016). Does Oil price Affect the Value of Firms? Evidence from Tunisian Listed Firms. *Front. Energ.* 10, 1–13. doi:10.1007/s11708-016-0396-8

Zhang, X., Lai, K. K., and Wang, S.-Y. (2008). A New Approach for Crude Oil price Analysis Based on Empirical Mode Decomposition. *Energ. Econ.* 30, 905–918. doi:10.1016/j.eneco.2007.02.012

Zhang, X., and Liang, H. (2011). Focused Information Criterion and Model Averaging for Generalized Additive Partial Linear Models. *Ann. Stat.* 39, 174–200. doi:10.1214/10-aos832

Zhang, X., Wan, A. T. K., and Zhou, S. Z. (2012). Focused Information Criteria, Model Selection, and Model Averaging in a Tobit Model with a Nonzero Threshold. *J. Business Econ. Stat.* 30, 132–142. doi:10.1198/jbes.2011.10075

Zhang, X., Yu, L., Wang, S., and Lai, K. K. (2009). Estimating the Impact of Extreme Events on Crude Oil price: An Emd-Based Event Analysis Method. *Energ. Econ.* 31, 768–778. doi:10.1016/j.eneco.2009.04.003

Zhang, Y., Li, J., Liu, H., Zhao, G., Tian, Y., and Xie, K. (2020). Environmental, Social, and Economic Assessment of Energy Utilization of Crop Residue in china. *Front. Energ.* 15, 308–319. doi:10.1007/s11708-020-0696-x

Zhao, L., Zhang, X., Wang, S., and Xu, S. (2016). The Effects of Oil price Shocks on Output and Inflation in china. *Energ. Econ.* 53, 101–110. doi:10.1016/j.eneco.2014.11.017

Zhao, Y., Li, J., and Yu, L. (2017). A Deep Learning Ensemble Approach for Crude Oil price Forecasting. *Energ. Econ.* 66, 9–16. doi:10.1016/j.eneco.2017.05.023

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Energy Research

ORIGINAL RESEARCH
published: 08 October 2021
doi: 10.3389/fenrg.2021.761567

The Construction and Empirical Study on Evaluation Index System of International Low-Carbon Economy Development

Zhang Zhongyu[1] and Zhang Zhongxiang[2]*

[1]College of Management and Economics, Tianjin University, Tianjin, China, [2]Ma Yinchu School of Economics, Tianjin University, Tianjin, China

Global climate change has become one of the core issues of world governance. Many countries have put forward the goal of carbon neutrality one after another, leading to the intensification of international low-carbon economy competition. To assess the current low-carbon competitiveness among countries, this article constructs an evaluation index system of international low-carbon economy development, and obtains the scores and rankings of countries in energy, society, economy and environment, as well as overall. Taking 20 countries with the highest carbon emissions in the world in 2019 as samples, starting from the concept of low-carbon economy and five evaluation principles, this article selects 40 low-carbon evaluation indicators from five aspects, including economy, society, science and technology, environment, and energy structure. By using the principal component factor analysis method to calculate and test, the four factors, energy factor, society factor, economy factor, and environment factor, are finally extracted to construct the evaluation index system. Results show that South Korea, France, China, Canada, and Germany are among the world's top five low-carbon economies. The overall competitiveness of China's low-carbon economy is in a relatively favorable position (3rd overall), with the most outstanding performance in terms of economic strength (1st), but poor performance in terms of social development (9th) and environmental carrying capacity (9th), and the biggest disadvantage in terms of energy structure (13th).

Keywords: low-carbon economy, evaluation index system, international competitive power, principal component, factor analysis, China

OPEN ACCESS

Edited by:
Jian Chai,
Xidian University, China

Reviewed by:
Ting Liang,
Hunan University, China
Quanying Lu,
Academy of Mathematics and Systems Science (CAS), China

****Correspondence:***
Zhang Zhongyu
jgxyzzy@tju.edu.cn

Specialty section:
This article was submitted to Smart Grids, a section of the journal Frontiers in Energy Research

Received: *20 August 2021*
Accepted: *14 September 2021*
Published: *08 October 2021*

Citation:
Zhongyu Z and Zhongxiang Z (2021) The Construction and Empirical Study on Evaluation Index System of International Low-Carbon Economy Development. Front. Energy Res. 9:761567. doi: 10.3389/fenrg.2021.761567

INTRODUCTION

With the release of the IPCC AR6 Synthesis Report on August 9, 2021, the world will fully enter the era of "carbon neutrality," and countries will strive to achieve carbon neutrality by the middle of the century. This ambitious goal will bring about the transformation of the whole social economy and the arrival of a new round of competition. Global competition for low-carbon economy is further intensified, and green and low-carbon development has become the focus of boosting global economic prosperity.

The European Union (EU), the US, and Japan were the first to peak carbon in 1979, 2007, and 2008 and will take 40–70 years to become carbon neutrality. As a developing country and the world's largest CO_2 emitter (accounted for 28.82% of the world's emissions in 2019), China will strive to achieve these two goals within 30 years. Then, compared with other countries, what advantages and obstacles does China have in developing a low-carbon economy? In what areas should China stick to its own development path, and in what areas should China learn from the experience of other countries? How does China fit into the global low-carbon economy?

However, there has not yet been an authoritative evaluation index system for low-carbon economy in the world, which makes it impossible to give guidance and suggestions on the development direction of low-carbon economy at the national level.

First, it is necessary to clarify the objective of the index system evaluation—low-carbon economy. The term “Low Carbon Economy” was first proposed by Kinzig and Kammen (1998), and was officially used as an official term in the UK Energy White Paper in 2003 (Vivid Economics 2009), the report (2009) on G20 countries’ low-carbon competitiveness defined the Low Carbon Economy was an economic form with a certain level of carbon productivity and sustainable development, which had characteristics about low energy consumption, low pollution, low emission and environment friendly, and global shared vision to control greenhouse gas emissions and develop social economy (Lu and Zhu, 2013).

As early as in 2005, the research on low-carbon economy in China started from Zhuang (2005, 2007), He et al. (2010), Fu and Liu (2010) et al. (e.g., Bao et al., 2008; Fu et al., 2008; Zhang et al., 2009). They took the lead in discussing low-carbon economy from the aspects of development form, development mode, and development process. Pan et al. (2010) believed that low-carbon economy had three core characteristics, namely “low-carbon emissions,” “high carbon productivity,” and “stages.”

Therefore, it can be seen that the essence of low-carbon economy is the efficient utilization of energy and the development of clean energy. Its core is technological and institutional innovation, and its goal is to control greenhouse gas emissions and promote the sustainable development of human beings (e.g., Yang, 2012; Xie et al., 2017; Zhong, 2018). In recent years, governments around the world have been racing to turn the development of a low-carbon economy from idea into practice. The EU took the lead in developing a number of low-carbon policies to change the traditional lifestyle of residents (Dagoumas and Barker, 2010; Hughes and Strachan, 2010; Government, 2009). The US paid more attention to technological innovation to solve environmental problems. Japan rapidly developed high and new technologies and applied them in the field of clean energy (Strachan et al., 2008; “2050 Japan Low-Carbon Society” project team, 2008). Countries in economic transition and developing countries, such as Russia, South Korea, China, South Africa, Brazil, and India, have joined the international competition led by low-carbon economy one after another.

Second, in terms of establishing the evaluation index system of low-carbon economy, Chinese scholars Fu and Zhuang (2010) were the first to set indicators with different linear weights (AHP and DEA) and rank them. Then, Zhuang and Pan et al. (2011) constructed an evaluation system by judging whether various indicators were within the preset threshold. To further refine the indicators, Fu and Zheng et al. (2011) designed an index system of evaluating the level of low-carbon economy development, which involved one target layer, five rule layers, and nineteen index layers, and then used the Analytic Hierarchy Process (AHP) method to carry out quantitative evaluation on the low-carbon economy at the provincial scale in China, and compared and contrasted some key indicators with those of other countries. Luo and Tong (2011) used the factor analysis method and the entropy weight method to calculate and rank the low-carbon economy development capacity of China’s provinces, and thus summarized the national low-carbon economy development capacity; Yan and Ma (2015) took Chongqing as the research object and comprehensively applied the expert scoring method (Delphi method), AHP, entropy weight method, and TOPSIS method (the superior and inferior solution distance method); Duan et al. (2016) took Dalian as the research object and adopted the AHP-entropy method; Azizalrahman and Hasyimi (2018) established a general multi-criteria evaluation model to evaluate ten cities around the world.

Most importantly, the existing literature fails to consider, from the nation level, to construct international low-carbon economy development indicators, and compare the low-carbon economy development level among counties. Most of current domestic and foreign research objects involve: first, the industry level, such as manufacturing (Wang and Pan, 2019), tourism (Tao, 2017), transportation (Fan et al., 2018), etc.; second, the city level (Xu and Liu, 2014; Pei and Tan, 2013; Yuan et al., 2017); third, the provincial level (Yang, 2012; Shi et al., 2018) and regional level (Xie et al., 2017; Zhong, 2018).

The main contributions of this article are as follows:

1) A new evaluation index system of international low-carbon economy development is designed and applicable to the national level.
2) Using principal component factor analysis, four principal factors are extracted (energy factor, social factor, economic factor, and environmental factor).
3) A clear list of four factor rankings and scores for 20 countries, as well as total scores and rankings, in which China presents clear strengths and weaknesses.

The rest of this article is organized as follows. Construction of evaluation index system is presented in *Construction of Evaluation Index System*. Empirical analysis is studied in *Empirical Analysis*. And conclusions and suggestions are drawn in *Conclusions and Suggestions*.

CONSTRUCTION OF EVALUATION INDEX SYSTEM

Significance, Theoretical Basis, and Principles of Index Construction

Reexamining the international low-carbon economy evaluation index system is of great theoretical and practical significance for further vigorously promoting global climate governance. To assess the main nations by multiple dimensions, we can understand the status quo of the world’s low-carbon economy development, identify the advantages and disadvantages of different countries, and put forward the universal evaluation

TABLE 1 | The evaluation index system of international low-carbon economy development.

Target layer		The international low-carbon economy development level	
Criterion layer		**Indicator layer**	**Direction**
Economy development indexes 8	Gross production	GDP per capita (constant 2010 US$)	+
		GDP growth (annual %)	+
	Industrial structure	Industry (including construction), value added (% of GDP)	–
		Services, value added (% of GDP)	+
		Gross fixed capital formation (% of GDP)	+
		External balance on goods and services (% of GDP)	–
		Foreign direct investment, net inflows (% of GDP)	–
		Energy imports, net (% of energy use)	–
Society development indexes 8	Social development	Population growth (annual %)	–
		Population density (people per sq. km of land area)	–
		Urban population (% of total population)	+
	Living standard	Gini index (World Bank estimate)	–
		Poverty headcount ratio at $5.50 a day (2011 PPP) (% of population)	–
		Consumer price index (2010 = 100)	+
		Labor force participation rate, total (% of total population ages 15–64)	+
		Unemployment, total (% of total labor force)	–
Technology development indexes 8	Technical level	Research and development expenditure (% of GDP)	+
		Researchers in R&D (per million people)	+
		Scientific and technical journal articles	+
		Patent applications, residents	+
		High-technology exports (% of manufactured exports)	+
		Electric power transmission and distribution losses (% of output)	–
	Education investment	Tertiary education enrollment (% gross)	+
		Government expenditure on education, total (% of GDP)	+
Environment development indexes 8	Air pollution	PM2.5 mean annual exposure (micrograms per cubic meter)	–
		CO_2 emissions (metric tons per capita)	–
		CO_2 emissions (kg per 2011 PP P $ of GDP)	–
		CO_2 intensity (kg per kg of oil equivalent energy use)	+
	Greening protection	Forest area (% of land area)	+
		Fertilizer consumption (kilograms per hectare of arable land)	–
		Renewable internal freshwater resources per capita (cubic meters)	+
		Disaster risk reduction progress score (1–5 scale; 5 = best)	+
Energy structure development indexes 8	Energy consumption	Energy use (kg of oil equivalent per capita)	–
		GDP per unit of energy use (constant 2011 PP P $ per kg of oil equivalent)	+
		Fossil fuel energy consumption (% of total)	–
		Alternative and nuclear energy (% of total energy use)	+
	Electricity production	from oil, gas, and coal sources (% of total)	–
		from hydroelectric sources (% of total)	+
		from nuclear sources (% of total)	+
		from other renewable sources (% of total)	+

standard. At the same time, for China, it can clearly identify the level and shortcoming, which is conducive to exploring excellent, replicable, and generalizable institutional achievements of other countries, learning effective major reform measures and successful experiences, and promoting the realization of carbon neutrality goals.

This evaluation system is a means and tool to objectively evaluate the level of low-carbon economy development of each country at the nation level. Its theoretical basis consists of the connotation and characteristics of core concepts, such as sustainable development, green economy, low-carbon economy, ecological civilization, new climate economics, carbon emission decoupling, and coping with climate change (Zhuang et al., 2020; Zhou et al., 2018).

The five principles of the index system construction are: 1) **Comprehensiveness**: the selected indicators should fully reflect the factors affecting the development of a country's low-carbon economy from multiple aspects; 2) **Effectiveness**: the selected indicators should have a high adoption rate in reflecting the low-carbon economy; 3) **Applicability**: the selected indicators should be applicable to the evaluation needs at the nation level, and the data should be available; 4) **Correlation**: the selected indicators should be representative, but the highly overlapping indicators with complete correlation (correlation coefficient 1, p value 0) should not be retained at the same time; 5) **Foresight**: the selected indicators should reflect both the current situation and the potential of low-carbon economy development in the future (Lan and Zheng, 2013; Lv et al., 2013; Cao, 2018).

Index Screening and Data Collection

Based on the policy evaluation of domestic low-carbon construction and the review of domestic and foreign low-

TABLE 2 | Countries rank of carbon dioxide.

		Million tonnes	Share (%) 2019	Growth rate per annum (%)	
				2019	2008–2018
1	China	9,825.8	28.8	3.4	2.6
2	US	4,964.7	14.5	**–3.0**	**–1.1**
3	India	2,480.4	7.3	1.1	5.3
4	Russian Federation	1,532.6	4.5	**–1.0**	**–0.03**
5	Japan	1,123.1	3.3	–3.5	**–1.1**
6	Germany	683.8	2.0	–6.5	**–1.0**
7	Iran	670.7	2.0	4.1	2.5
8	South Korea	638.6	1.9	–3.6	2.2
9	Indonesia	632.1	1.8	8.8	4.4
10	Saudi Arabia	579.9	1.7	1.1	3.0
11	Canada	556.2	1.6	–1.7	0.4
12	South Africa	478.8	1.4	1.8	**–0.1**
13	Mexico	455.0	1.3	–2.5	0.8
14	Brazil	441.3	1.3	–0.2	1.7
15	Australia	428.3	1.3	4.2	**–0.2**
16	United Kingdom	387.1	1.1	–2.5	**–3.4**
17	Turkey	383.3	1.1	-2.2	3.6
18	Italy	325.4	1.0	–2.0	**–2.8**
19	Poland	303.9	0.9	–4.9	0.0
20	France	299.2	0.9	–2.6	**–1.8**
	Total World	34,169.0	100	0.5	1.1

Data source: BP statistical review of world energy 2020.

carbon economy index system, we preliminarily established an evaluation index system according to the development and operability of low-carbon economy.

The index system includes five dimensions: economy, society, science and technology, environment, and energy structure. For screening indicators of dimension, we searched and collected relevant information **extensively**, drew on the low-carbon development indicator system of relevant regions, provinces, and industries. Then, we sorted, summarized, classified, and summarized nearly one hundred **effective** indicators with high adoption rate, and screened out suitable indicators at the national level and **available** data (2009–2019). In addition, we solicited the opinions of low-carbon economic experts. Finally, the **correlation** test was carried out on all the variables according to the correlation principle, and the indexes that were completely correlated with each other were eliminated. A total of 40 indexes covering 5 dimensions were retained to ensure that the above five principles were met.

As shown in **Table 1**, the index system mainly includes three levels: target layer, criterion layer, and indicator layer (Fu et al., 2011). The target layer is the international low-carbon economy development level, the criterion layer includes five dimensions of economy, society, science and technology, environment, and energy structure, and the index layer includes 11 first-level indicators and 40 second-level indicators.

It is worth noting that the positive and negative correlation (**Table 1**) between the indexes and low-carbon economy is limited to the general economic laws, and the specific and detailed change laws are not within the scope of this article.

In accordance with the principles of openness, reliability, and consistency in the process of data collection, basic data from open channels were used as much as possible in this article. The data of the 40 development indexes of the above 20 countries were mainly derived from the World Bank database, International Energy Agency (IEA), U.S. Energy Information Administration (EIA), the World Economic Yearbook, the BP Statistical Review of World Energy, the report of the United Nations Food and Agriculture Organization, and other relevant statistics.

To maintain the authenticity, accuracy, and availability of the data, for the difference in the updating time of different indexes in the statistical data, data of 2019 were selected uniformly in this article for comparative analysis. For statistical data differences caused by different statistical calibers, the World Bank database shall prevail in this article. For the default values, the method of substitution of adjacent years, or averaging or substitution of similar countries were adopted.

EMPIRICAL ANALYSIS

Selection of Representative Countries

This article selects the world's top 20 CO_2 emitters in 2019, and the total carbon emissions of these countries reach nearly 80% of the world's total carbon emissions. **Table 2** lists the proportion and growth rate of CO_2 emissions in 20 countries in 2019, as well as in the past 10 years.

In **Table 2**, the global CO_2 emission in 2019 reaches 34.17 billion tons, among which China (28.76%), the United States (14.53%), and India (7.26%) account for nearly half of the global carbon emissions. In 2019, global carbon emissions grow by 0.5%, less than half the average growth rate of 1.1% over the past decade. Nine countries, including the United States, Russia, Japan, the United Kingdom, and some EU countries, have experienced long-term negative growth in their carbon emissions, which means that they have reached "carbon peak." With economic growth slowing in 2019 and some of the one-off factors driving energy demand easing in 2018, the growth of energy markets across the world has slowed, especially in the US and Russia, where carbon emissions' growth has fallen back from 2.61 and 4.19% in 2018 to −2.97 and −1.02% in 2019, respectively. With the exception of China, its carbon emissions are still growing at a faster pace (3.4%) in 2019, indicating a good economic growth.

Principal Component Factor Analysis

As we know, there are numerous indicators or variables that reflect the low-carbon economy development of a nation. It is necessary to reduce the data dimension and the complexity of problem analysis. Moreover, there are some structures or dimensions in the data that exist but cannot be observed directly, or variables that have, between themselves, relatively high correlation coefficients, so new variables that capture the joint features of the original variables are desired to be established for subsequent multivariate analyses.

Factor analysis is a multivariate technique that tries to identify a relatively small number of factors that represent the joint behavior of interdependent original variables. Each one of these new variables is called **common factor**, which can be understood as the cluster of variables from the previously established criteria (Fávero and Belfiore, 2019). Among the methods used to determine factors, the one known as **principal components** is, without a doubt, the most widely used in factor analysis, because it is based on the assumption that uncorrelated factors can be extracted from linear combinations of the original variables.

The principal component factor analysis has four main objectives: 1) to identify correlations between the original variables to create factors that represent the linear combination of those variables (structural reduction); 2) to verify the validity of the previously established constructs, bearing in mind the allocation of the original variables to each factor; 3) to prepare rankings by generating performance indexes from the factors; and 4) to extract orthogonal factors for future use in confirmatory multivariate techniques that need the absence of multicollinearity (Fávero and Belfiore, 2019).

The Model

Let us assume a dataset that has$n$ countries, and for each country$i\,(i = 1, \ldots, n)$, values corresponding to each one of the$p$metric variables$X$. And there is a strong correlation between these variables, then the basic matrix form of the factor model can be expressed as:

$$\mathbf{X} - \boldsymbol{\mu} = \mathbf{L}\mathbf{F} + \boldsymbol{\varepsilon}, \tag{1}$$

where$\mathbf{X} = (X_1, X_2, \cdots, X_p)'$, $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_p)'$, $\mathbf{L} = \begin{bmatrix} l_{11} & l_{12} & \ldots & l_{1m} \\ l_{21} & l_{22} & \ldots & l_{2m} \\ \ldots & \ldots & & \ldots \\ l_{p1} & l_{p2} & \ldots & l_{pm} \end{bmatrix}$, $\mathbf{F} = (F_1, F_2, \ldots, F_m)'$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_i)'$,

and where $X_i - \mu_i$represents the i^{th} standardized variable$X$, $i = 1, 2, \ldots, p$. F_j represents thej^{th} extracted **principal factor**, $j = 1, 2, \ldots, m$, and usually m is much less than p. l_{ij} is the coefficient value of factor F_j, which represents the load of the i^{th} variable on thej^{th} factor (**factor loading**). ε_i represents the **special factor** or error of the i^{th} variable, and is the part that cannot be explained by principal factors.

Hypotheses are that

$$E(\mathbf{F}) = \mathbf{0}, \operatorname{cov}(\mathbf{F}, \mathbf{E}) = E\left(\mathbf{F}\mathbf{F}'\right) = \mathbf{I}, \tag{2}$$

$$E(\varepsilon) = \mathbf{0}, \operatorname{cov}(\varepsilon, \varepsilon) = E\left(\varepsilon\varepsilon'\right) = \boldsymbol{\Psi} = \begin{bmatrix} \psi_1 & 0 & \ldots & 0 \\ 0 & \psi_2 & \ldots & 0 \\ \ldots & \ldots & & \ldots \\ 0 & 0 & \ldots & \psi_p \end{bmatrix}, \tag{3}$$

$$\operatorname{cov}(\varepsilon, \mathbf{F}) = E\left(\varepsilon\mathbf{F}'\right) = \mathbf{0}. \tag{4}$$

In this article, we discuss the orthogonal factor model (1), which satisfies hypothesis (2)–(4).

Next, when choosing the number of factors, only the factors that correspond to eigenvalues greater than one are considered. This criterion is often used and known as the latent root criterion or Kaiser criterion. Also, these extracted factors have respective proportions of variance shared by the original variables and the first factorF_1, formed by the highest proportion, is also called principal factor. In general, when the cumulative proportion of variance reaches more than 80%, it can be thought that these extracted factors are enough to explain the original variables.

Next, the **principal components method** is used to calculate the factor loadings, which simply are Pearson correlations between the original variables and each one of the factors. This method expresses the factor **F** (Expression 5) in the linear form of the variable**X**, so that the variance of the variable can be explained by the principal component, which is suitable for the situation where the least variable is used to explain as much variance as possible. Moreover, the total shared variance of each variable in all the extracted factors is also calculated, which is defined as **Communality**.

$$\begin{gathered} \begin{bmatrix} F_1 \\ F_2 \\ \ldots \\ F_p \end{bmatrix} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \ldots & \alpha_{1p} \\ \alpha_{21} & \alpha_{22} & \ldots & \alpha_{2p} \\ \ldots & \ldots & & \ldots \\ \alpha_{p1} & \alpha_{p2} & \ldots & \alpha_{pp} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \ldots \\ X_p \end{bmatrix}, \\ \operatorname{var}(F_i) = \boldsymbol{\alpha}_i'\boldsymbol{\Sigma}\boldsymbol{\alpha}_i, \quad i = 1, 2, \ldots, p, \\ \operatorname{cov}\left(F_i, F_j\right) = \boldsymbol{\alpha}_j'\boldsymbol{\Sigma}\boldsymbol{\alpha}_j, \quad i, j = 1, 2, \ldots, p. \end{gathered} \tag{5}$$

where $\boldsymbol{\alpha}_i = (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{ip})'$and $\boldsymbol{\alpha}_\mathbf{i}'\boldsymbol{\alpha}_\mathbf{i} = 1$. Under this constraint condition, the principal component is solved by maximizing the variance of the linear function $F_1 = \boldsymbol{\alpha}_1'\mathbf{X}$.

Next, to better visualize the variables represented by a certain factor, we can think about a rotation around the origin of the originally extracted factor **F**, so that we can bring the points corresponding to variable **X** closer to one of the new factors. Even though there are several factor rotation methods, the orthogonal rotation method, also known as Varimax, whose main purpose is to minimize the number of variables that have high loadings on a certain factor through the redistribution of the factor loadings and maximization of the variance shared in factors that correspond to lower eigenvalues, is the most frequently used (Fávero and Belfiore, 2019).

Based on Expression 1, the new rotated factor is expressed as a linear combination of the original variables again, and **factor scores** are calculated.

$$\widehat{\mathbf{F}} = \mathbf{BX} = \mathbf{L}'\mathbf{R}^{-1}\mathbf{X}, \tag{6}$$

where $\mathbf{B}$ is the regression coefficient that needs to be estimated; $\mathbf{X}$ is the standardized variable; $\mathbf{L}'$ is the rotated factor loading; $\mathbf{R}$ is the correlation matrix.

The Results

The overall adequacy of the factor analysis needs to be evaluated based on the **KMO statistic** and, mainly using the result of **Bartlett's test of sphericity**. These 15 index variables (**Table 4**) are the optimal combination results of the factor analysis model. The KMO statistic provides the proportion of variance considered common to all the variables present in the analysis, and by calculating, KMO = 0.628, which suggests that the overall adequacy of the factor analysis is middling. On the other hand, $Sig.\chi^2_{Bartlett} < 0.05$ allows us to reject that correlation matrix is statistically equal to identity matrix with the same dimension, at a significance level of 0.05 and based on the hypotheses of Bartlett's test of sphericity. Thus, we can conclude that the factor analysis is adequate.

In **Table 3**, based on the Kaiser criterion, only four factors that correspond to eigenvalues greater than 1 are taken into consideration, formed by sharing 81.541% of the total variance of the original variables, that is, with a total variance loss of 18.459%. After factor rotation, 22.572, 21.929, 20.610, and 16.429% of the total variance are shared to form each factor respectively, representing the weight of each factor in the total score.

Further combined with **Table 4**, it is found that after rotation, variables X1-X4 have high loadings on the first factor, named as "**energy factor**," variables X5-X8 have high loadings on the second factor, named as "**society factor**," variables X9-X12 have high loadings on the third factor, named as "**economy factor**," variables X13-X15 have high loadings on the fourth factor, named as "**environment factor**."

Based on Expression (6), we can calculate the **factor scores** expressions from the **loadings**. The rotated factor scores can be obtained through the estimation of four multiple linear regression models, in which a certain factor is considered to be a dependent variable in each one of them, and as explanatory variables, the standardized variables. For example, we are able to write the expressions for factor F1 as follows:

$$\widehat{F}_1 = 0.298 \cdot X_1 - 0.33 \cdot X_2 + \cdots + 0.078 \cdot X_{15}$$

The four **factor scores** of each country are shown in **Table 5**, with higher scores leading to higher rankings. However, for the energy factor, the lower the score, the smaller the energy consumption, the more in line with the requirements of low-carbon economy, so the higher the ranking.

Finally, a well-accepted criterion that is used to form **integrated rankings** from factors is known as **weighted rank-sum criterion**. In this criterion, for each country, the values of all the extracted factors obtained weighted by the respective proportions of shared variance are added, with the subsequent ranking of the countries based on the results obtained. In **Table 5**, for example:

$$Score_{China} = 22.572 \cdot Score^{energy}_{China} + 21.929 \cdot Score^{society}_{China} + 20.610 \cdot Score^{economy}_{China} + 16.429 \cdot Score^{environment}_{China}$$

Result Analysis

From **Tables 4**, **5**, we can see that the first factor has a relatively high factor loading in the four indicators, such as energy consumption, power generation, and CO_2 intensity, which indicate that the low-carbon economy is first and most significantly affected by the energy consumption and structure. In the ranking of "energy factor," **France (−3.25), Brazil (−1.20), and Canada (−1.13)** rank the top three, indicating that these three countries have the best performance in energy factor. Combined with the statistics of the World Bank in 2019, the main reasons are as follows: Brazil and Canada have abundant water resources, and their hydropower generation accounts for about 60% of the total electricity generation; nuclear power accounts for 77.63% of France's electricity generation. Thus, these three countries are relatively low in fossil energy dependence and CO2 intensity. On the contrary, **Japan, Australia, South Africa, Poland, and Iran** rank at the bottom, whose fossil energy consumption accounts for about 90%, since they have basically given up nuclear power generation, or lack of domestic water resources or abundant fossil resources, respectively.

The second factor has a relatively high factor loading in the four indicators, such as R&D expenditure and researchers, school enrolment ratio, and poverty ratio, which indicates that the low-carbon economy is secondary affected by science and technology, education, social security, and other social factors. In the ranking of "society factor," **South Korea (1.64), Australia (1.37), and United States (0.82)** rank the top three, indicating that as developed countries, they have higher levels of science, technology, education, and income, and can realize low-carbon production and life style. For example, factories use more advanced low-carbon production technology and equipment, and residents generally accept low-carbon and environmentally friendly life style and have the economic ability to take actions and implement it. On the contrary, **Indonesia, India, and South Africa**, as developing countries, rank at the bottom. Since the government does not invest enough in research and education, citizens are too poor to attend higher education institutions, and factories are lack of high-tech talent and high-tech enterprises, which hinder the transition to low-carbon economy severely.

The third factor has a relatively high factor loading in the four indicators, such as GDP growth rate, services value added, gross fixed capital formation, and the urbanization ratio, which indicates that the low-carbon economy is based on the economic foundation and vitality. In the ranking of "economy factors," **China (2.34), India (2.08), and South Korea (1.20)** rank the top three, while **Brazil, the United Kingdom, and South**

TABLE 3 | Extracted principal components and total variance explained.

Component	Initial eigenvalues			Rotation sums of squared loadings		
	Total	% Of variance	Cumulative %	Total	% Of variance	Cumulative %
F_1	6.326	42.173	42.173	3.386	**22.572**	22.572
F_2	2.706	18.042	60.215	3.289	**21.929**	44.502
F_3	2.021	13.471	73.687	3.092	**20.610**	65.112
F_4	1.178	7.854	**81.541**	2.464	**16.429**	81.541

Africa rank the bottom three. In 2019, the world economy faces downward pressure due to a combination of major uncertainties, such as trade frictions, trade protection, geopolitics, and recession risks. Regardless of the backdrop of weak global business confidence and investment motivation, emerging economies, such as China, India, and South Korea, are likely to gain momentum (Zhi, 2020).

The fourth factor has a relatively high factor loading in the three indicators, such as forest coverage rate, PM2.5 concentration, and labor force participation ratio, which indicates that the low-carbon economy cannot be separated from a country's environmental carrying capacity. In the ranking of "environment factors," **Japan (1.94), South Korea (1.19), and Indonesia (1.14)** rank the top three, mainly due to their extremely high proportion of forest area, which reaches 68.5, 63.4, and 49.9%, respectively. In addition, the three countries are all island countries or peninsulas, where the air is highly mobile and the PM2.5 concentration is relatively low. On the contrary, **Saudi Arabia and Iran** are at the bottom, with weak environmental carrying capacity due to their arid deserts, low forest cover, oil production, and high levels of PM2.5. In addition, the labor force participation rate is included as an environmental factor. According to the test, the labor force participation rate is significantly negatively correlated with the PM2.5 concentration at the level of 1%, and significantly positively correlated with the forest coverage rate at the level of 5%.

In terms of overall scores, the top five countries are **South Korea (84.51), France (75.19), China (49.73), Canada (29.04), and Germany (24.13)**, the last three are **Mexico, Iran, and South Africa**. Combined with the statistical data of 2019 and the factor scores in **Table 5**, every country has different status quo and advantages of the low-carbon economy.

South Korea, which ranks first overall, also ranks among the world's top three in terms of technology, environment, and economy. While its population is just 51million, the per capita income is $31,400. In 2019, industrial output ranked the sixth in the world, with manufacturing and service industries as the main industries, particularly shipbuilding, automobile, electronics, steel, textile, and other industries ranked among the world's top 10 in output, and semiconductor sales ranked the first in the world, and tourism was also relatively developed. Moreover, South Korea attaches great importance to the development of education and science and technology, such as high-speed Internet services, the aerospace industry, robot, and biotechnology, which are highly competitive in the world. However, South Korea is at a relative disadvantage in terms of energy factors in developing low-carbon economy, due to its small land area, few mineral resources, lack of natural resources, and dependence on imports of major industrial raw materials.

Interestingly, **France**, which ranks second overall, ranks first in energy factor, in complete contrast to South Korea. France has low CO_2 intensity, due to its high use of clean energy. It has closed all iron and coal mines, fully exploited hydropower and geothermal resources, and even approximately 78% of electricity is provided by nuclear power. Its GDP ranks seventh in the world; the service sector employees account for approximately 77% of the total labor force; it is the world's largest tourist reception country, but also the world's consumption center, due to developed business. In addition, in the world, it ranks second in nuclear power equipment capacity, petroleum and petroleum processing technology, third in aviation and aerospace industry, and sixth in steel and textile industry. It is also a high-welfare country with a well-developed social insurance system.

China's Ranking

China, which ranks third overall, ranks first in **economy factor.** The rapid accumulation of capital has become the most important factor for a country's economic growth, since natural resources are limited by land area and labor force is restricted by population growth rate. 1) Based on these three factors of production, China is the fourth largest in land area and the first largest in population in the world. 2) China's GDP growth rate is 6.81%, and gross fixed capital formation accounts for 42.29% of GDP, compared with 2 and 20% in most developed countries, respectively.

The disadvantages of China's low-carbon economy are also obvious. In terms of the other two indicators of "economy factors," China's added value of services and the urbanization rate, rank fourth from the bottom and third from the bottom among the 20 countries, respectively. 1) Although China has advantages in production factors, it obviously does not allocate the factors in the tertiary industry with higher added value, and the industrial structure is unreasonable. 2) In addition, although the urbanization rate has increased from 10.64% in 1949 to 59.58% in 2019, with an average annual increase of 0.71 percentage points, making it the largest and fastest urbanization in the history of the world, there is insufficient support in basic areas and frequent problems in urban development.

China ranks 9th in **society factor**. In the process of economic growth, technological progress can break the law of declining returns on capital while accumulating capital and maintaining

TABLE 4 | Rotated factor loadings matrix and component score coefficient matrix.

Indexes		**F_1**		**F_2**		**F_3**		**F_4**	
		Load	**Coef**	**Load**	**Coef**	**Load**	**Coef**	**Load**	**Coef**
X1	Fossil fuel energy consumption	**0.897**	0.298	0.260	0.129	–0.056	–0.069	–0.243	–0.039
X2	Alternative and nuclear energy consumption	**–0.894**	–0.330	0.293	0.127	–0.094	0.092	0.081	–0.174
X3	Electricity production from oil, gas and coal	**0.884**	0.305	–0.183	–0.051	0.241	–0.011	–0.101	0.138
X4	CO_2 intensity	**0.852**	0.265	–0.012	0.051	0.240	0.019	–0.221	0.024
X5	R&D expenditure	–0.144	0.002	**0.674**	0.200	0.069	0.152	0.569	0.182
X6	Researchers in R&D	–0.198	0.002	**0.757**	0.224	–0.127	0.087	0.491	0.108
X7	Tertiary education enrollment	0.048	–0.002	**0.885**	0.382	–0.076	0.103	0.001	–0.173
X8	Poverty headcount ratio	0.072	–0.001	**–0.809**	–0.280	0.453	0.058	–0.063	0.143
X9	GDP growth rate	0.164	–0.036	–0.189	0.072	**0.852**	0.318	–0.130	–0.007
X10	Services, value added	-0.262	-0.002	0.324	–0.004	**-0.653**	–0.202	0.258	0.041
X11	Gross fixed capital formation	0.053	–0.037	0.085	0.144	**0.890**	0.389	0.163	0.095
X12	Urban population ratio	–0.070	0.044	0.586	0.130	**–0.664**	–0.184	0.142	–0.049
X13	Labor force participation ratio	–0.180	0.066	0.349	–0.010	–0.128	0.020	**0.737**	0.345
X14	PM2.5 mean annual exposure	0.186	–0.093	–0.150	0.148	0.575	0.202	**–0.626**	–0.316
X15	Forest area ratio	0.194	0.078	–0.020	–0.174	0.014	0.024	**0.840**	0.482

TABLE 5 | Rankings and scores of low-carbon economy development level.

Country	**Low-carbon**		**Energy factor**		**Society factor**		**Economy factor**		**Environment factor**	
	Ranking	**Score**	**Ranking**	**Score**	**Ranking**	**Score**	**Ranking**	**Score**	**Ranking**	**Score**
South Korea	1	84.51	7	–0.18	**1**	**1.64**	3	1.20	2	1.19
France	2	75.19	**1**	**–3.25**	7	0.51	8	-0.04	16	–0.51
China	**3**	**49.73**	**13**	**0.47**	**9**	**0.32**	**1**	**2.34**	**9**	**0.32**
Canada	4	29.04	3	–1.13	10	0.30	13	–0.42	8	0.33
Germany	5	24.13	6	–0.19	6	0.61	10	–0.22	4	0.66
United States	6	14.35	10	0.04	3	0.82	14	–0.48	7	0.43
Australia	7	7.21	19	0.93	2	1.37	9	–0.17	10	0.10
Turkey	8	3.90	11	0.12	4	0.80	5	0.56	18	–1.36
Indonesia	**9**	3.79	9	–0.11	**20**	**–1.92**	4	1.19	3	1.14
Japan	10	2.46	**20**	**1.12**	8	0.40	15	–0.63	**1**	**1.94**
India	11	1.87	4	–0.33	19	–1.65	2	2.08	17	–0.75
Russia	12	–2.23	12	0.36	12	0.15	11	–0.22	6	0.45
United Kingdom	13	–12.22	5	–0.26	11	0.27	19	–1.09	13	–0.08
Brazil	14	–12.47	2	–1.20	17	–1.18	**20**	**–1.10**	5	0.55
Poland	15	–18.82	17	0.90	14	–0.11	6	0.11	11	0.10
Italy	16	–27.73	8	-0.13	15	–0.37	17	-0.99	14	–0.13
Saudi Arabia	17	–34.46	14	0.58	5	0.69	7	–0.01	**20**	**–2.22**
Mexico	18	–51.24	15	0.59	16	–1.04	16	–0.77	12	0.04
Iran	19	–54.46	16	0.77	13	0.00	12	–0.34	19	–1.84
South Africa	20	-82.54	18	0.92	18	–1.60	18	–1.00	15	–0.36

high enthusiasm for capital accumulation, which is also conducive to product innovation and industrial upgrading. 1) China spends 2.13% of its GDP on R&D, much less than South Korea (4.5%), Japan (3.2%), and Germany (3.04%). 2) China has 1,234 R&D researchers per million people, compared with 7,514 in South Korea and 5,304 in Germany. 3) As the support of education is undoubtedly behind the talents, the enrollment rate of Chinese colleges and universities is 50.6%, while that of South Korea is 94.35%. 4) In addition to the urgent need for national investment in education, poverty is a top priority. According to the World Bank's standard of $5.50/day, China's poverty rate is 27.2%, while that of developed countries is only 0.2–3.5%. China's GDP per capita is $7,752, only 14% of that of the United States. Therefore, compared with developed countries, China still has a higher proportion of poor population, insufficient development of higher education, and shortage of research funds and researchers, which leads to the relatively backward pace of low-carbon economy development.

China ranks 9th in **environment factor**. 1) China's forest coverage rate is only 22.35%. On the one hand, because of the serious desertification, rocky desertification, and soil erosion in northwest China, on the other hand, because of the high proportion of domestic agriculture, there is a great demand for water resources and arable land. 2) Meanwhile, China's rich coal, poor oil, a little gas, and energy-intensive industries have resulted in a PM2.5 concentration of 52.66, compared with 7.41 in the United States.

China ranks 13th in **energy factor**. China is the world's largest energy consumer, accounting for 24% of global energy consumption and 34% of global growth in energy consumption in 2019. 1) In the primary energy consumption, fossil fuels account for 87.67% of energy consumption, among which coal account for 58%. 2) Fossil fuels account for 72.96% of electricity generation. 3) Due to China's heavy reliance on fossil fuels, CO_2 intensity is as high as 3.37, compared with 1.25 in France. 4) China has been optimizing its energy structure for many years, with coal consumption accounting for 58%, down from 72% a decade ago. In 2019, renewable energy consumption grows 29%, accounting for 45% of global growth.

CONCLUSIONS AND SUGGESTIONS

Research Conclusions

This article focuses on the low-carbon economy development evaluation indicators at the nation level, taking the world's top 20 countries in CO_2 emissions as the research observations, closely concentrating on the concept of low-carbon economy, based on the five principles of index evaluation (comprehensiveness, effectiveness, applicability, correlation, and foresight), 40 indicators were selected from five dimensions of economy, society, science and technology, environment, and energy structure.

Since there are so many indicators to measure the development of a country's low-carbon economy, we need to reduce the data dimension. Factor analysis is a multivariate technique that tries to identify a relatively small number of factors that represent the joint behavior of interdependent original variables. Thus, by using correlation coefficients to group variables, four factors, energy factor, society factor, economy factor, and environment factor, are generated and extracted. Then, based on the factor scores, the ranking of the four factors and the total score of 20 countries are given. In the end, South Korea, France, China, Canada, and Germany ranked among the world's top five countries in terms of low-carbon economy development and competitiveness.

Furthermore, through the evaluation index system of international low-carbon economy development, we have clearly identified the strengths and weaknesses of the 20 countries in developing a low-carbon economy, which will help China to define its own position, discover its own problems, identify the right development direction, learn useful experience, draw lessons from the experience, and avoid repeating the same mistakes.

Measures and Suggestions

Overall, China's low-carbon economy development is in a relatively favorable position, ranking the third in the world. It is most prominent in terms of economic strength (No. 1), but underperforms in terms of social development (No. 9) and environmental carrying capacity (No. 9). The biggest weakness is in the energy structure (No. 13).

Taking into account China's national conditions, development stage, sustainable development strategy, and international responsibility, China should accelerate the development of low-carbon economy, focus on key points, strengthen weak areas, and refine various indicators and tasks.

1) Promoting high-quality, efficient, and steady economic development. We will accelerate the development of advanced manufacturing and modern service industries, and apply the concept of low-carbon development to the whole process of urban planning, construction and management, and raise people's income through targeted poverty alleviation and full employment.
2) Strengthening support for science, technology, and human resources. For developing countries, in a relatively short period of time and at a lower cost, for realizing low-carbon technological innovation, we can not only introduce, imitate, and purchase patents, but also need to strengthen the research and development new technologies such as energy saving and consumption reduction, renewable energy and advanced nuclear energy, carbon capture, utilization, and storage. And most importantly, giving priority to education.
3) Increasing carbon sink and reducing environmental pollution. We will continue to take action to prevent and control air pollution from its source with all the people, and build an environmental governance system in which the government plays the leading role, enterprises play the main role, and social organizations and the public participate.
4) Building a low-carbon energy system and forming an energy-saving and low-carbon industrial system. In 2019, China's energy structure continued to improve: the proportion of coal consumption in the primary energy reached a record low (57.7%); renewable energy consumption grew 14.2%, accounting for 26% of global growth. China's electricity generation accounted for 96% of net global growth. Compared with 2018, solar power generation increased by 26.5%, wind power by 10.9%, biomass and geothermal energy by 9.7%, and water power by 5.9%. Nuclear power generation grew by 18.2%, higher than the 10 years average growth rate (+15%), and China accounted for 56% of the global increase (Dudley Bob, 2020). China will continue to follow a new path of industrialization, develop the circular economy, improve the industrial structure, strictly control the expansion of industries that have high emissions and energy intensive, speed up the phasing out of backward production facilities, and vigorously develop the service sector and strategic emerging industries.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

REFERENCES

Azizalrahman, H., and Hasyimi, V. (2018). Towards a Generic Multi-Criteria Evaluation Model for Low Carbon Cities. *Sustain. Cities Soc.* 39, 275–282. doi:10.1016/j.scs.2018.02.026

Bao, J. Q., Miao, Y., and Chen, F. (2008). Low Carbon Economy: Revolution in the Way of Human Economic Development. *China Ind. Econ.* 241 (4), 153–160. doi:10.19581/j.cnki.ciejournal.2008.04.018

Cao, Y. P. (2018). Research on the Evaluation Method of Low Carbon Economy index System under the Background of Supply-Side Reform. *Knowledge Economy* 10, 18–19. doi:10.15880/j.cnki.zsjj.2018.10.009

Dagoumas, A. S., and Barker, T. S. (2010). Pathways to a Low-Carbon Economy for the UK with the Macro-Econometric E3MG Model. *Energy Policy* 38 (6), 3067–3077. doi:10.1016/j.enpol.2010.01.047

Duan, Y., Mu, H., Li, N., Li, L., and Xue, Z. (2016). Research on Comprehensive Evaluation of Low Carbon Economy Development Level Based on AHP-Entropy Method: a Case Study of Dalian. *Energ. Proced.* 104, 468–474. doi:10.1016/j.egypro.2016.12.079

Dudley, B. (2020). *BP Statistical Review of World Energy*, 69th ed.; London, UK: BP.

Fan, H. M., Xu, Z. L., and Zhang, R. (2018). The Urban Low-Carbon Traffic Evaluation index System Based on DPSIR Model: the Case of Dalian. *Ecol. Economy* 34 (04), 64–69.

Fávero, L. P., and Belfiore, P. (2019). "Principal Component Factor Analysis," in *Data Science for Business and Decision Making*. Editors L. P. Fávero and P. Belfiore, Cambridge, MA: Academic Press, 383–438. doi:10.1016/B978-0-12-811216-8.00012-4

Fu, J. F., and Liu, X. M. (2010). A Framework for China's Low Carbon Economy on the Basis of Scenario Analysis and Discussion on Relevant Issues. *Resour. Sci.* 32 (2), 205–210.

Fu, J. F., Zheng, L. C., and Cheng, X. L. (2011). China's Low-Carbon Economic Development: an Inter-provincial and International Comparison. *Resour. Sci.* 33 (04), 664–674.

Fu, J. F., Zhuang, G. Y., and Gao, Q. X. (2010). Conceptual Identification and Evaluation index System for Low Carbon Economy. *China Popul. Resour. Environ.* 20 (8), 38–43. doi:10.3969/j.issn.1002-2104.2010.08.007

Fu, Y., Ma, Y. H., Liu, Y. J., and Niu, W. Y. (2008). Development Patterns of Low Carbon Economy. *China Popul. Resour. Environ.* 18 (3), 14–18.

Government, H. M. (2009). *The UK Low Carbon Transition Plan: National Strategy for Climate and Energy*. London: Department of Energy & Climate Change. TSO (The Stationery Office) http://www.decc.gov.uk/.

He, J. K., Zhou, J., Liu, B., and Sun, Z. Q. (2010). Global Trends of Low Carbon Economy and China's Responses. *World Econ. Polit.* 9 (04), 18–35+156.

Hughes, N., and Strachan, N. (2010). Methodological Review of UK and International Low Carbon Scenarios. *Energy policy* 38 (10), 6056–6065. doi:10.1016/j.enpol.2010.05.061

Kinzig, A. P., and Kammen, D. M. (1998). National Trajectories of Carbon Emissions: Analysis of Proposals to foster the Transition to Low-Carbon Economies. *Glob. Environ. Change* 8 (3), 183–208. doi:10.1016/S0959-3780(98)00013-2

Lan, Q. X., and Zheng, X. D. (2013). Study on index System and International Evaluation of China's Low-Carbon Economy Development Level: an Observation Based on the Comparison of G20 Countries. *J. Beijing Normal Univ. (Social Sciences)* 58 (02), 135–144.

Lu, S. H, and Zhu, Q. G. (2013). Review on the Evaluation System of China's Low Carbon Economy. *Mod. Manag. Sci.* (12), 12–14. doi:10.4236/lce.2013.41002

Luo, Z.-q., and Tong, X.-f. (2011). "Evaluation on Development Capability of Low-Carbon Economy and Countermeasures in China," in Procedia Environmental Sciences in Proceedings of 2011 3rd International Conference on Environmental Science and Information Application Technology (ESIAT 2011). Xian, China, August 20-21, 2011, 902–907. doi:10.1016/j.proenv.2011.09.144

Lv, X. D., WangHuang, Y. P. C., and Sun, J. (2013). Research on Assessment Method on index System of Low-Carbon Economy. *China Popul. Resour. Environ.* 23 (07), 27–33. doi:10.3969/j.issn.1002-2104.2013.07.005

Pan, J. H., Zhuang, G. Y., Zheng, Y., Zhu, S. X., and Xie, Q. Y. (2010). Clarification of the Concept of Low-Carbon Economy and Analysis of its Core Elements. *Int. Econ. Rev.* 18 (04), 88–101+5.

Pei, X. J., and Tan, Y. (2013). Research Progress of Urban Low Carbon Economic Development Evaluation. *Stat. Decis.* 29 (24), 30–34. doi:10.13546/j.cnki.tjyjc.2013.24.018

Shi, X. F., Sun, Y., and Cui, Y. (2018). Evaluation of Low Carbon Economic Development Level in Tianjin Based on Entropy Principal Component Analysis. *Sci. Tech. Manag. Res.* 38 (03), 247–252. doi:10.3969/j.issn.1000-7695.2018.03.037

Strachan, N., Foxon, T., and Fujino, J. (2008). *Modelling Long-Term Scenarios for Low Carbon Societies (Climate Policy)*. London: Earths can Publications Ltd.

Tao, W. (2017). Multi-point Cooperative Multicast Video Design and Research. *Coop. Economy Sci.* 33 (20), 36–38. doi:10.1109/iccnea.2017.36

Vivid Economics (2009). *G20 Low Carbon Competitiveness*. London: The Climate Institute. 1–52.

Wang, X. Y., and Pan, J. Y. (2019). Evaluation of Development Level of Low Carbon Economy in Manufacturing Industry of Shaanxi Province Based on Drift. *Sci. Tech. Manag. Res.* 39 (24), 240–246. doi:10.3969/j.issn.1000-7695.2019.24.032

Xie, Z. X., QinShen, Y. C. W., and Rong, P. J. (2017). Efficiency and Impact Factors of Low Carbon Economic Development in China. *Econ. Geogr.* 37 (03), 1–9. doi:10.15957/j.cnki.jjdl.2017.03.001

Xu, X., and Liu, C. Y. (2014). Construction and Demonstration of Urban Low Carbon Competitiveness Evaluation index System. *Stat. Decis.* 12 (21), 60–61. doi:10.13546/j.cnki.tjyjc.2014.21.016

Yan, Y. G., and Ma, M. (2015). Research on the Evaluation Indicator System of Regional Low Carbon Competitiveness Based on AHP-EM-TOPSIS Resultant Evaluation Methods-Based on the Empirical Evaluation Study of Chongqing. *Sci. Tech. Manag. Res.* 35 (7), 39–45+57. doi:10.3969/j.issn.1000-7695.2015.07.008

Yang, Y. (2012). Research on Evaluation of Sichuan Low-Carbon Economy Efficiency. *China Popul. Resour. Environ.* 22 (06), 52–56. doi:10.3969/j.issn.1002.2104.2012.06.009

Yuan, H. C., Chen, Z. M., Wang, M., Li, W. W., and Liu, X. Y. (2017). Research on the Evaluation Index System of International Low-Carbon Cities. *Sci. Tech. Economy Market* (08), 76–77.

Zhang, K. M., Pan, J. H., and Cui, D. P. (2009). *Low Carbon Development Theory*. Beijing: China Environmental Science Press.

Zhi, Y. (2020). Weak and Stable World Economy: New Variables, New Drivers and New Opportunities. *World Economy Stud.* 39 (01), 3–10. doi:10.13516/j.cnki.wes.2020.01.001

Zhong, Y. Y. (2018). Construction and Empirical Analysis of Regional Low-Carbon Economic Evaluation index System in China. *J. Nanjing Univ. Posts Telecommunications (Social Sci. Edition)* 20 (01), 93–102. doi:10.14132/j.cnki.nysk.2018.01.013

Zhou, Z. G., Zhuang, G. Y., and Chen, Y. (2018). Assessment of Low-Carbon City Development: Theoretical Basis,analysis Framework and Policy Implications. *China Popul. Resour. Environ.* 28 (6), 160–169. doi:10.12062/cpre.20180317

Zhuang, G. Y. (2005). Analysis of the Ways and Potential of China's Low Carbon Economic Development. *Stud. Int. Tech. Economy* 8 (3), 79–87.

Zhuang, G. Y. (2020). *Evaluation of Urban Low-Carbon Construction in China: Methods and Empirical Studies*. Beijing: Social Sciences Academic Press.

Zhuang, G. Y. (2007). *Low Carbon Economy: China's Development Road under the Background of Climate Change*. Beijing: China Meteorological Press.

Zhuang, G. Y., Pan, J. H., and Zhu, S. X. (2011). The Connotation of Low Carbon Economy and the Construction of Comprehensive Evaluation index System. *Econ. Perspect.* 52 (1), 132–136.

2050 Japan Low-Carbon Society Project Team (2008). Japan Scenarios and Actions Towards Low-Carbon Societies (LCS). Available at: http://2050.nies.go.jp/LCS/jpn/japan.html (Accessed Mar 3, 2021).

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Energy Research

ORIGINAL RESEARCH
published: 14 October 2021
doi: 10.3389/fenrg.2021.771433

Knowledge Mapping in Electricity Demand Forecasting: A Scientometric Insight

Dongchuan Yang[1], *Ju-e Guo*[1], *Jie Li*[2,3], *Shouyang Wang*[4,5,6] *and Shaolong Sun*[1]*

[1]School of Management, Xi'an Jiaotong University, Xi'an, China, [2]National Science Library, Chinese Academy of Sciences, Beijing, China, [3]College of Safety Science and Engineering, Liaoning Technical University, Fuxin, China, [4]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China, [5]School of Economics and Management, University of Chinese Academy of Sciences, Beijing, China, [6]Center for Forecasting Science, Chinese Academy of Sciences, Beijing, China

Electricity demand forecasting plays a fundamental role in the operation and planning procedures of power systems, and the publications related to electricity demand forecasting have attracted more and more attention in the past few years. To have a better understanding of the knowledge structure in the field of electricity demand forecasting, we applied scientometric methods to analyze the current state and the emerging trends based on the 831 publications from the Web of Science Core Collection during the past 20 years (1999–2018). Employing statistical description analysis, cooperative network analysis, keyword co-occurrence analysis, co-citation analysis, cluster analysis, and emerging trend analysis techniques, this study gives a comprehensive overview of the most critical countries, institutions, journals, authors, and publications in this field, cooperative networks relationships, research hotspots, and emerging trends. The results can provide meaningful guidance and helpful insights for researchers to enhance the understanding of crucial research, emerging trends, and new developments in electricity demand forecasting.

Keywords: electricity demand forecasting, scientometric, visualization, citespace, knowledge mapping

OPEN ACCESS

Edited by:
Jian Chai,
Xidian University, China

Reviewed by:
Wendong Yang,
Shandong University of Finance and Economics, China
Tong Niu,
Zhengzhou University, China

****Correspondence:***
Shaolong Sun
sunshaolong@xjtu.edu.cn

Specialty section:
This article was submitted to Smart Grids, a section of the journal Frontiers in Energy Research

Received: *06 September 2021*
Accepted: *27 September 2021*
Published: *14 October 2021*

Citation:
Yang D, Guo J-e, Li J, Wang S and Sun S (2021) Knowledge Mapping in Electricity Demand Forecasting: A Scientometric Insight.
Front. Energy Res. 9:771433.
doi: 10.3389/fenrg.2021.771433

INTRODUCTION

Nowadays, electricity is the most critical energy and plays an indispensable role in many fields. In recent years, a large number of researchers have proved that the accuracy of electricity demand forecasting is the basis of power system planning and operation (Raza and Khosravi, 2015; Kuster et al., 2017). Accurate electricity demand forecasting can not only ensure the reliable operation of power systems but also have an excellent cost-saving potential for power corporations (Al-Ghandoor et al., 2009).

With the increase of electricity demand and the rapid development of artificial intelligence, electricity demand forecasting has attracted more and more attention, and new research methods, emerging trends, and new developments have emerged at the same time (Alfares and Nazeeruddin, 2002). A lot of forecasting techniques and researches have been proposed and applied in electricity load forecasting (Hippert et al., 2001; Bourdeau et al., 2019), and support vector regression (Mohandes, 2002; Sousa et al., 2014) and ANN (Bhattacharyya and Thanh, 2004; Cavallaro, 2005) are widely used in recent years. In addition, more and more hybrid models are applied in electricity load forecasting. Mohan et al. (2018) applied dynamic mode decomposition (DMD) to extract the spatiotemporal dynamic characteristics of power loads that change with time and forecasted future electric load. Al-Musaylh et al. (2019) presented a hybrid model that including multivariate adaptive regression, and multiple linear regression, artificial neural network models to forecast short-term electricity demand in Australia.

In the past, many scholars had reviewed the methods, techniques, and methods of evaluation in the field of electricity demand forecasting. Shao et al. (2017) conducted decomposition methods for electricity demand forecasting and presented that Empirical mode decomposition and wavelet decomposition are the most popular technique. Kuster et al. (2017) presented a review that revealed that artificial neural networks, multivariate regression, time series analysis, and multiple linear regression are popular and effective methods for electricity and electricity forecasting. Hong et al. (2016) offered a summary of the recent research progress about probabilistic energy forecasting and introduced the Global Energy Forecasting Competition 2014 with load forecasting. However, previous review studies focused on the techniques and methods already used in power load forecasting and very little research has analyzed the collaborative relationship, new developments, and emerging trends of electricity demand prediction and visualized the knowledge map of the field.

Scientometrics is a crucial method to explore the scientific research rules, identify research trends, and evaluate the development of the field (Kim and Chen, 2015; Olawumi and Chan, 2018). Yu and Xu (2017) analyzed the current status of carbon emissions trading and discussed future research trends by the scientometric method. Olawumi and Chan (2018) evaluated the research development status of institutions, countries, and journals in the research field. Niazi and Hussain (2011) evaluated all sub-domains of agent-based computing and found agent-based computing extensive in other dominos.

With the rapid growth of attention and publications for electricity demand forecasting, it is necessary and urgent to summarize the current situation and analyze the collaborative relationship, new developments, and emerging trends of electricity demand forecasting. According to Web of Science (WoS), about 831 papers related to electricity demand forecasting have been published in the last 20 years (1999–2018), but no research has been performed to analyze and visualize the overall knowledge structure of this topic. Therefore, the purpose of this study is to assess the research on electricity demand forecasting and seek an overview of the structure of the relevant information. In this study, scientometrics analysis is performed in the electricity demand forecasting domain, and software named CiteSpace is utilized to analyze and visualize the emerging trends. CiteSpace, invented by Chen Chaomei, is a particularly popular software of scientometrics that can be used to identify knowledge areas and emerging trends in a visual form (Lairmore et al., 2000; Chen, 2006). In recent years, CiteSpace has attracted the interest of many scholars and has been applied to many fields. Chen et al. (2014) used published literature to investigate new developments and emerging trends in the field of regenerative medicine. Yang et al. (2018) comprehensively analyzed the status of PM2.5 research and found the frontiers of research in this field. Fang et al. (2018) examined the interaction between climate change and tourism and described the research characteristics of the field in the past 25 years.

The structure of this article is as follows: *Methodology* gives the source and search strategy of publications. *Basic summary of electricity demand forecasting research* introduces the basic summary of electricity demand forecasting research. In *Cooperative structure in the field of power demand forecasting*, this study visualizes the cooperation network of authors, institutions, and countries/regions. *Active topics and emerging trends* analyzes the active topics and emerging trends in electricity demand forecasting, including keyword analysis and co-citation analysis. *Conclusions* gives comprehensive conclusions and discussions.

METHODOLOGY

This section provides the search strategy of data. For the searched phrase in Web of Science (WoS), some articles perform an exact search on a certain phrase, such as Yu and Xu (2017), and some articles perform an exact search on multiple phrases and merge the results, such as Chen (2017). Searching with inexact themes requires that the query words do not have to appear consecutively, which gets a large number of publications that are not related to the search subject. It is worth noting that this article searches precise themes and non-precise titles. This article focuses on a more subdivided field, and the number of related articles is little. Searching with precise themes will ignore indispensable publications in this field and affect the conclusion of this article seriously. To improve the recall rate and avoid retrieving a large number of irrelevant publications, this article adopts the strategies of searching with precise themes and inexact titles. To ensure the accuracy of publications being retrieved, this study culled out irrelevant publications through means of manual screening.

The data used for analysis in our research is downloaded from WoS, and the search strategy followed is below:

1) TS=("electric* load forecast*" OR "electric* load predict*" OR "electric* demand forecast*" OR "electric* demand predict*" OR "electric* consumption forecast*" OR "electric* consumption predict*" OR "power load forecast*" OR "power load predict*" OR "power demand forecast*" OR "power demand predict*" OR "power consumption forecast*" OR "power consumption predict*" OR "grids load forecast*" OR "grids load predict*") OR TI=(electric* load forecast* OR electric* load predict* OR electric* demand forecast* OR electric* demand predict* OR electric* consumption forecast* OR electric* consumption predict* OR power load forecast* OR power load predict* OR power demand forecast* OR power demand predict* OR power consumption forecast* OR power consumption predict* OR grids load forecast* OR grids load predict*)
2) Databases = Science Citation Index Expanded (SCI-EXPANDED) and Social Sciences Citation Index (SSCI)

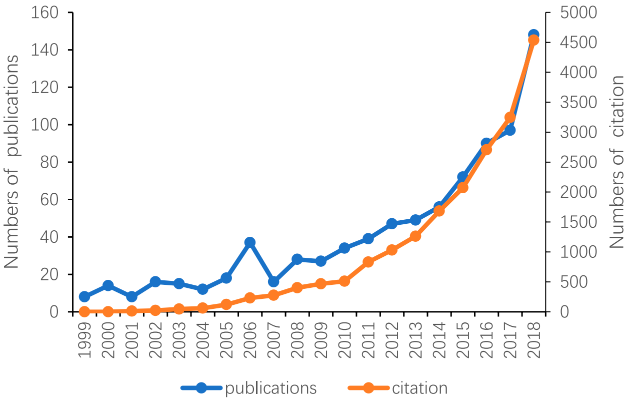


FIGURE 1 | Numbers of publications and citations in electricity demand forecasting, 1999–2018.

3) Timespan = "1999–2018"
4) Document types = "article" or "review"
5) Literature type = "English"; 901 publications are retrieved, and 70 publications that were not related to electricity demand forecasting were deleted through means of manual screening. Finally, 831 publications were downloaded on October 18, 2019.

BASIC SUMMARY OF ELECTRICITY DEMAND FORECASTING RESEARCH

This section provides statistical analysis from five parts, including distribution of time, subject categories, high-yield journals, high-yield institutions, high-yield authors, and highly cited publications in electricity demand forecasting.

The Distribution of Publications

Figure 1 shows that the number of publications in electricity demand forecasting is increasing over the past 20 years, from eight publications in 1999 to 148 publications in 2018, with steady growth in 199–2009 and rapid growth in 2010–2018. The publications have been cited 19,506 times from 1999 to 2018. The number of citations is increasing, year by year, and has similar growth trends with the numbers of publications. From this, it can be seen that electricity demand forecasting has received more and more attention, especially in the last decade.

Figure 2 shows that China, the United States, Iran, and the United Kingdom are the main countries publishing papers in this field. China is the country with the most publications, especially after 2015, the number of publications in China exceeds the sum of the United Kingdom, the United States, and Iran. It should be noted that the publications of Taiwan and Hong Kong are included in China. The numbers of publications in the United States and the United Kingdom are both fluctuating. Iran has published its first publication in 2007, and Iran has published more than three papers each year. In 1999–2018, China published 33.81% (281) of the total publications in electricity demand forecasting, the US for 9.99% (83), Iran for 6.74% (56), and the United Kingdom for 6.14% (51).

Subject Categories

Figure 3 shows that electricity demand forecasting is a cross-disciplinary research area, including energy fuels accounting for 36.82% (306), engineering electrical electric accounting for 26.23% (218), computer science artificial intelligence accounting for 16.49% (137), thermodynamics accounting for 13.48% (112) and economics accounting for 6.38% (53).

High-Yield Journals

199 journals published papers in electricity demand forecasting from 1999 to 2018 in our dataset. **Table 1** lists the top 10 journals, and it can be seen that energy and power are areas of most significant concern to the top 10 journals. "Energy" is the highest yield journal with 81 publications, followed by "Energies", "International Journal of Electrical Power Energy Systems", "Applied Energy", "Energy Conversion and Management", "Electric Power Systems Research", "Energy and Buildings", "International Journal of Forecasting", "IEEE Transactions on Power Systems", and "Lecture Notes in Computer Science". In the top 10 journals, the impact factor of "Energy", "Applied Energy", "Energy Conversion and Management", and "IEEE Transactions on Power Systems" are all more than 5.

Figure 4 shows the distribution of leading journals in electricity demand forecasting. There are 29 journals in **Figure 4** and each of them published at least five publications. We denoted with NP_j the number of publications for the journal j, $T_{i,j}$ the publication year of publication i in the journal j, $NC_{i,j}$ the number of citations for publication i in journal j from 1999 to 2018. And $AY_j = \sum_{i=1}^{NP_j} T_{i,j}/NP_j$ represents the average year of publication in the journal j, $AAC_j = \sum_{i=1}^{NP_j} \frac{NC_{i,j}}{2019-T_{i,j}}/NP_j$ represents the average annual citation for the journal j. The black horizontal dashed line in **Figure 4** represents the average annual citation of all publications in this field, and the number of an average annual citation for each journal above this line is higher than the average

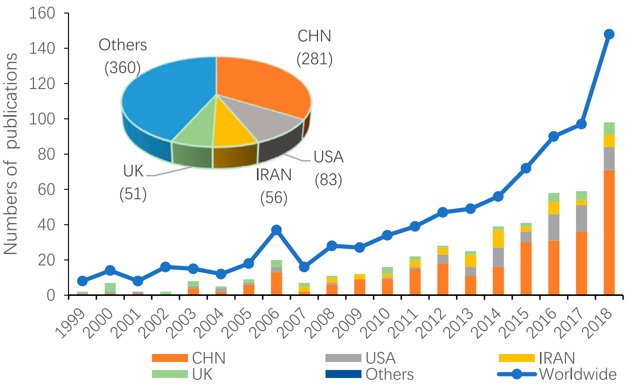


FIGURE 2 | Distribution of publications in electricity demand forecasting, 1999–2018.

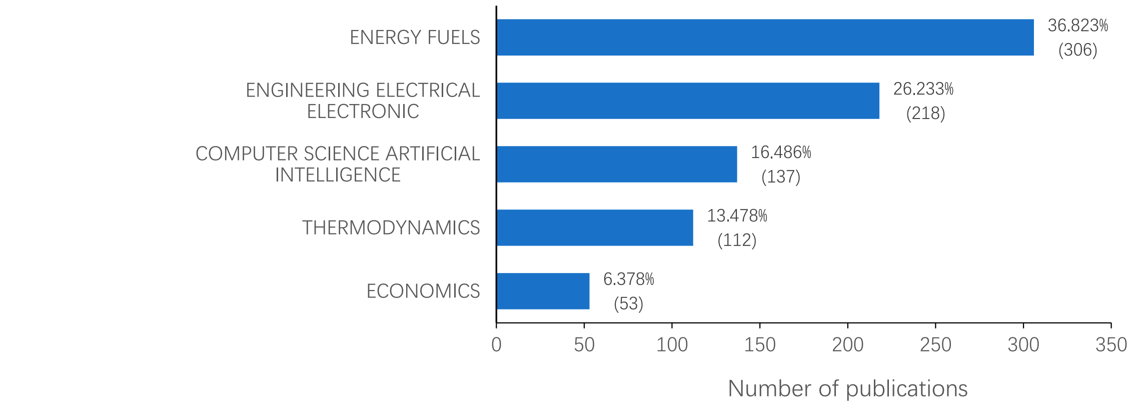


FIGURE 3 | Distribution of the main Subjects in electricity demand forecasting, 1999–2018.

TABLE 1 | High-yield journals in electricity demand forecasting.

Num	Journal	TP	Proportion (%)	If	Country
1	Energy	81	9.75	5.537	England
2	Energies	63	7.58	2.707	Switzerland
3	International Journal of Electrical Power Energy Systems	46	5.54	4.418	England
4	Applied Energy	39	4.69	8.426	England
5	Energy Conversion and Management	29	3.49	7.181	England
6	Electric Power Systems Research	26	3.13	3.022	Switzerland
7	Energy and Buildings	23	2.77	4.495	Switzerland
8	International Journal of Forecasting	23	2.77	3.386	Netherlands
9	IEEE Transactions on Power Systems	22	2.65	6.807	United States
10	Lecture Notes in Computer Science	19	2.29	0.402	United States

Note: *TP: the total number of publications; IF: Impact Factor.*

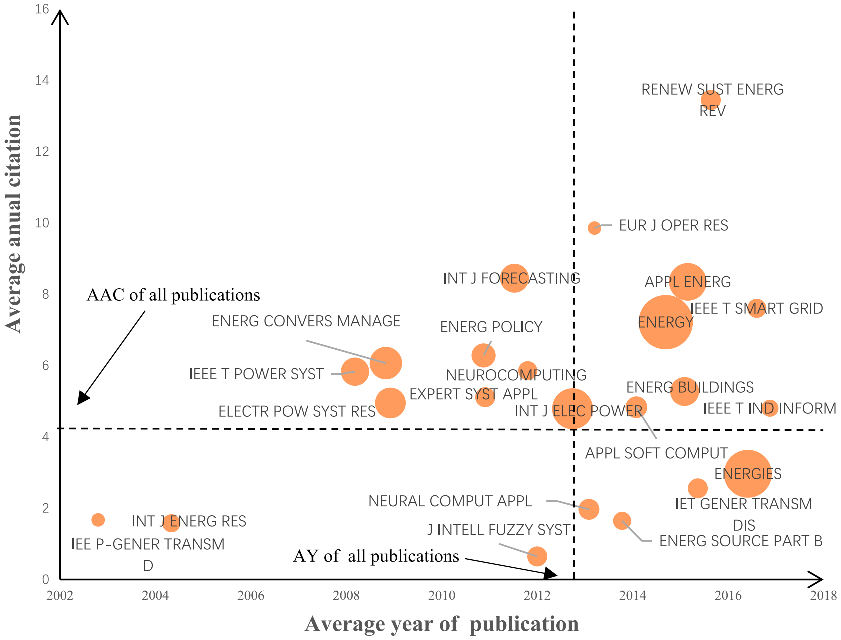


FIGURE 4 | Distribution of main journals in electricity demand forecasting, 1999–2018.

TABLE 2 | High-yield institutions in electricity demand forecasting.

Institution	Country	TP	TPC (%)	TPW (%)
North China Electric Power University	CHINA	57	20.28	6.86
Lanzhou University	CHINA	35	12.46	4.21
University of North Carolina	United States	24	29.27	2.89
Islamic Azad University	IRAN	18	32.14	2.17
Oriental Inst Technol	CHINA	15	5.34	1.81
Electricite De France Edf	FRANCE	14	56.00	1.69
University of Tehran	IRAN	14	25.00	1.69
University of Oxford	UK	13	25.49	1.56
Dongbei University of Finance Economics	CHINA	13	4.63	1.56
Hefei University of Technology	CHINA	13	4.63	1.56

Note: *TPC: publications share in its country; TPW: publications share in the world.*

TABLE 3 | High-yield authors in electricity demand forecasting.

Author	Country	TP	TC	TC/TP	MC	H-index*
Wang JZ	China	24	864	35.21	95	16
Hong WC	China	21	1,473	70.14	254	17
Niu DX	China	16	370	23.13	181	10
Hong T	United States	11	611	55.55	180	11
Azadeh A	Iran	11	490	44.55	141	8
Amjady N	Iran	10	604	60.4	181	9
Taylor JW	England	10	1,334	133.4	269	9
Yang SL	China	10	123	12.3	36	6
Goude Y	France	8	275	34.38	75	6
Che JX	China	7	199	28.43	71	6

Note: *TC: the total citations of TP; MC: the max citations of his/her one publication.*

in this field. The black vertical dashed line in **Figure 4** represents the average year of publications in this field, and the average year of publication for each journal on the right of this line is closer. The intersection of the horizontal and vertical dashed lines is (2012.76, 4.18), which means that the average year of publication for 831 publications is 2012.76, and the average number of citations for 831 publications is 4.18. The size of the dot in **Figure 4** represents the number of publications for a journal from 1999 to 2018, which means that the larger the dot, the greater the number.

Figure 4 shows the number of publications, publication time, and citations of significant journals in this field. The journals in the 1, 2, and 4 quadrants are worthy of our attention, especially the journals in the first quadrant, whose publications had been cited more in recent years (such as "Renewable and Sustainable Energy Reviews", "European Journal of Operational Research", "Applied Energy", "IEEE Transactions on Smart Grid", "Energy", "Energy and Buildings", "Applied Soft Computing", "IEEE Transactions on Industrial Informatics"). The journals in the second quadrant are likely to publish much-watched publications by 2012. Journals in the fourth quadrant published articles with low citations recently, but their articles may become hotspots in the future. There are some journals, such as "Energy" and "Energies", had published the most publications in this field.

High-Yield Institutions

848 institutions published papers in electricity demand forecasting from 1999 to 2018. **Table 2** lists the top 10 institutions, it can be seen that five of the top 10 institutions come from China and China also published the largest number of articles, which is also the same conclusion as **Figure 2**. North China Electric Power University is the highest yield Institution with 57 publications, followed by Lanzhou University, University of North Carolina, Islamic Azad University, Oriental Institute of Technology, Electricite de France edf, University of Tehran, University of Oxford, Dongbei University of Finance Economics, and Hefei University of Technology.

High-Yield Authors

Table 3 shows high-yield authors, published the most publications in this field, mainly from China, the United States, Iran, England, and France. **Table 3** shows that Wang JZ had published the most articles in this field, with 24 publications. Hong WC is the most cited author, with a total of 1,473 citations, and had 17 publications that had been cited more than 17 times. The publications of Taylor JW have cited an average of 133.4 times, and the maximum number of citations of his publications was cited 269 times.

Figure 5 shows the distribution of leading authors in electricity demand forecasting. There are 32 authors in **Figure 5** and each of them published at least five publications. Similar to **Figure 4**, *AY* in **Figure 5** represents the average year of publication for an author, *AAC* in **Figure 5** represents the average annual citation for an author, and the size of the dot represents the number of publications for an author from 1999 to 2018. The intersection of the horizontal and vertical dashed lines is (2012.76, 4.18) too. The number of average annual citations for each author above this line is higher than the average in this field, and the average year of publication for each author on the right of this line is closer.

Figure 5 shows the number of publications, publication time, and citations of leading authors in this field. The authors in the 1, 2, and 4 quadrants are worthy of our attention, especially the authors in the first quadrant, whose articles have received extensive attention in recent years, such as Zareipourh, Khosravia, Hong T, Abediniao,

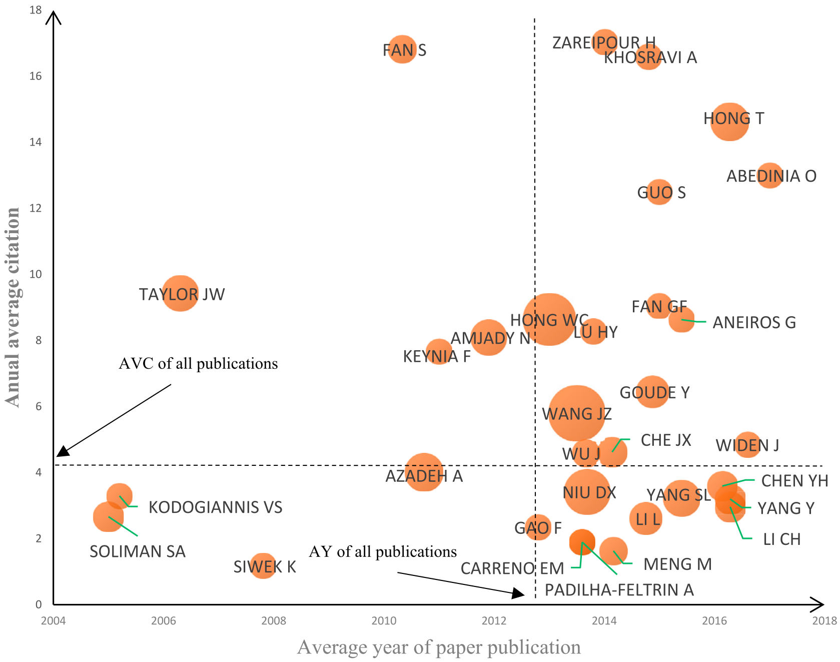


FIGURE 5 | Distribution of main authors in electricity demand forecasting, 1999–2018.

TABLE 4 | Highly cited publications in electricity demand forecasting.

Authors	Year	Journal	TC	ACY
Alfares and Nazeeruddin, (2002)	2002	International Journal of Systems Science	272	15.11
Taylor, (2003)	2003	Journal of the Operational Research Society	265	15.59
Pai and Hong, (2005)	2005	Electric Power Systems Research	239	15.93
Bunn, (2000)	2000	Proceedings of the IEEE	236	11.8
Hahn et al. (2009)	2009	European Journal of Operational Research	233	21.18
Taylor and Buizza, (2002)	2002	IEEE Transactions on Power Systems	231	12.83
Akay and Atak, (2007)	2007	Energy	223	17.15
Hsu and Chen, (2003)	2003	Energy Conversion and Management	220	12.94
Taylor et al. (2006)	2006	International Journal of Forecasting	219	15.64
Li et al. (2013)	2013	Knowledge-Based Systems	218	31.14

Note: *ACY: average citations per year.*

TABLE 5 | Highly average annual cited publications in electricity demand forecasting.

Authors	Year	Journal	TC	ACY
Hong and Fan, (2016)	2016	International Journal of Forecasting	159	39.75
Hong et al. (2016)	2016	International Journal of Forecasting	143	35.75
Ahmad et al. (2014)	2014	Renewable and Sustainable Energy Reviews	207	34.5
Raza and Khosravi, (2015)	2015	Renewable and Sustainable Energy Reviews	157	31.4
Li et al. (2013)	2013	Knowledge-Based Systems	218	31.14
Quan et al. (2014)	2014	IEEE Transactions on Neural Networks and Learning Systems	172	28.67
Mohammadi et al. (2018)	2018	Neural Processing Letters	48	24
Kaboli et al. (2017)	2017	Energy	68	22.67
Boroojeni et al. (2017)	2017	Electric Power Systems Research	66	22
Hahn et al. (2009)	2009	European Journal of Operational Research	233	21.18

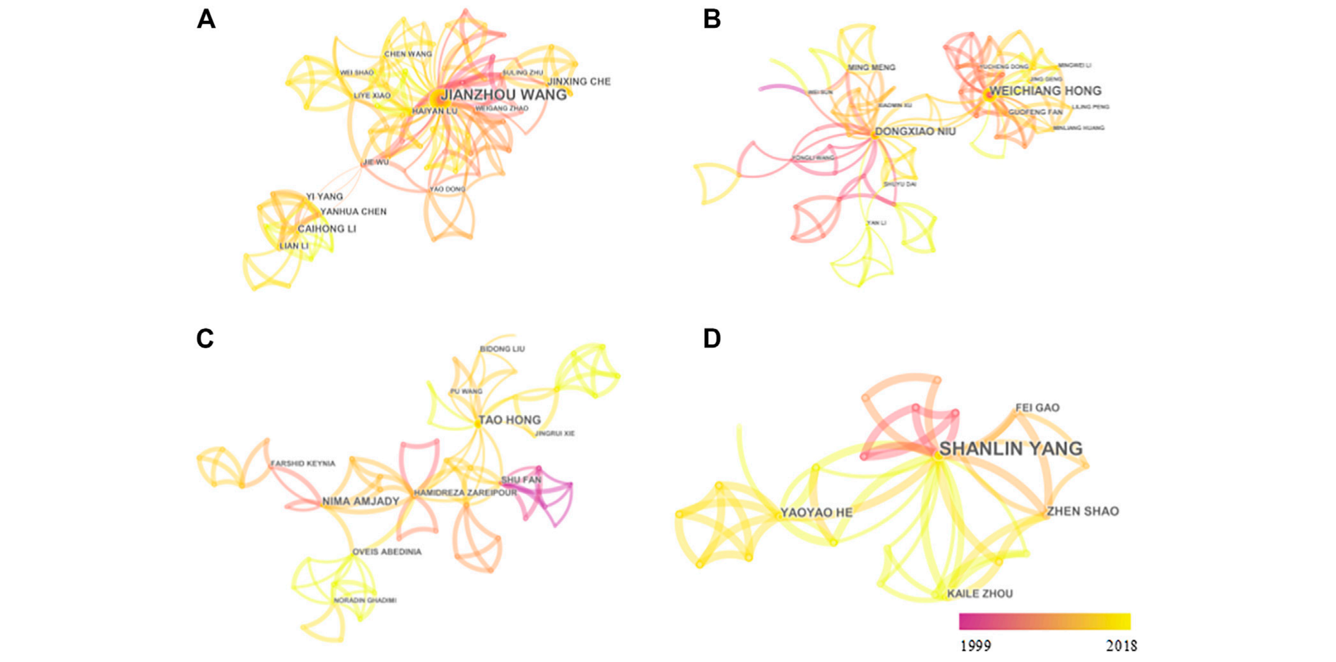


FIGURE 6 | Author cooperation network in electricity demand forecasting (**A**, **B**, **C**, **D** represent the four largest cooperative networks).

and Guo S. The authors of the second quadrant is likely to publish a much-watched article by 2012. Authors of the fourth quadrant recently published articles with low attention, but their articles may become hotspots in the future. There are also some authors, such as Hong WC, Wang JZ, and Hong T, who published the most publications in this field.

Highly Cited Publications

The top 10 highly cited publications are shown in **Table 4**. Only the article published by Li et al. (2013) on "Knowledge-Based Systems" was published after 2009. Others were published before 2009. Among them, Taylor JW, Hong WC, and other authors are shown in **Figure 5**. It is worth noting that Alfares HK; Nazeeruddin M and Taylor JW had paid little attention to this field after 2009. These articles are an essential foundation in this field and are helpful for researchers to understand the important basics of this field. Alfares and Nazeeruddin (2002) offered a review and categorization of electricity demand forecasting techniques. They classified these techniques into nine categories and discussed these technique's advantages and disadvantages.

Table 5 shows the top 10 publications with the highest average citations per year. Only the article published by Hahn et al. (2009) in "European Journal of Operational Research" was published in 2009. Others were published after 2013. Among them, Hong Tao, Guo Sen, Fan Shu, and other authors are shown in **Figure 5**. It is worth noting that the publications with the highest average citations per year are mainly published in the past 5 years, indicating that the electricity demand forecasting may have received more attention in the near future, or new developments have appeared. Hong and Fan (2016) offered a review of probabilistic power load forecasting and introduced the methodologies, techniques, applications, evaluation methods, and future research needs.

COOPERATIVE STRUCTURE IN THE FIELD OF POWER DEMAND FORECASTING

Author Cooperation Network

The author's cooperative network shows the cooperation of all authors in the 831 papers in the field of electricity demand forecasting. The node size is proportional to the number of the author's publications, and the connection between nodes represents the author's cooperative relationship. The thickness of the connection represents the strength of the cooperation between the authors. The color of the connection between nodes and nodes corresponds to the time when the cooperation first appeared. The change of connection's color from cool, such as blue and green, to warm, such as yellow, indicates the change of time from early to recent.

There are 2,143 nodes and 3,561 edges in the author's cooperation network. Obviously, there are many scholars involved in the field of electricity demand forecasting, but most of them only cooperate in a small scope. Many small independent networks of cooperation have not formed

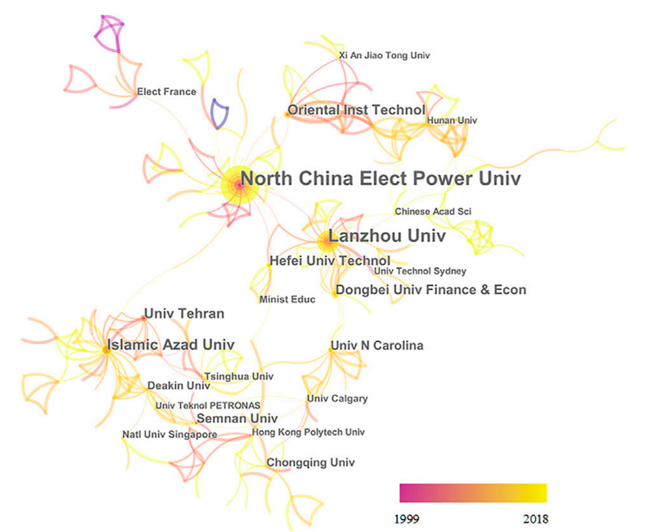


FIGURE 7 | Institutional cooperation network in electricity demand forecasting.

extensive cooperation. There is also a large independent cooperation network in the cooperative network. This study extracted the four largest cooperative networks from the author's cooperation network, as shown in **Figure 6**.

Figure 6A shows the four largest cooperative networks, with 74, 70, 53, and 24 nodes in each of the four networks. Part a of **Figure 6B** is the largest cooperative network with 74 nodes. Among them, Jianzhou Wang, Caihong Li, Yang Yi, Lian Li, Chen Wang, Weigang Zhao, Jie Wu, and other authors constitute a research cluster based on Lanzhou University. Part b of **Figure 6C** is the second network with 70 nodes. Among them, Dongxiao Niu, Ming Meng, and other authors constitute the research cluster based on North China Electric Power University. Weichang Hong, Guofeng Fan, Liling Peng, and other authors. It constitutes a research cluster based on the Pingdingshan Normal University of Jiangsu Normal University. Part c of **Figure 6D** is the third network with 53 nodes. Among them, Tao Hong, Bidong Liu, Pu Wang, Jingrui Xie, and other authors, who are researchers at the University of South Carolina, constitute a research cluster. Nima Amjady, Oveis Abedinia, and other authors, who are researchers at the University of Semnan, constitute a research cluster, and authors such as Shu Fan and Hamidreza Zareipour, who are researchers at the University of Calgary, constitute a research cluster. Authors such as Fei Gao, Shanlin Yang, Yaoyao He, Zhen Shao, and Kaile Zhou, who are researchers at Hefei University of Science and Technology, constitute a research cluster. Other scholars with a large number of publications and extensive cooperation include Azadeh A of Tehran University; Goude Y of the University of Paris-Sud; Taylor JW of Oxford University. Although there are many participants, there are more networks of less than 10 partners in the cooperation network, indicating that cooperation in the field of electricity demand forecasting is lack.

Institutional Cooperation Network

The institutional cooperative network in electricity demand forecasting has 859 nodes and 894 edges. **Figure 7** shows the largest independent network in the institutional cooperation network, with 217 nodes. North China Electric Power University and Lanzhou University are the leading contributors to the cooperation in this field and have published the most articles. 10 institutions that have more than 10 connections were listed: North China Electric Power University (57), Lanzhou University (35), Islamic Azad University (18), Oriental Institute of Technology (15), University of Tehran (14), Hefei University of Technology (13), Dongbei University of Finance and Economics (13), University of Oxford (12), Semnan University (11), University of North Carolina (10). Most institutions in this field are located in China, and there is more cooperation between domestic institutions and transnational institutions. The cooperators of Lanzhou University in China are mainly the University of Chinese Academy of Sciences, Hefei University of Science and Technology, Dongbei University of Finance and Economics. It is worth noting that Wang Jinzhou is a highly productive author in this field, and he has worked at Lanzhou University and Dongbei University of Finance and Economics, which also indicates the cooperative relationship between Lanzhou University and Dongbei University of Finance and Economics. The University of Tehran and Islamic Azad University are the main partners of each other, and both of them are located in Iran. Besides, cross-border cooperation is widespread, with Lanzhou University, University of Technology, Tsinghua University, and Semnan University cooperating with each other. At the same time, it is evident that compared with the author's cooperation network, and the institutional cooperation network is closer.

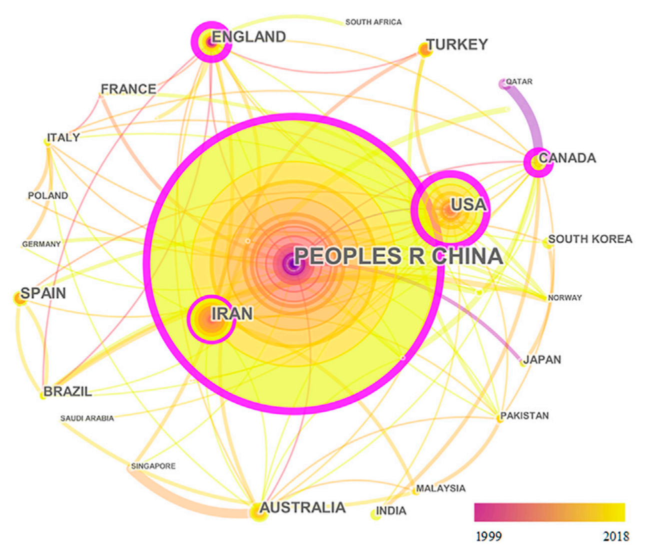


FIGURE 8 | Countries/regions cooperation network in electricity demand forecasting.

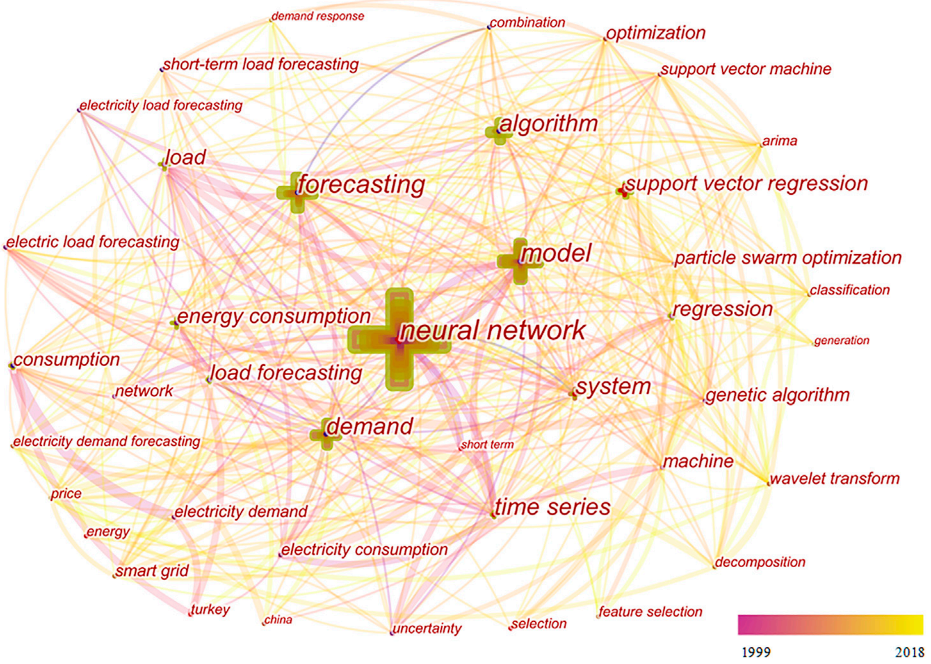


FIGURE 9 | Keyword co-occurrence network in electricity demand forecasting.

Country/Region Cooperation Network

The countries cooperative network for electricity demand forecasting has 37 nodes and 89 edges (deleting links with fewer than two). **Figure 8** shows the largest connected network that contains 30 countries/regions. The top 10 countries are the People's Republic of China, the USA, Iran, England, Turkey, Spain, Australia, Brazil, Canada. From **Figure 8**, it can be found that China is the largest contributor to the country's cooperation network in this field. The main partners of China are the United States (16), Australia (8) and Canada (8), the United Kingdom (7), and Japan (4). The figures in parentheses indicate the number of articles published in cooperation between the two countries. The main partners of the United States are China (15), Italy (3), Pakistan (3), and Poland (3). The main partners of Iran are Australia (3), Malaysia (3), Canada (2), United Kingdom (2), and Hungary (2). The main partners of the UK are China (7), France (3), Brazil (2), Singapore (2), and the United States (2).

ACTIVE TOPICS AND EMERGING TRENDS

Co-Occurrence Network

Keywords are a clear sign of the critical content of research. Co-occurrence analysis is used to analyze the number of occurrences of a pair of words within the same literature and measure the relationship between different publications. The burst detection of keywords is often applied to reveal the emergence of hotspots and active topics. **Figure 9** shows a keyword co-occurrence network for electricity demand forecasting. For ease of observation, **Figure 9** only retains nodes where the co-occurrence frequency is greater than 10. The keyword co-occurrence network is intricate and complex, and the nodes are closely related. It mainly presents the nouns and methods used in this field. The keywords with occurrence frequency higher than 100 are neural network (351), model (190), forecasting (170), demand (127), system (125), algorithm (123), and time series (109). It can be found that the left part of **Figure 9** mostly refers to the main terms related to electricity demand forecasting such as forecasting, load, demand, consumption, etc., and the main methods involved in electricity demand forecasting are the neural network, model, algorithm, support vector regression, regression, etc.

Table 6 shows 17 keywords with the highest bursts, and their strength, begin time and end time. The red line in **Table 6** represents the specific time period for keywords burst. This study found that the neural network is the keyword with the most strength (9.4035), and its duration is as long as 9 years (1999–2007), which indicates that the neural network is one of the most essential basic methods in this field. At the same time, the larger keywords detected by burst are mostly methods. The reason is that the field is too narrow, and the research is more concentrated in this field. On the other hand, the new hotspots in this field are mostly referred to as the improvement of methods.

Co-Citation Network Analysis

Co-citation network analysis is an analysis tool, usually used to examine a large number of documents and reveal the knowledge map of a scientific discipline. In co-citation networks, some key nodes are easily identified because of their prominent structure and characteristics.

The publication with citation bursts represents it has attracted special attention in this field for a period of time. **Table 7** shows

TABLE 6 | Top 17 keywords with bursts during 1999–2018.

Keywords	Year	Strength	Begin	End	1999–2018
Neural network	1999	9.4035	1999	2007	🟥🟥🟥🟥🟥🟥🟥🟥🟥🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦
System	1999	4.277	1999	2003	🟥🟥🟥🟥🟥🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦
Implementation	1999	4.9452	2002	2012	🟦🟦🟦🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟦🟦🟦🟦🟦🟦
Short term	1999	4.8666	2005	2011	🟦🟦🟦🟦🟦🟦🟥🟥🟥🟥🟥🟥🟥🟦🟦🟦🟦🟦🟦🟦
Load forecasting	1999	5.5095	2005	2006	🟦🟦🟦🟦🟦🟦🟥🟥🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦
Time-series	1999	7.3654	2006	2011	🟦🟦🟦🟦🟦🟦🟦🟥🟥🟥🟥🟥🟥🟦🟦🟦🟦🟦🟦🟦
Turkey	1999	4.9077	2009	2011	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟥🟦🟦🟦🟦🟦🟦🟦
Electricity demand	1999	3.3649	2010	2012	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟥🟦🟦🟦🟦🟦🟦
Short-term load forecasting	1999	4.4801	2012	2014	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟥🟦🟦🟦🟦
Particle swarm optimization	1999	4.2798	2013	2016	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟥🟥🟦🟦
Combination	1999	3.8789	2013	2016	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟥🟥🟦🟦
Network	1999	3.8621	2014	2016	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟥🟦🟦
Intelligence	1999	3.6923	2014	2016	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟥🟦🟦
Selection	1999	5.4265	2015	2018	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟥🟥
Energy	1999	3.0846	2016	2018	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟥
Support vector regression	1999	3.7342	2016	2018	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟥
Wavelet transform	1999	3.4861	2016	2018	🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟥🟥🟥

TABLE 7 | Top 10 references with the strongest citation bursts during 1999–2018.

References	Year	Strength	Begin	End	1999–2018
Bakirtzis et al. (1996)	1996	11.1269	1999	2006	🟥🟥🟥🟥🟥🟥🟥🟥🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦
Ramanathan et al. (1997)	1997	7.1414	2000	2007	🟦🟥🟥🟥🟥🟥🟥🟥🟥🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦🟦
Hippert et al. (2001)	2001	20.7847	2002	2011	⬜⬜🟦🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟦🟦🟦🟦🟦🟦🟦
Darbellay and Slama, (2000)	2000	8.8169	2003	2010	⬜🟦🟦🟦🟥🟥🟥🟥🟥🟥🟥🟥🟦🟦🟦🟦🟦🟦🟦🟦
Pai and Hong, (2005)	2005	6.1568	2006	2013	⬜⬜⬜⬜⬜⬜🟦🟥🟥🟥🟥🟥🟥🟥🟥🟦🟦🟦🟦🟦
Taylor, (2003)	2003	5.2395	2006	2013	⬜⬜⬜⬜🟦🟦🟦🟥🟥🟥🟥🟥🟥🟥🟥🟦🟦🟦🟦🟦
Shyh-Jier and Kuang-Rong, (2003)	2003	9.3716	2006	2013	⬜⬜⬜⬜🟦🟦🟦🟥🟥🟥🟥🟥🟥🟥🟥🟦🟦🟦🟦🟦
Cottet and Smith, (2003)	2003	4.4897	2006	2013	⬜⬜⬜⬜🟦🟦🟦🟥🟥🟥🟥🟥🟥🟥🟥🟦🟦🟦🟦🟦
Amjady, (2007)	2007	3.9629	2008	2015	⬜⬜⬜⬜⬜⬜⬜⬜🟦🟥🟥🟥🟥🟥🟥🟥🟥🟦🟦🟦
Fan and Chen, (2006)	2006	6.8201	2008	2015	⬜⬜⬜⬜⬜⬜⬜🟦🟦🟥🟥🟥🟥🟥🟥🟥🟥🟦🟦🟦

that the top-ranked references by bursts were published by Hippert et al. (2001), with bursts strength of 20.7847. The second one was Bakirtzis et al. (1996), with bursts strength of 11.1269.

There are 875 nodes and 4,429 edges in **Figure 10**. The development history and research Frontier in this field, and the crucial articles in this field are mainly concentrated after 2009, which also shows that the articles in this field have experienced explosive growth in recent 10 years. The authors of crucial articles overlap with a large number of high-yield and high-cited authors in the field, such as Hong Tao, Fan S, Taylor JW, and so on. Hippert et al. (2001) reviewed articles published from 1991 to 1999 to assess the practical application of neural networks in short-term electricity load forecasting, and evaluate the design and testing of the neural networks presented in these papers critically. So it becomes a key node in the network. Taylor et al. (2006) assessed the forecast accuracy of short-term electricity demand forecasting

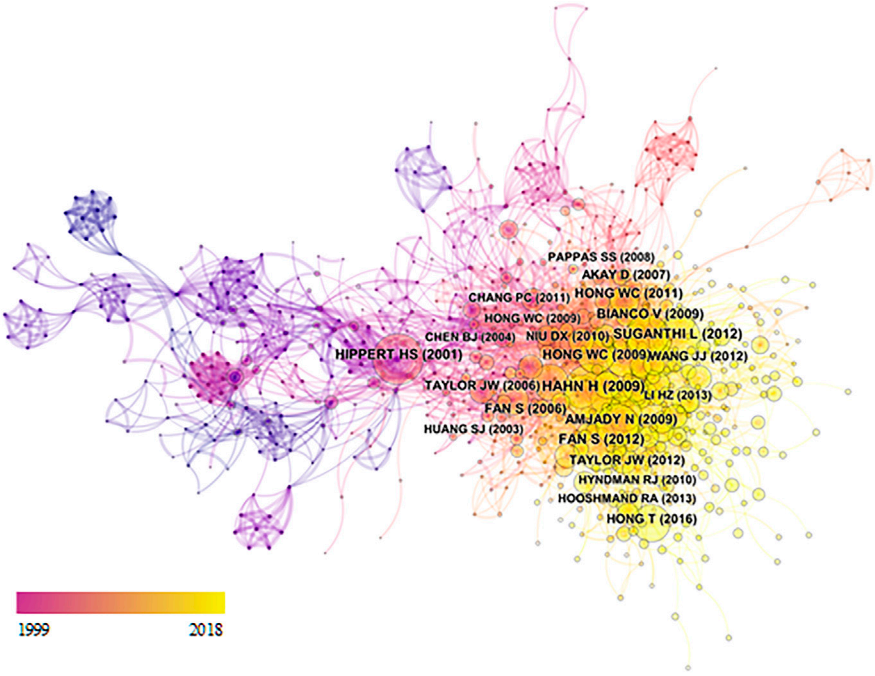


FIGURE 10 | Reference co-citation network in electricity demand forecasting.

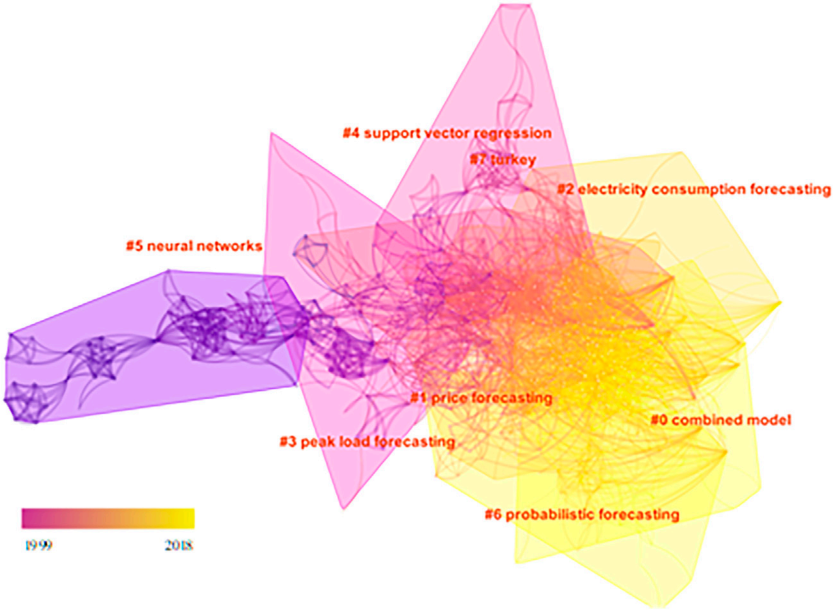


FIGURE 11 | Main co-cited references cluster in electricity demand forecasting.

with six univariate methods. Fan and Hyndman (2012) proposed a semiparametric model to estimate the impact of electricity demand data on model variables. Hong (2011) proposed an electricity demand forecasting method combining chaotic artificial bee colony algorithm and a seasonal recursive support vector regression model. Suganthi and Samuel (2012) reviewed various energy demand prediction models, such as regression, time series, fuzzy logic, ARIMA, support vector regression, and neural networks.

It is obvious that the network has six clusters of a combined model, price forecasting, electricity consumption forecasting, peak load forecasting, support vector regression,

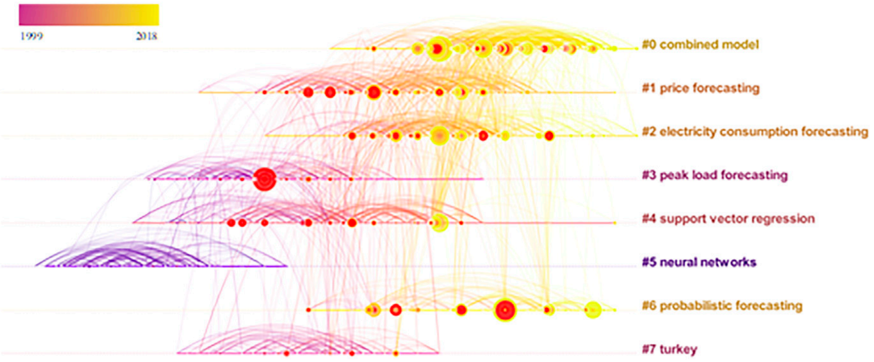


FIGURE 12 | A timeline visualization for the main references cluster.

TABLE 8 | Largest clusters of co-cited references, 1999–2018.

Cluster ID	Size	Silhouette	Mean (Year)	Label (LLR)
0	140	0.995	2013	combined model
1	98	0.996	2008	price forecasting
2	90	0.995	2010	electricity consumption forecasting
3	72	0.997	2001	peak load forecasting
4	58	0.998	2004	support vector regression (svr)
5	57	0.998	1995	neural networks
6	56	0.998	2011	probabilistic forecasting
7	53	0.999	2003	turkey

neural networks, probabilistic forecasting, and turkey in **Figure 11**. **Figure 12** is a line graph of the co-citation of the literature, similar to the keyword timeline. It shows the time evolution process of the six clusters. From the results, the earliest cluster is neural networks, and it's also the same as the keyword timeline visualization results, which illustrates the importance of neural networks in this field. The combined model is the largest cluster, which has been continuously appearing since 2004, indicating that the research Frontier is a hybrid model. Price forecasting, electricity consumption forecasting, peak load forecasting, and probabilistic forecasting reflect the main content of this field. Price forecasting and peak load forecasting are the contents of early attention. Probabilistic forecasting and electricity consumption forecasting are more concerned in recent years. Support vector regression shows that it is one of the important methods in the field. Turkey was the main cluster between 1998 and 2008, indicating that during this period turkey's power forecasting was an area of concern. From the clustering results, the changes in the method of the field and changes in the content of the research. In particular, it is pointed out that the recent research method hotspot is the combined model.

In **Table 8**, Size represents the number of articles in a cluster, and there are 140 articles in the cluster (#0). Silhouette is a measure of a cluster's homogeneity, and the closer its value is to 1, the more homogeneous it is. Mean (Year) represents the average year of publications in a cluster, and it is used to evaluate the average time when the cluster appears. All silhouettes of eight clusters are greater than 0.99, which means the clustering results are reliable.

CONCLUSION

This study offered a bibliometric and visualization analysis on electricity demand forecasting based on 831 publications retrieved from the Web of Sciences. A basic summary, integrated knowledge maps, hot topics, and emerging trends of electricity demand forecasting are presented by statistical description analysis, cooperative network analysis, keyword co-occurrence analysis, co-citation analysis, cluster analysis, and emerging trend analysis techniques. Some interesting and useful conclusions are as follows.

First, electricity demand forecasting has received more and more attention, the numbers of citations and publications are increasing rapidly, especially in the last decade. "Energy fuels", which account for 36.82%, is the largest subject category in the electricity demand forecasting research area. "Energy" is the highest yield journal with 81 publications, followed by "Energies", "International Journal of Electrical Power Energy Systems" and "Applied Energy". "Renewable and Sustainable Energy Reviews", "European Journal of Operational Research" and "Applied Energy", are of constant interest to researchers in this field recently. North China Electric Power University, Lanzhou University, University of North Carolina, Islamic Azad University and Oriental Institute of Technology is the top five yield Institution. Wang Jianzhou, which publishing 24 articles in the field, is the most high-yield author, followed by Hong Weichang, Niu Dongxiao, and Hong Tao. In recent years, the publications of Zareipourh, Khosravia, Hong, Abediniao, and Guo have attracted much attention, and the publications of Fan have attracted lots of attention all the time. The top 10 highly cited publications were mainly published before 2009 but the top 10 publications with the highest average citations per year were mainly published after 2013. The reason may be that earlier publications were the essential foundation of this field and got a lot of citations, and recent publications were the current research focus and got more citations recently.

Second, there are 2,143 scholars involved in the field of electricity demand forecasting but most of them only cooperate in a very small scope and almost cooperate with authors in the same institution. The largest cooperative networks were formed with Wang Jianzhou, who is the largest Structural hole in the cooperative networks. North China Electric Power University, Lanzhou University, Islamic Azad University, Oriental Institute of Technology, and the University of Tehran are the five most irreplaceable productive institutions and contributors in the field of electricity demand forecasting. The People's Republic of China (including Taiwan, Hong Kong, and Macau), United States, Iran, England, and Turkey are the five most significant contributors to country/region cooperation networks.

Third, combined model, neural network, and support vector regression are the main methods in electricity load forecasting, and support vector regression, combined model, and wavelet transform are hotspots methods. Price forecasting, electricity consumption forecasting, peak load forecasting, and probabilistic forecasting are primary researches in electricity demand forecasting. Probabilistic forecasting and electricity consumption forecasting are hotspots.

The basic situation of subject classification, journals, authors, institutions, countries, and highly cited papers in the electricity demand forecasting can be figured out based on the research of this study. At the same time, the collaboration of countries/regions, institutions, and authors is also studied in our study. Furthermore, emerging trends and new developments in this field are also discussed in this research. The results of this research provide a comprehensive description of electricity demand forecasting and are helpful for scholars to maintain the development of this field.

The limitations of our study are that, due to the limits of co-citation analysis in citespace, the literature in our paper are only retrieved from the core database of WoS, document types are limit in "article" or "review", and literature type are limit in "English", which may make some significant literature have been overlooked.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

DY: Conceptualization, Methodology, Software, Formal analysis, Writing-original draft preparation. JG: Conceptualization, Resources, Funding acquisition. JL: Methodology, Software, Formal analysis. SW: Supervision, Project administration, Funding acquisition. SS: Conceptualization, Methodology, Writing–original draft.

FUNDING

This research work was partly supported by the National Natural Science Foundation of China under Grants No. 72101197 and No. 71988101, and by the Fundamental Research Funds for the Central Universities under Grant No. SK2021007.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fenrg.2021.771433/full#supplementary-material.

REFERENCES

Ahmad, A. S., Hassan, M. Y., Abdullah, M. P., Rahman, H. A., Hussin, F., Abdullah, H., et al. (2014). A Review on Applications of ANN and SVM for Building Electrical Energy Consumption Forecasting. *Renew. Sustainable Energ. Rev.* 33, 102–109. doi:10.1016/j.rser.2014.01.069

Akay, D., and Atak, M. (2007). Grey Prediction With Rolling Mechanism for Electricity Demand Forecasting of Turkey. *Energy.* 32, 1670–1675. doi:10.1016/j.energy.2006.11.014

Al-Ghandoor, A., Jaber, J. O., Al-Hinti, I., and Mansour, I. M. (2009). Residential Past and Future Energy Consumption: Potential Savings and Environmental Impact. *Renew. Sustainable Energ. Rev.* 13, 1262–1274. doi:10.1016/j.rser.2008.09.008

Al-Musaylh, M. S., Deo, R. C., Adamowski, J. F., and Li, Y. (2019). Short-Term Electricity Demand Forecasting Using Machine Learning Methods Enriched With Ground-Based Climate and ECMWF Reanalysis Atmospheric Predictors in Southeast Queensland, Australia. *Renew. Sustainable Energ. Rev.* 113, 109293. doi:10.1016/j.rser.2019.109293

Alfares, H. K., and Nazeeruddin, M. (2002). Electric Load Forecasting: Literature Survey and Classification of Methods. *Int. J. Syst. Sci.* 33, 23–34. doi:10.1080/00207720110067421

Amjady, N. (2007). Short-term Bus Load Forecasting of Power Systems by a New Hybrid Method. *IEEE Trans. Power Syst.* 22, 333–341. doi:10.1109/Tpwrs.2006.889130

Bakirtzls, A. G., Petridls, V., Klartzis, S. J., Alexladls, M. C., and Malssls, A. H. (1996). A Neural Network Short Term Load Forecasting Model for the Greek Power System. *IEEE Trans. Power Syst.* 11, 858–863. doi:10.1109/59.496166

Bhattacharyya, S. C., and Thanh, L. T. (2004). Short-Term Electric Load Forecasting Using an Artificial Neural Network: Case of Northern Vietnam. *Int. J. Energ. Res.* 28, 463–472. doi:10.1002/er.980

Boroojeni, K. G., Amini, M. H., Bahrami, S., Iyengar, S. S., Sarwat, A. I., and Karabasoglu, O. (2017). A Novel Multi-Time-Scale Modeling for Electric Power Demand Forecasting: From Short-Term to Medium-Term Horizon. *Electric Power Syst. Res.* 142, 58–73. doi:10.1016/j.epsr.2016.08.031

Bourdeau, M., Zhai, X. q., Nefzaoui, E., Guo, X., and Chatellier, P. (2019). Modeling and Forecasting Building Energy Consumption: A Review of Data-Driven Techniques. *Sustainable Cities Soc.* 48, 101533. doi:10.1016/j.scs.2019.101533

Bunn, D. W. (2000). Forecasting Loads and Prices in Competitive Power Markets. *Proc. IEEE.* 88, 163–169. doi:10.1109/5.823996

Cavallaro, F. (2005). Electric Load Analysis Using an Artificial Neural Network. *Int. J. Energ. Res.* 29, 377–392. doi:10.1002/er.1054

Chen, C. (2006). CiteSpace II: Detecting and Visualizing Emerging Trends and Transient Patterns in Scientific Literature. *J. Am. Soc. Inf. Sci.* 57, 359–377. doi:10.1002/asi.20317

Chen, C., Dubin, R., and Kim, M. C. (2014). Emerging Trends and New Developments in Regenerative Medicine: a Scientometric Update (2000 - 2014). *Expert Opin. Biol. Ther.* 14, 1295–1317. doi:10.1517/14712598.2014.920813

Chen, C. (2017). Science Mapping: A Systematic Review of the Literature. *J. Data Info Sci.* 2, 1–40. doi:10.1515/jdis-2017-0006

Cottet, R., and Smith, M. (2003). Bayesian Modeling and Forecasting of Intraday Electricity Load. *J. Am. Stat. Assoc.* 98, 839–849. doi:10.1198/016214503000000774

Darbellay, G. A., and Slama, M. (2000). Forecasting the Short-Term Demand for Electricity. *Int. J. Forecast.* 16, 71–83. doi:10.1016/S0169-2070(99)00045-X

Fan, S., and Chen, L. (2006). Short-term Load Forecasting Based on an Adaptive Hybrid Method. *IEEE Trans. Power Syst.* 21, 392–401. doi:10.1109/Tpwrs.2005.860944

Fan, S., and Hyndman, R. J. (2012). Short-Term Load Forecasting Based on a Semi-Parametric Additive Model. *IEEE Trans. Power Syst.* 27, 134–141. doi:10.1109/Tpwrs.2011.2162082

Fang, Y., Yin, J., and Wu, B. (2018). Climate Change and Tourism: a Scientometric Analysis Using CiteSpace. *J. Sustainable Tourism.* 26, 108–126. doi:10.1080/09669582.2017.1329310

Hahn, H., Meyer-Nieberg, S., and Pickl, S. (2009). Electric Load Forecasting Methods: Tools for Decision Making. *Eur. J. Oper. Res.* 199, 902–907. doi:10.1016/j.ejor.2009.01.062

Hippert, H. S., Pedreira, C. E., and Souza, R. C. (2001). Neural Networks for Short-Term Load Forecasting: A Review and Evaluation. *IEEE Trans. Power Syst.* 16, 44–55. doi:10.1109/59.910780

Hong, T., and Fan, S. (2016). Probabilistic Electric Load Forecasting: A Tutorial Review. *Int. J. Forecast.* 32, 914–938. doi:10.1016/j.ijforecast.2015.11.011

Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., and Hyndman, R. J. (2016). Probabilistic Energy Forecasting: Global Energy Forecasting Competition 2014 and beyond. *Int. J. Forecast.* 32, 896–913. doi:10.1016/j.ijforecast.2016.02.001

Hong, W.-C. (2011). Electric Load Forecasting by Seasonal Recurrent SVR (Support Vector Regression) With Chaotic Artificial Bee colony Algorithm. *Energy.* 36, 5568–5578. doi:10.1016/j.energy.2011.07.015

Hsu, C.-C., and Chen, C.-Y. (2003). Applications of Improved Grey Prediction Model for Power Demand Forecasting. *Energ. Convers. Management.* 44, 2241–2249. doi:10.1016/S0196-8904(02)00248-0

Kaboli, S. H. A., Fallahpour, A., Selvaraj, J., and Rahim, N. A. (2017). Long-Term Electrical Energy Consumption Formulating and Forecasting via Optimized Gene Expression Programming. *Energy.* 126, 144–164. doi:10.1016/j.energy.2017.03.009

Kim, M. C., and Chen, C. (2015). A Scientometric Review of Emerging Trends and New Developments in Recommendation Systems. *Scientometrics.* 104, 239–263. doi:10.1007/s11192-015-1595-5

Kuster, C., Rezgui, Y., and Mourshed, M. (2017). Electrical Load Forecasting Models: A Critical Systematic Review. *Sustainable Cities Soc.* 35, 257–270. doi:10.1016/j.scs.2017.08.009

Lairmore, M. D., Albrecht, B., D'souza, C., Nisbet, J. W., Ding, W., Bartoe, J. T., et al. (2000). In Vitroandin VivoFunctional Analysis of Human T Cell Lymphotropic Virus Type 1 pX Open Reading Frames I and II. *AIDS Res. Hum. Retroviruses.* 16, 1757–1764. doi:10.1089/08892220050193272

Li, H.-z., Guo, S., Li, C.-j., and Sun, J.-q. (2013). A Hybrid Annual Power Load Forecasting Model Based on Generalized Regression Neural Network With Fruit Fly Optimization Algorithm. *Knowledge-Based Syst.* 37, 378–387. doi:10.1016/j.knosys.2012.08.015

Mohammadi, M., Talebpour, F., Safaee, E., Ghadimi, N., and Abedinia, O. (2018). Small-Scale Building Load Forecast Based on Hybrid Forecast Engine. *Neural Process. Lett.* 48, 329–351. doi:10.1007/s11063-017-9723-2

Mohan, N., Soman, K. P., and Sachin Kumar, S. (2018). A Data-Driven Strategy for Short-Term Electric Load Forecasting Using Dynamic Mode Decomposition Model. *Appl. Energ.* 232, 229–244. doi:10.1016/j.apenergy.2018.09.190

Mohandes, M. (2002). Support Vector Machines for Short-Term Electrical Load Forecasting. *Int. J. Energ. Res.* 26, 335–345. doi:10.1002/er.787

Niazi, M., and Hussain, A. (2011). Agent-Based Computing From Multi-Agent Systems to Agent-Based Models: a Visual Survey. *Scientometrics.* 89, 479–499. doi:10.1007/s11192-011-0468-9

Olawumi, T. O., and Chan, D. W. M. (2018). A Scientometric Review of Global Research on Sustainability and Sustainable Development. *J. Clean. Prod.* 183, 231–250. doi:10.1016/j.jclepro.2018.02.162

Pai, P.-F., and Hong, W.-C. (2005). Forecasting Regional Electricity Load Based on Recurrent Support Vector Machines With Genetic Algorithms. *Electric Power Syst. Res.* 74, 417–425. doi:10.1016/j.epsr.2005.01.006

Quan, H., Srinivasan, D., and Khosravi, A. (2014). Short-term Load and Wind Power Forecasting Using Neural Network-Based Prediction Intervals. *IEEE Trans. Neural Netw. Learn. Syst.* 25, 303–315. doi:10.1109/TNNLS.2013.2276053

Ramanathan, R., Engle, R., Granger, C. W. J., Vahid-Araghi, F., and Brace, C. (1997). Short-run Forecasts of Electricity Loads and Peaks. *Int. J. Forecast.* 13, 161–174. doi:10.1016/S0169-2070(97)00015-0

Raza, M. Q., and Khosravi, A. (2015). A Review on Artificial Intelligence Based Load Demand Forecasting Techniques for Smart Grid and Buildings. *Renew. Sustainable Energ. Rev.* 50, 1352–1372. doi:10.1016/j.rser.2015.04.065

Shao, Z., Chao, F., Yang, S.-L., and Zhou, K.-L. (2017). A Review of the Decomposition Methodology for Extracting and Identifying the Fluctuation Characteristics in Electricity Demand Forecasting. *Renew. Sustainable Energ. Rev.* 75, 123–136. doi:10.1016/j.rser.2016.10.056

Shyh-Jier, H., and Kuang-Rong, S. (2003). Short-Term Load Forecasting via ARMA Model Identification Including Non-Gaussian Process Considerations. *IEEE Trans. Power Syst.* 18, 673–679. doi:10.1109/tpwrs.2003.811010

Sousa, J. C., Jorge, H. M., and Neves, L. P. (2014). Short-Term Load Forecasting Based on Support Vector Regression and Load Profiling. *Int. J. Energ. Res.* 38, 350–362. doi:10.1002/er.3048

Suganthi, L., and Samuel, A. A. (2012). Energy Models for Demand Forecasting-A Review. *Renew. Sustainable Energ. Rev.* 16, 1223–1240. doi:10.1016/j.rser.2011.08.014

Taylor, J. W., and Buizza, R. (2002). Neural Network Load Forecasting With Weather Ensemble Predictions. *IEEE Trans. Power Syst.* 17, 626–632. doi:10.1109/Tpwrs.2002.800906

Taylor, J. W., De Menezes, L. M., and Mcsharry, P. E. (2006). A Comparison of Univariate Methods for Forecasting Electricity Demand up to a Day Ahead. *Int. J. Forecast.* 22, 1–16. doi:10.1016/j.ijforecast.2005.06.006

Taylor, J. W. (2003). Short-term Electricity Demand Forecasting Using Double Seasonal Exponential Smoothing. *J. Oper. Res. Soc.* 54, 799–805. doi:10.1057/palgrave.jors.2601589

Yang, S., Sui, J., Liu, T., Wu, W., Xu, S., Yin, L., et al. (2018). Trends on PM2.5 Research, 1997-2016: a Bibliometric Study. *Environ. Sci. Pollut. Res.* 25, 12284–12298. doi:10.1007/s11356-018-1723-x

Yu, D., and Xu, C. (2017). Mapping Research on Carbon Emissions Trading: a Co-Citation Analysis. *Renew. Sustainable Energ. Rev.* 74, 1314–1322. doi:10.1016/j.rser.2016.11.144

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Energy Research

ORIGINAL RESEARCH
published: 14 October 2021
doi: 10.3389/fenrg.2021.752593

Forecasting of Steam Coal Price Based on Robust Regularized Kernel Regression and Empirical Mode Decomposition

Xiangwan Fu[1†], *Mingzhu Tang*[1*], *Dongqun Xu*[2], *Jun Yang*[3], *Donglin Chen*[1] *and Ziming Wang*[4†]

[1]School of Energy and Power Engineering, Changsha University of Science and Technology, Changsha, China, [2]China Datang Corporation Ltd, Beijing, China, [3]Hunan Datang Xianyi Technology Co., Ltd., Changsha, China, [4]School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin, China

Aiming at the problem of difficulties in modeling the nonlinear relation in the steam coal dataset, this article proposes a forecasting method for the price of steam coal based on robust regularized kernel regression and empirical mode decomposition. By selecting the polynomial kernel function, the robust loss function and L2 regular term to construct a robust regularized kernel regression model are used. The polynomial kernel function does not depend on the kernel parameters and can mine the global rules in the dataset so that improves the forecasting stability of the kernel model. This method maps the features to the high-dimensional space by using the polynomial kernel function to transform the nonlinear law in the original feature space into linear law in the high-dimensional space and helps learn the linear law in the high-dimensional feature space by using the linear model. The Huber loss function is selected to reduce the influence of abnormal noise in the dataset on the model performance, and the L2 regular term is used to reduce the risk of model overfitting. We use the combined model based on empirical mode decomposition (EMD) and auto regressive integrated moving average (ARIMA) model to compensate for the error of robust regularized kernel regression model, thus making up for the limitations of the single forecasting model. Finally, we use the steam coal dataset to verify the proposed model and such model has an optimal evaluation index value compared to other contrast models after the model performance is evaluated as per the evaluation index such as RMSE, MAE, and mean absolute percentage error.

Keywords: the steam coal price forecasting, kernel function, empirical mode decomposition, Huber loss function, L2 regular term

Edited by:
Jian Chai,
Xidian University, China

Reviewed by:
Shaolong Sun,
Jiangsu University, China
Chengyuan Zhang,
Xidian University, China

***Correspondence:**
Mingzhu Tang
tmz@csust.edu.cn

[†]These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Smart Grids, a section of the journal Frontiers in Energy Research

Received: *03 August 2021*
Accepted: *27 August 2021*
Published: *14 October 2021*

Citation:
Fu X, Tang M, Xu D, Yang J, Chen D and Wang Z (2021) Forecasting of Steam Coal Price Based on Robust Regularized Kernel Regression and Empirical Mode Decomposition. Front. Energy Res. 9:752593. doi: 10.3389/fenrg.2021.752593

INTRODUCTION

Accurate forecasting of the steam coal price can provide a certain basis for enterprises related to steam coal to formulate the procurement plan. The steam coal price indicates the general performance of supply and demand on the steam coal market. China is a big consumer of coal (Xiong and Xu, 2021; Wang and Du, 2020). Forecasting the steam coal price accurately can help in the analysis of the steam coal market, grasp the implied law in the steam coal market, and improve the steam coal market's efficiency.

In recent years, many references have proposed many different methods for coal price forecasting. The time series model can mine the implicit law in the time series. Matyjaszek et al. removed the effect of abnormal fluctuations in prices on the forecasting model by using the transgenic time series (Matyjaszek et al., 2019). Ji et al. effectively improved the forecasting accuracy of the forecasting model by using the ARIMA model and neural network model (Ji et al., 2019). Wu et al. decomposed the price series into several components, used the ARIMA model and SBL model for coal price forecasting, and added up the forecasted values of all the components as final forecasting result. Compared to the contrast model adopted, this model can effectively improve the model forecasting precision (Wu et al., 2019). Chai et al. combined STL decomposition method with ETS model. The experimental results show that it has the best forecasting performance compared with benchmark models and neural network (Chai et al., 2021). It is difficult to learn the implicit nonlinearity law in the data by using the time series model, which is sensitive to the abnormal value in the data and only considers a single variable factor other than other influence factors.

A neural network model can learn the nonlinearity law in the data which is studied based on the interconnection between nerve cells. Alameer et al. effectively improved the forecasting accuracy of coal price based on LSTM model and DNN model (Alameer et al., 2020). Lu et al. adopted the full empirical mode to decompose and preprocess the original dataset, and then chose the radial basis function neural network model for model training and forecasting. The results show higher stability (Lu et al., 2020). Yang et al. adopted the improved whale optimization algorithm to optimize the decomposition and LSTM combined model based on the improved integration empirical model, which has a better model forecasting performance compared to other reference models (Yang et al., 2020). Zhang et al. decomposed the original data series by multi-resolution singular value decomposition method and forecasted the coal price by using MFO-optimized ELM model. Experimental results show the forecasting performance of the proposed model was superior to that of the contrast model (Zhang et al., 2019). However, the neural network model is a black box model which is difficult to interpret.

The steam coal market is a complex nonlinear system, containing influence factors such as economy, steam coal transportation, steam coal supply, and steam coal demand. The influence factors involve a wide range and many feature data and contain some noise data. This method improves the model interpretability by using linear model and reduces the adverse impact of noise data on the forecast model by using Huber loss function (Gupta et al., 2020). We use the kernel function to mine the implicit nonlinearity law in the steam coal data (Li and Li, 2019; Vu et al., 2019; Ye et al., 2021). The combined model can improve the model performance based on the advantages of the sub-model (Wang et al., 1210; Zhou et al., 2019; Wang et al., 2020a; Wang et al., 2020b; Qiao et al., 2021; Zhang et al., 2021). This method can decompose the forecasting error of the forecasting model into multiple modal components by using the EMD method (Yu et al., 2008; Xu et al., 2019; Wang and Wang, 2020; Xia and Wang, 2020), build the ARIMA model (Conejo et al., 2005; Karabiber and Xydis, 2019) for each modal component for forecasting, and add up the forecasted values of all the modal components to compensate error for the original forecasting model.

For the problem of difficulties in modeling the nonlinear relation in the steam coal dataset, this article proposes a forecast method for the price of steam coal based on robust regularized kernel regression and empirical mode decomposition. The second part introduces the used algorithm theory content; the third part states the data preprocessing steps, the selection of features, and the whole process of model training and forecasting; the fourth part shows the model comparison and experimental results; and the fifth part contains conclusion and prospect.

METHODOLOGY

Huber–Ridge Model

The Huber function (Huber et al., 1992) has great robustness, which can effectively reduce the negative influence of abnormal data on model performance. The Huber loss function is shown in **Eq. 1**:

$$\phi_{\text{hub}}(u) = \begin{cases} u^2 & |u| \le M \\ M(2|u| - M) & |u| > M \end{cases}, \tag{1}$$

where u is the residual value and M is the threshold value of the Huber function. The Huber function imposes the punishment which is larger than the threshold value residual to effectively lower the influence of abnormal sample points on the model training.

The Ridge model is added with L2 penalty term based on the objective function of the linear regression model. The objective function of the model is shown in **Eq. 2**:

$$\min_{\omega} \frac{1}{2}\|X\omega - Y\|_2^2 + \alpha_{l2}\|\omega\|_2^2, \tag{2}$$

where X refers to the set of feature parameters, ω refers to the weight coefficient vector, Y refers to the forecasted target quantity, $\|\omega\|_2^2$ refers to the L2 penalty term, and α_{l2} refers to the regular coefficient.

L2 regular term compresses the feature weight value adversely to the model forecasting and makes it approximate to 0 in order to reduce the impact of features with low correlation. When the regular coefficient of α_{l2} is large, the L2 regular term makes more parameters' weight in the parameter weight vector approximate to 0 to screen out main features and mitigate the degree of model overfitting to some extent.

The Huber loss function and L2 regular term are combined to construct the Huber–Ridge model (Owen, 2006), improving the model robustness and lowering the overfitting risk. Its objective function is shown in **Eq. 3**:

$$\widehat{w}_j = \operatorname{argmin}_w \left(\phi_{\text{hub}}(u) + \frac{\lambda}{2}\sum_{j=1}^{k}\left(w_j\right)^2\right). \tag{3}$$

Polynomial Kernel Huber–Ridge Model

The T.M. Cover theorem (Cover, 1965) points out that the data in the high-dimensional space can show the linearity law more easily. The kernel function maps the vector of low-dimensional feature space to the high-dimensional feature space, and transforms the nonlinearity law in the low-dimensional feature space into linearity law in the high-dimensional space to learn the linearity law in the high-dimensional space by using the linear model and indirectly learn the nonlinearity law in the original feature space based on the model. Due to the high-dimensional feature space having high dimensionality, the dimension disaster may happen if the model is directly used for fitting in the high-dimensional space. The introduced kernel function can effectively solve the above problem, and the kernel function can represent inner product value in the high-dimensional space with the inner product value in the low-dimensional space. Thus, it can avoid the inner product calculation in the high-dimensional space and greatly reduce the calculation of the model.

The regular risk functions have a unified expression mode (Schölkopf et al., 2001), as shown in **Eq. 4**:

$$f(x) = \sum_{i=1}^{n} w_i k(x, x_i) + b. \tag{4}$$

The kernel function is introduced to the Huber–Ridge model (Jianke Zhu et al., 2008). Thus, the model can learn the implicit nonlinearity law in the data. The objective function of Huber–Ridge kernel model is shown in **Eq. 6**:

$$u = y_i - f(x), \tag{5}$$

where y_i is the actual value, $f(x)$ is the forecasting value, and u is the forecasting error

$$\phi_{\text{hub}}(y_i, f(x)) = \begin{cases} M(2u - M) & A_1 = \{x|u > M\} \\ u^2 & A_2 = \{x| - M \le u \le M\} \\ -M(2u - M)A_3 & A_3 = \{x|u > M\} \end{cases}, \tag{6}$$

$$\widehat{w}_j = \text{argmin}_{\text{w}} \sum_{j=1}^{n} \phi_{\text{hub}}\left(y_i, \sum_{i=1}^{n} k\left(x_i, x_j\right) w_j + b \right) + \lambda \sum_{i,j=1}^{n} w_i w_j k\left(x_i, x_j\right), \tag{7}$$

$$T = \text{argmin}_{\text{w}} \sum_{i=1}^{n} \phi_{\text{hub}}\left(y, \sum_{i=1}^{n} k_i^T w + b \cdot I \right) + \lambda w^T K w. \tag{8}$$

Eqs. 9–12 can be obtained, respectively, by getting the partial derivative of w and b:

$$\frac{\partial \text{T}}{\partial w} = 2\left(\lambda K w + K I^0 K w + K q\right) = 0, \tag{9}$$

$$\text{q} = -I^0 y + me + b \cdot I^0, \tag{10}$$

$$\text{e}_i = \begin{cases} 1 & x_i \in A_1 \\ 0 & x_i \in A_2 \\ -1 & x_i \in A_3 \end{cases}, \tag{11}$$

$$\frac{\partial T}{\partial b} = 2\left(I^0 \text{Kw} + q\right) = 0, \tag{12}$$

where I^0 is a diagonal matrix of shape (n, n), the values of its elements in the A domain are 1, and the remaining values are 0.

The basic Newton method is used to iteratively update w and b, as shown in **Eq. 13**:

$$\begin{bmatrix} w' \\ b' \end{bmatrix} = \begin{bmatrix} w \\ b \end{bmatrix} - \gamma H^{-1} \nabla, \tag{13}$$

where H^{-1} is the first derivative of the gradient matrix as shown in **Eq. 14**, γ is the length of the step, usually with a value of 1, and ∇ is a Hessen matrix as shown in **Eq. 15**:

$$H^{-1} = \frac{1}{2} \begin{bmatrix} \lambda K + K I^0 K & I^0 K \\ K I^0 & I^0 \end{bmatrix}^{-1}, \tag{14}$$

$$\nabla = 2 \begin{bmatrix} \lambda K w + K I^0 K w + K q \\ I^0 K w + q \end{bmatrix}. \tag{15}$$

Simplify **Eq. 13** to obtain the final computational equation of objective function of the kernel Huber–Ridge model, as shown in **Eq. 16**:

$$\begin{bmatrix} w' \\ b' \end{bmatrix} = \begin{bmatrix} 0 \\ b \end{bmatrix} + \begin{bmatrix} \lambda I + I^0 K & 1 \\ 1 & 0 \end{bmatrix}^{-1} \begin{bmatrix} q \\ 0 \end{bmatrix}. \tag{16}$$

The polynomial kernel is a commonly used kernel function, and the polynomial kernel function is shown in **Eq. 17**:

$$\mathbf{K}(\mathbf{x}, \mathbf{x}) = (\mathbf{x} \cdot \mathbf{x} + 1)^d, \tag{17}$$

where $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ is the feature vector and d is the kernel parameter. The polynomial kernel function (17) is substituted into **Eq. 16** to obtain the final computational equation of polynomial kernel Huber–Ridge model objective function.

EMD Model

The empirical mode decomposition is a signal decomposition technology and decomposes the original signal into a series of components which are the intrinsic mode functions. The empirical mode decomposition is often used to handle the time series data and decomposes the original time series into a series of different components to explore the implicit law in the time series data.

The intrinsic mode function should meet the two conditions below:

1. In the data interval, difference between numbers of extreme points and zero points is at most one.
2. The average value of the upper envelope and the lower envelope is zero.

The EMD model is adaptive and can decompose the original series for a time series data without the number of components specified till the standard of stopping decomposition is met. The relationship between the original series and the decomposed components is shown in **Eq. 18**:

$$\text{X}(\text{t}) = \sum_{i=1}^{n} imf_i + r, \tag{18}$$

where $\text{X}(\text{t})$ refers to the original time series, $\sum_{i=1}^{n} imf_i$ refers to the sum of the components, and r refers to the residual. When the residual series is a monotonic function, the decomposition stopped.

The decomposition step of empirical mode decomposition is shown as follows:

STEP 1: Identify all the maximum points and minimum points in the time series, and fit the upper envelope e_u and the lower envelope e_l by using the cubic spline finite difference method according to maximum points and minimum points.

STEP 2: Calculate the average value of the upper envelope e_u and the lower envelope e_l, and obtain the mean envelope e_{mean} $\left(e_{mean}=\frac{e_u+e_l}{2}\right)$.

STEP 3: Calculate the difference between the original series $X(t)$ and the mean envelope e_{mean}, and obtain the intermediate time series$(e_i = X(t) - e_{mean})$.

STEP 4: Judge whether the intermediate time series e_i can be an intrinsic mode function according to the constraint condition of intrinsic mode functions. If satisfied, the intermediate time series shall be used as the imf_i component. If unsatisfied, such intermediate time series shall be used as the basis to execute the steps 1–4.

STEP 5: Subtract the component imf_i from the original time series $X(t)$ and execute the steps 1–4 again. If the standard of stopping decomposition is satisfied, the decomposition process will end.

ARIMA Model

The autoregressive integrated moving average (ARIMA) model is defined in **Eq. 19**:

$$y^t = \varphi_1 y^{t-1} + \varphi_2 y^{t-2} + \cdots + \varphi_p y^{t-p} + \varepsilon^t - \theta_1 \varepsilon^{t-1} - \theta_2 \varepsilon^{t-2} - \cdots - \theta_q \varepsilon^{t-q} + \theta_0, \tag{19}$$

where y^t and ε^t indicate the actual value and residual value at the time point t, respectively, and $\varphi = (\varphi_1, \varphi_2, \cdots, \varphi_p)$ and $\theta = (\theta_1, \theta_2, \cdots, \theta_q)$ refer to the weight vectors. p and q are the model orders. The historical time series data and historical white noise error data of the variable are used to forecast the current value.

The prerequisite of using ARIMA model is to use stationary data. The non-stationary data can be handled by combining the autoregressive integrated moving average (ARIMA) model and different methods (Gilbert, 2005). The ARIMA model has three parameters, (p, d, q), in which d refers to the differential order of the data series.

Evaluation Indexes

The model performance is evaluated by the mean absolute error (MAE) and the definition of MAE is shown in **Eq. 20**:

$$\mathrm{MAE} = \frac{1}{n}\sum\nolimits_{i=1}^{n} \left| y_i - \widehat{y}_i \right|. \tag{20}$$

The root-mean-square error (RMSE) is used for model performance assessment and the definition of RMSE is shown in **Eq. 21**:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n} \left(y_i - \widehat{y}_i \right)^2}. \tag{21}$$

The mean absolute percentage error (MAPE) is used for model performance assessment and the definition of MAPE is shown in **Eq. 22**:

$$\mathrm{MAPE} = \frac{1}{n}\sum\nolimits_{i=1}^{n} \left| \frac{y_i - \widehat{y}_i}{y_i} \right|, \tag{22}$$

where n is the number of samples in the set verified, $\widehat{y}_i$ is the forecasted value of the model, and y_i is the true value. The closer the MAE, RMSE, and MAPE values are to 0, the better the model performance will be.

The Training and Forecasting Process of Polynomial Kernel Huber–Ridge–EMD–ARIMA Model

The training and forecasting process for the forecast framework of polynomial kernel Huber–Ridge–EMD–ARIMA model (PK–Huber–Ridge–EMD–ARIMA) proposed is shown in **Figure 1**. Its process steps are as follows.

STEP 1: Data preprocessing and feature selection: Screen the correlation features according to Pearson correlation coefficient and Spearman correlation coefficient after the data preprocessing and divide training dataset and test dataset.

STEP 2: Model training: The model parameter M, α_{l2}, and training dataset are input to the polynomial kernel Huber–Ridge for model training.

STEP 3: Model forecasting: The test dataset is input to the trained polynomial kernel Huber–Ridge model. The model outputs the time series $\{y_1', \cdots, y_k', y_{k+1}'\}$ of coal price forecasting data.

STEP 4: Forecast the forecasting error of steam coal price at the next time point: The forecasting error series $\{\varepsilon_1, \cdots \varepsilon_k\}$ of coal price is input to the EMD model, and the EMD model outputs j modal components $(IMF_1, \cdots, IMF_j)$. Each modal component is subject to training and forecasting by the ARIMA model; the ARIMA model $(ARIMA_1, \cdots, ARIMA_j)$ corresponding to each modal component outputs the coal price forecasting error $(\varepsilon_{k+1}^1, \cdots, \varepsilon_{k+1}^j)$ of each modal component at the next time point, respectively. The $(\varepsilon_{k+1}^1, \cdots, \varepsilon_{k+1}^j)$ series accumulation is conducted, and the accumulation result $\{\varepsilon_{k+1}^{'}\}$ is used as the forecasted value of the coal price forecasting error at a time point k+1.

STEP 5: Obtain the final forecasted value of steam coal price at the next time point: The forecasted value $\{\varepsilon_{k+1}^{'}\}$ of coal price forecasting error at the time point k+1 is used for $\{y_{k+1}^{'}\}$ correction of the forecasted value of coal price at a time point k+1. The forecasted value $\{y_{k+1}^{''}\}$ of coal price after correction is used as the forecasted value of coal price at the final time point k+1.

EMPIRICAL STUDY

Data Description

Qinhuangdao steam coal price data that are 4500-kilocalorie, 5000-kilocalorie, and 5500-kilocalorie steam coal exit price data of Qinhuangdao Port from Jan. 2017 to Jul. 2021, are used for experimental study. The sampling is conducted once a week, which is the weekly frequency data.

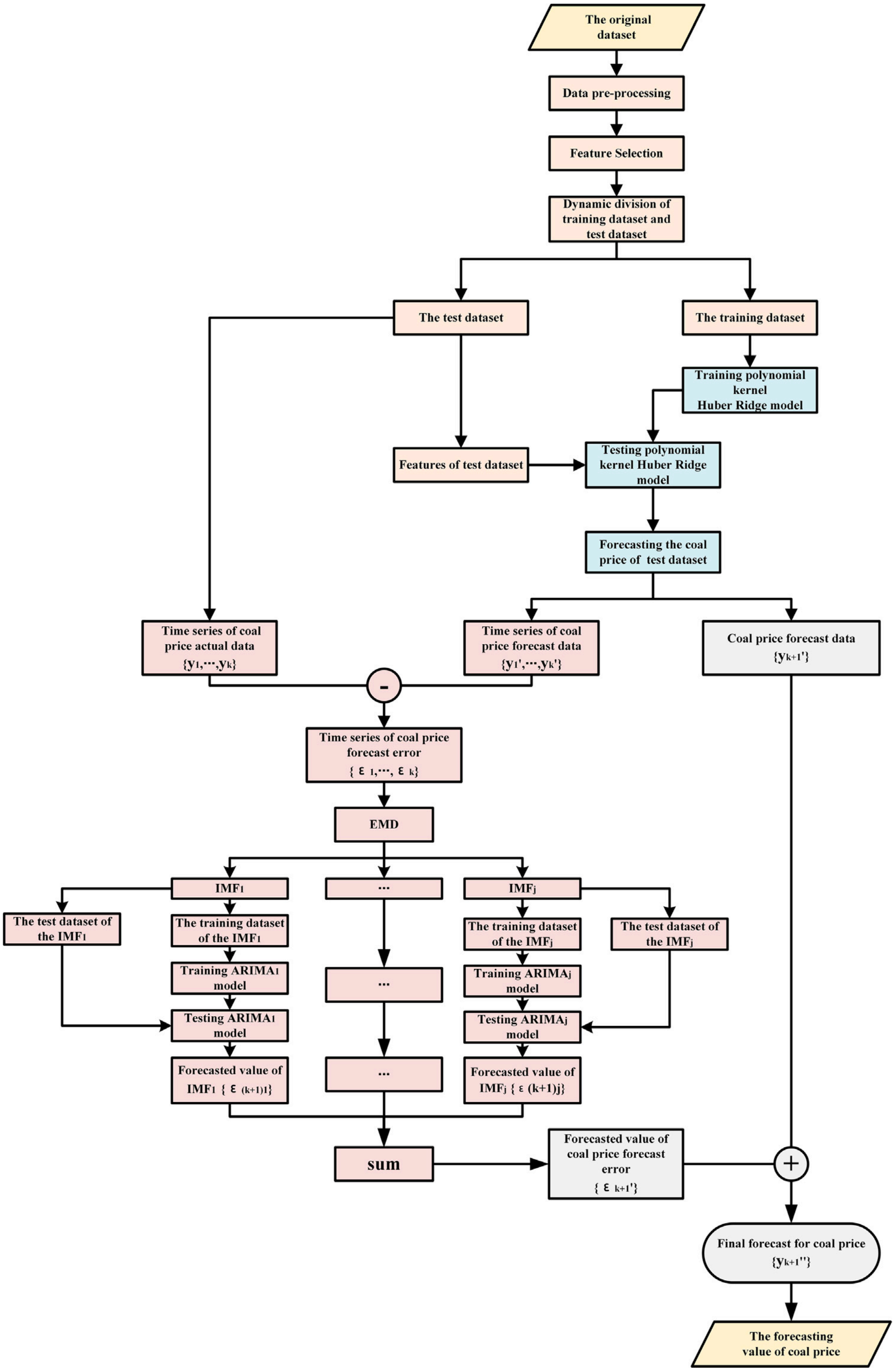

The original dataset
Data pre-processing
Feature Selection
Dynamic division of training dataset and test dataset
The test dataset
The training dataset
Training polynomial kernel Huber Ridge model
Features of test dataset
Testing polynomial kernel Huber Ridge model
Forecasting the coal price of test dataset
Time series of coal price actual data {y_1,···,y_k}
Time series of coal price forecast data {y_1',···,y_k'}
Coal price forecast data {y_{k+1}'}
-
Time series of coal price forecast error {$ε_1$,···,$ε_k$}
EMD
IMF1
···
IMFj
The test dataset of the IMF1
The training dataset of the IMF1
Training ARIMA1 model
Testing ARIMA1 model
Forecasted value of IMF1 {$ε_{(k+1)1}$}
···
···
The training dataset of the IMFj
The test dataset of the IMFj
Training ARIMAj model
Testing ARIMAj model
Forecasted value of IMFj {$ε_{(k+1)j}$}
sum
Forecasted value of coal price forecast error {$ε_{k+1}'$}
+
Final forecast for coal price {y_{k+1}''}
The forecasting value of coal price

FIGURE 1 | Flowchart for training and forecasting of the PK–Huber–Ridge–EMD–ARIMA model.

Data Preprocessing

The original data about steam coal price and its features have the disadvantage of missing value and inconsistent data time sampling frequency; so, such original data shall be pre-processed and the data processed can be brought into the model for model training and forecasting.

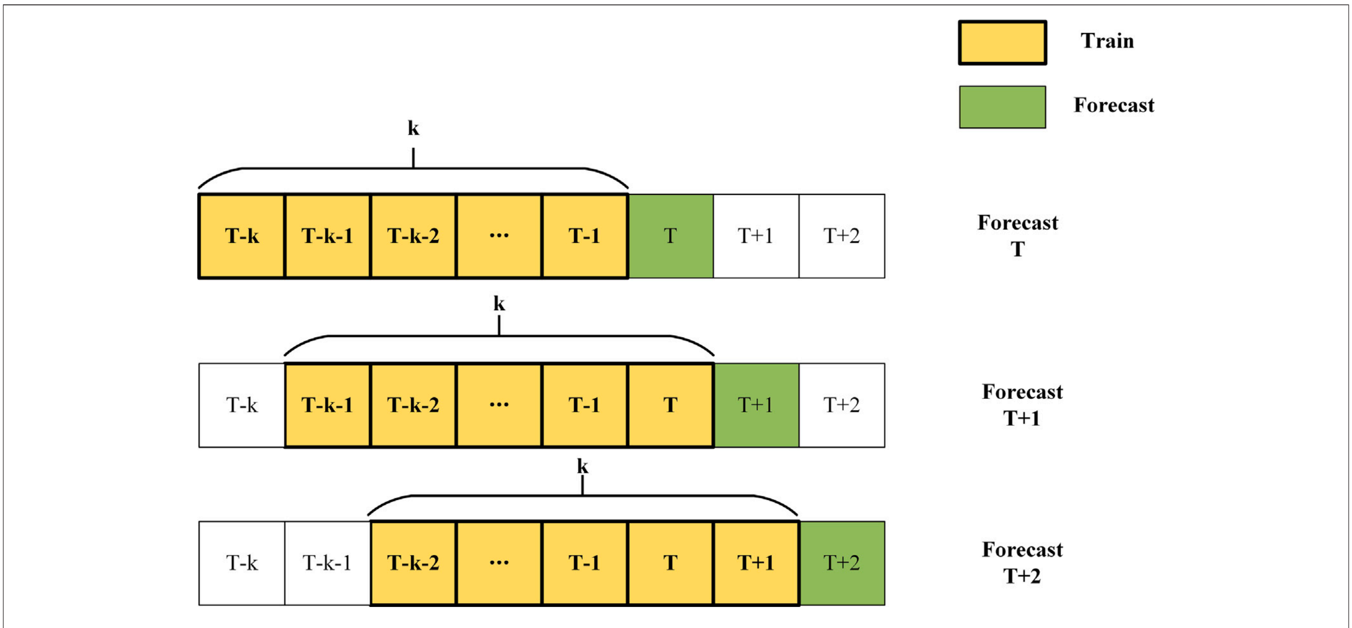


FIGURE 2 | Dynamic division of training dataset and test dataset.

TABLE 1 | Selection of steam coal feature data.

Types of the feature	Feature set
Coal production	Raw coal production of key state-owned coal mines, coal production of large coal enterprises, ... , and raw coal production of non–state-owned coal mines
Coal supply	National coal import quantity, coal inventory and dispatching data of Qinhuangdao Port, cargo ship ratio of four ports around Bohai Sea, and historical data of total coal storage in five ports around Bohai Sea
Coal transportation	Coal transportation quantity of Daqin Line, coal sales volume of national key coal mines, ..., and daily average number of coal railway loading vehicles
Coal consumption	National industrial power consumption, national social electricity consumption, power generation in coastal provinces, and coal consumption in power grid
Macroeconomic	Coal future price, added value of national secondary industry, GDP, consumer price index, the producer price index, and investment in fixed assets of the whole society

TABLE 2 | Parameters setting of the feature selection process.

Types of parameter	r_{mp}	r_{sp}	$delay_{max}$
Parameter value	0.5	0.5	5

TABLE 3 | Linear features with optimal delay.

Features	Optimal delay order (delay)	Pearson correlation coefficient (r_p)
Quantity of anchored vessels in Caofeidian Port	1	0.589,706,909
Volume of ships anchored at Caofeidian Port Phase II	1	0.594,892,547
Total coal stock in ports around the Bohai Sea	1	0.552,136,323
Power coal future closing price	1	0.895,980,863
………	………	………
Total coal stock in Fangcheng Port	4	0.575,740,516

TABLE 4 | Nonlinear features with optimal delay.

Features	Optimal delay order (delay)	Pearson correlation coefficient (r_s)
Ship ratio of four ports around the Bohai Sea	3	0.74,183,513
Quantity of anchored vessels in Qinhuangdao Port	2	0.743,160,533
Quantity of anchored vessels in CIT Jingtang Port	2	0.552,341,897
Total coal stock in mainstream ports	5	0.625,224,659
………	………	………
Total coal stock at coastal ports	5	0.589,975,915

TABLE 5 | Value of hyper-parameters for five different models.

Forecasting model	Value of hyper-parameters for forecasting model
Lasso	$\alpha_{l1} = 0.1$
Ridge	$\alpha_{l2} = 0.2$
Huber–Ridge	$M = 1.35; \alpha_{l2} = 0.2$
PK–Huber–Ridge	$d = 2; M = 1.35; \alpha_{l2} = 0.2$
PK–Huber–Ridge–EMD–ARIMA	$d_k = 2; M = 1.35; \alpha_{l2} = 0.2; p \in [1, 2, 3, 4];$ $d_a \in [1, 2, 3]; q \in [1, 2, 3, 4]$

TABLE 6 | Evaluation index value of forecasting results of five forecasting models.

Dataset	Forecasting model	MAE	RMSE	MAPE(%)
Dataset 1	Lasso	36.0138	53.6206	6.6257
	Ridge	28.6532	44.1649	5.1159
	Huber–Ridge	30.7005	50.0595	5.6320
	PK–Huber–Ridge	26.8503	40.3592	4.8223
	PK–Huber–Ridge–EMD–ARIMA	**19.2267**	**26.0293**	**3.4813**
Dataset 2	Lasso	38.6772	60.7722	6.2920
	Ridge	29.8993	47.9723	4.5484
	Huber–Ridge	32.3476	55.5871	5.1023
	PK–Huber–Ridge	30.4047	45.9770	4.6723
	PK–Huber–Ridge–EMD–ARIMA	**18.9126**	**26.3342**	**2.9432**
Dataset 3	Lasso	37.6168	63.9781	5.3979
	Ridge	34.8572	57.777	4.7187
	Huber–Ridge	34.0952	61.9462	4.8548
	PK–Huber–Ridge	33.828	56.22	4.7179
	PK–Huber–Ridge–EMD–ARIMA	**22.9183**	**37.6673**	**3.1237**

The data preprocessing step is shown as follows:

1. Unify the data sampling frequency: The data input to the model have the features of one-to-one relationship between coal feature data and coal price, that is, the sampling frequency of coal feature data and coal price data is the same. When the sampling frequency of the original coal price data and related feature data is inconsistent, such original data shall be operated at the unified data sampling frequency; the frequency of the data higher than the specified sampling frequency shall be reduced and the frequency of the data lower than the specified sampling frequency shall be raised. The daily frequency data are reduced to weekly frequency data. The quarterly and monthly data are raised to the weekly data and the missing value arises after the low-frequency data are raised to the high-frequency data. The raised data are processed by ascending order as per the date, and then the missing value is filled up by linear difference filling.
2. Fill up the missing value: There are some missing values and non-numerical parts in the original coal data which need to be filled up to better utilize the dataset. The missing part in the data is filled up by linear difference, and the non-numerical part is deleted and then the missing part deleted is filled up by linear difference. The equation of missing value between filling points (x_0, y_0) and (x_k, y_k) is shown in **Eq. 23**:

$$\phi(x) = \frac{x - x_k}{x_0 - x_k} y_0 + \frac{x - x_0}{x_k - x_0} y_k. \tag{23}$$

3. Standardize the dataset: There is a dimensional difference between different types of data. To avoid the dimensional error and lower the model performance, the standardized equation is used to transform the data distribution into standard distribution with the mean value of 0 and variance of 1. The standardized equation is shown as follows:

$$\left\{ x_{ki} = \frac{X_{ki} - \overline{X_i}}{\sigma_i} \sigma_i = \left(X_{1i} - \overline{X_i}\right)^2 + \frac{\left(X_{2i} - \overline{X_i}\right)^2 + \ldots + \left(X_{ni} - \overline{X_i}\right)^2}{n}. \right. \tag{24}$$

After transformation as per **Eq. 24**, the distribution of original feature data is transformed into a standard normal distribution with the mean value of 0 and variance of 1. x_{ki} indicates the kth numerical value of the ith feature index. $\overline{X_i}$ indicates the mean value of the ith feature index data, σ_i indicates the standard deviation of the ith feature index, and n indicates the sample size of the ith feature index.

4) Divide training dataset and test dataset: The training dataset and test dataset are not fixed, but they change dynamically. In the energy market, the influencing factors of energy indicators change with time (Liang et al., 2019). The coal data and related feature data at the time point within the sliding window of k width are used as the training dataset for

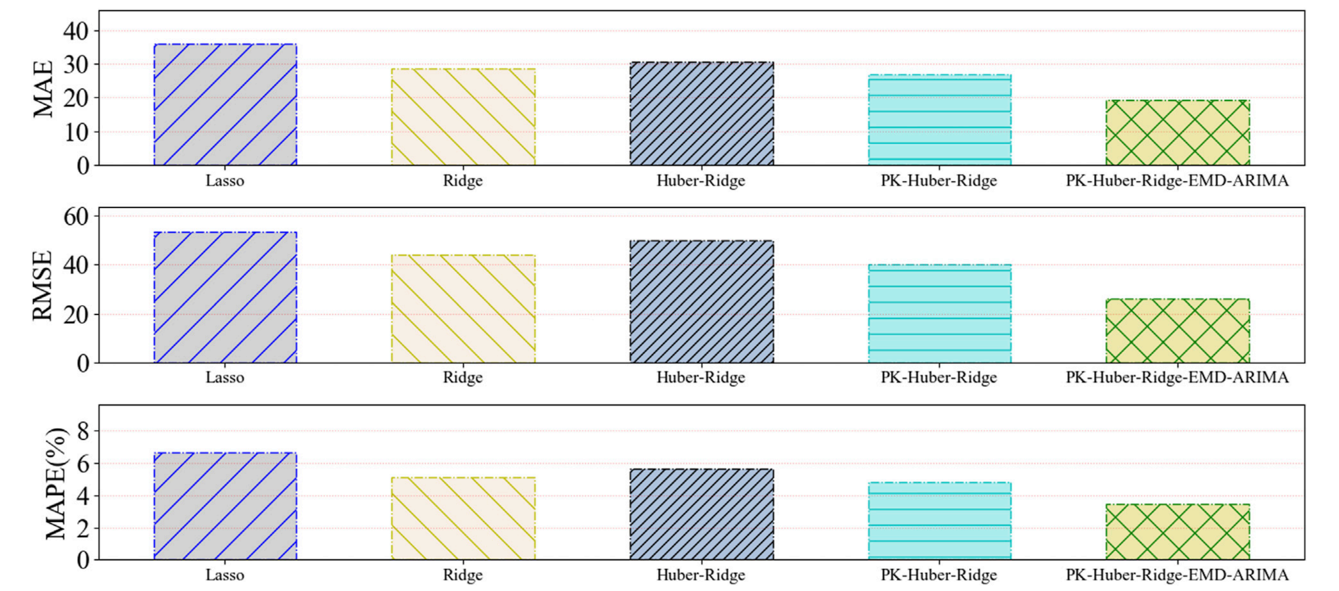


FIGURE 3 | A histogram of the predictive evaluation index values based on dataset 1 for five types of forecasting models.

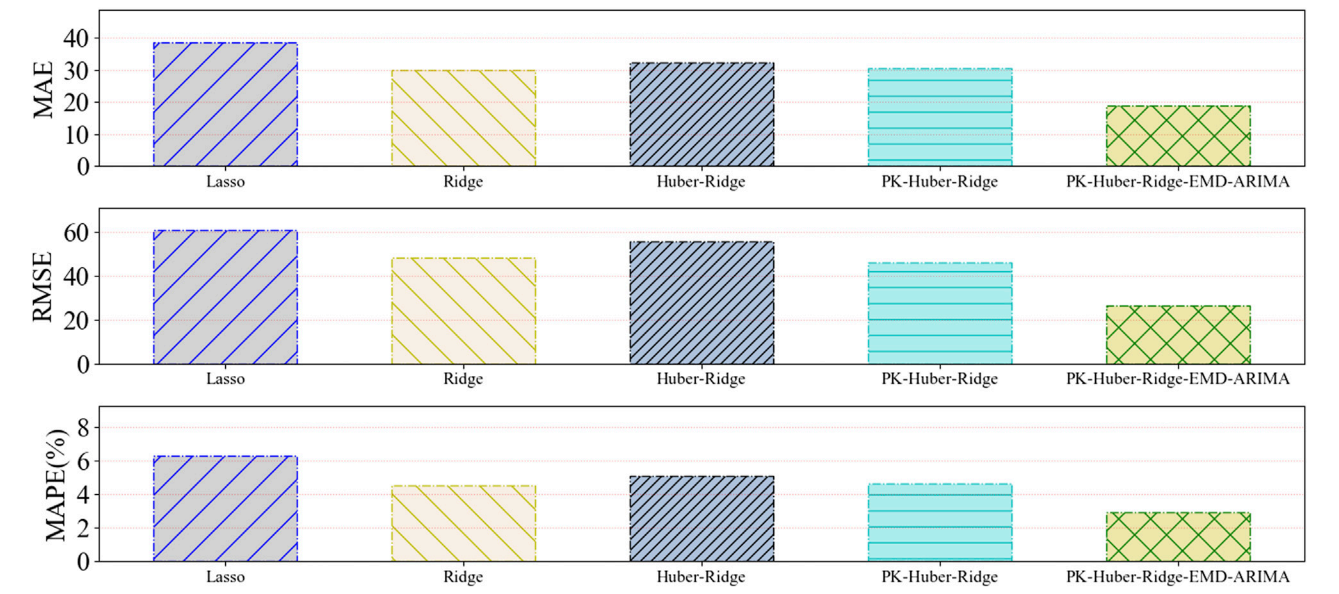


FIGURE 4 | A histogram of the predictive evaluation index values based on dataset 2 for five types of forecasting models.

training the price forecasting model. The price forecasting model outputs the forecasted coal price at a time point according to the coal-related feature data after sliding the window. The corresponding original price data and relevant feature data are used as the test dataset to verify the model forecasting performance. The dynamic division of the training dataset and test dataset progresses over time, as shown in **Figure 2**.

Feature Selection

Selecting comprehensive and relevant features can greatly improve the performance of the forecasting model. All feature data are presented as a data matrix, and the optimal feature variable is chosen according to the feature type and feature optimal time interval.

Feature type: The coal price pertains to many factors; there are many factors influencing coal market price, and the main

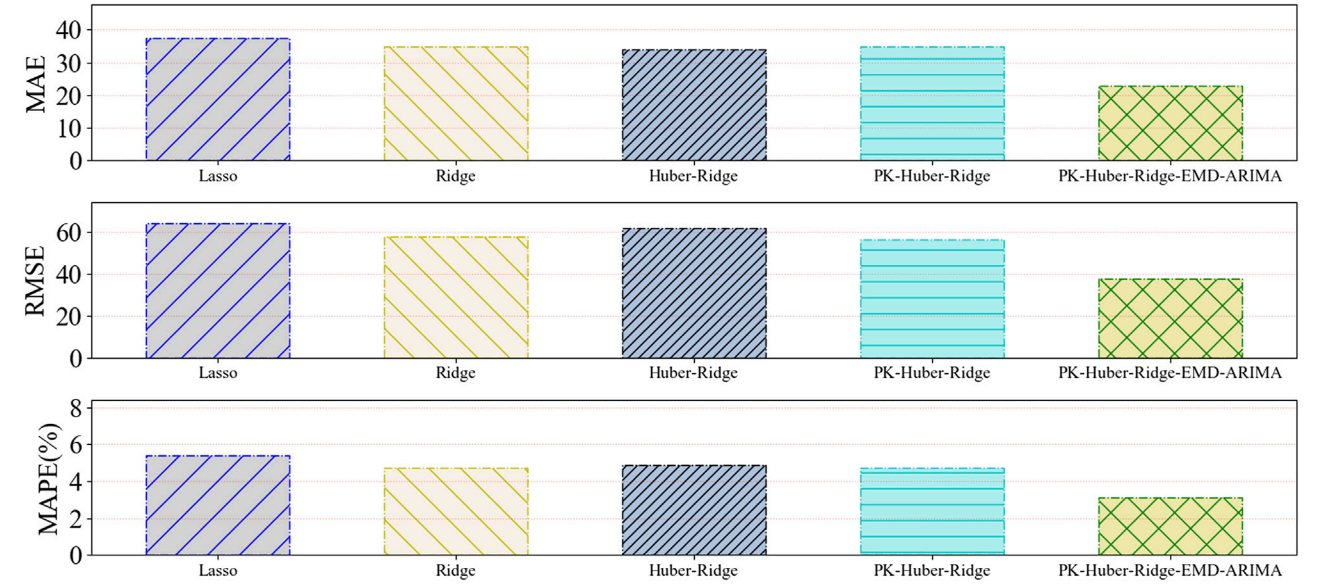


FIGURE 5 | A histogram of the predictive evaluation index values based on dataset 3 for five types of forecasting models.

influence factors cover coal supply coal consumption, coal transportation, and economic factor. The feature indexes chosen are shown in **Table 1**.

There are many initial feature indexes, so the feature screening needs to be performed. The feature variable at the same time point as the forecasted target variable does not necessarily have the highest correlation, and it is necessary to find out the optimal time interval, that is, optimal delay order, for each feature variable. The optimal delay linear feature and the optimal delay nonlinear relevant features are screened as per Pearson correlation coefficient and Spearman correlation coefficient.

The value range of Pearson correlation coefficient r_p Spearman and correlation coefficient r_s is [-1, 1]. The closer the absolute value of r_p and r_s is to 1, the stronger the correlation will be; the closer the absolute value of r_p and r_s is to 0, the weaker the correlation will be.

Given the corresponding threshold values of the Pearson correlation coefficient and Spearman correlation coefficient are r_{mp} and r_{sp}, respectively, the feature indexes whose correlation coefficient exceeds the threshold value r_{mp} and r_{sp} are screened. Given the maximum delay order of the feature is $delay_{max}$, the parameter setting is shown in **Table 2**.

The chosen feature variables and steam coal price are input to the forecasting model mentioned, and the model outputs the steam coal price at the next time point. **Table 3** and **Table 4** show the selection of the optimal delay feature variable when forecasting the steam coal price on Jul. 6, 2021. The feature variable selected as per this method changes over time.

Experiment Result

In this article, Lasso, Ridge, Huber–Ridge, PK–Huber–Ridge, and PK–Huber–Ridge–EMD–ARIMA models are used for comparison. One-step forecasting is used for empirical test. Qinhuangdao thermal coal data and feature data at the first 120 time points are used as the data variables of the forecasting model, and the forecasting model outputs the thermal coal price data at the 121st time point.

The set values of hyperparameters of Lasso, Ridge, Huber–Ridge, PK–Huber–Ridge, and PK–Huber–Ridge–EMD–ARIMA models are shown in **Table 5**.Here, α_{l1} is the coefficient of L1 regular term; α_{l2} is the coefficient of L2 regular term; M is the threshold of Huber loss function; d_k is the kernel parameter of the polynomial kernel and represents the order of the polynomial; p is the autoregressive order of ARIMA model; d_a represents the difference order; and q represents the moving average order. BIC criterion (Burnham and Anderson, 2004) is used to select the optimal ARIMA model hyperparameters p, d_a, and q.

The forecasting model is used to forecast 4500-kilocalorie steam coal price data (Dataset 1), 5000-kilocalorie steam coal price data (Dataset 2), and 5500-kilocalorie steam coal price data (Dataset 3) of Qinhuangdao Port from March 17, 2020 to July 6, 2021. **Table 6** and **Figure 3** and **Figure 4** and **Figure 5** show the evaluation index results of the forecasting model.

Through the comparison of the experimental results of five thermal coal price forecasting models, the following conclusions can be obtained.

Compared with the single model, the proposed combination model has a better forecasting performance. In dataset 1, dataset 2, and dataset 3 experiments, the forecasting performance of PK–Huber–Ridge–EMD–ARIMA model is better than the Lasso model, Ridge model and Huber–Ridge model, and PK–Huber–Ridge model. The thermal coal price dataset is complex, and the forecasting performance of a single forecasting model is very limited. The combination model can better deal with complex datasets. PK–Huber–Ridge–EMD–ARIMA model adopts the method of decomposition integration and time series forecasting model to compensate for the error of single model. We consider the residual error rule of model forecasting to complement the hidden rules that the original single model does not learn.

Compared with the ordinary model, the robust kernel function model has better performance. In dataset 1, dataset 2, and dataset 3 experiments, the forecasting performance of PK–Huber–Ridge–EMD–ARIMA model is better than the Lasso model, Ridge model, and Huber–Ridge model. Thermal coal dataset has nonlinear law. PK–Huber–Ridge–EMD–ARIMA model uses polynomial kernel function to map nonlinear features into high-dimensional space, so that the linear model can learn the nonlinear law in the original feature space, so as to further improve the forecasting performance of the forecast model.

RESULT AND DISCUSSION

For the nonlinearity law in the steam coal dataset and limitations of the single forecasting model, this article proposes the forecast method for the price of steam coal based on robust regularized kernel regression and empirical mode decomposition. The robust regularized kernel regression model learns the nonlinearity law in the original data by using the kernel function. This model selects the Huber loss function to enhance the robustness of the forecasting model. We select the L2 regular term to lower the risk of model overfitting. The combined model based on EMD and ARIMA is used for error compensation against the Huber–Ridge polynomial kernel model, further improving the forecasting performance of the forecasting model. Compared to Lasso, Ridge, Huber–Ridge, and PK–Huber–Ridge, the proposed forecasting model (PK–Huber–Ridge–EMD–ARIMA) has the minimum value of MAE, RMSE, and MAPE.

The influence factors of steam coal price are complex which are easily affected by national policies. How to quantify policy factors and input them into the forecasting model for model training and model forecasting is the next work.

DATA AVAILABILITY STATEMENT

The data analyzed in this study are subject to the following licenses/restrictions: The data used to support the findings of this study are currently under embargo, while the research findings are commercialized. Requests to access these datasets should be directed to tmz@csust.edu.cn.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 62173050 and 61403046), the Natural Science Foundation of Hunan Province, China (Grant No. 2019JJ40304), Changsha University of Science and Technology “The Double First Class University Plan” International Cooperation and Development Project in Scientific Research in 2018 (Grant No. 2018IC14), the Research Foundation of the Education Bureau of Hunan Province (Grant No.19K007), Hunan Provincial Department of Transportation 2018 Science and Technology Progress and Innovation Plan Project (Grant No. 201843), Energy Conservation and Emission Reduction Hunan University Student Innovation and Entrepreneurship Education Center, Innovative Team of Key Technologies of Energy Conservation, Emission Reduction and Intelligent Control for Power-Generating Equipment and System, CSUST, Hubei Superior and Distinctive Discipline Group of Mechatronics and Automobiles(XKQ2021003 and XKQ2021010), Major Fund Project of Technical Innovation in Hubei (Grant No. 2017AAA133), and Guangxi Key Laboratory of Trusted Software (No.201728), Graduate Scientific Research Innovation Project of Changsha University of Science and Technology (No. 2021-89).

REFERENCES

Alameer, Z., Fathalla, A., Li, K., Ye, H., and Jianhua, Z. Multistep-ahead Forecasting of Coal Prices Using a Hybrid Deep Learning Model. *Resour. Pol.* 65, 2020.

Burnham, K. P., and Anderson, D. R. (2004). Multimodel Inference. *Sociological Methods Res.* 33 (2), 261–304. doi:10.1177/0049124104268644

Chai, J., Zhao, C., Hu, Y., and Zhang, Z. G. (2021). Structural Analysis and Forecast of Gold price Returns. *J. Manag. Sci. Eng.* 6 (2), 135–145. doi:10.1016/j.jmse.2021.02.011

Conejo, A. J., Plazas, M. A., Espinola, R., and Molina, A. B. (2005). Day-ahead Electricity price Forecasting Using the Wavelet Transform and ARIMA Models. *IEEE Trans. Power Syst.* 20 (2), 1035–1042. doi:10.1109/tpwrs.2005.846054

Cover, T. M. (1965). Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition. *IEEE Trans. Electron. Comput.* EC-14 (3), 326–334. doi:10.1109/pgec.1965.264137

Gilbert, K. (2005). An ARIMA Supply Chain Model. *Manag. Sci.* 51 (2), 305–310. doi:10.1287/mnsc.1040.0308

Gupta, D., Hazarika, B. B., and Berlin, M. (2020). Robust Regularized Extreme Learning Machine with Asymmetric Huber Loss Function. *Neural Comput. Applic* 32 (16), 12971–12998. doi:10.1007/s00521-020-04741-w

Huber, P. J. (1992). “Robust Estimation of a Location Parameter,” in *Breakthroughs in Statistics: Methodology and Distribution*. Editors S. Kotz and N. L. Johnson (New York, NY: Springer New York), 492–518. doi:10.1007/978-1-4612-4380-9_35

Ye, J., He, L., and Jin, H. (2021). A Denoising Carbon price Forecasting Method Based on the Integration of Kernel Independent Component Analysis and Least Squares Support Vector Regression. *Neurocomputing* 434, 67–79.

Ji, L., Zou, Y., He, K., and Zhu, B. (2019). Carbon Futures price Forecasting Based with ARIMA-CNN-LSTM Model. *Proced. Comp. Sci.* 162, 33–38. doi:10.1016/j.procs.2019.11.254

Jianke Zhu, J., Hoi, S., and Lyu, M. R. T. (2008). Robust Regularized Kernel Regression. *IEEE Trans. Syst. Man. Cybern. B* 38 (6), 1639–1644. doi:10.1109/tsmcb.2008.927279

Karabiber, O. A., and Xydis, G. (2019). Electricity Price Forecasting in the Danish Day-Ahead Market Using the TBATS, ANN and ARIMA Methods. *Energies* 12 (5). doi:10.3390/en12050928

Li, Y., and Li, Z. (2019). Forecasting of Coal Demand in China Based on Support Vector Machine Optimized by the Improved Gravitational Search Algorithm. *Energies* 12 (12). doi:10.3390/en12122249

Liang, T., Chai, J., Zhang, Y. J., and Zhang, Z. G. (2019). Refined Analysis and Prediction of Natural Gas Consumption in China. *J. Manag. Sci. Eng.* 4 (2), 91–104. doi:10.1016/j.jmse.2019.07.001

Lu, H., Ma, X., Huang, K., and Azimi, M. Carbon Trading Volume and price Forecasting in China Using Multiple Machine Learning Models. *J. Clean. Prod.* 249, 2020.

Matyjaszek, M., Riesgo Fernández, P., Krzemień, A., Wodarski, K., and Fidalgo Valverde, G. (2019). Forecasting Coking Coal Prices by Means of ARIMA Models and Neural Networks, Considering the Transgenic Time Series Theory. *Resour. Pol.* 61, 283–292. doi:10.1016/j.resourpol.2019.02.017

Owen, A. B. (2006). A Robust Hybrid of Lasso and ridge Regression. *Contemp. Math.* 443, 59–72.

Qiao, W., Liu, W., and Liu, E. A Combination Model Based on Wavelet Transform for Predicting the Difference between Monthly Natural Gas Production and Consumption of U.S. *Energy* 235, 2021.

Schölkopf, B., Herbrich, R., and Smola, A. J. (2001).A Generalized Representer Theorem. In Computational Learning Theory. Berlin, Heidelberg, 416–426. doi:10.1007/3-540-44581-1_27

Vu, D. H., Muttaqi, K. M., Agalgaonkar, A. P., and Bouzerdoum, A. (2019). Short-Term Forecasting of Electricity Spot Prices Containing Random Spikes Using a Time-Varying Autoregressive Model Combined with Kernel Regression. *IEEE Trans. Ind. Inf.* 15 (9), 5378–5388. doi:10.1109/tii.2019.2911700

Wang, B., and Wang, J. Energy Futures and Spots Prices Forecasting by Hybrid SW-GRU with EMD and Error Evaluation. *Energ. Econ.*, 90, 2020.

Wang, J., Cao, J., Yuan, S., and Cheng, M. Short-term Forecasting of Natural Gas Prices by Using a Novel Hybrid Method Based on a Combination of the CEEMDAN-SE-And the PSO-ALS-Optimized GRU Network. *Energy* 233, 121082.

Wang, J., Lei, C., and Guo, M. Daily Natural Gas price Forecasting by a Weighted Hybrid Data-Driven Model. *J. Pet. Sci. Eng.* 192, 2020.

Wang, J., Zhou, H., Hong, T., Li, X., and Wang, S. A Multi-Granularity Heterogeneous Combination Approach to Crude Oil price Forecasting. *Energ. Econ.* 91, 2020.

Wang, K., and Du, F. (2020). Coal-gas Compound Dynamic Disasters in China: A Review. *Process Saf. Environ. Prot.* 133, 1–17. doi:10.1016/j.psep.2019.10.006

Wu, J., Chen, Y., Zhou, T., and Li, T. (2019). An Adaptive Hybrid Learning Paradigm Integrating CEEMD, ARIMA and SBL for Crude Oil Price Forecasting. *Energies* 12 (7). doi:10.3390/en12071239

Xia, C., and Wang, Z. Drivers Analysis and Empirical Mode Decomposition Based Forecasting of Energy Consumption Structure. *J. Clean. Prod.* 254, 2020.

Xiong, J., and Xu, D. (2021). Relationship between Energy Consumption, Economic Growth and Environmental Pollution in China. *Environ. Res.* 194.

Xu, W., Hu, H., and Yang, W. (2019). Energy Time Series Forecasting Based on Empirical Mode Decomposition and FRBF-AR Model. *IEEE Access* 7, 36540–36548. doi:10.1109/access.2019.2902510

Yang, S., Chen, D., Li, S., and Wang, W. (2020). Carbon price Forecasting Based on Modified Ensemble Empirical Mode Decomposition and Long Short-Term Memory Optimized by Improved Whale Optimization Algorithm. *Sci. Total Environ.* 716, 137117. doi:10.1016/j.scitotenv.2020.137117

Yu, L., Wang, S., and Lai, K. K. (2008). Forecasting Crude Oil price with an EMD-Based Neural Network Ensemble Learning Paradigm. *Energ. Econ.* 30 (5), 2623–2635. doi:10.1016/j.eneco.2008.05.003

Zhang, H., Yang, Y., Zhang, Y., He, Z., Yuan, W., Yang, Y., et al. (2021). A Combined Model Based on SSA, Neural Networks, and LSSVM for Short-Term Electric Load and price Forecasting. *Neural Comput. Applic* 33 (2), 773–788. doi:10.1007/s00521-020-05113-0

Zhang, X., Zhang, C., and Wei, Z. (2019). Carbon Price Forecasting Based on Multi-Resolution Singular Value Decomposition and Extreme Learning Machine Optimized by the Moth–Flame Optimization Algorithm Considering Energy and Economic Factors. *Energies* 12 (22). doi:10.3390/en12224283

Zhou, J., Huo, X., Xu, X., and Li, Y. (2019). Forecasting the Carbon Price Using Extreme-Point Symmetric Mode Decomposition and Extreme Learning Machine Optimized by the Grey Wolf Optimizer Algorithm. *Energies* 12 (5). doi:10.3390/en12050950

Conflict of Interest: DX was employed by China Datang Corporation Ltd., Beijing, China.

JY was employed by Hunan Datang Xianyi Technology Co., Ltd., Changsha, China.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Energy Research

ORIGINAL RESEARCH
published: 08 November 2021
doi: 10.3389/fenrg.2021.763977

Municipal Solid Waste Forecasting in China Based on Machine Learning Models

Liping Yang[1,2], Yigang Zhao[3], Xiaxia Niu[1], Zisheng Song[4], Qingxian Gao[5] and Jun Wu[1]*

[1]School of Economics and Management, Beijing University of Chemical Technology, Beijing, China, [2]School of Management, University of Science and Technology of China, Anhui, China, [3]Beijing Institute of Petrochemical Technology, Beijing, China, [4]Department of International Exchange and Cooperation, Beijing University of Chemical Technology, Beijing, China, [5]Chinese Research Academy of Environmental Sciences, Beijing, China

As the largest producing country of municipal solid waste (MSW) around the world, China is always challenged by a lower utilization rate of MSW due to a lack of a smart MSW forecasting strategy. This paper mainly aims to construct an effective MSW prediction model to handle this problem by using machine learning techniques. Based on the empirical analysis of provincial panel data from 2008 to 2019 in China, we find that the Deep Neural Network (DNN) model performs best among all machine learning models. Additionally, we introduce the *SHapley Additive exPlanation* (SHAP) method to unravel the correlation between MSW production and socioeconomic features (e.g., total regional GDP, population density). We also find the increase of urban population and agglomeration of wholesales and retails industries can positively promote the production of MSW in regions of high economic development, and vice versa. These results can be of help in the planning, design, and implementation of solid waste management system in China.

Keywords: municipal solid waste, influencing factors, machine learning, deep learning, SHAP value

OPEN ACCESS

Edited by:
Xun Zhang,
Academy of Mathematics and Systems Science (CAS), China

Reviewed by:
Kenneth E. Okedu,
National University of Science and Technology (Muscat), Oman
Bo Liu,
King Fahd University of Petroleum and Minerals, Saudi Arabia

****Correspondence:***
Jun Wu
wujun@mail.buct.edu.cn

Specialty section:
This article was submitted to Smart Grids, a section of the journal Frontiers in Energy Research

Received: 24 August 2021
Accepted: 19 October 2021
Published: 08 November 2021

Citation:
Yang L, Zhao Y, Niu X, Song Z, Gao Q and Wu J (2021) Municipal Solid Waste Forecasting in China Based on Machine Learning Models. Front. Energy Res. 9:763977. doi: 10.3389/fenrg.2021.763977

INTRODUCTION

Over the past decade, the urban population in China has reached up to 900 million residents with an urbanization rate of over 60% (NBSC, 2021), which significantly challenges the existing urban sources (e.g., water, air, and energy) related to residents' life quality (Hoornweg and Bhada-Tata, 2012). The municipal solid waste (MSW), as renewable energy, is considered an essential part of the Waste-to-Energy (WtE) system (Ouda et al., 2013; Kuznetsova et al., 2019; Mukherjee et al., 2020). It is reported that the production of MSW in China was around 242 million tons in 2020 compared with that of 8.17 million tons in 2008 (NBSC, 2020). In other words, the efficient management of municipal solid waste is becoming an important concern for urban sustainability governance. However, the utilization efficiency of MSW was merely about 45% in China, which was much lower than that in other advanced countries, such as over 80% in Japan (Ding et al., 2021). Therefore, how to increase the utilization efficiency of MSW would impact both central and local governments in China to promote urban sustainable development (He and Lin, 2019).

In general, an integrated decision-support methodology for waste-to-energy management systems (WtEMS) design is mainly composed of three modules: 1) the waste modeling and prediction, 2) optimization of WtEMS, and 3) a multi-dimensional assessment, as shown in **Figure 1** (Kuznetsova et al., 2019). Among these three modules, waste modeling and its prediction of MSW play a fundamental role in effectively conducting urban planning and energy management. Many international scholars have carried out extensive studies on this

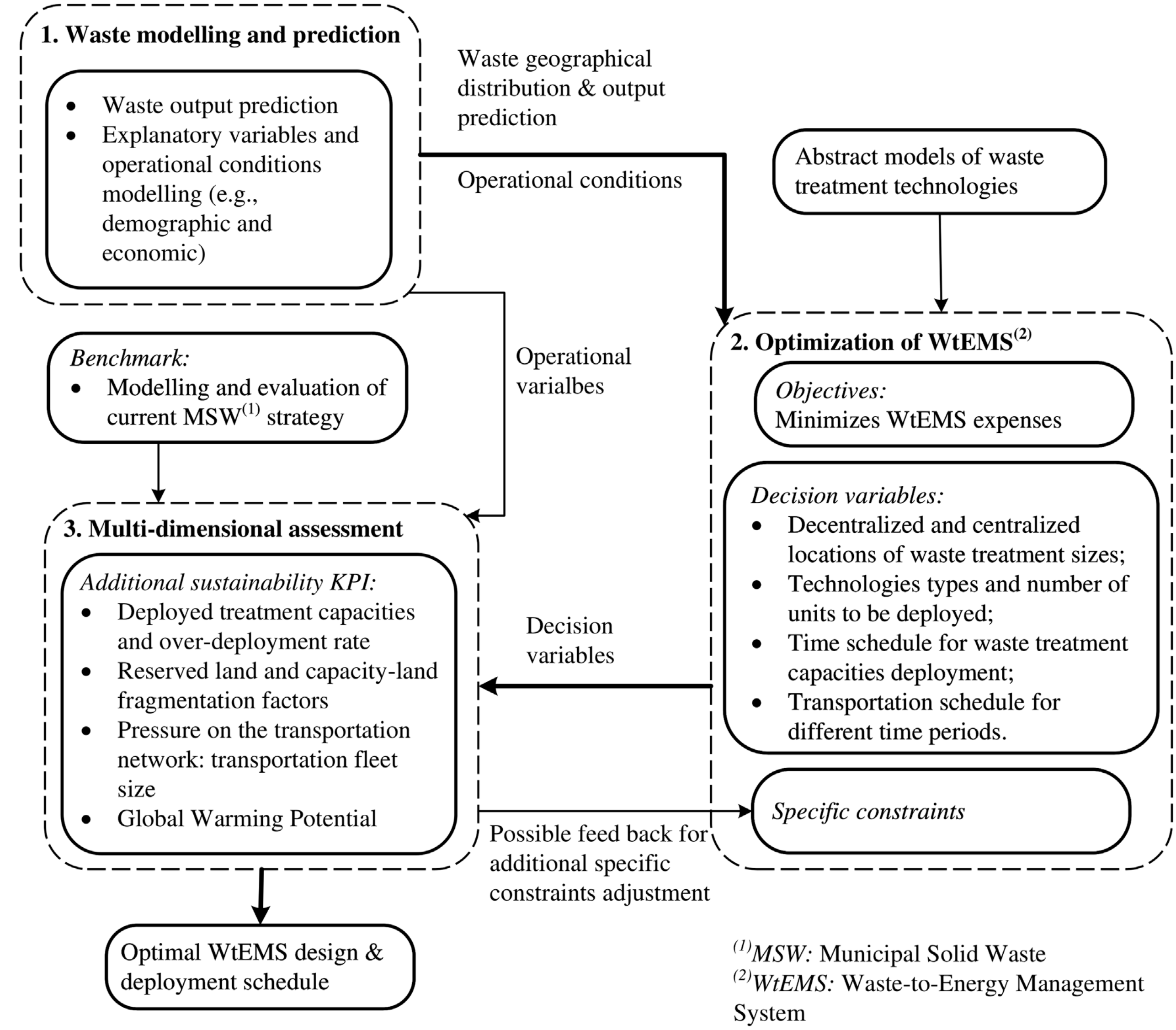

1. Waste modelling and prediction
• Waste output prediction
• Explanatory variables and operational conditions modelling (e.g., demographic and economic)
Waste geographical distribution & output prediction
Operational conditions
Abstract models of waste treatment technologies
Benchmark:
• Modelling and evaluation of current MSW(1) strategy
Operational varialbes
2. Optimization of WtEMS(2)
Objectives:
Minimizes WtEMS expenses
Decision variables:
• Decentralized and centralized locations of waste treatment sizes;
• Technologies types and number of units to be deployed;
• Time schedule for waste treatment capacities deployment;
• Transportation schedule for different time periods.
3. Multi-dimensional assessment
Additional sustainability KPI:
• Deployed treatment capacities and over-deployment rate
• Reserved land and capacity-land fragmentation factors
• Pressure on the transportation network: transportation fleet size
• Global Warming Potential
Decision variables
Specific constraints
Possible feed back for additional specific constraints adjustment
Optimal WtEMS design & deployment schedule
(1)MSW: Municipal Solid Waste
(2)WtEMS: Waste-to-Energy Management System

FIGURE 1 | Integrated decision support method for WtEMS design: methodology flowchart.

module by using group comparisons, time series analysis, and system dynamics (Beigl et al., 2008). Recently, with the popularity of machine learning (ML) methods, alternative methods were put forward to forecast the quantity of generated municipal solid waste effectively (Guo et al., 2021). For instance, based on the example of Suzhou (Niu et al., 2021), constructed the long short-term memory (LSTM) neural network, autoregressive integrated moving average (ARIMA), and traditional neural network to predict the MSW production. They found that the LSTM played a vital role in predicting MSW production but did not reveal the correlation between the production of MSW and socio-economic variables. Nguyen et al. (2021) selected residential areas in Vietnam as a case of study and figured out that both the random forest (RF) and the k-nearest neighbor (KNN) approaches performed effectively in predicting the amount of urban waste. Birgen et al. (2021) developed a Gaussian Processes Regression (GPR) method to predict the daily lower heating value of MSW by combining the historical data of a WtE plant and the weather and calendar data. In addition, other ML methods, such as the support vector machine (SVM) (Kumar et al., 2018) and decision tree (Kannangara et al., 2018) have also been employed to predict the MSW production.

Similar to other energy forecasting research topics (e.g., crude oil prices, gas consumption), MSW production is also was highly influenced by various socio-economic factors (Zhang et al., 2009; Liang et al., 2019; Huang et al., 2021a). However, previous studies neither revealed the correlation between each factor and MSW production nor identified their interaction in different socio-economic circumstances (Kannangara et al., 2018; Niu et al., 2021; Nguyen et al., 2021). In the context of China, existing

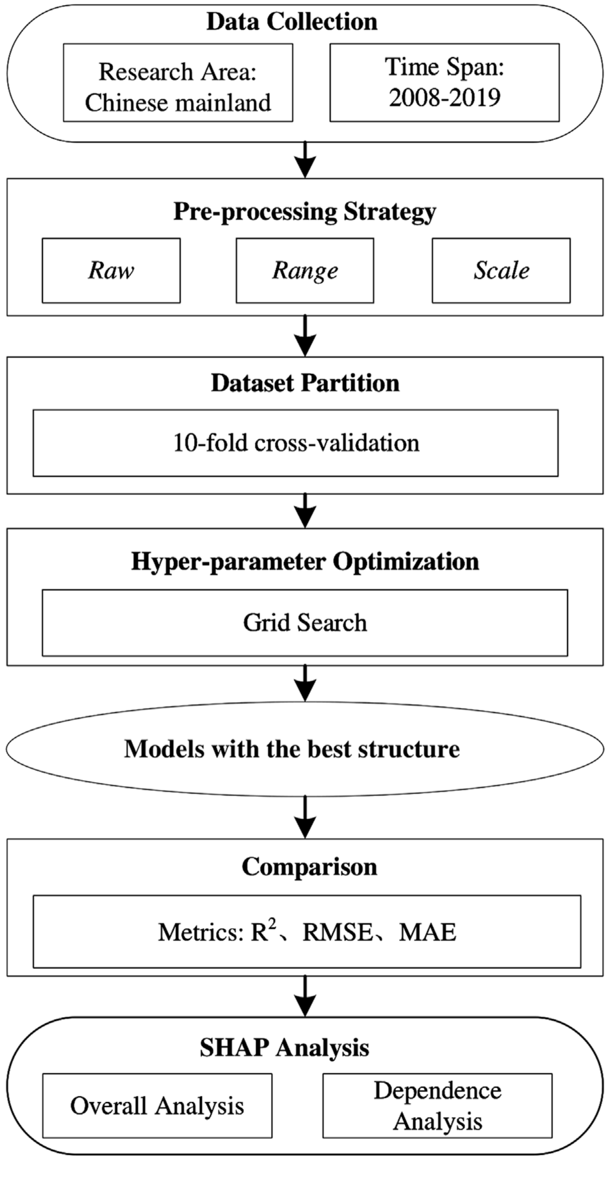


FIGURE 2 | Procedures of methodology.

studies scarcely discussed the performances and applications of different ML methods in predicting MSW. Therefore, this paper mainly aimed to construct a prediction model by using machine learning models by using provincial panel data of 2008–2019 in China. Besides, it also discussed the comparison of the performances of six different ML models in predicting China's municipal solid waste generation. Considering that data input form and model hyperparameters have a great influence on prediction results, we tested different preprocessing strategies to ensure robust estimation and prediction of the ML model. Finally, this paper provided some potential implications for both policy-makers and other industry stakeholders in terms of convincing evidence concluded from the ML prediction model.

The initial contributions of this paper are threefold. First, it emphasized the good performance of machine learning approaches in predicting MSW production and extended the existing literature to construct a prediction model by comparing six supervised learning algorithms. These models varied from linear, non-linear to ensemble methods and artificial neural network methods, including a body of discussions on data preprocessing, resampling, model training, testing, and interpretation steps. Therefore, the constructed prediction model of MSW would theoretically shed light on other similar research related to prediction issues in the future. Second, this paper estimated the impacts of diverse socio-economic factors on MSW production, such as the regional economic development level (e.g., regional GDP, population density, per capita disposable income), industrial structure (e.g., wholesale and retail values added), and waste generation characteristics. Third, to improve the interpretations of ML models, this paper employed the *SHapley Additive exPlanation* (SHAP) approach and visualized the SHAP value of each explanatory variable. This technique would also provide good evidence to explain the outcomes of ML models for other researchers in the future.

The remaining sections of this paper are organized as follows: *Materials and Methods* describes the models adopted in this paper and the process of data acquisition. *Results* reports the results of comparison among six ML models, *via* presenting the predictive capability and SHAP analysis. *Conclusion* provides conclusions and some implications.

MATERIALS AND METHODS

Figure 2 outlines the main steps of the methodology used in this study. In this paper, we first preprocessed the original database and selected critical variables for MSW prediction. Second, this paper focused on comparing with six ML models, including the multiple linear regression (MLR), support vector regression (SVR), Random Forest, extreme gradient boosting (XGBoost), k-nearest neighbor, and deep neural network (DNN). Thirdly, three evaluation metrics are used to compare the prediction performance of each algorithm. Finally, the SHAP method is employed to analyze and discuss the output.

ML-Based Models and Applications for Waste Prediction

The Multiple Linear Regression Liner Model

The multiple linear regression is a commonly used ML method to estimate the marginal effects of independent variables (or called feature vector in machine learning techniques) on the dependent variable. It is widely applied to waste prediction of desirable explanatory power in different regions and countries (Beigl et al., 2008). In China, this approach is also employed to predict the MSW production in "Calculation and Prediction Method of Municipal Solid Waste Production (CJ/T 106-1999)", which is the official guide compiled by the Ministry of Construction, China.

The model can be expressed as **Eq. 1**:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \epsilon, \quad (1)$$

where Y is MSW generation in this paper, β_0 denotes regression constant, $\beta_1 \sim \beta_k$ are regression coefficients, $X_1 \sim X_k$ are explanatory variables, ϵ marks the regression residuals.

Usually, MLR uses the ordinary least squares (OLS) method to estimate the parameters that can achieve the lowest sum-of-squared errors between the observed and predicted responses. Under the OLS estimation, MLR's results could be easily interpreted. However, some drawbacks have to be considered in MLR. For instance, the multicollinearity among the predictors can result in estimation errors, as well as the omitted variables could induce a biased estimation. In this paper, we mainly concentrated on the performance of each ML model and considered the variables selection based on earlier studies (Kannangara et al., 2018; Namlis and Komilis, 2019; Niu et al., 2021; Nguyen et al., 2021). The multicollinearity and omitted variables problems are not our concerns.

Support Vector Regression

SVM was originally used to deal with pattern recognition problems, and recently extended to estimate regression models due to its properties of the sparse solution and good generalization (Demir and Bruzzone, 2014). By introducing an ε-tube to reformulate the optimization problem, the SVM model could be transformed to an SVR model and finds the optimal approximation of the continuous-valued function while balancing the complexity and prediction error of the prediction model (Huang et al., 2021b). In addition, the accuracy of an SVR model heavily relies on three parameters: a penalty parameter (C), the kernel width (γ) and the precision parameter (ε) (Abbasi and El Hanandeh, 2016; Li et al., 2021). Specifically, the smaller C is, the smaller the fitting error and the weaker the generalization ability would be. The larger γ is, the more support vectors; and vice versa. ε is a precision parameter representing the tube's radius located around the regression function. In other words, the choice of ε donates the magnitude of errors that can be neglected. Since the above three parameters are critical to the adaptability of the model, we will tune them using a grid optimization approach in *Results* to optimize the SVR model.

A great body of literature has discussed the SVR and SVM models in predicting the generation of MSW. For example (Abbasi and El Hanandeh, 2016), adjusted the hyper-parameters of SVR by combining the grid search method and applying the model with the optimal parameters to the monthly prediction of MSW in Logan City, Australia. They found that SVR can effectively reduce the mean absolute error (*MAE*) and root-mean-square error (*RMSE*), and improve prediction performance (*R-square*). Besides (Nguyen et al., 2021), applied SVM to the prediction of MSW production in Vietnam with an *MAE* of 131.07, which confirmed that the SVM model performed a better prediction. Kumar et al. (2018) applied it to the prediction of the production rate of plastic waste, and found that the prediction result of SVM (R^2=0.74) is better than RF (R^2=0.66) and lower than artificial neural network (ANN) (R^2=0.75). Mehrdad et al. (2021) argued that SVM was superior to both the adaptive neuro-fuzzy inference system and artificial neural network models in predicting methane generation.

Random Forest

Random Forest is an evolution of Bagging which aims to reduce the variance of a statistical model, simulates the variability of data through the random extraction of bootstrap samples from a single training set and aggregates predictions on a new record (see Breiman, 1996). It performs a more stable and better prediction of explained variables than other machine learning models (Huang et al., 2021b). Generally, the RF algorithm implementation can be expressed as follows:

1) Bagging is used to randomly generate sample subsets;
2) Use the idea of random subspace by randomly extracting features, splitting nodes, and building a regression sub-decision tree;
3) Repeat the above steps to construct T (the number of decision trees) regression decision subtrees to form a random forest;
4) Take the predicted values of T sub-decision trees and take the mean as the final prediction result.

The RF model was widely used in the prediction of waste. Kumar et al. (2018) used RF for the prediction of plastic waste generation rate that showed an R-square of 0.66. The size of the random forest, that is, the number of decision trees ($Ntrees$) and the number of features tried in each segmentation ($Nfeatures$) have a significant impact on the predictive ability of the RF model (Hariharan, 2021). When $Ntrees$ exceed a certain value, the prediction performance of the model converges. In this case, increasing the number of decision trees will not improve the model performance, but will result in model redundancy. In addition, using a smaller number of $Ntrees$ reduces the similarity in the forest, but also reduces the complexity and strength of the model. Conversely, the increase in $Ntrees$ can make each tree more powerful, but also increase the correlation between the trees. Therefore, in the following section, we will optimize these two hyper-parameters to acquire better results.

Extreme Gradient Boosting

XGBoost algorithm, proposed in 2016, is a relatively new approach (Chen and Guestrin, 2016). Different from RF model using bagging integration method, XGBoost model is an integration tree model using boosting method to integrate classification and regression tree (CART). It has the advantages of fast training speed and high prediction accuracy. The result of XGBoost is the sum of prediction scores of all CARTs (Chen and Guestrin, 2016) as formed in **Eq. 2**:

$$\hat{y} = \sum_{n=1}^{N} f_m(X), \quad (2)$$

where N represents the number of trees in the model, f_m represents each CART tree and $\hat{y}$ is predicted result.

Since its introduction, the XGBoost model has been widely used in the prediction of oil price (Costa et al., 2021) and energy usage (Feng et al., 2021). However, up to date, XGBoost model has not been applied to the research of MSW generation prediction. Similar to RF, the number of integrated CARTs (*Ntrees*) in XGBoost has a great influence on the prediction performance. Therefore, in order to increase the model's performance in predicting the MSW generation, it is necessary to optimize this hyper-parameter. In *Results*, we also use the grid search method to confirm the different combinations of these two parameters to obtain the optimal model structure.

K-Nearest Neighbor

KNN algorithm is a non-parametric learning method first proposed by Cover and Hart (Cover and Hart, 1967). Since its introduction, it has been widely used in regression and classification due to its simple and intuitive mathematical form (Wu et al., 2008). It is essentially a supervised learning technique that *via* the clustering algorithm classify the similarity between the test sample and *K* nearest training samples (Zheng et al., 2020). Here, *K* is a user-defined number, normally an odd number, and the similarity is measured by the commonly used Euclidean distance. The test sample is classified based on the most frequent classification among the training samples. The mean value of the *K* nearest training samples is regarded as the predicted value. The mathematical measurement of Euclidean distance is expressed in **Eq. 3**:

$$\begin{aligned} d(x,y) &= \sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\ldots+(x_n-y_n)^2} \\ &= \sqrt{\sum\nolimits_{i=1}^{n}(x_i-y_i)^2} \end{aligned} \tag{3}$$

One drawback of KNN approach is the pre-selected number of K, a hyperparameter, because it would greatly influence the numbers of nearest samples (Wu et al., 2008; Zheng et al., 2020). In the following section, we first limit K to positive integers between 1 and 30, and then cross-verify them on a 10-fold sample to avoid this drawback.

Several studies applied the KNN approach into the prediction of MSW. For example, (Abbasi and El Hanandeh, 2016) first attempt to evaluate the ability of KNN to forecast MSW generation. They concluded that KNN can give good prediction performance and may be applied to establish the forecasting models that could provide accurate and reliable MSW generation prediction. Nguyen et al. (2021) predicted the MSW production in Vietnam and the R-square was over 0.96, which indicated that more than 96% of MSW production would be explained by the KNN model.

Artificial Neural Network

The ANN model is a computational system composed of multiple layers of neurons (input-hidden-output) (Al-Dahidi et al., 2019). This model is widely used in waste management because of its strong fault-tolerant ability to describe the complex relationship between variables in a multivariate system. (Abbasi and El Hanandeh, 2016; Mehrdad et al., 2021; Nguyen et al., 2021; Niu et al., 2021). The deep neural network is a branch of ANN based on a perceptron model. Indeed, an ANN model with multiple hidden layers is called a DNN since it has to train and process through multiple layers (Liu et al., 2017). The structure of DNN also includes input layer, hidden layer, and output layer. In general, the structure of DNN and ANN is similar, and their training algorithm is not different. However, studies showed that DNN tends to provide better performance and accuracy than conventional ANN models (Yang et al., 2021).

In this paper, a DNN with four layers of structure is constructed, namely the input layer, the first hidden layer, the second hidden layer and the output layer with one neuron. The number of neurons in the hidden layer has a great influence on the prediction performance of DNN. The smaller the number of neurons, the more likely it is to lead to insufficient fitting. On the contrary, an excessive number of neurons may lead to over-fitting. Therefore, selecting the appropriate number of neurons for DNN is also one of the bases to improve the model performance. In this paper, the number of neurons in the first hidden layer ($Nh1$) and the number of neurons in the second hidden layer ($Nh2$) are optimized to gain better results. Specifically, we first specify the numerical space of the number of neurons, and then test on the train and test samples, taking the optimal result as the optimal network structure.

Data Collection

In this paper, we aim to construct a ML-based prediction model of MSW production that is the predictor in all ML models. However, because there are no relevant statistics of MSW production in China at present, we utilize a proxy indicator of the MSW removal volume (Niu et al., 2021; Namlis and Komilis, 2019). More specifically, we obtained this annual statistical data for all provinces in mainland China from 2008 to 2019 to support our research.

The input variables of this paper in predicting MSW production are collected from provincial panel databases of the China Statistical Yearbook 2008–2019. Nine diverse socio-economic factors on MSW production, such as the regional economic development level (e.g., regional GDP, population density, per capita disposable income), industrial structure (e.g., wholesale and retail values added), and waste generation characteristics are obtained (Nguyen et al., 2021). **Table 1** reported the variable definition and descriptive statistics. As plotted in **Figure 3**, the skewness and kurtosis of each variable existed noticeable differences. To mitigate the influences in predicting the MSW production, we employ three different data preprocessing methods and proceed to explore the model's performance under different circumstances in the following sub-sections.

Machine Learning Techniques

Data Preprocessing and Re-Sampling

The preprocessing methods adopted include linear normalization (*Range*) and standard deviation normalization (*Scale*), as shown in **Eq. 4** and **Eq. 5** respectively. For ML models (such as KNN) that need to calculate the distance between samples, different orders of magnitude between variables will greatly affect the performance of the model. We retained the original input data in this paper (*Raw*), and conducted two normalization strategies

TABLE 1 | Definition of variables and descriptive statistics.

Category	Variable	Description	Mean	Median	Maximum	Minimum	*Std. Dec*	Unit
Explained variable	*MSW*	Total solid waste collected amount	8343.85	6125.25	42951.80	130.00	7767.28	10,000 tons
Explanatory variables	*InGDP*	Total Regional GDP.	20265.24	14580.35	107986.90	398.20	18414.77	100 million RMB
	InTSP	Value added by transportation, warehousing, and postal services	932.68	727.80	3658.00	20.60	746.80	100 million RMB
	InWAR	Wholesale and retail value added	1955.78	1250.85	11000.20	23.40	2097.82	100 million RMB
	InAAM	Value added by the accommodation and catering industry	379.55	284.75	1880.50	13.10	339.47	100 million RMB
	Ca	City area	6065.09	4625.75	23206.32	295.00	5135.54	Square kilometers
	Upd	Urban population density	2788.65	2584.46	5967.00	515.00	1193.25	people/ square km
	Nup	The number of urban populations.	601.03	493.97	3347.32	16.30	477.08	10,000 people
	Dip	Urban per capita disposable income	2393.92	2112.35	8226.00	64.89	1601.20	RMB
	Scg	Total retail sales of consumer goods	26277.49	25027.32	73848.51	9746.80	11190.64	100 million RMB

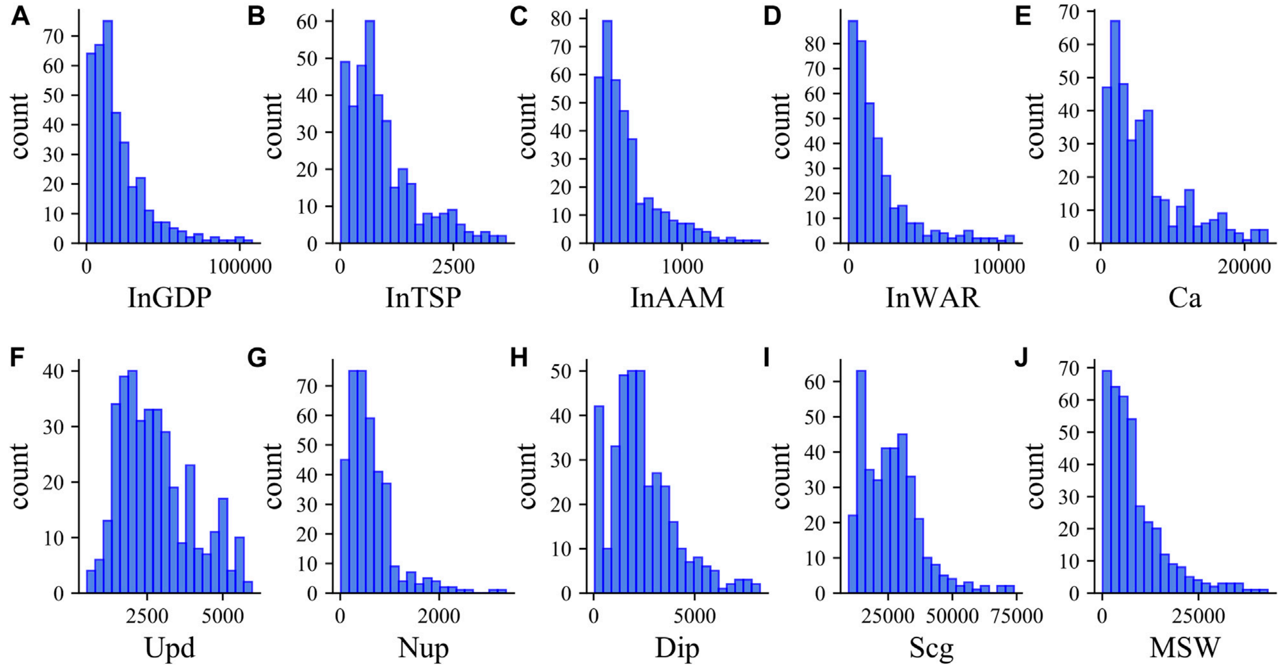


FIGURE 3 | Histogram plots for the different inputs and output variables used to train the ML methods. **(A)** is *InGDP*, **(B)** is *InTSP*, **(C)** is *InAAM*, **(D)** is *InWAR*, **(E)** is *Ca*, **(F)** is *Upd*, **(G)** is *Nup*, **(H)** is *Dip*, **(I)** is *Scg*, **(J)** is *MSW*.

of *Range* and *Scale* to reduce the influence of data's dimensions and skewness on the predictions. Thus, the results of the three preprocessing methods would be comparable.

$$x = \frac{x - x_{min}}{x_{max} - x_{min}}, \tag{4}$$

$$x = \frac{x - \bar{x}}{\sigma^2}, \tag{5}$$

where x_{min} represents the minimum value of variables while x_{max} represents the maximum value. $\bar{x}$ represents the numerical average value and σ^2 is the variance of each variable.

To minimize the deviation caused by sampling and prevent the model from over-fitting, we adopted the 10-folds cross validation method of resampling technique to create a random sample subset of input data as a training set. The remaining data was used as test set to obtain the generalization ability of the algorithms.

Metrics of the Model

To evaluate the performance of each machine learning algorithm, we use three metrics of the *MAE*, *RMSE* and the coefficient of determination (R^2) (Chai et al., 2021;

TABLE 2 | Hyper-parameters optimization.

Algorithm	Hyper-parameters	Other parameter settings
SVR	(C, γ, ε)	Kernel = Gaussian Kernel
KNN	K	Using Default Parameters
RF	$(Ntree, Nfeatures)$	Using Default Parameters
XGBoost	$Ntree$	Learning Rate = 0.05
DNN	$(Nh1, Nh2)$	Activation Function = Relu

Nguyen et al., 2021). These measurements are formulated as **Eqs 6–8**.

$$MAE = \frac{\sum_{i=1}^{n}\left|y_i - x_i\right|}{n}, \tag{6}$$

$$RMSE = \sqrt{\sum\nolimits_{i=1}^{n}\frac{\left(y_i - x_i\right)^2}{n}}, \tag{7}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - x_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{x}\right)^2}, \tag{8}$$

where n is the number of samples, x_i is the predicted response by the model, y_i is the actual value of the response, $\overline{x_i}$ is average estimated value.

Model Interpretation

Model interpretability is a major challenge to applications of ML methods, which has not been given enough attention in the field of ML and MSW forecasting research. To improve the interpretations of machine learning models, this paper employed the SHAP method that assigned each input variable a value reflecting its importance to predictor (Lundberg and Lee, 2017).

For socio-economic factor subset $S \subseteq F$ (where F stands for the set of all factors), two models are trained to extract the effect of factor i. The first model $f_{S\cup\{i\}}(x_{S\cup\{i\}})$ is trained with factor i while the other one $f_S(x_S)$ is trained without it, where $x_{S\cup\{i\}}$ and x_S are the values of input features/socio-economic factors. Then $f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)$ is computed for each possible subset $S \subseteq F\backslash\{i\}$. The Shapley value of a risk factor i is calculated using **Eq. 9**.

$$\phi_i = \sum_{S\subseteq F\backslash\{i\}}\frac{|S|!\,(|F| - |S| - 1)!}{|F|!}\left(f_{S\cup\{i\}}\left(x_{S\cup\{i\}}\right) - f_S\left(x_S\right)\right), \tag{9}$$

However, a major limitation of **Eq. 9** is that as the number of features/socio-economic factors increases, the computation cost will grow exponentially. To solve this problem (Lundberg et al., 2020), proposed a computation-tractable explanation method, i.e., TreeExplainer, for decision tree-based ML models such as RF. The TreeExplainer method marks it much more efficient to calculate a risk factor's SHAP value both locally and globally (Ayoub et al., 2021).

The SHAP combines optimal allocation with local explanations using the classic Shapley values. It would help users to trust the predictive models, not only what the prediction is but also why and how the prediction is made (Ayoub et al., 2021). Thus, the SHAP interaction values can be calculated as the difference between the Shapley values of factor i with and without factor j in **Eq. 10**:

TABLE 3 | Hyper-parameters search space of SVR.

Strategy	C	γ	ε
Raw	(1, 4000)	$(Scaled, Auto)$	(0, 5000)
Range	(0.01, 10)	$(Scaled, Auto)$	(0.0001, 0.001)
Scale	(0.01, 10)	$(Scaled, Auto)$	(0.0001, 0.001)

$$\phi_{i,j} = \sum_{S\subseteq F\backslash\{i,j\}}\frac{|S|!\,(|F| - |S| - 2)!}{|F|!}\left(f_{S\cup\{i,j\}}\left(x_{S\cup\{i,j\}}\right)\right.$$
$$\left. - f_{S\cup\{i\}}\left(x_{S\cup\{i\}}\right) - f_{S\cup\{j\}}\left(x_{S\cup\{j\}}\right) - f_S\left(x_S\right)\right). \tag{10}$$

For this superiority, we employ it to explain RF models which is based on decision trees. Therefore, compared with the existing methods (Nguyen et al., 2021), SHAP can reflect the influence of features in each sample, show the positive and negative effects of the influence, and thereby improve the explanatory of the model output.

RESULTS

Comparison of Model Results

The programming environment used in this study is Python (version 3.8.3) with additional support packages namely scikit-learn (version 0.24.1), Tensorflow (version 2.2.2) to calculate and run the ML algorithms.

Tuning

In this section, parameters of machine learning models are tuned, excluding multiple linear regression approach because it doesn't involve any hyper-parameters. Specific adjustment for parameters is shown in **Table 2**.

In the tuning process of SVR, we conduct the aforementioned three data preprocessing strategies (the Raw, Range, and Scale) respectively. As shown in **Table 3**, in the Raw strategy, that is to retain the original form of input data, the penalty parameter (C) varies from 1 to 4000, compared with that in the Range strategy of 0.01–10. The precision parameter (ε) is an interval between 0.0001 and 0.001 in the Range and Scale strategies, compared with that of an interval from 0 to 5000. The kernel width (γ) doesn't show any differences among the three strategies. The processing strategies of Range and Scale can effectively improve the normalization and scaling of the distributions of input variables.where $Scaled$ and $Auto$ in γ represent the results of **Eq. 11** and **Eq. 12** as the γ value of the SVR.

$$Scaled: \gamma = \frac{1}{N_S \times S^2}, \tag{11}$$

$$Auto: \gamma = \frac{1}{N_S}, \tag{12}$$

where N_S represents the number of sample features and S^2 represents sample variance. The optimization results are shown in **Figure 4**.

The hyper-parameters in other ML models are also tuned. For RF, the number of variables tried in each

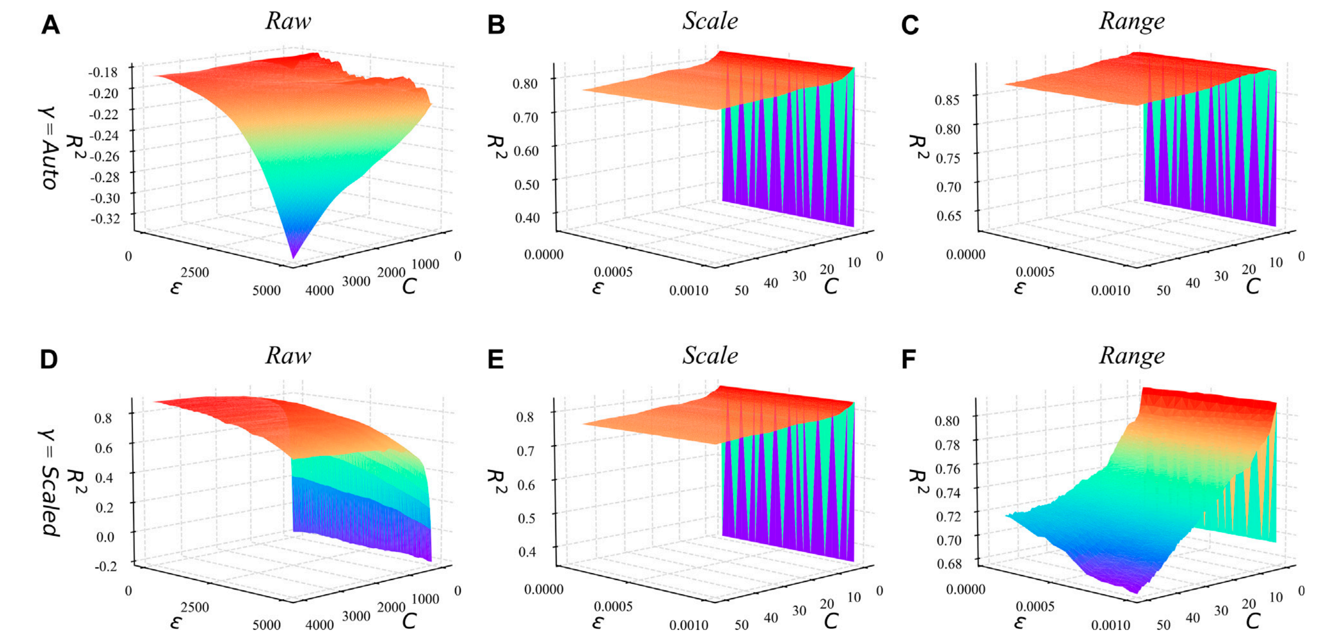


FIGURE 4 | Grid search results of SVR under different preprocess methods and different γ. **(A)** is *Raw* & γ = *Auto*, **(B)** is *Scale* & γ = *Auto*, **(C)** is *Range* & γ = *Auto*, **(E)** is *Raw* & γ = *Scaled*, **(F)** is *Scale* & γ = *Scaled*, **(G)** is *Range* & γ = *Scaled*.

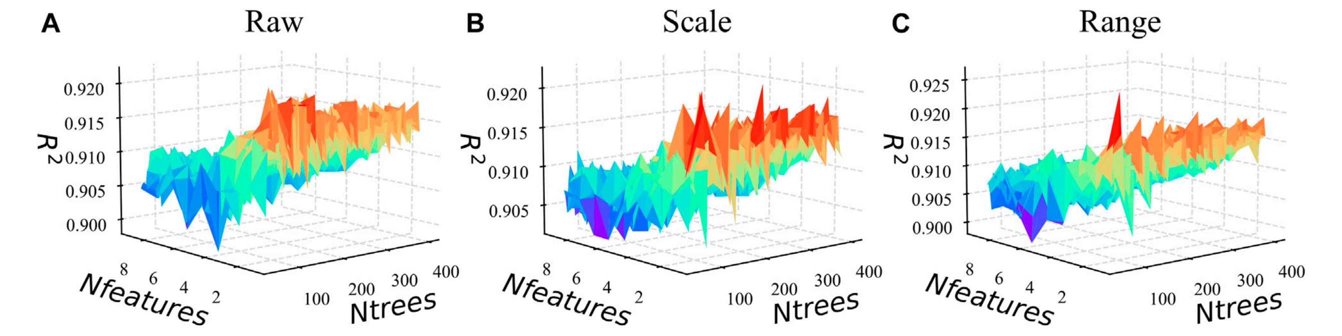


FIGURE 5 | Grid search results of RF under different preprocess methods. **(A)** is *Raw*, **(B)** is *Scale*, **(C)** is *Range*.

segmentation ($Nfeatures$) is set as positive integers between 1 and 9 in terms of nine input variables in this paper. The forest size ($Ntree$) is set as positive integers between (50,400). The optimization results of hyper-parameters are shown in **Figure 5**. In **Figures 4**, **5**, the redder the color is, the higher the R^2 of the parameter combination (therefore, the better the prediction), and vice versa. For KNN, the number of neighbors K is set as a positive integer between 1 and 29. For the XGBoost, the number of trees ($Ntree$) is set to 23 positive integers between 50 and 490. For DNN, the number of neurons in the first hidden layer ($Nh1$) is set as a positive integer increasing by 16 between (16,240), and the number of neurons in the second hidden layer ($Nh2$) is set as one half of the number of the first hidden layer.

Moreover, the Adma method is used as the optimization method, MAE is set as the loss function and the maximum number of epochs is set to 200. Meanwhile, to prevent over-fitting of the DNN, the EarlyStop mechanism is introduced, and the minimum learning rate is set as 0.003 and the tolerance is set as 20. The hyper-parameter selection results of KNN, XGBoost, and DNN are shown in **Figure 6**. The hyper-parameters adopted by each method are shown in **Table 4**.

Model Application and Generation Ability

Figure 7 presents the prediction performance of different ML models by using three preprocessing strategies. Several findings can conclude from the comparison among models. First, the prediction performance of MLR is the worst among all the methods because it doesn't involve hyper-parameter and responding adjustments. Second, the overall performances of SVR and KNN are similar, but the prediction ability of SVR is

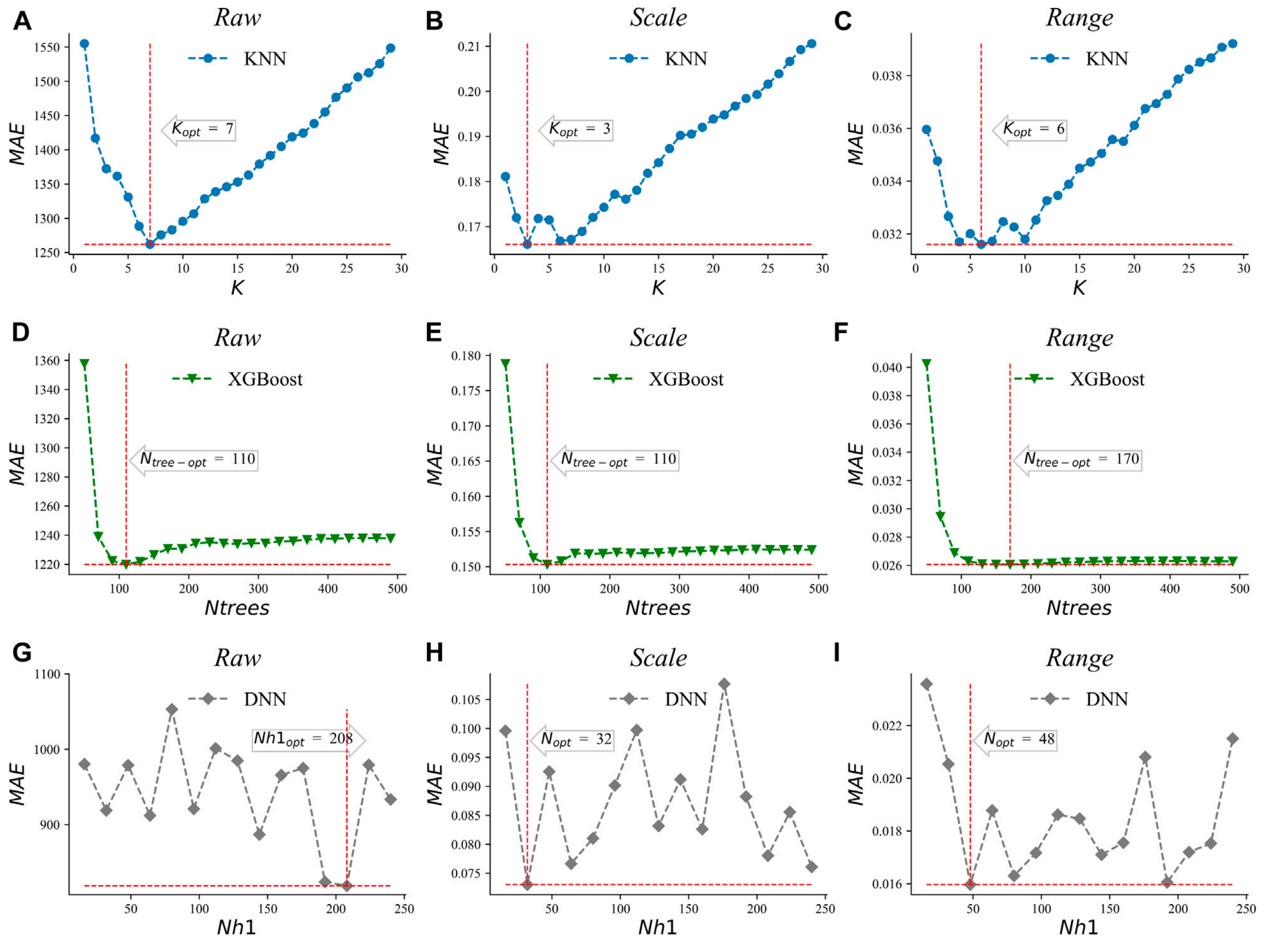


FIGURE 6 | Hyperparameter optimization results of different methods under different preprocess approaches. **(A)** is KNN and *Raw*, **(B)** is KNN and *Scale*, **(C)** is KNN and *Range*, **(D)** is XGBoost and *Raw*, **(E)** is XGBoost and *Scale*, **(F)** is XGBoost and *Range*, **(G)** is DNN and *Raw*, **(H)** is DNN and *Scale*, **(I)** is DNN and *Range*.

TABLE 4 | Hyper-parameter selection result for each algorithm.

Algorithm	Raw	Scale	Range
SVR	(4000,*Scaled*,202)	(1.019,*Auto*, 0.0001)	(4.049,*Auto*, 0.0001)
KNN	7	3	6
RF	(92,2)	(78,1)	(67,1)
XGBoost	110	110	170
DNN	(208,104)	(208,104)	(48,24)

slightly higher than that of KNN except for results in Scale processing. Normally, the conducting SVR model needs a more complex process than KNN. By inputting different forms of data, the KNN only needs to adjust one super parameter, which requires less work than SVR. Third, the RF and XGBoost models present significant and similar advantages in predicting MSW production compare with MLR, SVR, and KNN according to the performance measurement of R^2. Fourth, the DNN has the best predictive performance among all the algorithms.

In this study, the RF and DNN models showed high R^2 values (> 0.9) during all preprocessing methods. That means the developed ML models had a good power of explanation and were not over-fitted or over-trained. Compared with the ML method for MSW prediction developed in the earlier studies, our results were significantly better in prediction accuracy. For example (Niu et al., 2021), developed LSTM and ANN models for predicting MSW generation and during the testing phase, the R^2 value were 0.92 and 0.74, respectively (**Table 5**). In addition, (Nguyen et al., 2021), reported a DNN model with predictive performance (R^2) of 0.9 for MSW production projections in Vietnam. According to Kumar et al. (2018) and Kannangara et al. (2018) the ANN, SVM and other ML models for predicting MSW generation showed R^2 even lower than 0.8. Thus, the machine learning model developed in this paper promotes the effective prediction of MSW production.

SHAP Analysis

Overall Analysis

Figure 8 shows the SHAP summary plot that orders features based on their importance to predict MSW production.

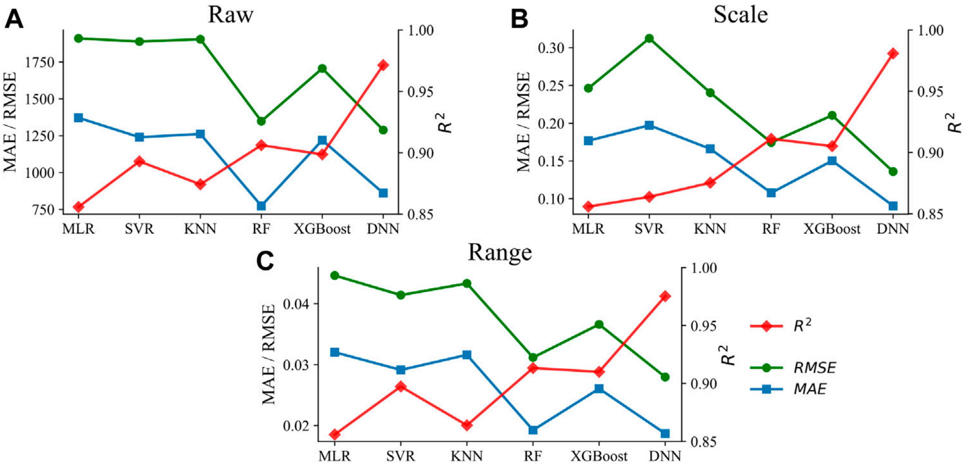


FIGURE 7 | Comparisons of algorithms predicts performance under different preprocess methods. **(A)** is *Raw*, **(B)** is *Scale*, **(C)** is *Range*.

TABLE 5 | Comparison of model performance for prediction of MSW generation.

Method	MAE	RMSE	R^2	References
DNN	861.03	1288.80	0.97	This study
RF	774.30	1348.63	0.91	
XGBoost	1219.91	1706.78	0.90	
LSTM	N/A	935.08	0.92	Niu et al. (2021)
ANN	N/A	547.14	0.74	
DNN	177.6	294.6	0.91	Nguyen et al. (2021)
ANN	N/A	9.53	0.75	Kumar et al. (2018)
SVM	N/A	9.88	0.74	
RF	N/A	9.88	0.66	
Decision Trees	N/A	23	0.54	Kannangara et al. (2018)
Neural Networks	N/A	16	0.72	

Specifically, a higher SHAP value of a feature indicates higher-ranked importance to the MSW production volume. For example, the difference in the region's GDP has the greatest impact on the model's prediction of MSW production. It is likely because waste production is highly related to the household wealth that directly influences one's daily consumption and potential production of MSW (Malinauskaite et al., 2017). Moreover, higher value of this feature result in higher SHAP values, which correspond to a higher output amount of MSW.

In addition, the industry structure presents a great influence on MSW production because of its indirect impacts on the citizens' consumption. For instance, a higher degree of the added value of wholesale and retail trade indicates higher

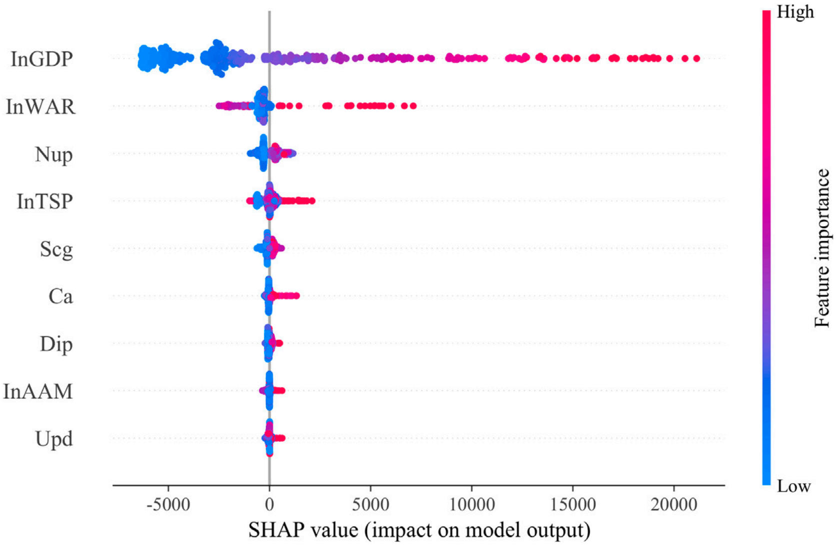


FIGURE 8 | SHAP summary plot.

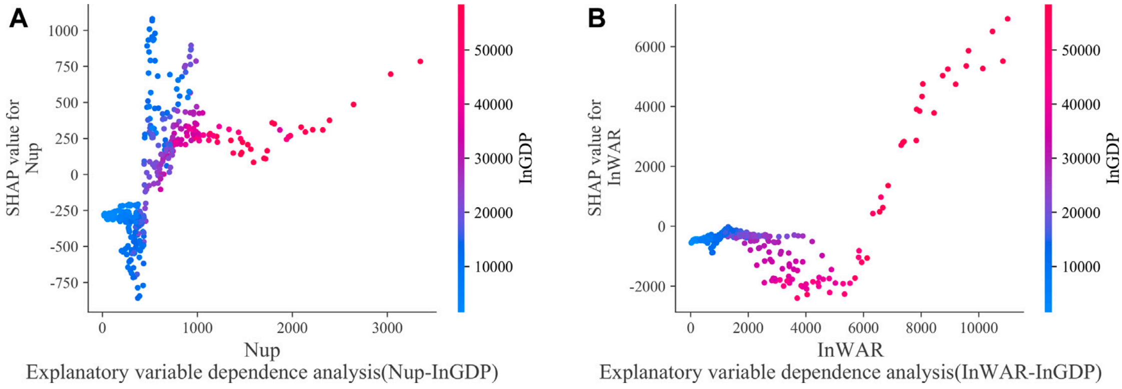


FIGURE 9 | Feature dependence analysis. **(A)** is *Nup* and *lnGDP*. **(B)** is *lnWAR* and *lnGDP*

production of MSW compared with other industries (e.g., transportation, warehousing, and postal services industries). Some studies have argued that consumption patterns and population increase are important factors that contribute to MSW production in developing countries (Liu et al., 2019; Nguyen et al., 2021). Besides, the urban population also shows a significant impact on MSW production, because of its functioning on the total amount of MSW production. In contrast, other socio-economic features have a relatively insignificant impact on MSW in China. In the following paper, we will continue to analyze the dependency among these three features to discover the generation mode of MSW in China.

Dependence Analysis

Figure 9 plots the relationship between a feature and its SHAP value dependent on another feature in the RF model. We select *Nup* and *lnWAR* as the features to discuss and identify their variation as changes of *lnGDP*. As shown in **Figures 9A,B**, the red points represent a higher value of *lnGDP*, and the blue points represent the lower one.

Figure 9A plots the moderating effects of GDP on the impacts of urban population on MSW production. It shows that under the condition of a low *Nup* and a low *lnGDP*, the SHAP value of *Nup* is below zero, which indicates that the impact of *Nup* would negatively impact the MSW production under these circumstances. In other words, the less developed region might undermine the impact of the urban population on MSW production, although the local urban population increases. In contrast, with the economic growth, the increase of the urban population will promote the production of MSW. It could be recognized by the red color of the SHAP value in this figure.

Figure 9B reflects the interaction between GDP and the added value of wholesale and retail industries on MSW production. For example, before *lnWAR* reached 600 billion, its SHAP value is always negative. However, if *lnWAR* exceeds 600 billion yuan as the increase of total GDP, the increase of the added value of wholesale and retail trade plays a positive role in promoting the production of MSW. It means that if the added value of the wholesale and retail industry remains at a low level (less than 6,000 billion yuan), these industries have little effect on MSW production. However, if the added value is more than the threshold of 6000 billion yuan, the regional GDP would promote the impact of the WAR industry added value. Correspondingly, the SHAP value of *lnWAR* indicates a significant promotion on MSW production.

CONCLUSION

To address the prediction in the production of municipal solid waste and support the WtE system design, we mainly constructed the MSW prediction method in China by using machine learning algorithms. In the comparisons of six ML models, we concentrated our attention on the predictive performances of each algorithm, particularly, by introducing three preprocessing strategies. As a result, SVR had the lowest hyperparameter consistency under different preprocessing strategies. Among the six ML methods established in this study, DNN has the best predictive ability, with an R-square of over 0.97 under all three data preprocessing strategies. The prediction performance of the machine learning methods developed in this paper is also significantly higher than the current standard (MLR) in China.

In addition, we find that the form of input hyper-parameter had a great influence on the models' performances. Specifically, the explanatory indicators of the regional GDP, urban population, the added values of wholesale and retail industries, are the most important variables that affect MSW production in different provinces of China. With the development of the urban economy, the urban population increase will promote the generation of municipal solid waste. Inversely, in less developed regions, the increase of the urban population will reduce the generation of MSW. Besides, the different stages of the development of the wholesale and retail industries also impact the production of MSW. It means that in the less developed regions, a less added value of the wholesale and retail industries indicates a weak impact on MSW production, and vice versa.

Our findings provide a reliable forecasting method for stakeholders. By increasing the prediction capability of MSW production, national and local policymakers could effectively conduct a series of governance policies to promote a friendly residential environment and urban sustainability. However, if given data from lower administrative, we can build even more powerful predictive models. Future studies can make effort on this to achieve more reliable and accurate results.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

JW and LY conceived, designed, and performed the experiments. YZ, XN, and ZS prepared, analyzed the data. QG contributes policy suggestions. LY and XN wrote the early version of the paper and all authors contributed discussion and revisions, all authors have read and approved the final manuscript.

FUNDING

This research is supported by Beijing Social Science Foundation (No. 17GLB014), National Key Research and Development Program of China (2018YFF0214804), BUCT Funds for First-Class Discipline Construction (XK1802-5), BUCT (G-JD202002).

REFERENCES

Abbasi, M., and El Hanandeh, A. (2016). Forecasting Municipal Solid Waste Generation Using Artificial Intelligence Modelling Approaches. *Waste Manag.* 56, 13–22. doi:10.1016/j.wasman.2016.05.018

Al-Dahidi, S., Ayadi, O., Adeeb, J., and Louzazni, M. (2019). Assessment of Artificial Neural Networks Learning Algorithms and Training Datasets for Solar Photovoltaic Power Production Prediction. *Front. Energ. Res.* 7, 130. doi:10.3389/fenrg.2019.00130

Ayoub, J., Yang, X. J., and Zhou, F. (2021). Combat COVID-19 Infodemic Using Explainable Natural Language Processing Models. *Inf. Process. Manag.* 58, 102569. doi:10.1016/j.ipm.2021.102569

Beigl, P., Lebersorger, S., and Salhofer, S. (2008). Modelling Municipal Solid Waste Generation: A Review. *Waste Manag.* 28, 200–214. doi:10.1016/j.wasman.2006.12.011

Birgen, C., Magnanelli, E., Carlsson, P., Skreiberg, Ø., Mosby, J., and Becidan, M. (2021). Machine Learning Based Modelling for Lower Heating Value Prediction of Municipal Solid Waste. *Fuel* 283, 118906. doi:10.1016/j.fuel.2020.118906

Breiman, L. (1996). Bagging Predictors. *Mach Learn.* 24, 123–140. doi:10.1007/BF00058655

Chai, J., Zhao, C., Hu, Y., and Zhang, Z. G. (2021). Structural Analysis and Forecast of Gold price Returns. *J. Manag. Sci. Eng.* 6, 135–145. doi:10.1016/j.jmse.2021.02.011

Chen, T., and Guestrin, C. (2016). Xgboost: A Scalable Tree Boosting System. in Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, August 13-17, 2016. San Francisco, CA, USA, 785–794.

Costa, A. B. R., Ferreira, P. C. G., Gaglianone, W. P., Guillén, O. T. C., Issler, J. V., and Lin, Y. (2021). Machine Learning and Oil price point and Density Forecasting. *Energ. Econ.* 102, 105494. doi:10.1016/j.eneco.2021.105494

Cover, T., and Hart, P. (1967). Nearest Neighbor Pattern Classification. *IEEE Trans. Inform. Theor.* 13, 21–27. doi:10.1109/TIT.1967.1053964

Demir, B., and Bruzzone, L. (2014). A Multiple Criteria Active Learning Method for Support Vector Regression. *Pattern Recognition* 47, 2558–2567. doi:10.1016/j.patcog.2014.02.001

Ding, Y., Zhao, J., Liu, J.-W., Zhou, J., Cheng, L., Zhao, J., et al. (2021). A Review of China's Municipal Solid Waste (MSW) and Comparison with International Regions: Management and Technologies in Treatment and Resource Utilization. *J. Clean. Prod.* 293, 126144. doi:10.1016/j.jclepro.2021.126144

Feng, Y., Duan, Q., Chen, X., Yakkali, S. S., and Wang, J. (2021). Space Cooling Energy Usage Prediction Based on Utility Data for Residential Buildings Using Machine Learning Methods. *Appl. Energ.* 291, 116814. doi:10.1016/j.apenergy.2021.116814

Guo, H.-n., Wu, S.-b., Tian, Y.-j., Zhang, J., and Liu, H.-t. (2021). Application of Machine Learning Methods for the Prediction of Organic Solid Waste Treatment and Recycling Processes: A Review. *Bioresour. Tech.* 319, 124114. doi:10.1016/j.biortech.2020.124114

Hariharan, R. (2021). Random forest Regression Analysis on Combined Role of Meteorological Indicators in Disease Dissemination in an Indian City: A Case Study of New Delhi. *Urban Clim.* 36, 100780. doi:10.1016/j.uclim.2021.100780

He, J., and Lin, B. (2019). Assessment of Waste Incineration Power with Considerations of Subsidies and Emissions in China. *Energy Policy* 126, 190–199. doi:10.1016/j.enpol.2018.11.025

Hoornweg, D., and Bhada-Tata, P. (2012). *What a Waste: A Global Review of Solid Waste Management*. Urban development series; knowledge papers no. 15. Washington, DC: World Bank.

Huang, B., Sun, Y., and Wang, S. (2021). A New Two-Stage Approach with Boosting and Model Averaging for Interval-Valued Crude Oil Prices Forecasting in Uncertainty Environments. *Front. Energ. Res.* 9, 707937. doi:10.3389/fenrg.2021.707937

Huang, Q., Yu, Y., Zhang, Y., Pang, B., Wang, Y., Chen, D., et al. (2021). Data-driven-based Forecasting of Two-phase Flow Parameters in Rectangular Channel. *Front. Energ. Res.* 9, 10. doi:10.3389/fenrg.2021.641661

Kannangara, M., Dua, R., Ahmadi, L., and Bensebaa, F. (2018). Modeling and Prediction of Regional Municipal Solid Waste Generation and Diversion in Canada Using Machine Learning Approaches. *Waste Manag.* 74, 3–15. doi:10.1016/j.wasman.2017.11.057

Kumar, A., Samadder, S. R., Kumar, N., and Singh, C. (2018). Estimation of the Generation Rate of Different Types of Plastic Wastes and Possible Revenue Recovery from Informal Recycling. *Waste Manag.* 79, 781–790. doi:10.1016/j.wasman.2018.08.045

Kuznetsova, E., Cardin, M.-A., Diao, M., and Zhang, S. (2019). Integrated Decision-Support Methodology for Combined Centralized-Decentralized Waste-To-Energy Management Systems Design. *Renew. Sust. Energ. Rev.* 103, 477–500. doi:10.1016/j.rser.2018.12.020

Li, R., Li, W., Zhang, H., Zhou, Y., and Tian, W. (2021). On-Line Estimation Method of Lithium-Ion Battery Health Status Based on PSO-SVM. *Front. Energ. Res.* 9, 693249401. doi:10.3389/fenrg.2021.693249

Liang, T., Chai, J., Zhang, Y.-J., and Zhang, Z. G. (2019). Refined Analysis and Prediction of Natural Gas Consumption in China. *J. Manag. Sci. Eng.* 4, 91–104. doi:10.1016/j.jmse.2019.07.001

Liu, J., Li, Q., Gu, W., and Wang, C. (2019). The Impact of Consumption Patterns on the Generation of Municipal Solid Waste in China: Evidences from Provincial Data. *Ijerph* 16, 1717. doi:10.3390/ijerph16101717

Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., and Alsaadi, F. E. (2017). A Survey of Deep Neural Network Architectures and Their Applications. *Neurocomputing* 234, 11–26. doi:10.1016/j.neucom.2016.12.038

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., et al. (2020). From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat. Mach Intell.* 2, 56–67. doi:10.1038/s42256-019-0138-9

Lundberg, S. M., and Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. In 31st conference on neural information processing systems, 4768–4777.

Malinauskaite, J., Jouhara, H., Czajczyńska, D., Stanchev, P., Katsou, E., Rostkowski, P., et al. (2017). Municipal Solid Waste Management and Waste-To-Energy in the Context of a Circular Economy and Energy Recycling in Europe. *Energy* 141, 2013–2044. doi:10.1016/j.energy.2017.11.128

Mehrdad, S. M., Abbasi, M., Yeganeh, B., and Kamalan, H. (2021). Prediction of Methane Emission from Landfills Using Machine Learning Models. *Environ. Prog. Sust. Energ.* 40, e13629. doi:10.1002/ep.13629

Mukherjee, C., Denney, J., Mbonimpa, E. G., Slagley, J., and Bhowmik, R. (2020). A Review on Municipal Solid Waste-To-Energy Trends in the USA. *Renew. Sust. Energ. Rev.* 119, 109512. doi:10.1016/j.rser.2019.109512

Namlis, K.-G., and Komilis, D. (2019). Influence of Four Socioeconomic Indices and the Impact of Economic Crisis on Solid Waste Generation in Europe. *Waste Manag.* 89, 190–200. doi:10.1016/j.wasman.2019.04.012

NBSC (2020). *China Statistical Yearbook 2020*. Beijing, China: Transport and Disposal of Consumption Wastes in Cities by Region. (in Chinese).

NBSC (2021). *Urban and Rural Population and Floating Population*. Beijing, China: Bulletin of the Seventh National Census. (No. 7) (in Chinese).

Nguyen, X. C., Nguyen, T. T. H., La, D. D., Kumar, G., Rene, E. R., Nguyen, D. D., et al. (2021). Development of Machine Learning - Based Models to Forecast Solid Waste Generation in Residential Areas: A Case Study from Vietnam. *Resour. Conservation Recycling* 167, 105381. doi:10.1016/j.resconrec.2020.105381

Niu, D., Wu, F., Dai, S., He, S., and Wu, B. (2021). Detection of Long-Term Effect in Forecasting Municipal Solid Waste Using a Long Short-Term Memory Neural Network. *J. Clean. Prod.* 290, 125187. doi:10.1016/j.jclepro.2020.125187

Ouda, O. K. M., Cekirge, H. M., and Raza, S. A. R. (2013). An Assessment of the Potential Contribution from Waste-To-Energy Facilities to Electricity Demand in Saudi Arabia. *Energ. Convers. Manag.* 75, 402–406. doi:10.1016/j.enconman.2013.06.056

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., et al. (2008). Top 10 Algorithms in Data Mining. *Knowl Inf. Syst.* 14, 1–37. doi:10.1007/s10115-007-0114-2

Yang, L., Nguyen, H., Bui, X.-N., Nguyen-Thoi, T., Zhou, J., and Huang, J. (2021). Prediction of Gas Yield Generated by Energy Recovery from Municipal Solid Waste Using Deep Neural Network and Moth-Flame Optimization Algorithm. *J. Clean. Prod.* 311, 127672. doi:10.1016/j.jclepro.2021.127672

Zhang, X., Yu, L., Wang, S., and Lai, K. K. (2009). Estimating the Impact of Extreme Events on Crude Oil price: An EMD-Based Event Analysis Method. *Energ. Econ.* 31, 768–778. doi:10.1016/j.eneco.2009.04.003

Zheng, J., Lai, C. S., Yuan, H., Dong, Z. Y., Meng, K., and Lai, L. L. (2020). Electricity Plan Recommender System with Electrical Instruction-Based Recovery. *Energy* 203, 117775. doi:10.1016/j.energy.2020.117775

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Energy Research

ORIGINAL RESEARCH
published: 26 November 2021
doi: 10.3389/fenrg.2021.720406

Forecasting Electricity Load With Hybrid Scalable Model Based on Stacked Non Linear Residual Approach

Ayush Sinha[1], Raghav Tayal[1], Aamod Vyas[2], Pankaj Pandey[3] and O. P. Vyas[1]*

[1]CPSEC Lab, Indian Institute of Information Technology Allahabad, Department of IT, Prayagraj, India, [2]Department of Business Informatics, University of Mannheim, Mannheim, Germany, [3]Norwegian University of Science and Technology (NTNU), Gjøvik, Norway

Power has totally different attributes than other material commodities as electrical energy stockpiling is a costly phenomenon. Since it should be generated when demanded, it is necessary to forecast its demand accurately and efficiently. As electrical load data is represented through time series pattern having linear and non-linear characteristics, it needs a model that may handle this behavior well in advance. This paper presents a scalable and hybrid approach for forecasting the power load based on Vector Auto Regression (VAR) and hybrid deep learning techniques like Long Short Term Memory (LSTM) and Convolutional Neural Network (CNN). CNN and LSTM models are well known for handling time series data. The VAR model separates the linear pattern in time series data, and CNN-LSTM is utilized to model non-linear patterns in data. CNN-LSTM works as CNN can extract complex features from electricity data, and LSTM can model temporal information in data. This approach can derive temporal and spatial features of electricity data. The experiment established that the proposed VAR-CNN-LSTM(VACL) hybrid approach forecasts better than more recent deep learning methods like Multilayer Perceptron (MLP), CNN, LSTM, MV-KWNN, MV-ANN, Hybrid CNN-LSTM and statistical techniques like VAR, and Auto Regressive Integrated Moving Average (ARIMAX). Performance metrics such as Mean Square Error, Root Mean Square Error, and Mean Absolute Error have been used to evaluate the performance of the discussed approaches. Finally, the efficacy of the proposed model is established through comparative studies with state-of-the-art models on Household Power Consumption Dataset (UCI machine learning repository) and Ontario Electricity Demand dataset (Canada).

Keywords: vector auto regression, convolutional neural network, long short term memory, electrical load forecasting, time series

OPEN ACCESS

Edited by:
Jian Chai,
Xidian University, China

Reviewed by:
Karthik Balasubramanian,
Keppel Offshore and Marine Limited,
Singapore
Hongxun Hui,
University of Macau, China

****Correspondence:***
Ayush Sinha
pro.ayush@iiita.ac.in

Specialty section:
This article was submitted to
Smart Grids,
a section of the journal
Frontiers in Energy Research

Received: *04 June 2021*
Accepted: *06 October 2021*
Published: *26 November 2021*

Citation:
Sinha A, Tayal R, Vyas A, Pandey P and Vyas OP (2021) Forecasting Electricity Load With Hybrid Scalable Model Based on Stacked Non Linear Residual Approach.
Front. Energy Res. 9:720406.
doi: 10.3389/fenrg.2021.720406

1 INTRODUCTION

As an option of petroleum products to create power, elective asset like sunlight based, wind and so on have become , quite possibly, the most encouraging sustainable power sources within the presence of greenhouse effect and polluted environment (Miller et al., 2009). The electric grid framework is complex since it should keep up the equilibrium among production, transmission and distribution of power. Taking into account the yield power from an alternate source is trademark in instability and discontinuity, presenting incredible difficulties to load dispatching, exact electrical load estimating assumes a significant part in soothing the pressing factor of managing top load and improving robustness limit with respect to electrical load demand. Electricity demand forecasting plays an important role as it enables the electric industry to make informed decisions in planning power system demand and supply. Moreover, accurate power demand forecasting is necessary as energy must be utilized as it is produced due to its physical characteristics (Ibrahim et al., 2008). Albeit ample studies have been dedicated to building powerful models to predict accurate electrical load (Du et al., 2019), the greater part of them are utilized for producing deterministic point prediction with single-variable yield each time. Generally applied point estimating models for electrical load can be partitioned into two classes: statistical models and machine learning models. Statistical models exploit as completely as conceivable the past records by giving attention to connections and patterns between the old and future exhibition of power load data dependent on the development of mathematical models (Ma et al., 2017). Nevertheless, statistical strategies can diminish the anticipating mistakes when the data features are under ordinary conditions, having high prerequisite for simple time series. Work like ARMA (Bikcora et al., 2018) and ARIMA (Wu et al., 2020) address traditional time series prediction strategies, however they ordinarily neglect to consider the impact of other covariate factors (Wu et al., 2020). Therefore, to counter the weaknesses of statistical models, machine learning models, known as artificial neural networks (ANN), are deployed for power load forecasting (Khwaja et al., 2020), (Wu et al., 2019) and (Xiao et al., 2016).

As a promising part of AI strategies, deep learning, mostly referring to multi-layer network having feature learning potential, has acquired a wide recognition for power load prediction due to three significant properties: solid generalization ability, large scale data processing and unsupervised way for the feature learning. From the work (Bedi and Toshniwal, 2019), it is widely perceived that deep learning models exhibit good performance in terms of precision, scalability and stability. Nonetheless, one of significant criticisms of picking up deep learning algorithms is, it lacks strong theoretical foundation and mathematical induction. This is additionally an effectively a disregarded issue in the viable use of electrical load prediction. To keep away from that issue, this paper presents a mathematical form of problem formulation followed by the proposed solution as VACL model which is a combination of statistical model VAR and Deep Learning methods CNN,LSTM. The present work is an extension of our previous work (Sinha et al., 2021).

Electricity demand forecasting can be of multiple types: short term (day), medium term (week to month) and long term (year). These forecasts are necessary for the proper operation of electric utilities. Precise power load forecasting can be helpful in financing planning to make a strategy of power supply, management of electricity, and market search (Stoll and Garver, 1989). It is a time series problem that is multivariate as electrical energy depends on many characteristics that use temporal data for the prediction. Temporal data depends on time and represented using time stamps. Prediction using classical load forecasting methods is challenging as power consumption can have a uniform seasonal pattern but an irregular trend component. To continue the discussions, the rest of the paper is organized as: **Section 2**,literature review of existing state-of-the-art models and issues relatingto them that will lead to the problem statement as presented in **Section 3**. To understand the basics about the multivariate time series analysis and deep learning forecasting strategies, **section 4** is presented. In continuation to existing approach, **Section 5** presents the detail about proposed methodology followed by **Section 6** which consists of experimental studies and discussion of application of proposed model on two large datasets. Finally,**Section 7** is about conclusion and states the future scope of the proposed method.

2 LITERATURE REVIEW

The new improvement of deep learning models, like Deep Neural Network (DNN), Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), has had an incredible impact in the fields of Natural Language Processing (NLP), computer vision, and recognition of speech. DNN can exhibit to model a function which is complex in nature and can efficiently mine important features of a dataset. Many researchers have explored these techniques for the multivariate time-series forecasting. Some of the recent advancement in this area is summarized as:

Authors in (Choi, 2018) discussed the ARIMA-LSTM hybrid model for time series forecasting. They used LSTM for temporal dependencies and their long-term predictive properties. To circumscribe linear properties, ARIMA is used, and for residuals that contain non-linear and temporal properties, LSTM is used. This hybrid model is compared with other methods, and it gave better results for evaluation metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). In (Kim and Cho, 2019), authors proposed a hybrid CNN-LSTM model that is evaluated on power consumption data. It is proposed that CNN can extract temporal and spatial features between several variables of data. In contrast, LSTM takes data returned by CNN as input and models temporal data and irregular trends. The proposed model is compared with other models like GRU, Bi-LSTM, etc., and it performed better on evaluation metrics such as MSE, RMSE, MAE, MAPE, etc. While Mahalakshmi et al. surveyed various methods for forecasting time series data and also discussed various types of time-series data that are being forecasted (Mahalakshmi et al., 2016), research has been done on various types of data such as electricity data, stock market data, etc. The performance evaluation parameter such as MAE, MSE proves that the hybrid forecasting model yields good results compared to other models. To investigate the forecasting outcome for non-linear data, Gasperin et al. discussed the problem of accurately predicting power load forecast owing to its non-linear nature (Gasparin et al., 2019). The authors worked on two power load forecast datasets and applied state-of-the-art deep learning techniques to short-term prediction data. Most relevant deep learning models applied to the short-term load forecasting problem are surveyed and experimentally evaluated. The focus has been given to these three main models: Sequence to Sequence Architectures, Recurrent Neural Networks, and recently developed Temporal Convolutional Neural Networks. LSTM performed better as compared to other traditional models. In continuation to use the deep learning models for forecasting no-linear data, authors in (Erica, 2021) propose a novel short-term load forecasting approach with Deep Neural Network architecture,CNN components to learn complex feature representation from historical load series,then the LSTM based Recurrent Neural component models the variability and dynamics in historical loading.

Siami-Namini et al. in their proposed work (Siami-Namini et al., 2018) compared deep learning methods such as LSTM with the traditional statistical methods like ARIMA for financial time series dataset. According to them, a forecasting algorithm based on LSTM improves the prediction by reducing the error rate by 85% when compared to ARIMA. On the similar lines, Wang et al. (Wang et al., 2016) worked on CNN-LSTM consisting of two parts: regional CNN and for predicting the VA rating method used is LSTM. According to their evaluation, regional CNN-LSTM outperformed regression and traditional Neural Network-based methods. Authors in (Sherstinsky, 2018) explained the essential fundamentals of RNN and CNN. They also discussed "Vanilla LSTM" and discussed the problems faced when training the standard RNN and solved that by RNN to "Vanilla LSTM" transformation through a series of logical arguments. The work done in (Hartmann et al., 2017) adopted the Cross-Sectional Forecasting approach on the AutoRegression model. It consumes available data from multiple same domain time series in a single model, covering a wide domain of data that also compensates missing values and quickly calculates accurate forecast results. This model can only deal with linear data but with multiple time series simultaneously while in (Choi and Lee, 2018), authors presented a novel LSTM ensemble forecasting algorithm that can combine many forecast results from a set of individual LSTM networks. The novel method can capture non-linear statistical properties and is easy to implement and is computationally efficient. In another domain with similar characteristics, Chniti et al. (Chniti et al., 2017) presented robust forecasting methods for phone price prediction using Support Vector Regression (SVR) and LSTM. Models have been compared for both univariate and multivariate data. In the multivariate model, LSTM performed better as compared to others. Another work like (Yan et al., 2018) attempted short-term load forecasting (STLF) for the electric power consumption dataset. Due to the varying nature of data for electricity, traditional algorithms performed poorly as compared to LSTM. To increase further accuracy, the authors discussed a hybrid approach consisting of CNN on top of LSTM and experimented on five different datasets. It performed fairly better than ARIMA, SVR, and LSTM alone. As a more advanced hybrid model, authors in (Babu and Reddy, 2014) proposed a linear and non-linear models combination that is a combination of ARIMA and ANN models where ARIMA is used for linear component and ANN for a non-linear component. For further improvement, the authors proposed that the nature of time series should be taken into account so volatile nature is taken into account by moving average filter, and then hybrid model applied; the proposed hybrid model is compared with these individual models and some other models, and it performed fairly well as compared to other models. While the work in (Shirzadi et al., 2021) showed that by utilizing deep learning, the model could foresee the load request more precisely than SVM and Random Forest (RF). However, it does not validate the result on more than one dataset. In (Bendaoud and Farah, 2020) another type of CNNN for one-day ahead load estimate utilizing a two-dimensional information layer (remembering the past states' utilizations for one layer and climatic and relevant contributions to another layer). They applied their model to a contextual analysis in Algeria and announced MAPE and RMSE of 3.16 and 270.60 (MW), individually. An approach based on clustering techniques, authors in (Talavera-Llames et al., 2019) introduced a clustering technique dependent on kNN to predict power price utilizing a multivariate dataset. The proposed model was applied on a power dataset in Spain (OMIE-Dataset, 2020) and the authors juxtaposed the outcome with existing state of art methods like MV-ANN (Hippert et al., 2001), MV-RF and traditional multivariate Box-Jenkins (Lütkepohl, 2013) model like ARIMAX (Box et al., 2011), autoregressive-moving-average (ARMAX) and autoregressive (ARX).

Coming to a more popular model, authors have proposed Elmann Recurrent Neural Networks (ERNN) in Elman (1990) to sum up feedforward neural network to better take care of ordered sequential data like time-series. Notwithstanding of the model simplicity, Elmann RNNs are difficult to prepare because of less efficiency of gradient (back) propagation. While forecasting the time series with Multi-Step Prediction method, authors in Sorjamaa and Lendasse (2006) proposed a DirRec strategy based on the combination of Recursive and Direct strategy. In this approach, a model is trained in a single mode to predict one next step of the time series data and combine it with a multiple model predictor with the same input. Authors in Bontempi (2008) presented a model as MIMO strategy where a single model is evolved to predict complete output sequence in a single effort. However the more advanced popular model known as DIRMO model Taieb et al. (2009) was proposed which is like a tradeoff with the MIMO and Direct approach. This model was proved to be more advanced in terms of multistep forecasting and computational time.

In a nut shell, the above literature survey generally centers around DNN, RNN and CNN models and shows that deep learning strategies can convey much better load forecasting precision than those accomplished by traditional models. Other deep learning models have not been investigated much for load forecastings, for example, attention model (Bourdeau et al., 2019), ConvLSTM and BiLSTM. Notwithstanding the works referred to, a different researchers have also centered around load anticipating at the structure scale, utilizing AI and deep learning strategies (Rashid et al., 2009), (Shi et al., 2017) and (Rahman et al., 2018). In any case, fewer investigations have analyzed the capacity of information digging methods for large-scale data and established their model's efficacy on multiple datasets with different characteristics.

3 LOAD FORECASTING INTRICACIES

Stemming out the research gap from the literature survey from **Section 2**, the present work aims at building a model that can accurately forecast power load data. The mathematical formulation and objectives of the problem is as follows:

1) Given fully observed time series data Y = {y_1, y_2,., y_T} where y_t belongs to R^n and n is the variable dimension, aim is to predict a series of future time series data
2) That is, assuming {y_1, y_2,., y_T} is available, then predicting y_{T+h} where h is the desirable time horizon ahead of the current timestamp (Chatfield, 1996).
3) The following constraints need to be satisfied by the model:
   a) Model should be able to handle numerous series data
   b) Model should be able to handle incomplete data
   c) Model should be able to handle noisy data

4 MULTIVARITE TIME SERIES ANALYSIS WITH DEEP LEARNING

4.1 Time Series

It is a series of discrete data points which are taken at fixed intervals of time (Wikipedia, 2021). An explicit order dependence is added between observations by time series via time dimension. Order of observations in time series gives a source of extra information which can be used in forecasting. There may be one or more variables in the time series. A time series that is having one variable changing over time is univariate time series. If greater than one variable varying with time, then that time series is multivariate.

It can have applications in many domains such as weather forecasting, power load forecasting, stock market prediction, signal processing, econometrics, etc.

4.2 Time Series Analysis

It constitutes methods for analyzing and drawing out meaningful information and patterns from data which can help in deciding the methods and getting better forecasting results (Cohen, 2021). It helps to apprehend the nature of the series that is needed to be predicted.

4.3 Time Series Forecasting

Time series forecasting involves creating a model and fitting it on a training set (historical data) and then using that model to make future predictions. In classical statistical handling, taking forecasts in the future is called extrapolation. A time series model can be evaluated by forecasting the future term and analyzing the performance by specific evaluation metrics like MSE, MAE, and RMSE.

4.4 Time Series Types

Time series forecasting techniques are inspired by various research on machine learning and have been changed from regression models to neural network-related models. There are multiple types of time series, of which two types are most common.

- STATIONARY: If statistical properties like mean, variance, autocorrelation, etc., of time series do not change with time, then that time series is stationary. As we know, stationary processes are easy to predict; we simply need to find out their statistical properties, which will remain the same over a while.
- NON-STATIONARY: In a non-stationary time series, data points have statistical properties like mean, variance, covariance, etc. and vary with time. There may be non-stationary behavior like trends, seasonality, and cycles that exists in the series data. Some of the most common patterns observed in non-stationary time series are(Erica, 2021):
- TREND: If there is a long duration increment or decrement in data, then trend exists. It need not be linear.
- SEASONALITY: When seasonal factors such as month of year, day of month etc. impact time series, then seasonal patterns are said to exist in time series with firm and known frequency.
- CYCLIC: When data exhibit rise and fall patterns without fixed period, then cyclic patterns occur.

4.5 Time Series Evaluation Metrics

The most commonly used error metrics for forecasting are:

- MEAN SQUARED ERROR: It is the average cumulative sum of the square of all prediction errors. It is formulated as:

$$MSE = \sum_{i=1}^{n} (y_i - \hat{y}i)^2/n \tag{1}$$

- MEAN ABSOLUTE ERROR: It is the average cumulative sum of the absolute value of all prediction errors. It is formulated as:

$$MAE = \sum_{i=1}^{n} \|y_i - \hat{y}i\|/n \tag{2}$$

- ROOT MEAN SQUARED ERROR: It is a square root of the mean of the cumulative sum of the square of all prediction error. It is formulated as:

$$RMSE = \sqrt{\sum_{i=1}^{n} (y_i - \hat{y}i)^2/n} \tag{3}$$

4.6 Terminology in Time Series Forecasting

- DIFFERENCING: It is a technique to transform non-stationary time series into a stationary one. In differencing, we take the difference of each value in the time series from its next value and continue until the new series become stationary.
- AIC: It refers to Akaike's Information Criterion. For models such as VAR, it provides information about how well a model can be fitted on the data by considering the terms count in the model.

- NOISE: The randomness in data series is frequently known as noise.
- TIME SERIES MODEL: It is a derived function that considers past observations of time series along with some parameters to predict the future.
- WEIGHT: Weights stipulate the importance given to individual parameters in forecasting, respectively. In order words, it decides the impact of each item on forecasting.
- DECOMPOSITION: It refers to splitting a time series into seasonal, trend, and cyclic components.

4.7 Artificial Neural Network

Artificial Neural Network (ANN) (Yao, 1993) consists of nodes that are interconnected, simulating neurons in the biological neural system. It can be utilized for various tasks such as regression, forecasting, and pattern recognition in circumstances of complex features such as seasonality and trends observed, handling linear and non-linear data, etc. ANN model that is being used is Multilayer Perceptron, as earlier ANNs consists of only a single layer with no hidden layers, which resulted in some limitations:

- Single neurons cannot solve complex tasks.
- The model cannot learn difficulty in learning non-linear features.

MLP is a feed-forward neural network that is comprised of inputs, many hidden layers, and an output layer (Shiblee et al., 2009). In MLP, every layer is connected fully to the next layer such that neurons between contiguous layers are fully connected while neurons between the same layers have no connection. Input is fed into the input layer, and output is extracted from the output layer. The number of the hidden layers can be increased to learn more complex features according to the task.

Input represents the data that is needed to be fed in the model. Data and weights are fed to next layer. Suppose X$(x_1, x_2, \ldots, x_n)$ be the input vector and w$(w_1, w_2, \ldots, w_n)$ are weights associated for a neuron,then input to neuron of hidden layer is Input:

$$f(X) = \sum_{i=1}^{n} (x_i . w_i) \tag{4}$$

Primary learning of the model takes place at the hidden layer (also known as the processing unit). Using the activation function, it remodels the value received from the input layer. Activation function is non-linear function applied on hidden layer input that enables the model to describe erratic relations. Sigmoid, ReLU, and tanh are the most widely used activation functions. Activation Functions mostly used are as (Yao, 1993):

- SIGMOID: It is formulated as:

$$\sigma = 1/(1 + e^{-x}) \tag{5}$$

- Rectified Linear Unit (ReLU): It is most extensively used activation function having a minimum 0 threshold and formulated as:

$$f(x) = max(0, x) \tag{6}$$

- tanh(x): Non-linear activation function with values lying between 0 and 1. It is formulated as:

$$tanh(x) = 2/(1 + e^{-2x}) - 1 \tag{7}$$

The main issue with MLP is the adjustment of its weights in the hidden layer, which is necessary to get better results as output, is dependent on these weights to minimize the error. Back propagation is used for the adjustment of weight parameters in the hidden layer. After loss calculation in the forward pass, the loss is backpropagated, and the model weights are updated via gradient descent. Backpropagation rule is given mathematically as:

$$\delta w = w - w_{\text{prev}} = -\eta * \frac{\delta E}{\delta w} \tag{8}$$

Where weights are represented by w, E(w) represents cost function, representing how far the predicted output is, from actual output, and η represents the learning rate.

4.8 Long Short Term Memory

RNN (Jordan, 1990), (Elman, 1990), (Chen and Soo, 1996) are types of neural networks where the goal is to predict the sequence's next step given previous steps in the sequence. In RNN, the basic idea is to learn information about the earlier state of sequence to predict the later ones. In RNN, hidden layers store the information captured about previous states of data. The same tasks (same weights and biases) are performed on every element of sequential data to capture information for the sequence to forecast future unseen data. The main challenge for RNN is the problem of Vanishing Gradients. To overcome the problem of Vanishing Gradients, a particular type of RNN is used, which is LSTM (Hochreiter and Schmidhuber, 1997), which is specifically designed to handle long-term dependency issues. The way LSTM achieves that, is by the use of a memory line. Remembering early data trend is made possible in LSTM via some gates which can control information flow through the memory line, LSTM consists of cells that capture and store the data streams. Adding some gates in each cell of LSTM enables us to filter, add or dispose of the data. It enables us to store the limited required data while forgetting the remainder. There are three types of gates that are used in LSTM. Gates are based on the sigmoid layer enabling LSTM cells to pass data or disposing of it optimally (Olah, 2013).

There are three types of gates mainly (Hochreiter and Schmidhuber, 1997):

- Forget Gate: This gate filters out the information cell state should discard. It considers previous hidden state (h_{t-1}) and input (x^t) and returns a vector consisting of values between zero and one for each number respectively in cell state C_{t-1} determining what to keep or discard. It is formulated as:

$$f_t = \sigma(W_f . [h_{t-1}, x_t] + b_f \tag{9}$$

- Input Gate: It decides new information that we need to put in a cell. It consists of a sigmoid-based layer that decides what values need to be updated. Moreover, it contains a tanh layer that creates a new candidate values vector, $\tilde{C}$ that is needed to be added to the state. We need to combine these two to define the update:

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i) \tag{10}$$

$$\widetilde{C}_t = tanh(W_c.[h_{t-1}, x_t] + b_c) \tag{11}$$

Now, the cell state will be updated by first forgetting the things from the previous state that was decided to be forgotten earlier and then adding $i_t * \widetilde{C}_t$. It is formulated as:

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t \tag{12}$$

- Output Gate: This gate decides the output out of each cell. To get output, we run a sigmoid layer on input data and a hidden layer that decides what will be output. Then cell state (C_t) is passed through the tanh layer and multiplied by the output gate such that we get the values that are decided as output:

$$o_t = \sigma(W_o.[h_{t-1}, x_t] + b_o) \tag{13}$$

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

4.9 CNN-Long Short Term Memory Neural Network

This model extracts temporal and spatial features for effectively forecasting time series data. It consists of a Convolutional Layer with a max-pooling layer on top of LSTM. CNN (Fukushima, 1980), (Rawat and Wang, 2017) consists of an input layer that accepts various correlated variables as input and an output layer that will send devised features to LSTM and other hidden layers. The convolution layer, ReLU layer, activation function, and pooling layer are types of hidden layers. The convolutional layer reads the multivariate input time series data, applies the convolution operation with filters, and sends results to the next layer, reducing the number of parameters and making the network deeper. If $x_i^0 = \{x_1, x2, \ldots, x_n\}$ is input vector, y_{ij}^1 output from first convolutional layer is (Fukushima, 1980), (Rawat and Wang, 2017):

$$y_{ij}^l = \sigma\left(b_j^1 + \sum_{m=1}^{M} w_{m,j}^1 x_{i+m-1,j}^0\right) \tag{15}$$

y_{ij}^1 is calculated by input x_{ij}^0 from previous layer and bias b_j^i represents bias for *jth* feature map, weights of kernel is represented as w and σ denotes the ReLU (Nair and Hinton, 2010) like activation function. Similarly resultant vector from *kth* convolutional layer is formulated as:

$$y_{ij}^l = \sigma\left(b_j^l + \sum_{m=1}^{M} w_{m,j}^1 x_{i+m-1,j}^0\right) \tag{16}$$

The convolution pooling layer is followed by a pooling layer that reduces the space size of the devised results from the convolutional layer, thereby reducing the number of parameters and computing costs. The most commonly used pooling approach is Max Pooling (Albawi et al., 2017) which uses the maximum value from previous neuron clusters. Suppose k is the stride and Z is the pooling cluster size. Max pooling operation is formulated as:

$$P_{ij}^l = \max_{z \in Z} y_{ixk+z,j}^{l-1} \tag{17}$$

After convolution operation, LSTM is used, which is the lower layer in CNN-LSTM neural network, which stores temporal information from features extracted from the convolution layer. It is well suited for forecasting as it reduces vanishing and exploding gradient, which is generally faced by Recurrent Neural Networks. Remembering early data trend is made possible in LSTM by gates which control the flow of information down the memory line.

LSTM consists of cells that capture and store the data streams. Adding some gates in each cell of LSTM enables us to filter, add or dispose of the data. Gates are based on the sigmoid layer, enabling LSTM cells to pass data or disposing it optimally.

Last unit of CNN-LSTM consists of dense layer (also known as fully connected layer) which can be used to generate the final output result. Here as we are forecasting for 1 h so no of the neuron units in dense layer is 1.

5 PROPOSED HYBRID MODEL FOR LOAD FORECASTING

The model which is best suited depends on historical data analysis and relationships between data to be forecasted. Neural networks can extract complex patterns from data thus are better suited as compared to statistical models. Among neural networks, RNNs are better suited for time series forecasting tasks. RNNs can remember the past inputs, thus improving the performance of sequential data, while neural network models like Multilayer Perceptron will treat the data like numerous inputs without considering the significance of time.

5.1 VAR-CNN-Long Short Term Memory Hybrid (VACL)

This model combines the ability of the statistical model to learn with combination with deep learning models. Time series data is known to be made of linear and non-linear segments which can be expressed as:
$d_t = N_t + L_t + \epsilon$

L_t is a linear component at time t, N_t is a component that is non-linear at time t and ϵ is the error component. VARector is a traditional statistical model for time series forecasting, which performs well on linear problems. On the other hand, neural network models like CNN-LSTM seem to work well on problems that have non-linearity in data. So, a combination of both models can identify both linear and non-linear patterns in data.

In this model, VAR can identify linear interdependence in data and residuals left from VAR used by CNN-LSTM to capture non-linear patterns in data. Now we will discuss each of these sectors used in the algorithm.

5.1.1 Vector Auto Regression Sector

When two or more time-series influence each other, then vector auto-regression can be used. This model is autoregressive, and in this model, each variable is formulated as a function of past values of variables (Prabhakaran, 2020). Compared to other models like ARIMA, the variable output is built as a linear combination of its past values and values of other variables in this model. In contrast, ARIMA output depends on the value of those particular variables on which we want to make predictions. A typical Auto Regression with order "p" can be formulated as:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} + \varepsilon \tag{18}$$

where α is a constant denoting the intercept, $\beta_1, \beta_2, \ldots, \beta_p$ are lag coefficients. To understand the equation for VAR (Biller and Nelson, 2003)let us assume there are two time-series Y_1 and Y_2 and have to be forecast at time t. We know that to calculate predicted values, VAR needs to consider past data of all related variables. So, equations of the value predicted at time t and order p become:

$$Y_{1,t} = \alpha_1 + \beta_{11,1}Y_{1,t-1} + \beta_{12,1}Y_{2,t-1} + \ldots + \beta_{11,p}Y_{1,t-p} + \beta_{12,p}Y_{2,t-p} \tag{19}$$

$$Y_{2,t} = \alpha_2 + \beta_{21,1}Y_{1,t-1} + \beta_{22,1}Y_{2,t-1} + \ldots + \beta_{21,p}Y_{1,t-p} + \beta_{22,p}Y_{2,t-p} \tag{20}$$

As a prerequisite, time series needs to be stationary to apply the VAR model. If it is stationary, we can directly predict using the VAR model; or else we need to make data differences to make it stationary. For checking the time-series stationarity, the Augmented Dickey-Fuller Test (ADF Test) can be used. It is a unit root stationarity test. The property of time series that makes it non-stationary is a unit root. The number of unit roots determines how many differencing operations are needed to make the series stationery. Consider the following equation (Biller and Nelson, 2003):

$$Y_t = \alpha + \beta t + \gamma Y_{t-1} + \delta_1 \Delta Y_{t-1} + \delta_2 \Delta Y_{t-2} + \ldots + \delta_p \Delta Y_{t-p} + \varepsilon \tag{21}$$

For the ADF Test, if the null hypothesis $\delta = 1$ in the model equation proves to be true, then the series is non-stationary; or else the series is stationary. Since the null hypothesis assumes the presence of unit root ($\delta = 1$), the value of p should be less than the significant level of 0.05 for rejecting the null hypothesis, hence proving that series is stationary.

After the series becomes stationary by differencing the series and verifying using ADF Test, we need to find the right order for VAR. For that purpose, we will iterate over different order values and fit the model. Then find out the order which gives us the least AIC.

AIC stands for Akaike Information Criterion, which is a method for selecting a model based on score. Suppose m be the no of parameters estimated for the model and L be the maximum likelihood. Then AIC value is the following:

$$AIC = 2 * m - 2ln(L) \tag{22}$$

We will select that model which has the least value of AIC. Though AIC rewards the goodness of fit, but the penalty function is implemented as increasing with an increase in several estimated parameters. After testing and getting all requisite parameters, forecasting can be performed on the data. The residual received after subtracting forecasted data from original test data is used as input to CNN, and that data contains non-linear patterns. It is formulated as:

$$d_t - L_t = N_t + \epsilon \tag{23}$$

5.1.2 CNN-Long Short Term Memory Sector

As we know, neural networks have a good performance on non-linear data primarily due to many versatile parameters. Moreover, due to non-linear activation functions in layers, they can quickly adapt to non-linear trends. They can model residuals received from VAR very effectively.

This model extracts temporal and spatial features for effectively forecast time series data. It consists of a convolutional layer with a max-pooling layer on top of LSTM. CNN (Fukushima, 1980), (Rawat and Wang, 2017) consists of an input layer that accepts various correlated variables as input and an output layer that will send devised features to LSTM. The convolution layer, ReLU layer, activation function, and pooling layer are types of hidden layers. The convolutional layer reads the multivariate input time-series data, applies the convolution operation with filters, and sends results to the next layer to reduce the number of parameters and make the network deeper. If $x_i^0 = \{x_1, x2, \ldots, x_n\}$ is input vector, y_{ij}^1 output from first convolutional layer is as from (Fukushima, 1980), (Rawat and Wang, 2017):

$$y_{ij}^1 = \sigma\left(b_j^1 + \sum_{m=1}^{M} w_{m,j}^1 x_{i+m-1,j}^0\right) \tag{24}$$

y_{ij}^1 is calculated by input x_{ij}^0 from previous layer and bias b_j^i represents bias for jth feature map, weights of kernel is represented as w and σ denotes the Rectified Linear Unit (ReLU) (Nair and Hinton, 2010) like activation function. Similarly resultant vector from kth convolutional layer is formulated as:

$$y_{ij}^l = \sigma\left(b_j^l + \sum_{m=1}^{M} w_{m,j}^1 x_{i+m-1,j}^0\right) \tag{25}$$

The convolution pooling layer is followed by a pooling layer that reduces the space size of the devised results from the convolutional layer, thereby reducing the number of parameters and computing costs. Max pooling (Albawi et al., 2017) operation is formulated as:

$$P_{ij}^l = \underset{z \in Z}{max} y_{ixk+z,j}^{l-1} \tag{26}$$

After convolution operation, LSTM is used, which is the lower layer in CNN-LSTM neural network, which stores temporal information from features extracted from the convolution layer. It is well suited for forecasting as it reduces the problem of vanishing and exploding gradient, which RNNgenerally face. Remembering early data trends is made possible in LSTM using some gates that control the flow of information through the memory line. LSTM consists of cells that capture and store the data streams. Adding some gates in each cell of LSTM enables us to filter, add or dispose of the data. Gates are based on a sigmoid layer that enables LSTM cells to pass data or dispose of it optimally. There are three types of gates mainly (Hochreiter and Schmidhuber, 1997):

- Forget Gate: This gate filters out the information that the cell state should discard. It is formulated as:

$$f_t = \sigma(W_f.[h_{t-1}, x_t] + b_f) \tag{27}$$

- Input Gate: It decides what new information should bein a cell. It consists of a sigmoid-based layer that decides which values need to be updated. Moreover, it contains a tanh layer that creates a new candidate values vector, $\tilde{C}$ that needs to be added to the state. We need to combine these two to define the update:

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i) \tag{28}$$

$$\tilde{C}_t = tanh(W_c.[h_{t-1}, x_t] + b_c) \tag{29}$$

Cell state is updated by disregarding the things from the previous state that was decided to be disregardedearlier and then adding $i_t * \tilde{C}_t$. It is formulated as:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{30}$$

- Output Gate: This gate decides the output out of each cell. To get output, we run a sigmoid layer on input data and a hidden layer for deciding what we are going to output. Then cell state (C_t) is passed through tanh layer and gets multiplied by the output gate such that we get the parameters to output:

$$o_t = \sigma(W_o.[h_{t-1}, x_t] + b_o) \tag{31}$$

$$h_t = o_t * tanh(C_t) \tag{32}$$

The last unit of CNN-LSTM consists of a dense layer (also known as a fully connected layer) which can be used to generate the final output result. As we are forecasting for 1 h, the number of neuron units in a dense layer is 1.

6 EXPERIMENTATION AND RESULT DISCUSSION

The experimentation has been done on two publicly available datasets:Household Electricity Consumption Dataset (Hebrail and Berard, 2012) and Ontario Electricity Demand Dataset (ontario Energy Price-Dataset, 2020) and (official website of the Government of Canada, 2020). The detail description of both the datasets and outcome of the proposed model using that dataset is presented in next two sections 6.1 and 6.2.

6.1 Discussion on Household Power Consumption Dataset

It is a multivariate time series dataset consisting of household energy consumption in a span of 4 years (2006–2010) at per minute sampling provided by UCI machine learning repository (Hebrail and Berard, 2012). It consists of seven time series namely:

1) **global active power**: total active power consumption by household (measured in kilowatt);
2) **global reactive power**: total reactive power consumption by household (in kilowatt);
3) **voltage**: average voltage of household (in Volts);
4) **global intensity**: average intensity of current (measured in amperes);
5) **sub metering 1**: active energy utilized for kitchen (watt-hours);
6) **sub metering 2**: active energy utilized for laundry (watt-hours);
7) **sub metering 3**: active energy utilized for climate control systems (watt-hours).

6.1.1 Preliminary Analysis

Preliminary analysis of data is being done, and patterns are evaluated, enabling us to make correct predictions. It is observed that given time series follow the seasonal pattern but with irregular trend components. We also performed correlation analysis and see there is a positive correlation between the two variables. Global Intensity has a significant impact on forecasting GAP value, and global active power and voltage do not have a strong correlation.

6.1.2 Performance Comparison of Models

The best-fitted model to be used depends on historical data availability and the relationship between variables to be forecast. Experiments are conducted for other neural network models consisting of MLP, LSTM, CNN-LSTM, etc., to establish the effectiveness of the proposed models, and results are evaluated with MSE and RMSE. Next, we will go through the architecture of each of these models and compare the results:

6.1.3 Multilayer Perceptron Model

Multilayer perceptron architecture is dependent on parameter adjustment and the number of hidden layers in the network. Multilayer perceptron consists of the input layer consisting of input neurons, hidden layers, and output layer. Hidden layers consist of dense layers. Parameters such as number of neurons in hidden layers, learning algorithm, and loss function can be optimized based on input data. Here input data is resampled to convert it into hour-based sampling. Input data consist of a sliding window of 24 data points for which we will predict the next hour of the result. Input is basically 24 × 7 size data where 24 is the number of time steps, and the number of variables is seven in each step. Adopted architecture has two hidden layers, each with 100 neurons used to extract patterns from the data. Model is trained with up to 50 epochs, and early stopping is used on data with a patience value of eight, which ensures if there is similar validation loss in each of eight consecutive epochs, then the model will stop running, and most optimal weights will be stored as output. ReLU activation is being used in the hidden layers, and for optimizing the weights **adam** optimizer is used. The result of

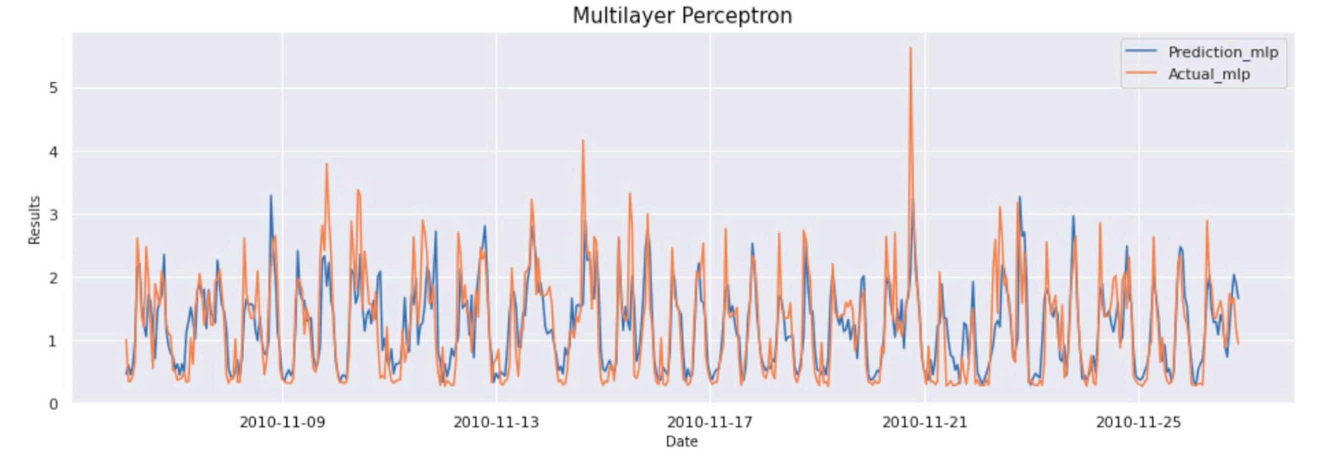


FIGURE 1 | Prediction vs Actual result of Multilayer Perceptron (Household Electricity Data).

this model is as MAE:0,395, MSE:0,303, and RMSE:551. The graph of actual vs predicted is as **Figure 1**:

6.1.4 Long Short Term Memory

The architecture of LSTM is dependent on the types of layers and parameters adjustment of layers in the network. It consists of the LSTM layer, Dropout Layer (to prevent overfitting), and Dense layer to predict the output. After preliminary analysis of data parameters such as number of layers, neurons in each layer, loss functions, and optimization, algorithms are adjusted to give the best possible outcome.

Input data consists of a sliding window consisting of 24 data points (resampled to an hour). So, the input to the LSTM is 24 × 7 size data. There are a total of seven variables used to make the prediction. The proposed architecture for the LSTM layer with 100 neurons has been used for extracting patterns from the data. Model is trained with up to 100 epochs, and early stopping is used on data with a patience value of eight. It ensures that if there is a similar validation loss in each of eight consecutive epochs, then the model will stop running, and the most optimal weights will be stored as output. For optimizing the weights **adam** optimizer with a learning rate 0.0001 is used with a batch size of 256.

The result of this model is as MAE:0,382, MSE:262, and RMSE: 512. The graph of actual vs predicted is as **Figure 2**.

6.1.5 CNN-Long Short Term Memory

The architecture of CNN-LSTM varies according to the number of layers, type of layers, and parameter adjustment in each layer. It consists of convolution layers, pooling layers, flatten layer, LSTM layers, and dense layer to predict the corresponding output. For convolution, the number of filters, size of the filter, and strides need to be adjusted. By adjustment of these parameters to an optimal level, accuracy can be significantly improved. To properly adjust the parameters of the model, data should be analyzed appropriately. As we already know that in CNN-LSTM, CNN layers use multiple variables and extract features between them hence improving time series forecasting significantly.

The correlation matrix shows a high correlation between different time-series variables with the variable we want to predict, i.e., Global Active Power (GAP). Input data consists of a sliding window consisting of 24 data points (resampled to an hour). So, the input to the CNN-LSTM is 24 × 7 size data. There are a total of seven variables used to make the prediction. The result of this model is as MAE:0,320, MSE: 221, and RMSE:470. The graph of actual vs predicted is as **Figure 3**.

6.1.6 VAR-CNN-Long Short Term Memory(VACL)

In this model architecture, first, we estimate VAR correctly on training data, and then we extract what VAR has learned and use it to refine the training of the CNN-LSTM process, giving better results. Firstly, to properly create a VAR model, data should be stationary. As already discussed, using ADFTest, it can be verified whether a time series is stationary or not. We applied the ADF Test on variables like global active power, global reactive power, voltage, global intensity, sub-metering 1, sub-metering 2, and sub-metering 3 with the null hypothesis that data has a unit root and is non-stationary. The ADF Test shows that all-time series are stationary, so differentiation is not needed for the series.

After doing this preliminary check, we need to find out the lag order, which can be calculated using AIC. All we need to do is to iterate through lag orders and find out the lag order with a minimum AIC score compared to its predecessors. In this case, 31 comes out to be the best lag order, as evident in **Table 1**. After getting the best order for VAR, we fit the VAR model on differentiated data. VAR can learn linear interdependencies in time series. This information is subtracted from raw data and gets the residuals that contain non-linear data.

The architecture of the above model varies according to the number of layers, type of layers, and parameter adjustment in each layer. It consists of convolution layers, pooling layers, flatten layer, LSTM layers, and dense layer to predict the corresponding output. For convolution operation, the number of filters, filter

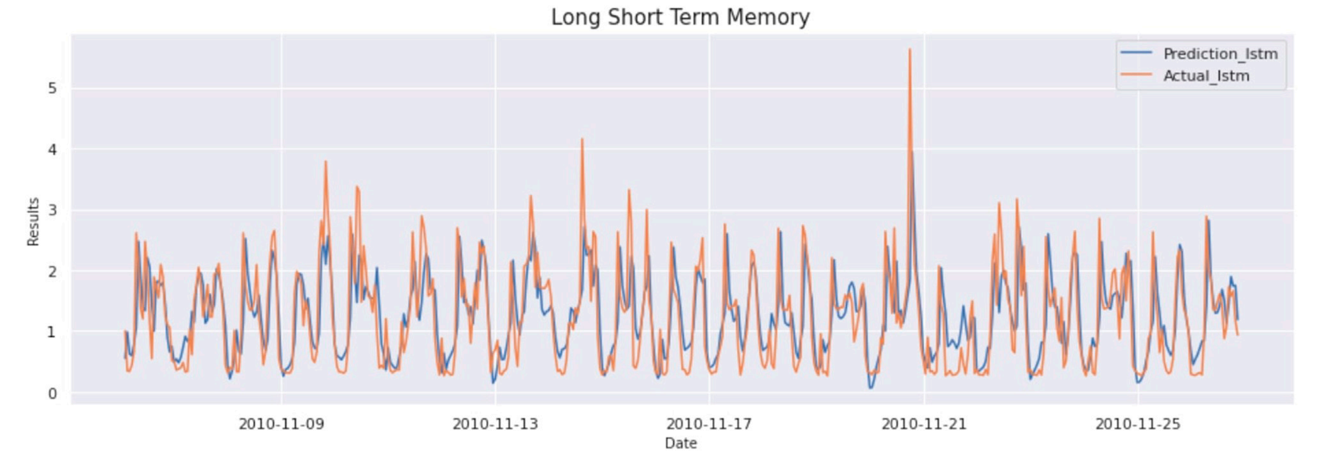


FIGURE 2 | Prediction vs Actual result of LSTM model (Household Electricity Data).

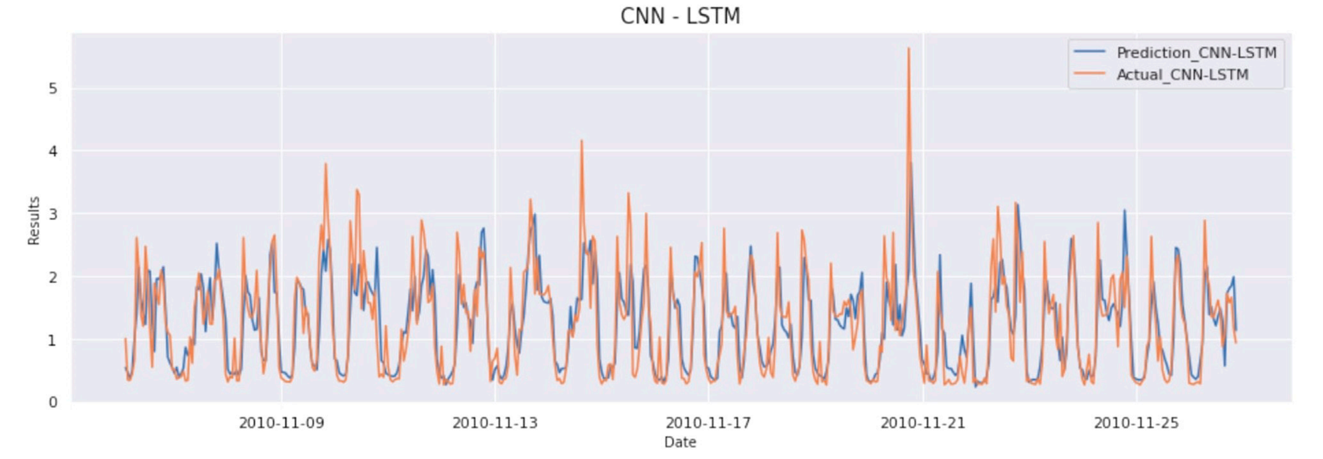


FIGURE 3 | Prediction vs Actual result of CNN-LSTM model (Household Electricity Data).

TABLE 1 | Akaike information criterion on HouseHold data (Hebrail and Berard, 2012).

Lag order	AIC	BIC
29	−5.3868	**−5.0376**
30	−5.3882	−5.0271
31	**−5.3893**	−5.0161
32	−5.3892	−5.0040

AIC, Akaike Information Criterion; BIC, Bayesian Information Criterion.

size, and strides need to be adjusted. By adjustment of these parameters to an optimal level, accuracy can be significantly improved. To properly adjust the parameters of the model, data should be analyzed appropriately. Input provided to the model consists of a sliding window of 24 data points (resampled to an hour). So, the input to CNN-LSTM is 24 × 7 size data. The result of this model is as MAE:0,317, MSE:210, and RMSE:458. The graph of actual vs predicted is as **Figure 4**.

The combined results of all the algorithms are displayed in **Table 2**. We can observe that from the above table that both CNN-LSTM and the proposed approach perform well for given data, but the proposed model performed slightly better in terms of error metrics.

6.2 Discussion on Ontario Power Demand Dataset

A multivariate time-series dataset consists of characteristics about Ontario Electricity Demand and corresponding Ontario Price and various other variables affecting these per 5-min sampling. It consists of ten time-series namely:

Ontario Price, Ontario Demand, Northwest, Northwest Temp, Northwest Dew Point Temp, Northwest Rel Hum, Northeast, Northeast Temp, Northeast Dew Point Temp, Northeast Rel Hum. The target is to forecast Ontario Price into the future by taking these variables. For the problem of price forecasting, two

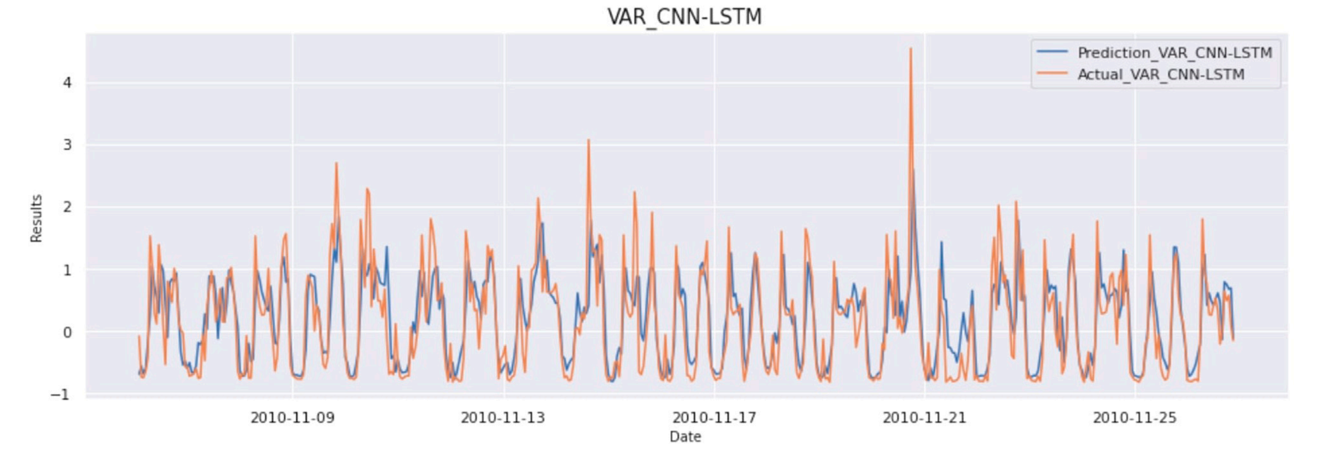


FIGURE 4 | Prediction vs Actual result of VAR CNN-LSTM model (Household Electricity Data).

TABLE 2 | Combined results of all the algorithms on Household Data (Hebrail and Berard, 2012).

	Mean absolute error	Mean squared error	Root mean squared error
VAR	0.698	0.654	0.865
MLP	0.395	0.303	0.551
ERNN-MIMO Bontempi (2008)	0.56	0.201	0.79
Seq2Seq Sutskever et al. (2014)	0.56	0.201	0.78
LSTM Hochreiter and Schmidhuber (1997)	0.57	0.221	0.512
Hybrid CNN-LSTM Alhussein et al. (2020)	0.310	0.220	0.462
CNN-LSTM	0.320	0.221	0.470
VAR-CNN-LSTM(VACL)	0.317	0.210	0.458

datasets from the Ontario region (Canada) are collected and combined from the following data sources:

1) ieso Power Data Directory. (ontario Energy Price-Dataset, 2020);
2) climate and weather data, Canada. (official website of the Government of Canada, 2020).

6.2.1 Preliminary Analysis

Preliminary analysis of data is being done, and patterns are evaluated, enabling us to make correct predictions. It is observed that time series follow seasonal patterns but there are irregular trend components. **Figure 5** depicts that only the Ontario Demand time series should be considered for forecasting the Ontario Price time series. The reason is the approximately minimum coefficient value of correlation should be 0.3 for having a constructive relationship between each of these time series.

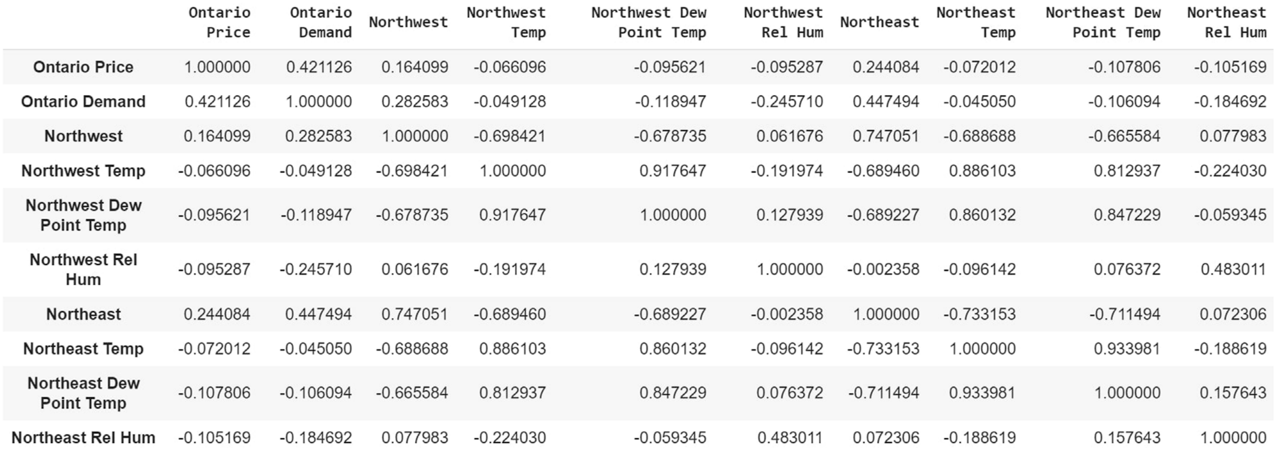
6.2.2 Performance Comparison of Models

The best-fitted model to be used depends on available historical data, and the relationship between variables to be forecast. Experiments have been conducted for other neural network models consisting of MLP, LSTM, CNN-LSTM, etc., to establish the effectiveness of the proposed models, and results are evaluated with MSE and RMSE. Next, we will go through the architecture of each of these models and compare the results:

6.2.3 Multilayer Perceptron Model

Multilayer perceptron architecture depends on parameter adjustment and the number of hidden layers in the network. Multilayer perceptron consists of input layer consisting of input neurons, hidden layers, and output layer. The hidden layers consist of dense layers. We can optimize the number of neurons in hidden layers, learning algorithm, and loss functions based on input data. Here input data is resampled to convert it into hour-based sampling. Input data consists of a sliding window of 24 data points for which we will predict the next hour of the result. Input is 24 × 2 size data where 24 is the number of time steps, and 2 is the number of variables in each step.

The used architecture has one hidden layer with 100 neurons used to extract patterns from the data. Model is trained with up to 80 epochs, and early stopping is used on data with a patience value of eight. It ensures, if there is a similar validation loss in each of eight consecutive epochs, that the model will stop running, and the most optimal weights will be stored as output. ReLU activation is being used in the hidden layers, and for optimizing the weights,



	Ontario Price	Ontario Demand	Northwest	Northwest Temp	Northwest Dew Point Temp	Northwest Rel Hum	Northeast	Northeast Temp	Northeast Dew Point Temp	Northeast Rel Hum
Ontario Price	1.000000	0.421126	0.164099	-0.066096	-0.095621	-0.095287	0.244084	-0.072012	-0.107806	-0.105169
Ontario Demand	0.421126	1.000000	0.282583	-0.049128	-0.118947	-0.245710	0.447494	-0.045050	-0.106094	-0.184692
Northwest	0.164099	0.282583	1.000000	-0.698421	-0.678735	0.061676	0.747051	-0.688688	-0.665584	0.077983
Northwest Temp	-0.066096	-0.049128	-0.698421	1.000000	0.917647	-0.191974	-0.689460	0.886103	0.812937	-0.224030
Northwest Dew Point Temp	-0.095621	-0.118947	-0.678735	0.917647	1.000000	0.127939	-0.689227	0.860132	0.847229	-0.059345
Northwest Rel Hum	-0.095287	-0.245710	0.061676	-0.191974	0.127939	1.000000	-0.002358	-0.096142	0.076372	0.483011
Northeast	0.244084	0.447494	0.747051	-0.689460	-0.689227	-0.002358	1.000000	-0.733153	-0.711494	0.072306
Northeast Temp	-0.072012	-0.045050	-0.688688	0.886103	0.860132	-0.096142	-0.733153	1.000000	0.933981	-0.188619
Northeast Dew Point Temp	-0.107806	-0.106094	-0.665584	0.812937	0.847229	0.076372	-0.711494	0.933981	1.000000	0.157643
Northeast Rel Hum	-0.105169	-0.184692	0.077983	-0.224030	-0.059345	0.483011	0.072306	-0.188619	0.157643	1.000000

FIGURE 5 | Correlation matrix (Ontario Demand Data).

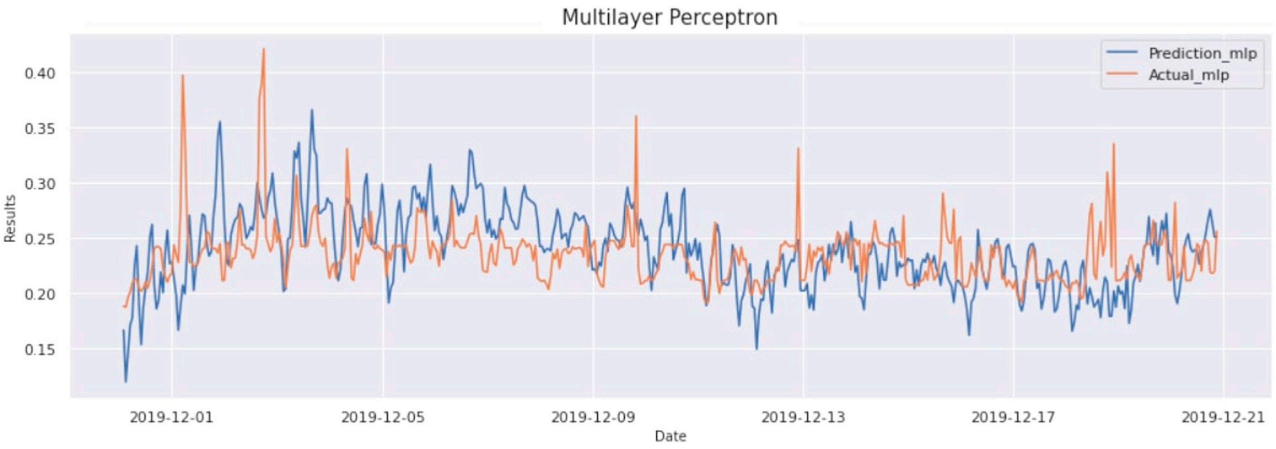


FIGURE 6 | Prediction vs Actual result of Multilayer Perceptron (Ontario Demand Data).

adam optimizer with learning rate, 0.000001 is used. The result of this model is as MAE:0,309, MSE:00204, and RMSE: 0452. The graph of actual vs predicted is as **Figure 6**.

6.2.4 Long Short Term Memory

The architecture of LSTM depends on the types of layers and parameters adjustment of layers in the network. It consists of the LSTM layer, Dropout Layer (to prevent overfitting), and Dense layer to predict the output. After preliminary analysis of data parameters such as the number of layers, neurons in each layer, loss functions, and optimization algorithms are adjusted to give the best possible outcome.

Input data consists of a sliding window consisting of 24 data points (resampled to an hour). So, the input to the LSTM is 24 × 2 size data. There are a total of two variables used to make the prediction. The proposed LSTM layer, each with 64 neurons, has been used for extracting patterns from the data. Model is trained with up to 100 epochs, and early stopping is used on data with a patience value of eight, which ensures if there is similar validation loss in each of eight consecutive epochs, then the model will stop running, and most optimal weights will be stored as output. The result of this model is as MAE:0,265, MSE:0015, and RMSE:0389. The graph of actual vs predicted is as **Figure 7**.

6.2.5 CNN-Long Short Term Memory

The architecture of CNN-LSTM varies according to the number of layers, layer type, and parameter adjustment in each layer. It consists of the convolution layers, pooling layers, flatten layer, LSTM layers, and dense layer to predict the corresponding output. For convolution operation, the number of filters, size of the filter, and strides need to be adjusted. By adjustment of these parameters to an optimal level, accuracy can be significantly improved. To properly adjust the parameters of the model, data should be analyzed appropriately.

It is known that in CNN-LSTM, CNN layers use multiple variables and extract features between them, improving time series forecasting significantly. As from the correlation matrix

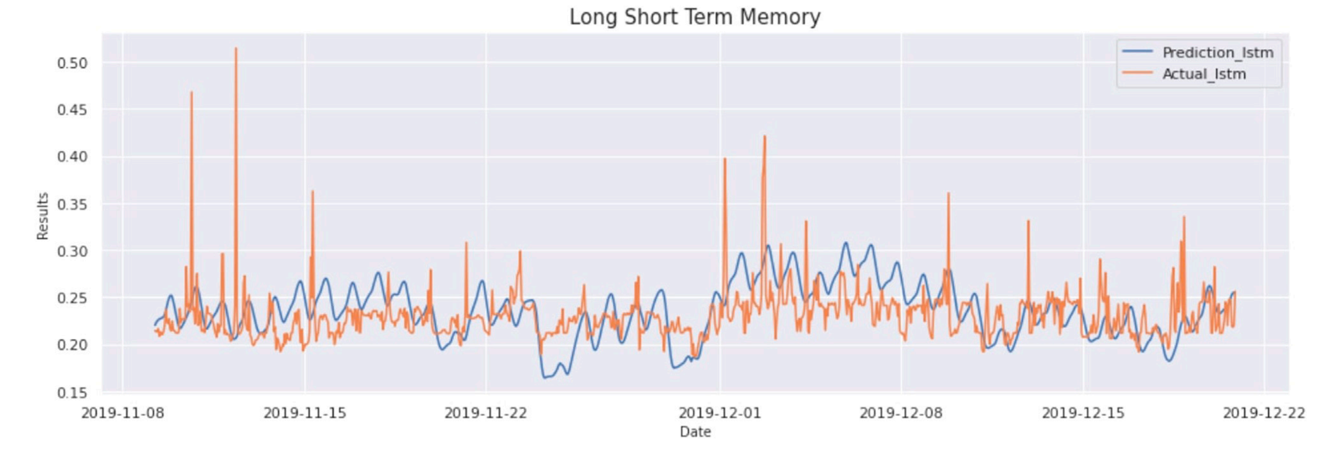


FIGURE 7 | Prediction vs Actual result of LSTM model (Ontario Demand Data).

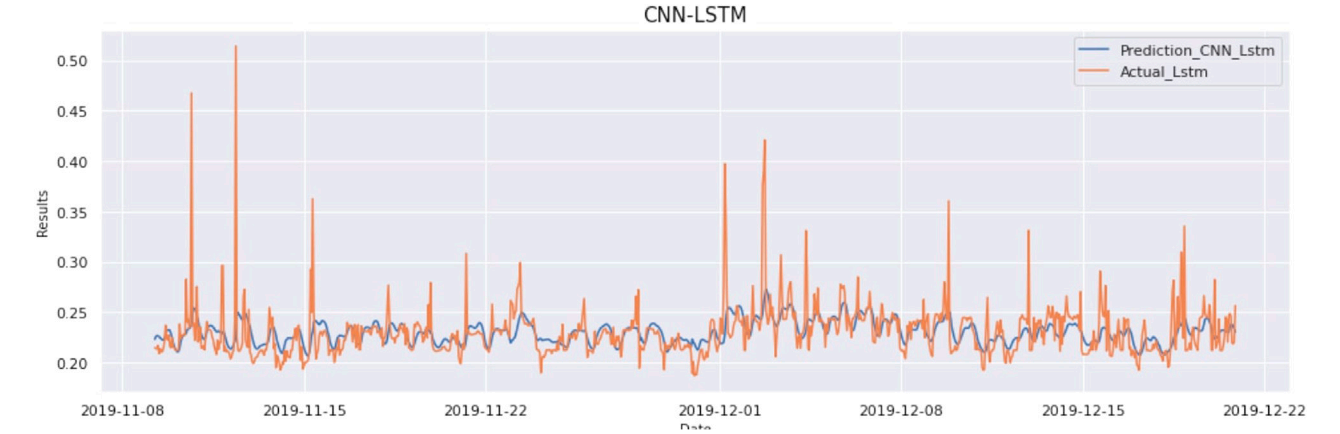


FIGURE 8 | Prediction vs Actual result of CNNLSTM model (Ontario Demand Data).

in **Figure 5**, it is observed that there is a high correlation between Ontario Price and Ontario Demand. Input data consists of a sliding window consisting of 24 data points (resampled to an hour). So, the input to the CNN-LSTM is 24 × 2 size data. There are a total of 2 variables used to make the prediction. The result of this model is as MAE:0,119, MSE: 00068, and RMSE:02616. The graph of actual vs predicted is as **Figure 8**.

6.2.6 VAR-CNN-Long Short Term Memory(VACL)

In this model architecture, we first estimate VAR correctly on training data, and then we extract what VAR has learned and use it to refine the training of the CNN-LSTM process. Firstly, to properly create a VAR model, we need to make data stationery, if not in the requisite format. As already discussed, using the ADFTest, we can check whether a time series is stationary or not. Results from the ADF Test show that every time-series are stationary, we do not need to differentiate the series.

TABLE 3 | Akaike information criterion on ontario demand data (ontario Energy Price-Dataset, 2020) (official website of the Government of Canada, 2020).

Lag order	AIC	BIC
28	17.5311	17.5587
29	17.5303	17.5589
30	17.5304	17.5598
31	17.5266	17.5571

After doing these preliminary checks, we need to find out the lag order, which can be calculated using AIC. All we need to do is to iterate through lag orders and find out the lag order with a minimum AIC score compared to its predecessors. In this case, 29 comes out to be the best lag order, as evident in this **Table 3**. After getting the best order for VAR, we fit the VAR model on differentiated data. VAR can learn linear interdependencies in time series. This information is subtracted from raw data to get the residuals which contain non-linear data as evident from **Figure 9**.

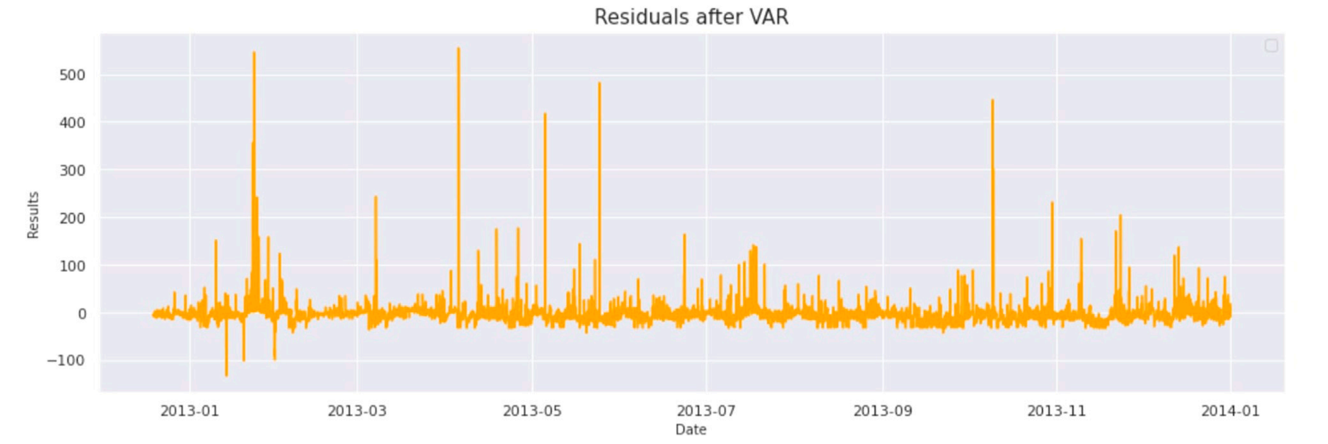
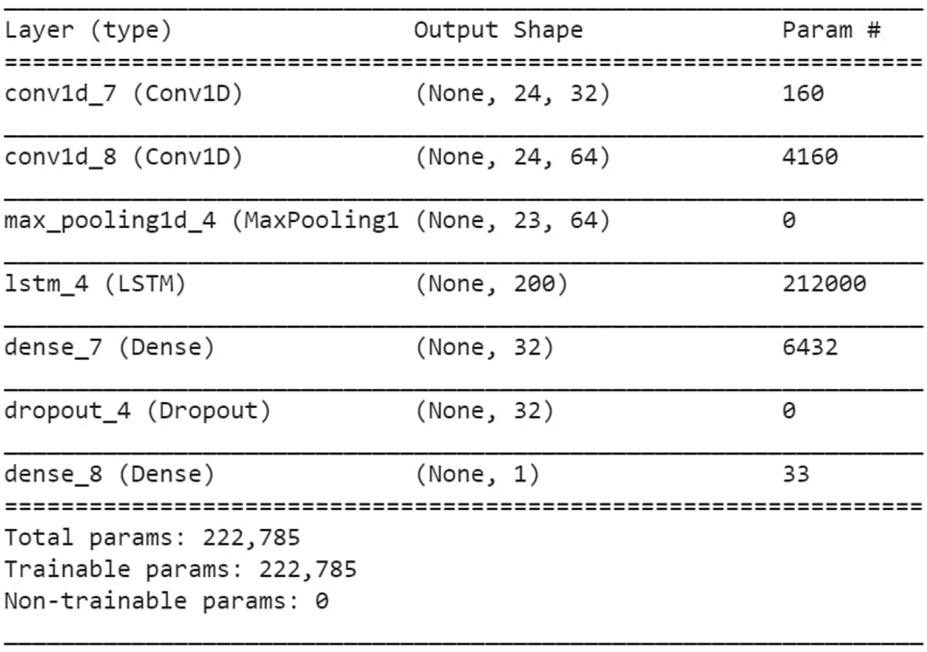


FIGURE 9 | Residuals left after applying VAR (Ontario Demand Data).



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv1d_7 (Conv1D)            (None, 24, 32)            160
_________________________________________________________________
conv1d_8 (Conv1D)            (None, 24, 64)            4160
_________________________________________________________________
max_pooling1d_4 (MaxPooling1 (None, 23, 64)            0
_________________________________________________________________
lstm_4 (LSTM)                (None, 200)               212000
_________________________________________________________________
dense_7 (Dense)              (None, 32)                6432
_________________________________________________________________
dropout_4 (Dropout)          (None, 32)                0
_________________________________________________________________
dense_8 (Dense)              (None, 1)                 33
=================================================================
Total params: 222,785
Trainable params: 222,785
Non-trainable params: 0
_________________________________________________________________
```

FIGURE 10 | VAR CNN-LSTM model summary (Ontario Demand Data).

After getting forecasting results from VAR, CNN-LSTM is trained on those forecasted results along with original data to learn all the intricacies from the data. The architecture of the CNN-LSTM model varies according to the number of layers, layer type, and parameter adjustment in each layer. It consists of convolution layers, pooling layers, flatten layer, LSTM layers, and dense layer to predict the corresponding output. For convolution operation, the number of filters, filter size, and strides need to be adjusted. By adjustment of these parameters to an optimal level, accuracy can be significantly improved. To properly adjust the parameters of the model, data should be analyzed appropriately. Input provided to the model consists of a sliding window of 24 data points (resampled to an hour). So, the input to CNN-LSTM is 24 × 2 size data as shown in **Figure 10**. The result of this model is as MAE:0,123, MSE: 00054, and RMSE:0233. The graph of actual vs predicted is as **Figure 11**.

The combined results of all the algorithms for Ontario Demand Data is displayed in **Table 4**. From the results of **Table 4**, it is clearly observed that the proposed VAR-CNN-LSTM(VACL) hybrid model has been compared with state of art models MV-KWNN(Talavera-Llames et al., 2019), MV-ANN(Hippert et al., 2001), ARIMAX (Box et al., 2011), VAR,

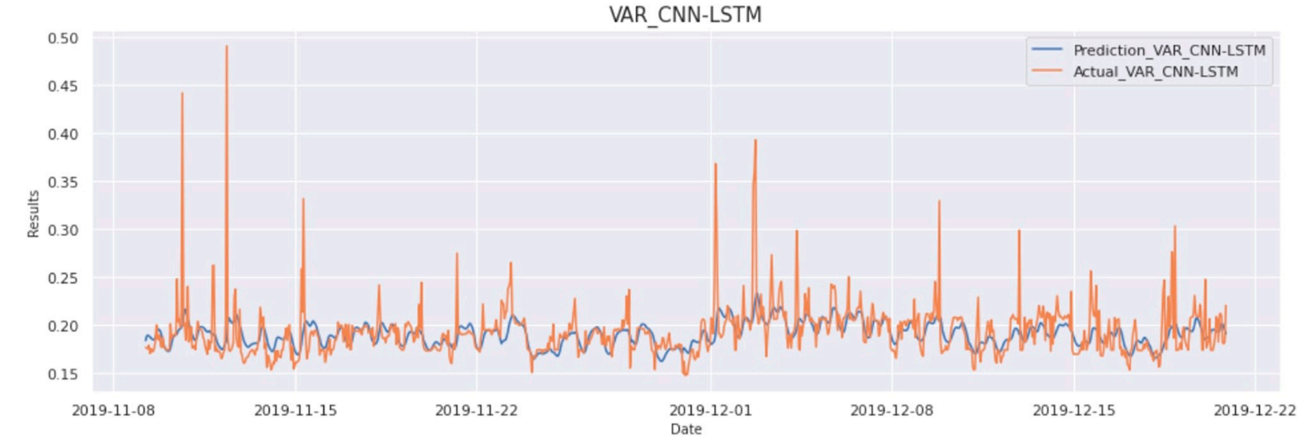


FIGURE 11 | Prediction vs Actual result of VAR CNN-LSTM model (Ontario Demand Data).

TABLE 4 | Combined results of all the algorithms on Ontario Demand Data (ontario Energy Price-Dataset, 2020) (official website of the Government of Canada, 2020).

	Mean absolute error	Mean squared error	Root mean squared error
VAR	0.651	0.521	0.569
MLP	0.0309	0.00204	0.0452
LSTM Hochreiter and Schmidhuber (1997)	0.0265	0.0015	0.0389
CNN-LSTM	0.0119	0.00068	0.02616
MV-kWNN Talavera-Llames et al. (2019)	0.0471	0.0421	0.0396
MV-ANN Hippert et al. (2001)	0.0596	0.0623	0.0696
ARIMAX Box et al. (2011)	0.0460	0.0583	0.0596
VAR-CNN-LSTM(VACL)	0.0123	0.00054	0.0233

MLP, LSTM and CNN-LSTM, and it outperforms all other models in terms of performance.

7 CONCLUSION AND FUTURE SCOPE

In this paper, the forecasting method for electricity load is investigated on a large dataset having linear and non-linear characteristics. We first formulated the problem as predicting the future term of multivariate time-series data, and then the proposed hybrid model VAR-CNN-LSTM(VACL) was deployed for efficient short-term power load forecasting. We have shown that the historical electrical load data is in the form of time series that consists of linear and non-linear components. Due to hybrid nature of the proposed model, the linear components were handled by VAR and residuals containing non-linear components by the combined CNN-LSTM layered architecture. The output efficiency was further enhanced by data preprocessing and analysis. With data preprocessing, the problem of missing values was solved, and data were normalized to bring values of the dataset to a common scale (Jaitley, 2019). From the data analysis, the correlation between variables have been discovered for, e.g., in household power consumption data, it was found that Global Active Power is correlated with all the variables in time series, so all variables wereused for forecasting. Since in Ontario Demand Dataset, only two variables were correlated, so all others were filtered out. The proposed method is modeled and tested on two publicly available datasets: Household Power Consumption Dataset and Ontario Demand dataset for short-term forecasting. The evaluation metrics used were MAE, MSE, and RMSE to show the effectiveness and errors respectively. From the results, it was established that the proposed hybrid VACL model performed better than other statistical and deep learning based techniques like VAR, CNN-LSTM, LSTM, MLP, and state-of-the-art model like MV-KWNN, MV-ANN and ARIMAX in all evaluation metrics.

One of the limitations of the proposed model was that determining all the hyperparameters like number of neurons, learning rate, number of epochs, batch size, etc., required great effort and time. As a future scope, more advanced hyperparameter optimization techniques may be used. Since the model has been tested for short-term load forecasting, the presented model will further analyze for the medium and long-term forecasting scenario.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://archive.ics.uci.edu/ml and https://www.ieso.ca/power-data.

AUTHOR CONTRIBUTIONS

Conceptualization and Investigation: AS. Software and validation: RT and AV. Supervision, Writing-review and editing: PP and OPV.

FUNDING

Funding received from Research Council of Norway Grant No.: 280617, CPSEC and joint collaboration with Department of Science and Technology (DST), India for the Cyber Physical Security in Energy Infrastructure for Smart Cities (CPSEC) project under Smart Environments theme of Indo-Norwegian Call.

ACKNOWLEDGMENTS

The work is supported by Department of Science and Technology(DST), India for the Cyber Physical Security in Energy Infrastructure for Smart Cities(CPSEC) project under Smart Environments theme of Indo-Norwegian Call.

REFERENCES

Albawi, S., Mohammed, T. A., and Al-Zawi, S. (2017). "Understanding of a Convolutional Neural Network," in 2017 International Conference on Engineering and Technology (ICET), Antalya, Turkey, August 21–23, 2017, 1–6. doi:10.1109/ICEngTechnol.2017.8308186

Alhussein, M., Aurangzeb, K., and Haider, S. I. (2020). Hybrid Cnn-Lstm Model for Short-Term Individual Household Load Forecasting. *IEEE Access* 8, 180544–180557. doi:10.1109/ACCESS.2020.3028281

Babu, C. N., and Reddy, B. E. (2014). A Moving-Average Filter Based Hybrid ARIMA-ANN Model for Forecasting Time Series Data. *Appl. Soft Comput.* 23, 27–38. doi:10.1016/j.asoc.2014.05.028

Bedi, J., and Toshniwal, D. (2019). Deep Learning Framework to Forecast Electricity Demand. *Appl. Energ.* 238, 1312–1326. doi:10.1016/j.apenergy.2019.01.113

Bendaoud, N. M. M., and Farah, N. (2020). Using Deep Learning for Short-Term Load Forecasting. *Neural Comput. Applic* 32, 15029–15041. doi:10.1007/s00521-020-04856-0

Bikcora, C., Verheijen, L., and Weiland, S. (2018). Density Forecasting of Daily Electricity Demand with Arma-Garch, Caviar, and Care Econometric Models. *Sustainable Energ. Grids Networks* 13, 148–156. doi:10.1016/j.segan.2018.01.001

Biller, B., and Nelson, B. L. (2003). Modeling and Generating Multivariate Time-Series Input Processes Using a Vector Autoregressive Technique. *ACM Trans. Model. Comput. Simul.* 13, 211–237. doi:10.1145/937332.937333

Bontempi, G. (2008). Long Term Time Series Prediction with Multi-Input Multi-Output Local Learning. *Proc. 2nd ESTSP*, 145–154.

Bourdeau, M., Zhai, X. q., Nefzaoui, E., Guo, X., and Chatellier, P. (2019). Modeling and Forecasting Building Energy Consumption: A Review of Data-Driven Techniques. *Sustain. Cities Soc.* 48, 101533. doi:10.1016/j.scs.2019.101533

Box, G. E., Jenkins, G. M., and Reinsel, G. C. (2011). *Time Series Analysis: Forecasting and Control*, Vol. 734. John Wiley & Sons.

Chatfield, C. (1996). *The Analysis of Time Series – an Introduction*. Chapman & Hall.

Chen, T.-B., and Soo, V.-W. (1996). "A Comparative Study of Recurrent Neural Network Architectures on Learning Temporal Sequences," in International Conference on Neural Networks, Washington, DC, June 3–6, 1996 (IEEE), 1945–1950.4.

Chniti, G., Bakir, H., and Zaher, H. (2017). "E-commerce Time Series Forecasting Using Lstm Neural Network and Support Vector Regression," in Proceedings of the International Conference on Big Data and Internet of Thing - BDIOT2017. doi:10.1145/3175684.3175695

Choi, H. K. (2018). *Stock price Correlation Coefficient Prediction with ARIMA-LSTM Hybrid Model. CoRR* abs/1808.01560.

Choi, J. Y., and Lee, B. (20182018). Combining Lstm Network Ensemble via Adaptive Weighting for Improved Time Series Forecasting. *Math. Probl. Eng.* 2018, 1–8. doi:10.1155/2018/2470171

[Dataset] Cohen, I. (2021). Time Series-Introduction. A Time Series Is a Series of data– by Idit Cohen – towards Data Science. Available at: https://towardsdatascience.com/time-series-introduction-7484bc25739a (Accessed on 04 06, 2021).

Du, P., Wang, J., Yang, W., and Niu, T. (2019). A Novel Hybrid Model for Short-Term Wind Power Forecasting. *Appl. Soft Comput.* 80, 93–106. doi:10.1016/j.asoc.2019.03.035

Elman, J. L. (1990). Finding Structure in Time. *Cogn. Sci.* 14, 179–211. doi:10.1207/s15516709cog1402_1

[Dataset] Erica (2021). Introduction to the Fundamentals of Time Series Data and Analysis - Aptech. Available at: https://www.aptech.com/blog/introduction-to-the-fundamentals-of-time-series-data-and-analysis/(Accessed on Jun 19, 2021).

Fukushima, K. (1980). Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position. *Biol. Cybernetics* 36, 193–202. doi:10.1007/bf00344251

Gasparin, A., Lukovic, S., and Alippi, C. (2019). *Deep Learning for Time Series Forecasting: The Electric Load Case. CoRR* abs/1907.09207.

Hartmann, C., Hahmann, M., Habich, D., and Lehner, W. (2017). "Csar: The Cross-Sectional Autoregression Model," in IEEE International Conference on Data Science and Advanced Analytics (DSAA), Tokyo, Japan, October 19–21, 2017, 232–241. doi:10.1109/DSAA.2017.27

Hebrail, G., and Berard, A. (2012). *Individual Household Electric Power Consumption Data Set*. Irvine: University of California.

Hippert, H. S., Pedreira, C. E., and Souza, R. C. (2001). Neural Networks for Short-Term Load Forecasting: a Review and Evaluation. *IEEE Trans. Power Syst.* 16, 44–55. doi:10.1109/59.910780

Hochreiter, S., and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Comput.* 9, 1735–1780. doi:10.1162/neco.1997.9.8.1735

Ibrahim, H., Ilinca, A., and Perron, J. (2008). Energy Storage Systems-Characteristics and Comparisons. *Renew. Sustain. Energ. Rev* 12 (5), 1221e50. doi:10.1016/j.rser.2007.01.023

[Dataset] Jaitley, U. (2019). Why Data Normalization Is Necessary for Machine Learning Models. Available at: https://medium.com/@urvashilluniya/why-data-normalization-is-necessary-for-machine-learning-models-681b65a05029".

Jordan, M. I. (1990). "Attractor Dynamics and Parallelism in a Connectionist Sequential Machine," in *Artificial Neural Networks: Concept Learning*. Artificial Neural Network:Concept Learning, 112–127.

Khwaja, A. S., Anpalagan, A., Naeem, M., and Venkatesh, B. (2020). Joint Bagged-Boosted Artificial Neural Networks: Using Ensemble Machine Learning to Improve Short-Term Electricity Load Forecasting. *Electric Power Syst. Res.* 179, 106080. doi:10.1016/j.epsr.2019.106080

Kim, T.-Y., and Cho, S.-B. (2019). Predicting Residential Energy Consumption Using Cnn-Lstm Neural Networks. *Energy* 182, 72–81. doi:10.1016/j.energy.2019.05.230

Lütkepohl, H. (2013). *Introduction to Multiple Time Series Analysis*. Springer Science & Business Media.

Ma, X., Jin, Y., and Dong, Q. (2017). A Generalized Dynamic Fuzzy Neural Network Based on Singular Spectrum Analysis Optimized by Brain Storm Optimization for Short-Term Wind Speed Forecasting. *Appl. Soft Comput.* 54, 296–312. doi:10.1016/j.asoc.2017.01.033

Mahalakshmi, G., Sridevi, S., and Rajaram, S. (2016). "A Survey on Forecasting of Time Series Data," in 2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16), Kovilpatti, India, January 7–9, 2016 (IEEE), 1–8. doi:10.1109/icctide.2016.7725358

Miller, N., Guru, D., and Clark, K. (2009). Wind Generation. *IEEE Ind. Appl. Mag.* 15, 54–61. doi:10.1109/mias.2009.931820

Nair, V., and Hinton, G. E. (2010). "Rectified Linear Units Improve Restricted Boltzmann Machines," in *Icml*.

[Dataset] official website of the Government of Canada (2020). *Historical Climate Data* (Accessed on Feburary 15, 2020).

[Dataset] Olah, C. (2013). *Understanding Lstm Networks*.

[Dataset] OMIE-Dataset (2020). Day-ahead Market Hourly Prices in spain. Available at: https://www.omie.es/en/file-access-list?parents%5B0%5D=/&parents%5B1%5D=Day-ahead%20Market&parents%5B2%5D=1.%20Prices&dir=%20Day-ahead%20market%20hourly%20prices%20in%20Spain&realdir=marginalpdbc (Accessed May 15, 2020).

[Dataset] ontario Energy Price-Dataset (2020). Hourly ontario Energy price (Hoep). Available at: https://www.ieso.ca/en/Power-Data/Data-Directory (Accessed on Feburary 15, 2020).

[Dataset] Prabhakaran, S. (2020). Vector Autoregression (Var) - Comprehensive Guide with Examples in python. Available at: https://www.machinelearningplus.com/time-series/vector-autoregression-examples-python/.

Rahman, A., Srikumar, V., and Smith, A. D. (2018). Predicting Electricity Consumption for Commercial and Residential Buildings Using Deep Recurrent Neural Networks. *Appl. Energ.* 212, 372–385. doi:10.1016/j.apenergy.2017.12.051

Rashid, T., Huang, B., Kechadi, T., and Gleeson, B. (2009). *Auto-regressive Recurrent Neural Network Approach for Electricity Load Forecasting 3*, 36–44.

[Dataset] Rawat, W., and Wang, Z. (2017). *Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review, Neural Computing*.

Sherstinsky, A. (2018). *Fundamentals of Recurrent Neural Network (Rnn) and Long Short-Term Memory (Lstm) Network*. New York: John Wiley & Sons.

Shi, H., Xu, M., and Li, R. (2017). Deep Learning for Household Load Forecasting—A Novel Pooling Deep Rnn. *IEEE Trans. Smart Grid* 9, 5271–5280. doi:10.1109/TSG.2017.2686012

Shiblee, M., Kalra, P. K., and Chandra, B. (2009). "Time Series Prediction with Multilayer Perceptron (Mlp): A New Generalized Error Based Approach," in Advances in Neuro-Information Processing Lecture Notes in Computer Science, Auckland, New Zealand, November 25–28, 2008, 37–44. doi:10.1007/978-3-642-03040-6_5

Shirzadi, N., Nizami, A., Khazen, M., and Nik-Bakht, M. (2021). Medium-Term Regional Electricity Load Forecasting through Machine Learning and Deep Learning. *Designs* 5, 27. doi:10.3390/designs5020027

Siami-Namini, S., Tavakoli, N., and Siami Namin, A. (2018). "A Comparison of Arima and Lstm in Forecasting Time Series," in 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, December 17–20, 2018, 1394–1401. doi:10.1109/ICMLA.2018.00227

Sinha, A., Tayal, R., Vyas, R., and Vyas, O. (2021). "Operational Flexibility with Statistical and Deep Learning Model for Electricity Load Forecasting," in *Lecture Notes in Electrical Engineering (LNEE)* (Springer).

Sorjamaa, A., and Lendasse, A. (2006). Time Series Prediction Using Dirrec Strategy. *Esann (Citeseer)* 6, 143–148.

Stoll, H. G., and Garver, L. J. (1989). *Least-cost Electric Utility Planning*. New York: Electrical Energy Systems; Electrical Companies.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). "Sequence to Sequence Learning with Neural Networks," in *Advances in Neural Information Processing Systems*, 3104–3112.

Taieb, S. B., Bontempi, G., Sorjamaa, A., and Lendasse, A. (2009). "Long-term Prediction of Time Series by Combining Direct and Mimo Strategies," in International Joint Conference on Neural Networks, Atlanta, GA, June 14–19, 2009 (IEEE), 3054–3061. doi:10.1109/ijcnn.2009.5178802

Talavera-Llames, R., Pérez-Chacón, R., Troncoso, A., and Martínez-Álvarez, F. (2019). Mv-kwnn: A Novel Multivariate and Multi-Output Weighted Nearest Neighbours Algorithm for Big Data Time Series Forecasting. *Neurocomputing* 353, 56–73. doi:10.1016/j.neucom.2018.07.092

Wang, J., Yu, L.-C., Lai, K. R., and Zhang, X. (2016). "Dimensional Sentiment Analysis Using a Regional CNN-LSTM Model," in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Berlin, Germany, August 7–12, 2016 (Berlin, Germany: Association for Computational Linguistics), 225–230. doi:10.18653/v1/P16-2037

[Dataset] Wikipedia (2021). Time Series Wikipedia. Available at: https://en.wikipedia.org/wiki/Time_series (Accessed on 0619, 2021).

Wu, F., Cattani, C., Song, W., and Zio, E. (2020). Fractional Arima with an Improved Cuckoo Search Optimization for the Efficient Short-Term Power Load Forecasting. *Alexandria Eng. J.* 59, 3111–3118. doi:10.1016/j.aej.2020.06.049

Wu, Z., Zhao, X., Ma, Y., and Zhao, X. (2019). A Hybrid Model Based on Modified Multi-Objective Cuckoo Search Algorithm for Short-Term Load Forecasting. *Appl. Energ.* 237, 896–909. doi:10.1016/j.apenergy.2019.01.046

Xiao, L., Shao, W., Wang, C., Zhang, K., and Lu, H. (2016). Research and Application of a Hybrid Model Based on Multi-Objective Optimization for Electrical Load Forecasting. *Appl. Energ.* 180, 213–233. doi:10.1016/j.apenergy.2016.07.113

Yan, K., Wang, X., Du, Y., Jin, N., Huang, H., and Zhou, H. (2018). Multi-step Short-Term Power Consumption Forecasting with a Hybrid Deep Learning Strategy. *MDPI J. Energies*. doi:10.3390/en11113089

Yao, X. (1993). A Review of Evolutionary Artificial Neural Networks. *Int. J. Intell. Syst.* 8, 539–567. doi:10.1002/int.4550080406

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Energy Research

ORIGINAL RESEARCH
published: 01 December 2021
doi: 10.3389/fenrg.2021.792358

A Novel Decomposition and Combination Technique for Forecasting Monthly Electricity Consumption

Xi Zhang[1] *and Rui Li*[2]*

[1]School of Economics, Beijing Technology and Business University, Beijing, China, [2]School of Economics and Management, University of Chinese Academy of Sciences, Beijing, China

With the share of electricity in total final energy consumption increasing quickly, the world is becoming increasingly dependent on electricity, which makes it more and more important to improve the forecasting accuracy of electricity consumption to ensure the normal operation of economic activities. In this paper, a novel decomposition and combination technique to forecast monthly electricity consumption is proposed. First, we use STL decomposition to obtain the trend, season, and residual components of the time series. Second, we use SARIMA, SVR, ANN, and LSTM to forecast trend, season, and residual component, respectively. Third, we use time correlation principle to improve the forecasting accuracy of season component. Fourth, we integrated the residual component predicted by SARIMA, SVR, ANN, and LSTM into a new sequence to improve the forecasting accuracy of residual component. In order to verify the performance of the proposed forecast model, monthly electricity consumption data in China is introduced as an example for empirical analysis. The results show that after STL decomposition, time correlation modification, and residual modification, the forecasting accuracy of each model has been gradually improved. We believe that the proposed forecast model in this paper can also be used to solve other mid- and long-term forecasting problems with obvious seasonal characteristics.

Keywords: monthly electricity consumption, STL, time correlation modification, ANN, LSTM

Edited by:
Jian Chai,
Xidian University, China

Reviewed by:
Hao Xiao,
Hunan University, China
Dongmei Guo,
Central University of Finance and Economics, China

Correspondence:
Rui Li
lirui21@ucas.ac.cn

Specialty section:
This article was submitted to Smart Grids, a section of the journal Frontiers in Energy Research

Received: *10 October 2021*
Accepted: *01 November 2021*
Published: *01 December 2021*

Citation:
Zhang X and Li R (2021) A Novel Decomposition and Combination Technique for Forecasting Monthly Electricity Consumption. Front. Energy Res. 9:792358. doi: 10.3389/fenrg.2021.792358

INTRODUCTION

Background

Resource depletion and global climate change are serious problems that human society is facing and will face for a long time. To escape from this dilemma, the global energy mix needs two transformations: clean energy substitution on the energy supply side and electric energy substitution on the energy consumption side. This paper focuses on electricity consumption. According to statistics, global electrification of the final consumption continues to follow an increasing trend, and the share of electricity in total final energy consumption is close to 20% in 2020.

As the world becomes more and more dependent on electricity, planning for electricity production is crucial. In addition, electricity is difficult to store, so it is usually used immediately after it is generated. This further increases the need for power companies to plan their electricity supply in a proactive manner. Reliable forecast of future electricity consumption level is the primary guiding principle of planning. In particular, high forecasting accuracy of medium- and long-term electricity consumption is the key to power system scheduling and planning. In contrast, inaccurate forecast of electricity consumption can backfire. Overestimation will waste scarce energy resources, huge capital investment, and long construction time. Underestimation will lead to more serious negative consequences, such as power shortage. Clearly, if effective early warning is given in advance based on high forecasting accuracy of electricity consumption, some measures can be adopted to avoid negative consequences. However, electricity consumption is uncertain, complex, and nonlinear, which depends on political conditions, economy (Lin and Liu, 2016), human activities, population behavior (Hussain et al., 2016), climate factors (Hernández, 2013), and other external factors affecting the forecasting accuracy of electricity consumption.

Literature Review and Motivation

At present, many techniques are used to forecast electricity consumption, which can be roughly divided into three categories: nonlinear intelligent model, statistical analysis model, and gray forecasting model. Nonlinear models mainly include the artificial neural network (Kandananond, 2011; Kaytez et al., 2015; Liu et al., 2017; Ghadimi et al., 2018; Bedi and Toshniwal, 2019; Hamzaçebi et al., 2019), support vector machine (Pai and Hong, 2005; Kavaklioglu, 2011; Cao and Wu, 2016), and Markov chain (Zhao et al., 2014). In addition to the nonlinear intelligent models mentioned above, statistical analysis models, such as regression analysis method (Mohamed and Bodger, 2005; Wang et al., 2018) and autoregressive integrated moving average (Yuan et al., 2016), have also been widely used in electricity consumption forecasting. The gray forecasting model proposed by Deng enjoys high popularity in many forecasting applications because it can describe the characteristics of uncertain systems even in the face of a small amount of data. Therefore, some literature forecast electricity consumption based on the gray model (Akay and Atak, 2007; Bahrami et al., 2014; Zhao and Guo, 2016; Xu et al., 2017; Ding et al., 2018; Wu et al., 2018).

These methods can generally provide good forecasts. However, the statistical analysis models have the limitation of linear (or near linear) assumption, the gray forecasting models are usually only suitable for time series that approximate exponential growth, and the nonlinear intelligent models often suffer from overfitting or the difficulty of parameter selection. To remedy these shortcomings, some decomposition and combination techniques have been proposed in recent years and achieve better performance: the SARIMA model with residual modification (Wang et al., 2012), wavelet transform combined with machine learning and time series models (Nguyen and Nabney, 2010), weighted hybrid model where trend and seasonal components are predicted by combined method, and SARIMA, respectively (Zhu, 2011), bagging ARIMA and exponential smoothing methods (de Oliveira and Cyrino Oliveira, 2018), convolutional neural networks and fuzzy time series (Sadaei et al., 2019), and structural combination of seasonal exponential smoothing forecasts (Rendon-Sanchez and de Menezes, 2019).

For the above existing researches, there are still some issues that need to be further studied. First, the statistical analysis models assume linearity and have good forecasting accuracy for periodic and regular sequences. The nonlinear intelligent model can forecast nonlinear and irregular time series better, but it has the problem of overfitting. How could the advantages of the two methods be combined to improve the forecasting accuracy? Second, except for the fluctuations of monthly electricity consumption affected by extreme weather changes, and sudden major economic and health events, the monthly electricity consumption also shows strong periodicity and regularity, so the comprehensive utilization of these two characteristics is meaningful to increase forecasting accuracy.

Contributions

To bridge the gap discussed above in the *Literature review and motivation* section, this paper develops a novel decomposition and combination forecasting technique. The primary research contents of this paper include three parts. First is the research on the monthly electricity consumption forecast based on STL decomposition. Second is the research on a time correlation modification based on annual periodicity and adjacent similarity to improve the forecasting accuracy of the season component. Third, considering the residual component has nonlinear and irregular characteristics, the individual model may only extract a certain feature of the sequence. Therefore, we integrate the residual component predicted by four models into a new sequence to improve the forecasting accuracy of the residual component. The main contributions of this paper are as follows:

1) A novel decomposition and combination forecasting model utilizing STL decomposition, time correlation principle (embodied as annual periodicity and adjacent similarity), and hybrid forecasting principle is proposed.
2) The monthly electricity consumption data of China are applied to evaluate the performance of the proposed model.

The remainder of the paper is organized as follows. The *Electricity consumption month-ahead forecasting model* section introduces the proposed forecasting model. The *Case study* section presents the simulation results and discussion, in which the performance of the proposed forecasting model is evaluated. Finally, conclusions are drawn in the *Conclusion* section.

ELECTRICITY CONSUMPTION MONTH-AHEAD FORECASTING MODEL

This section first briefly introduces individual models, including the STL algorithm, SARIMA, SVR, ANN, and LSTM model. Then the operation process of the proposed decomposition and combination method is described.

Seasonal–Trend Decomposition Using Loess Decomposition

For seasonal time series, academics generally use STL decomposition proposed by Cleveland et al. (1990) to obtain trend, season, and residual components. STL is a decompose model in the form of addition. In STL, loess is used to divide the time series into trend component, seasonal component, and residual component. Division is addition, that is, adding up the parts to get the original series. Specifically, the steps of STL decomposition are 1) detrending; 2) periodic subsequence smoothing: establish a sequence for each seasonal component and smooth it separately; 3) smoothing periodic substring low-pass filtering: recombine substring to smooth; 4) detrending the seasonal series; 5) detrending the original series using the seasonal components calculated in the previous steps; and 6) smoothing the de-seasonal sequence to obtain the trend component.

Seasonal Autoregressive Integrated Moving Average

SARIMA is one of the most widely used linear models for time series prediction. The general equation of this model is given by **Eq. 1**.

$$\phi_p(\mathrm{B})(1-B)^d\Phi_P(B^s)(1-B^s)^D y_t = \theta_q(\mathrm{B})\Theta_Q(B^s)a_t. \tag{1}$$

Here y_t is time series, a_t is white noise, and B is the lag operator. D represents the seasonal differentiation order, and d represents the regular differentiation order.

$$\phi_p(\mathrm{B}) = 1-\varphi_1 B-\varphi_2 B^2-\ldots-\varphi_p B^p. \tag{2}$$

$$\theta_p(\mathrm{B}) = 1-\theta_1 B-\theta_2 B^2-\ldots-\theta_q B^q. \tag{3}$$

Eqs. 2, **3** represent the autoregressive and moving average polynomial, respectively. They represent the dependence of future values of time series on past values as well as errors.

$$\Phi_P(\mathrm{B}) = 1-\mu_1 B^S-\mu_2 B^{2S}-\ldots-\mu_P B^{PS}. \tag{4}$$

$$\Theta_Q(\mathrm{B}) = 1-\upsilon_1 B^S-\nu_2 B^{2S}-\ldots-\nu_P B^{QS}. \tag{5}$$

Similarly, **Eqs. 4**, **5** represent the seasonal autoregressive and seasonal moving average polynomials, respectively. Addition of these polynomials to the ARIMA equation helps in capturing the seasonal variation in time series. Differentiation is necessary for converting the nonstationary time series to a stationary one. S represents the order of seasonality.

Support Vector Machine

SVM was first proposed by Vapnik (1963) based on the statistical learning theory and principle of structural risk minimization, which possess good performance even for small samples. The basic idea of support vector regression is to map original data to high-dimensional feature space and perform linear regression in the space. It can be formulated into:

$$f(x) = w^T\varphi(x)+b, \tag{6}$$

where $\varphi(x)$ is a nonlinear mapping function, $f(x)$ is the estimation value, and w^T and b are weights. It can be translated into an optimization problem:

$$\mathrm{Min}\,\frac{1}{2}w^T w + C\sum\nolimits_{t=1}^{T}(\xi_t+\xi_t^*),$$
$$s.t\begin{cases} w^T\varphi(x_t)+b-y_t\le\varepsilon+\xi_t,\ (t=1,2,\ldots,T) \\ y_t-\left(w^T\varphi(x_t)+b\right)\le\varepsilon+\xi_t^*,\ (t=1,2,\ldots,T), \\ \xi_t,\xi_t^*\ge 0\end{cases} \tag{7}$$

where C is the penalty parameter, and ξ_t and ξ_t^* are the nonnegative slack variables. Generally speaking, the parameters of SVR have a great influence on the accuracy of the regression estimation. Thereby, the grid search method is employed to automatically choose the optimal parameters of SVR in this paper.

Artificial Neural Network

ANN is an information processing method based on the biological neural network. Neural networks can theoretically simulate any complex nonlinear relationship through nonlinear units (neurons) and have been widely used in the field of forecast. The structure of artificial neural network consists of input layer, hidden layer, and output layer. The most widely used ANN model is the BP neural network model based on the BP algorithm. The neural network is determined by determining the weight between each layer. Therefore, the neural network is trained to set all the weights before being used for prediction. The initial weights are set randomly, and the output data can be obtained according to certain rules when the training process is going forward. The weights are modified based on the difference between the output data and the expected data during the fallback process. The forward and backward process is repeated until the difference between the output data and the required data is small enough.

Long Short-Term Memory

Traditional artificial neural networks (ANN) attempt to establish direct mapping between input historical data and output forecast data to achieve prediction methods. However, due to the absence of time correlation in data series, the neural network model cannot capture the relationship between data and time, which limits its application in time series prediction methods. Therefore, recursive neural network (RNN) is proposed to overcome this shortcoming. By adding cyclic connections on neurons, RNN can establish sequence-to-sequence mappings between input and output data. Therefore, the output of each time step is affected by the input of the previous time step. Therefore, RNN is used to realize the memory feature (Sutskever et al., 2014; LeCun et al., 2015).

The structure of RNN is shown in **Figure 1**. Each node represents a single time-step neuron. The connection weight

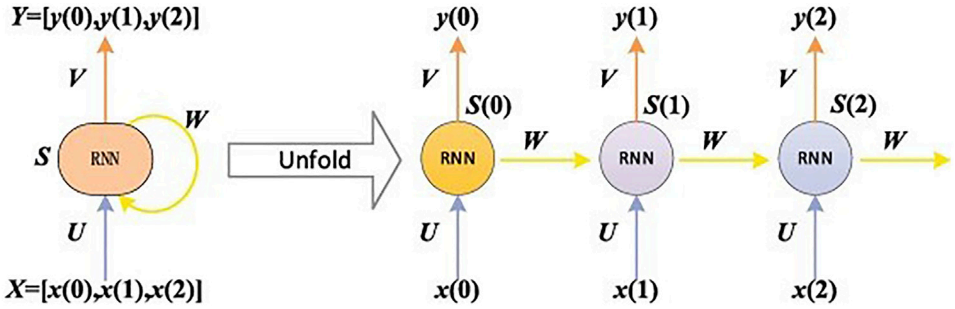


FIGURE 1 | The structure of recursive neural network (RNN) (Wang et al., 2020).

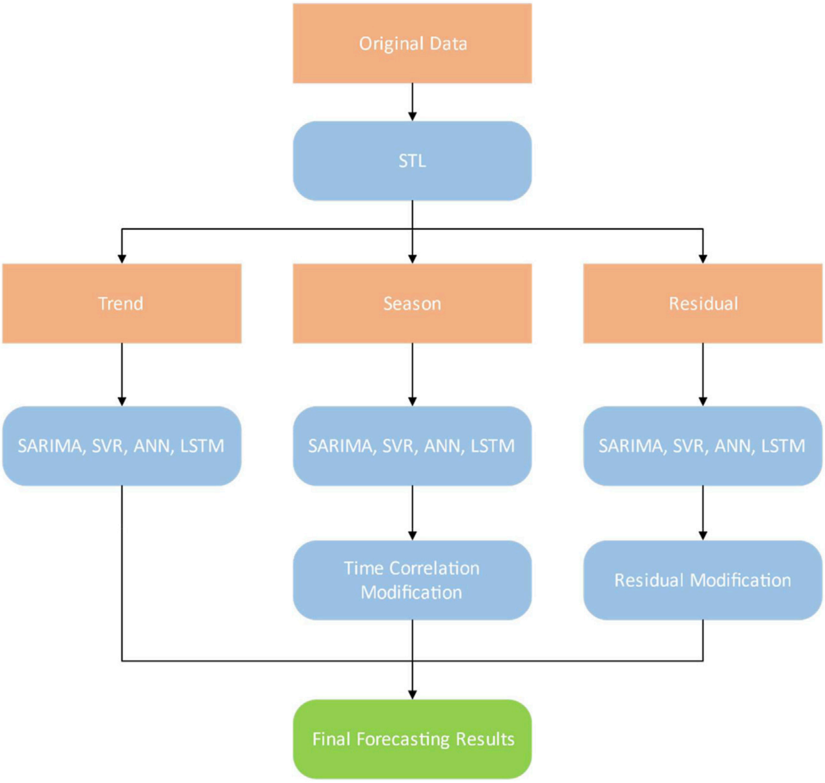


FIGURE 2 | The proposed forecast framework.

of input neuron is W1, the self-connection weight of each neuron is W2, and the connection weight of output neuron is W3. The input data sequence enters the network in turn according to the time step, and the weight coefficient is recycled.

The Proposed Forecast Framework

The proposed forecast framework utilizing STL decomposition, time correlation modification (embodied as annual periodicity and adjacent similarity) and residual modification is illustrated in **Figure 2**.

In **Figure 2**, The proposed forecast framework consists of four steps:

In the first step, we use the Seasonal–Trend decomposition using Loess (STL decomposition) to obtain the trend, season, and residual components of the time series.

In the second step, we use SARIMA, SVR, ANN, and LSTM to forecast trend, season, and residual component, respectively.

In the third step, we use the time correlation principle to improve the forecasting accuracy of the season component. The season component presents time correlation characteristics, which embodies as annual periodicity and adjacent similarity. Here the annual periodicity means that data from the same month in the next year are

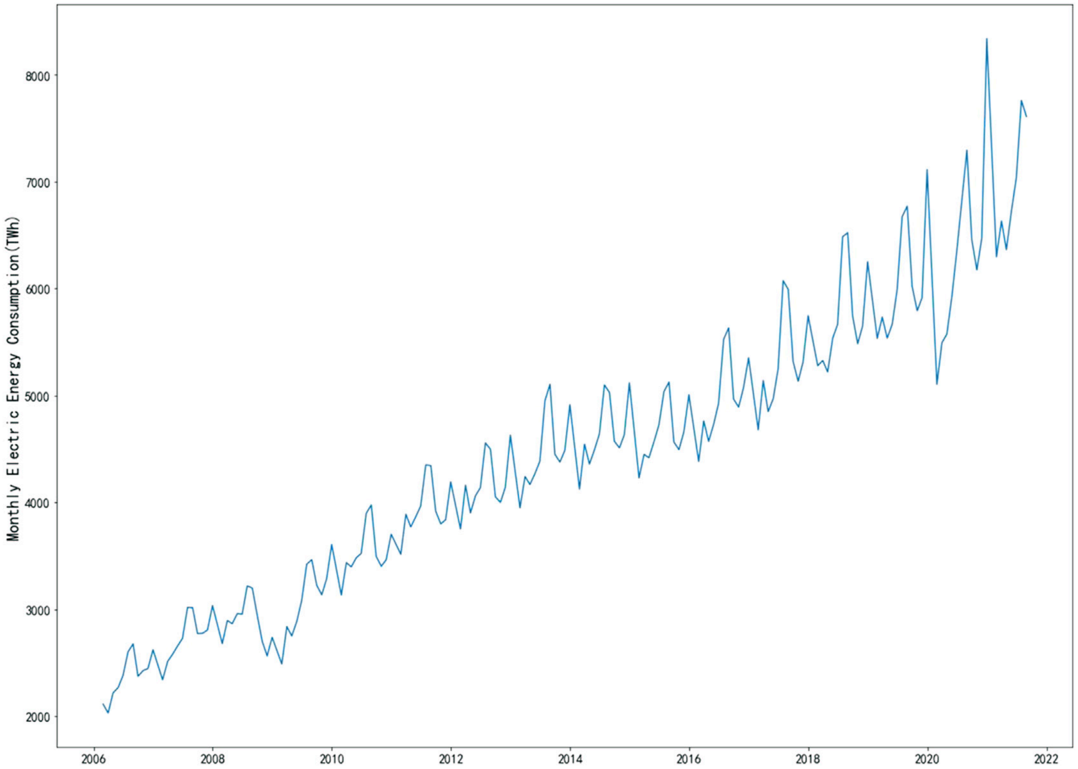


FIGURE 3 | The monthly electricity consumption of China from 1 and 2 2006 to August 2021.

similar. The adjacent similarity means that data are close to each other in adjacent months. In the second step, only adjacent similarity is used. We divide the season component into 11[2] subsequences, each of which represents a certain month. Then the exponential smoothing method is used to forecast each subsequence. The forecasting results are weighted with the season component predicted by each model (SARIMA, SVR, ANN, LSTM) to improve the forecasting accuracy of the season component. The weight is calculated based on the last forecasting error of the model.

In the fourth step, because the residual component has nonlinear and irregular characteristics, the individual model may only extract a certain feature of the sequence, so the forecasting accuracy is low. In fact, it is rare that a single forecasting model is always best in all cases. Each model has its own unique strengths and weaknesses. When multiple forecasting models are available, consider a combined approach, which is a good way to take full advantage of the strengths of each model. Therefore, we integrate the residual component predicted by SARIMA, SVR, ANN, and LSTM into a new sequence, and replace the residual component predicted by the above four methods with the new sequence to improve the forecasting accuracy of the residual component.

CASE STUDY

Data Collection

We evaluate the performance of the proposed forecasting method using the monthly electricity consumption data of China. However, these figures cannot be used directly as Chinese New Year always lasts for a few days in January or February. Almost all companies and factories have stopped operating. As a result, electricity consumption in January and February is sometimes abnormal. To avoid this problem, we treat the January and February averages as observations of a new month 1 and 2 each year, i.e., each year has 11 monthly values with a period length of 11. This study collects electricity consumption data from the beginning of 1 and 2 2006 to the end of August 2021 to keep relevant to the current situation of electricity development. These original data are shown in **Figure 3**.

Experimental Design

We select the data from 1 and 2 2006 to December 2018 as the training dataset (i.e., the first 143 data points) and the remaining data as the test dataset (i.e., the last 29 data points). The training data set is further divided into the optimization training data set and the verification data

[2] In the later empirical research, we treat the January and February averages as observations of a new month 1 and 2 each year.

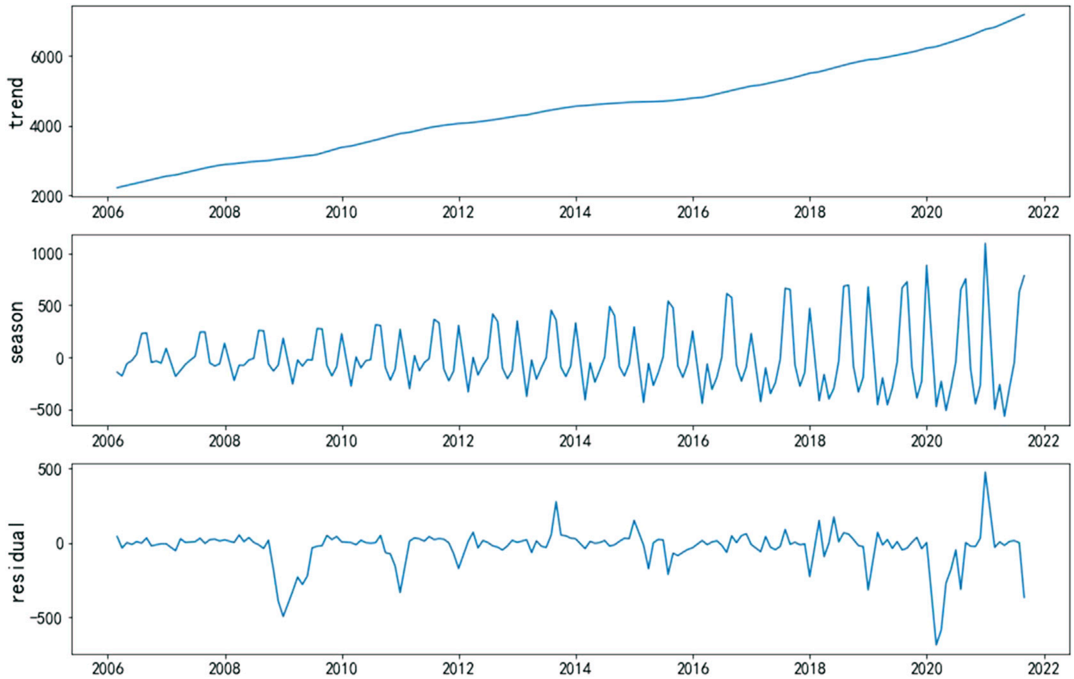


FIGURE 4 | The trend, seasonal, and residual for monthly electricity consumption data decomposed by Seasonal–Trend decomposition using Loess (STL).

TABLE 1 | Performance evaluations of different models with or without Seasonal–Trend decomposition using Loess (STL).

Horizons		**One-step ahead**			**Two-step ahead**			**Three-step ahead**		
Indices		**Mean absolute error (MAE) (TWh)**	**Mean absolute percentage error (MAPE) (%)**	**Correlation coefficient(COR)**	**MAE (TWh)**	**MAPE (%)**	**COR**	**MAE (TWh)**	**MAPE (%)**	**COR**
Without STL	SARIMA	230	3.58	0.88	259	4.06	0.86	280	4.40	0.85
	SVR	346	5.59	0.79	346	5.59	0.79	383	6.22	0.76
	Artificial neural network (ANN)	266	4.22	0.87	266	4.22	0.87	317	5.07	0.86
	LSTM	336	5.36	0.78	336	5.36	0.78	406	6.35	0.71
With STL	STL-SARIMA	126	1.98	0.96	121	1.93	0.96	131	2.10	0.96
	STL-SVR	145	2.23	0.95	145	2.23	0.95	197	3.04	0.94
	STL-ANN	150	2.34	0.95	150	2.34	0.95	189	2.99	0.94
	STL-LSTM	169	2.78	0.96	173	2.83	0.96	211	3.41	0.93

set. The optimization training data set contains the first 121 data points, and the verification training data set contains the last 22 data points. Optimization training and validation data sets were used to determine the hyperparameters of SVR, ANN, and LSTM models, while test data sets are used to evaluate forecasting performance.

Three error indices, mean absolute error (MAE), mean absolute percentage error (MAPE), and correlation coefficient (COR), are applied to evaluate the model performance according to forecast results. The official functions of the three error indices are:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} \left| y_i - \hat{y}_i \right|}{n}.$$

$$\mathrm{MAPE} = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|.$$

$$\mathrm{COR} = \frac{Cov\left(y_i\,,\ \hat{y}_i\right)}{\sigma_y \sigma_{\hat{y}}},$$

where y is the actual value, $\hat{y}$ is the forecasted value, and i is the index value of the data.

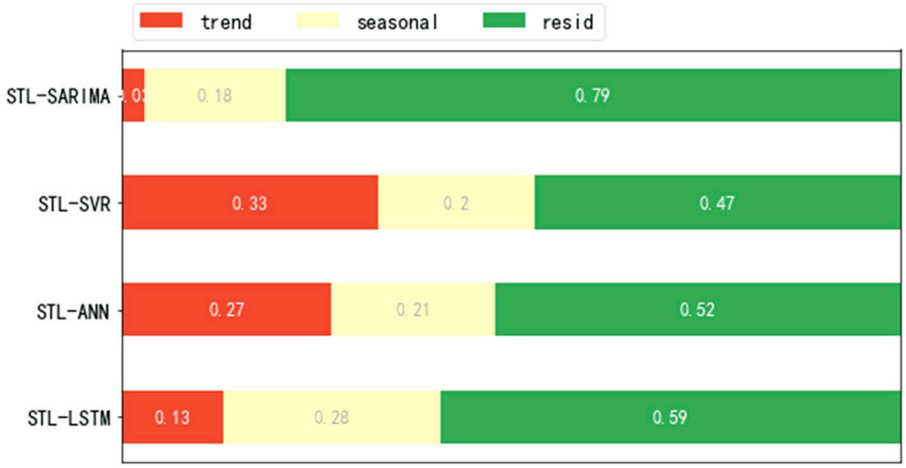


FIGURE 5 | Trend, season, random error ratio.

TABLE 2 | Performance evaluations of different models with time correlation modification and residual modification.

Horizons		**One-step ahead**			**Two-step ahead**			**Three-step ahead**		
Indices		**MAE (TWh)**	**MAPE (%)**	**COR**	**MAE (TWh)**	**MAPE (%)**	**COR**	**MAE (TWh)**	**MAPE (%)**	**COR**
TCM	STL-SARIMA-TCM	118	1.86	0.97	115	1.83	0.96	127	2.04	0.96
	STL-SVR-TCM	131	2.05	0.97	131	2.05	0.97	182	2.84	0.96
	STL-ANN-TCM	134	2.11	0.97	136	2.14	0.96	168	2.68	0.96
	STL-LSTM-TCM	158	2.61	0.97	165	2.71	0.97	167	2.70	0.95
RM	STL-SARIMA-TCM-RM	113	1.82	0.97	113	1.80	0.97	123	1.98	0.96
	STL-SVR-TCM-RM	135	2.14	0.97	136	2.15	0.97	187	2.92	0.96
	STL-ANN-TCM-RM	129	2.06	0.97	132	2.10	0.96	156	2.48	0.96
	STL-LSTM-TCM-RM	129	2.10	0.97	131	2.11	0.97	141	2.27	0.96

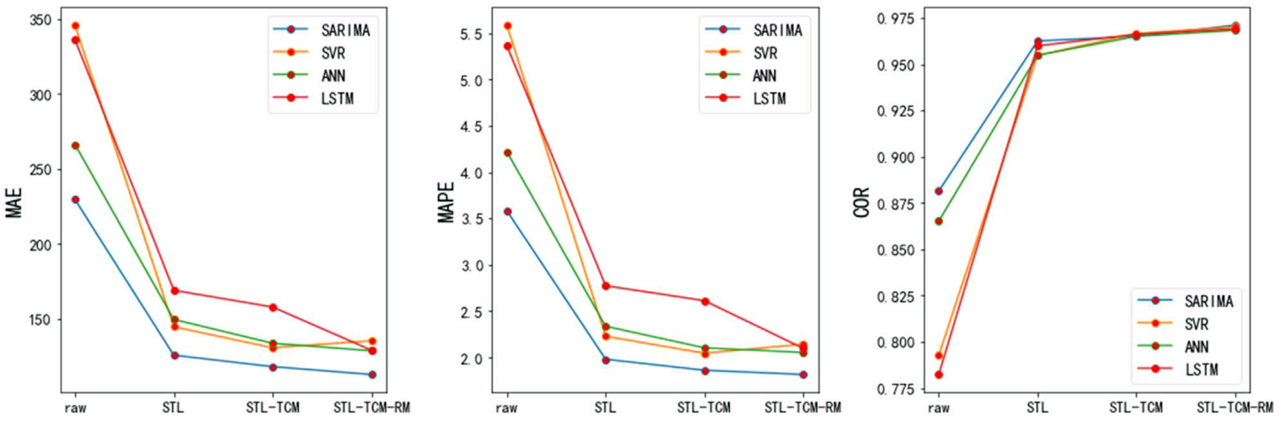


FIGURE 6 | Comparison between different models for 1-month ahead forecasting.

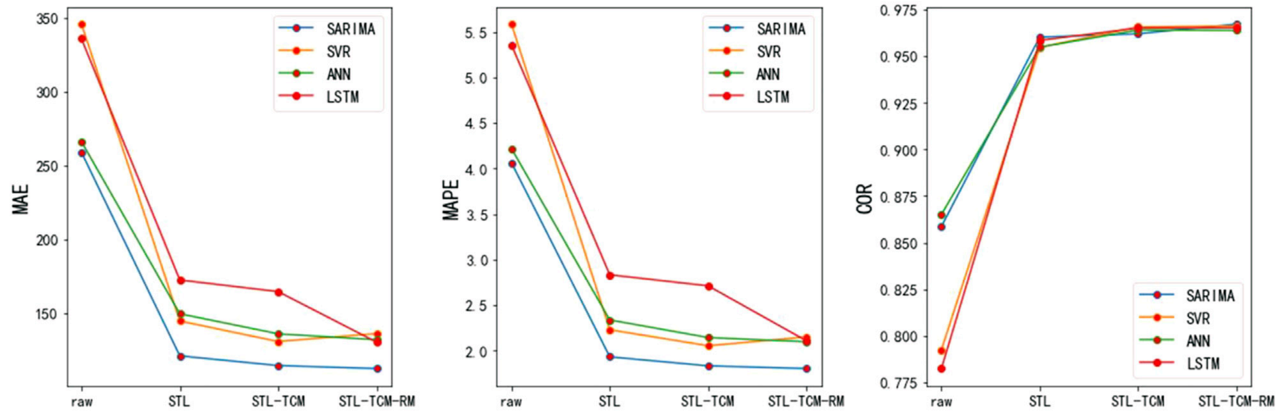


FIGURE 7 | Comparison between different models for 2-month ahead forecasting.

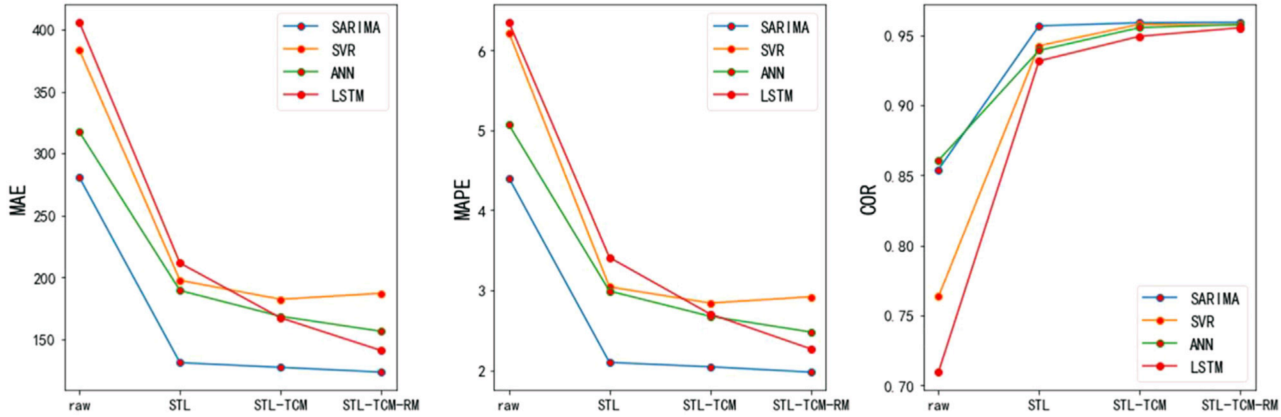


FIGURE 8 | Comparison between different models for 3-month ahead forecasting.

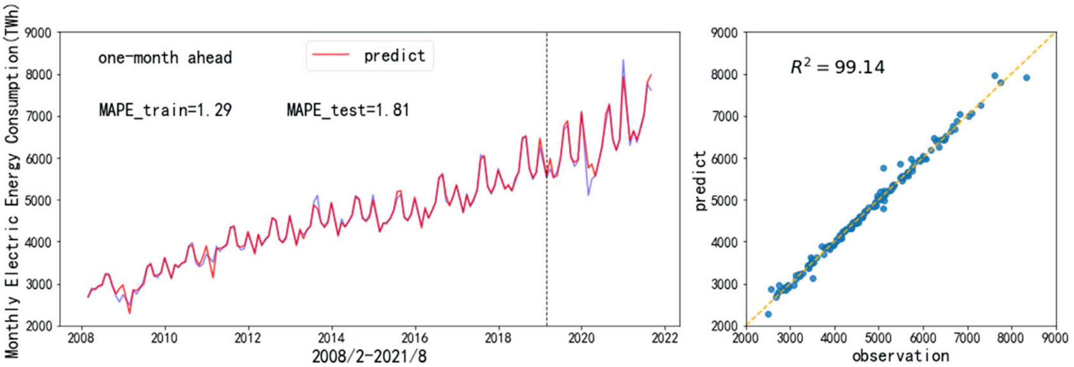


FIGURE 9 | STL-SARIMA–time correlation modification (TCM)–residual modification (RM) for 1-month ahead forecasting.

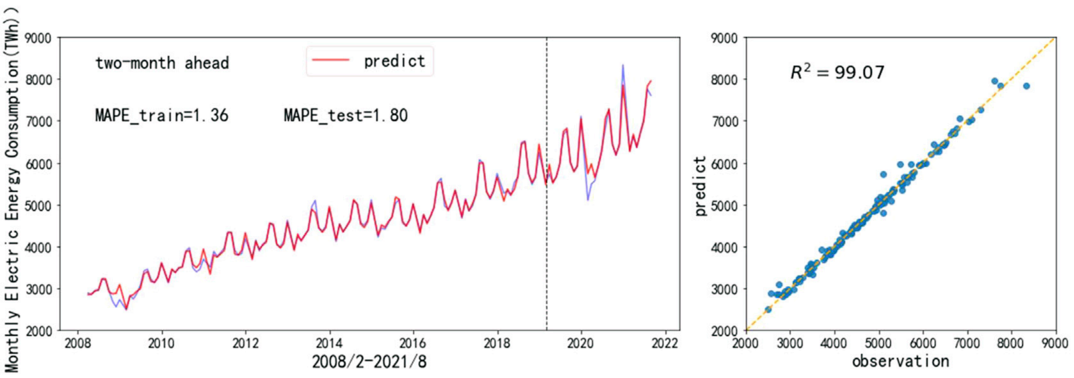


FIGURE 10 | STL-SARIMA-TCM-RM for 2-month ahead forecasting.

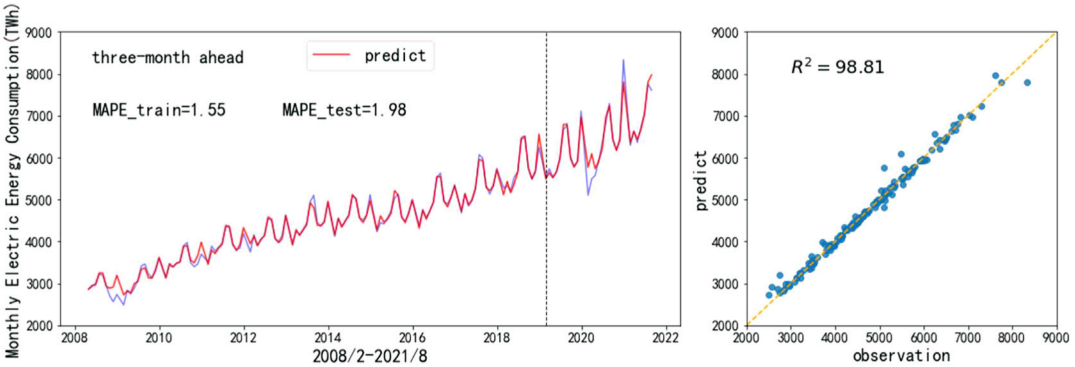


FIGURE 11 | STL-SARIMA-TCM-RM for 3-month ahead forecasting.

TABLE 3 | Performance evaluation of STL-SARIMA-TCM-RM in 2019, 2020, and 2021.

Horizons	One-step ahead			Two-step ahead			Three-step ahead		
Indices	MAE (TWh)	MAPE (%)	COR	MAE (TWh)	MAPE (%)	COR	MAE (TWh)	MAPE (%)	COR
2019	70	1.16	0.99	52	0.86	0.99	48	0.76	0.99
2020	166	2.77	0.97	193	3.19	0.97	220	3.66	0.96
2021	97	1.36	0.98	82	1.11	0.99	89	1.24	0.98

Results

Seasonal–Trend decomposition using Loess Decomposition

Figure 4 shows the STL decomposition results of the monthly electricity consumption. The trend component of the electricity consumption of China is increasing year by year, and the growth trend has accelerated since 2016. This is mainly because in 2016, eight departments in China jointly issued *The Guidelines on Promoting the Substitution of Electric Energy*, with a view to increasing the proportion of electric energy in the final energy consumption to 27%. Electric energy substitution is an important way to achieve carbon peak and carbon neutrality by replacing coal, oil, gas, and wood with electricity in energy consumption. The season component vibrates more and more. Due to financial crisis, extreme weather events, and epidemic, there are several relatively large negative and positive shocks on the residual component. If the original sequence is directly used, these huge shocks will seriously threaten the forecasting accuracy of the model.

Table 1 shows the performance evaluation results of four models without STL decomposition and with STL decomposition. The model comparisons demonstrate that STL decomposition is effective in boosting the forecasting accuracy of monthly electricity consumption. Compared with any single model (SARIMA, SVR, ANN, LSTM), the models with STL decomposition leads to reductions in all of the evaluation indices (MAE, MAPE and COR).

According to **Table 1**, the divide-and-conquer strategy improves the forecasting accuracy. Next, we analyze the source of errors, that is, the percentage of trend, season, and residual component forecasting errors to the total errors. As shown in **Figure 5**, for any model, most of the errors come from residual component forecast. SARIMA, in particular, was the least effective. This is because the residual component has nonlinear and irregular characteristics, and SARIMA is not good at forecasting these kinds of sequences. In addition, a single model may only extract a certain feature of the sequence, so the forecasting accuracy is low. For trend component, it can be seen that SARIMA has the highest forecasting accuracy, while for machine learning algorithms, such as SVR, ANN, and LSTM, the accuracy is not high. Therefore, it can be concluded that the traditional statistical method is better for simple sequence like trend component. The forecasting errors of season component also account for a large part.

Time Correlation Modification

As shown in **Figure 5**, the errors caused by season component account for 18%–28%. In this section, we use the periodicity of the seasonal series to improve the forecasting accuracy of the season component. As we can see, the season component presents time correlation characteristics, which embodies as annual periodicity and adjacent similarity. We divide the season component into 11 subsequences, each of which represents a certain month. Then exponential smoothing method is used to forecast each subsequence. The forecasting results are weighted with the season component predicted by each model (SARIMA, SVR, ANN, and LSTM) to improve the forecasting accuracy of the season component. The weight is calculated based on the last forecast error of the model. Rows 3–6 in **Table 2** show that the forecasting accuracy has been improved after time correlation modification (TCM).

Residual Modification

Figure 5 shows that most of the errors come from a residual component. This is because the residual component has nonlinear and irregular characteristics; a single model may only extract a certain feature of the sequence, so the forecasting accuracy of a single model is low. Therefore, we need to improve the forecasting accuracy of the residual component. We integrate the residual component predicted by SARIMA, SVR, ANN, and LSTM into a new sequence, and replace the residual component predicted by the above four methods with the new sequence to optimize each model. Rows 7–10 in **Table 2** show that the forecasting accuracy has been improved after residual modification (RM).

Comparison Between Different Models

Figures 6–8 show that after STL decomposition, time correlation modification, and residual modification, the forecasting accuracy of each model has been gradually improved. Among them, the forecasting accuracy improved the most after STL decomposition. This is mainly because there are many random disturbances in the original sequence, and the model will be affected by these disturbances if it is not decomposed.

DISCUSSION

According to **Figures 6–8**, STL–SARIMA–TCM–RM is the most accurate forecasting model.[3]. As we can see, on the one hand, compared with machine learning, SARIMA is better at forecasting trend, season, and other sequences with clear patterns. That is why it is so accurate. On the other hand, SARIMA is not good at forecasting an irregular random term. Therefore, residual modification can improve the forecasting accuracy of SARIMA most significantly. **Figures 9–11** show the STL–SARIMA–TCM–RM forecasting performance in the data set, as well as a scatter plot of forecasting results and actual values.

Considering that the test set includes COVID-19, we divide the test set into 2019, 2020, and 2021. **Table 3** shows that the 2019 forecast results are significantly better than that for 2020 and 2021.

CONCLUSION

This paper provides a novel decomposition and combination method to forecast electricity consumption. This approach first uses STL to decompose the sequence into trend, season, and residual components. Then the three decomposed subsequences are forecasted, and the season component forecasting results are modified according to the annual periodicity, and the forecasting results of the residual component of each model are integrated. The results show that STL-SARIMA-TCM-RM is the most accurate forecasting model.

In addition to electricity forecasting, we believe that the forecasting method proposed in this paper can also be used to solve other mid- and long-term forecasting problems with obvious seasonal characteristics, including tourist flow forecasting, energy consumption forecasting, traffic flow forecasting, and so on. Furthermore, this paper only focuses on univariate time series analysis and does not consider other factors affecting electricity consumption. If these factors can be introduced into the proposed learning method, the predictive performance may be better.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

[3] The STL-SARIMA-TCM-RM model is also superior to the existing models (Zhao et al., 2014; Cao and Wu, 2016; Jiang et al., 2020; Liu et al., 2020). Their models' MAPE all above 2, and the STL-SARIMA-TCM-RM model is less than 2.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

REFERENCES

Akay, D., and Atak, M. (2007). Grey Prediction with Rolling Mechanism for Electricity Demand Forecasting of Turkey. *Energy* 32 (9), 1670–1675. doi:10.1016/j.energy.2006.11.014

Bahrami, S., Hooshmand, R.-A., and Parastegari, M. (2014). Short Term Electric Load Forecasting by Wavelet Transform and Grey Model Improved by PSO (Particle Swarm Optimization) Algorithm. *Energy* 72, 434–442. doi:10.1016/j.energy.2014.05.065

Bedi, J., and Toshniwal, D. (2019). Deep Learning Framework to Forecast Electricity Demand. *Appl. Energ.* 238, 1312–1326. doi:10.1016/j.apenergy.2019.01.113

Cao, G., and Wu, L. (2016). Support Vector Regression with Fruit Fly Optimization Algorithm for Seasonal Electricity Consumption Forecasting. *Energy* 115, 734–745. doi:10.1016/j.energy.2016.09.065

Cleveland, R. B., Cleveland, W., McRae, J. E., and Terpenning, I. J. (1990). STL: A Seasonal-Trend Decomposition Procedure Based on Loess (With Discussion). *J. Off. Stat* 6, 3–73.

de Oliveira, E. M., and Cyrino Oliveira, F. L. (2018). Forecasting Mid-long Term Electric Energy Consumption through Bagging ARIMA and Exponential Smoothing Methods. *Energy* 144, 776–788. doi:10.1016/j.energy.2017.12.049

Ding, S., Hipel, K. W., and DangDang, Y.-g. (2018). Forecasting China's Electricity Consumption Using a New Grey Prediction Model. *Energy* 149, 314–328. doi:10.1016/j.energy.2018.01.169

Ghadimi, N., Akbarimajd, A., Shayeghi, H., and Abedinia, O. (2018). Two Stage Forecast Engine with Feature Selection Technique and Improved Meta-Heuristic Algorithm for Electricity Load Forecasting. *Energy* 161, 130–142. doi:10.1016/j.energy.2018.07.088

Hamzaçebi, C., Hüseyin Avni., Es., Es, H. A., and Çakmak, R. (2019). Forecasting of Turkey's Monthly Electricity Demand by Seasonal Artificial Neural Network. *Neural Comput. Applic* 31 (7), 2217–2231. doi:10.1007/s00521-017-3183-5

Hernández, D. (2013). Energy Insecurity: A Framework for Understanding Energy, the Built Environment, and Health Among Vulnerable Populations in the Context of Climate Change. *Am. J. Public Health* 103 (4), e32–e34. doi:10.2105/ajph.2012.301179

Hussain, A., Rahman, M., and Memon, J. A. (2016). Forecasting Electricity Consumption in Pakistan: The Way Forward. *Energy Policy* 90, 73–80. doi:10.1016/j.enpol.2015.11.028

Jiang, W., Wu, X., Gong, Y., Yu, W., and Zhong, X. (2020). Holt-Winters Smoothing Enhanced by Fruit Fly Optimization Algorithm to Forecast Monthly Electricity Consumption. *Energy* 193, 116779. doi:10.1016/j.energy.2019.116779

Kandananond, K. (2011). Forecasting Electricity Demand in Thailand with an Artificial Neural Network Approach. *Energies* 4 (8), 1246–1257. doi:10.3390/en4081246

Kavaklioglu, K. (2011). Modeling and Prediction of Turkey's Electricity Consumption Using Support Vector Regression. *Appl. Energ.* 88 (1), 368–375. doi:10.1016/j.apenergy.2010.07.021

Kaytez, F., Taplamacioglu, M. C., Cam, E., and Hardalac, F. (2015). Forecasting Electricity Consumption: A Comparison of Regression Analysis, Neural Networks and Least Squares Support Vector Machines. *Int. J. Electr. Power Energ. Syst.* 67, 431–438. doi:10.1016/j.ijepes.2014.12.036

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep Learning. *Nature* 521 (7553), 436–444. doi:10.1038/nature14539

Lin, B., and Liu, C. (2016). Why Is Electricity Consumption Inconsistent with Economic Growth in China. *Energy Policy* 88, 310–316. doi:10.1016/j.enpol.2015.10.031

Liu, Y., Wang, W., and Ghadimi, N. (2017). Electricity Load Forecasting by an Improved Forecast Engine for Building Level Consumers. *Energy* 139, 18–30. doi:10.1016/j.energy.2017.07.150

Liu, Y., Zhao, J., Liu, J., Chen, Y., and Ouyang, H. (2020). Regional Midterm Electricity Demand Forecasting Based on Economic, Weather, Holiday, and Events Factors. *IEEJ Trans. Elec Electron. Eng.* 15, 225–234. doi:10.1002/tee.23049

Mohamed, Z., and Bodger, P. (2005). Forecasting Electricity Consumption in New Zealand Using Economic and Demographic Variables. *Energy* 30 (10), 1833–1843. doi:10.1016/j.energy.2004.08.012

FUNDING

The research was supported by National Natural Science Foundation of China (72003009).

Nguyen, H. T., and Nabney, I. T. (2010). Short-Term Electricity Demand and Gas Price Forecasts Using Wavelet Transforms and Adaptive Models. *Energy* 35 (9), 3674–3685. doi:10.1016/j.energy.2010.05.013

Pai, P.-F., and Hong, W. C. H. (2005). Support Vector Machines with Simulated Annealing Algorithms in Electricity Load Forecasting. *Energ. Convers. Manag.* 46 (17), 2669–2688. doi:10.1016/j.enconman.2005.02.004

Rendon-Sanchez, J. F., and de Menezes, L. M. (2019). Structural Combination of Seasonal Exponential Smoothing Forecasts Applied to Load Forecasting. *Eur. J. Oper. Res.* 275 (3), 916–924. doi:10.1016/j.ejor.2018.12.013

Sadaei, H. J., de Lima e SilvaSilva, P. C., Guimarães, F. G., and Lee, M. H. (2019). Short-Term Load Forecasting by Using a Combined Method of Convolutional Neural Networks and Fuzzy Time Series. *Energy* 175, 365–377. doi:10.1016/j.energy.2019.03.081

Sutskever, I., Vinyals, O., and QuocLe, v. (2014). Sequence to Sequence Learning with Neural Networks. *Adv. Neural Inf. Process. Syst.* 4 (January), 3104–3112.

Zhu, L., Huang, X., Shi, H., Cai, X., and Song, Y. (2011). Transport Pathways and Potential Sources of PM10 in Beijing. *Atmos. Environ.* 45 (3), 594–604. doi:10.1016/j.atmosenv.2010.10.040

Vapnik, V. (1963). Pattern Recognition Using Generalized Portrait Method. *Autom. Remote Control* 24, 774–780.

Wang, F., Xuan, Z., Zhen, Z., Li, K., Wang, T., and Shi, M. (2020). A Day-Ahead PV Power Forecasting Method Based on LSTM-RNN Model and Time Correlation Modification under Partial Daily Pattern Prediction Framework. *Energ. Convers. Manag.* 212, 112766. doi:10.1016/j.enconman.2020.112766

Wang, Y., Wang, J., Zhao, G., and Dong, Y. (2012). Application of Residual Modification Approach in Seasonal ARIMA for Electricity Demand Forecasting: A Case Study of China. *Energy Policy* 48, 284–294. doi:10.1016/j.enpol.2012.05.026

Wang, Z.-X., Li, Q., Pei, L.-L., and Pei, L. L. (2018). A Seasonal GM(1,1) Model for Forecasting the Electricity Consumption of the Primary Economic Sectors. *Energy* 154, 522–534. doi:10.1016/j.energy.2018.04.155

Wu, L., Gao, X., Xiao, Y., Yang, Y., and Chen, X. (2018). Using a Novel Multi-Variable Grey Model to Forecast the Electricity Consumption of Shandong Province in China. *Energy* 157, 327–335. doi:10.1016/j.energy.2018.05.147

Xu, N., Dang, Y., and Gong, Y. (2017). Novel Grey Prediction Model with Nonlinear Optimized Time Response Method for Forecasting of Electricity Consumption in China. *Energy* 118, 473–480. doi:10.1016/j.energy.2016.10.003

Yuan, C., Liu, S., and Fang, Z. (2016). Comparison of China's Primary Energy Consumption Forecasting by Using ARIMA (The Autoregressive Integrated Moving Average) Model and GM(1,1) Model. *Energy* 100, 384–390. doi:10.1016/j.energy.2016.02.001

Zhao, H., and Guo, S. (2016). An Optimized Grey Model for Annual Power Load Forecasting. *Energy* 107, 272–286. doi:10.1016/j.energy.2016.04.009

Zhao, W., Wang, J., and Lu, H. (2014). Combining Forecasts of Electricity Consumption in China with Time-Varying Weights Updated by a High-Order Markov Chain Model. *Omega* 45, 80–91. doi:10.1016/j.omega.2014.01.002

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Energy Research

ORIGINAL RESEARCH
published: 10 December 2021
doi: 10.3389/fenrg.2021.802697

Urban Household Energy Consumption Forecasting Based on Energy Price Impact Mechanism

*Zhang Lianwei** and *Xiaoni Wen*

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

The energy price influence system is one of the key mechanisms in the study of energy consumption. China's household energy consumption has obvious regional differences, and rising income levels and urbanisation have changed the willingness and ability of households to make energy consumption choices. Based on the linear price effect of household energy consumption, this paper explores the scenario characteristics of energy prices affecting energy consumption, taking electricity and natural gas consumption as examples. Based on household energy consumption statistics from 2005 to 2018 in 36 major cities across China, the accuracy and change trends of household energy consumption forecasts are investigated through the decision tree-support vector machine (DT-SVR) non-linear forecasting technique. The study shows that the non-linear forecasting technique accurately portrays the predicted trends of changes in total urban household electricity and natural gas consumption. Within the less developed regions of economic development, income levels are still the main constraint on changes in urban household energy consumption, and the stimulating effect of income levels on household energy consumption has not been seen in the process of economic development in these less developed regions. Urbanisation as an important factor in examining household energy consumption, different development patterns and development processes will gradually be reflected in scenario aspects such as the choice of urban household energy consumption and changes in total consumption.

OPEN ACCESS

Edited by:
Lean Yu,
Beijing University of Chemical Technology, China

Reviewed by:
Quanying Lu,
Institute of Systems Science, Academy of Mathematics and Systems Science (CAS), China
Shaolong Sun,
Jiangsu University, China

Correspondence:
Zhang Lianwei
lwzhang2017@163.com

Specialty section:
This article was submitted to Sustainable Energy Systems and Policies, a section of the journal Frontiers in Energy Research

Received: *28 October 2021*
Accepted: *22 November 2021*
Published: *10 December 2021*

Citation:
Lianwei Z and Wen X (2021) Urban Household Energy Consumption Forecasting Based on Energy Price Impact Mechanism. Front. Energy Res. 9:802697. doi: 10.3389/fenrg.2021.802697

Keywords: household energy consumption, electricity, natural gas, DT-SVR, energy forecasting

INTRODUCTION

Globally, household energy consumption has reached nearly 35% of energy end-use consumption; the actual figures for China reflect this proportion to be over 10%, making it the second largest energy consuming sector in addition to industrial energy consumption. With the rapid economic and social development of China, household energy consumption has been growing at a relatively fast rate, with an average annual growth rate of 8% over the last two decades (Zheng et al., 2014), and this growth trend will continue to accelerate in the future (Yuan et al., 2015).

In terms of the main types of energy consumed by households in China, the use of fossil fuels is still the main source of household energy consumption. Under the constraints of global warming and environmental pollution, policy changes in household energy consumption will face reconciliation of accounts in terms of consumption structure and consumption patterns. There are many factors influencing household energy consumption, including per capita income, urbanisation and climatic conditions, when examined in terms of economic, social and environmental factors. Among these, there are inconsistent findings on the impact of urbanisation on household energy consumption. As the role of the population changes with urbanisation, the increased demand for electricity and the consumption of new electrical products are the main reasons for the increase in energy consumption due to the shift from rural to urban households (Fan et.al., 2017) Another type of study suggests that the efficient use of public facilities due to dense urban populations will help reduce the growth of household energy consumption (Han et al., 2016), etc. In addition, the main factor influencing household energy consumption is temperature conditions, which is important for the characteristics of household energy consumption in China. Energy consumption in the country is concentrated in a number of typical economic regions. Overall, there is a significant spatial distribution of household energy consumption in China. Beijing, Tianjin and Hebei are the main centres of energy consumption. In terms of changes in total per capita energy consumption, the main regional characteristics are "high in the north and low in the south", and the national trend of household coal consumption is "high in the west and low in the east". In the area of household electricity consumption, the nationwide pattern is "high in the south and low in the north".

REVIEW OF THE LITERATURE

Macro Factors of Household Energy Consumption

In terms of the influence of key factors on changes in total household energy consumption, income levels and population size are the main drivers of changes in household energy consumption. Increasing per capita income is positively correlated with household energy consumption; urbanisation has a typical "U" shaped non-linear relationship with household energy consumption; and energy prices have a significant negative relationship in influencing the change in total household energy consumption. At the same time, the trend of the influence of regional temperature on household energy consumption varies from region to region. In terms of household electricity consumption, per capita income, energy price, regional temperature, and urbanisation all show non-linear relationships on changes in household electricity consumption (Ding and Peng, 2020).

The consumption of energy in households has become a major sector contributing to the main growth in energy consumption. In this context, domestic and international research on household energy consumption is increasing year by year. In terms of exploring the factors influencing household energy consumption, Barnes et al. found, based on the energy ladder theory, that the structure of energy consumption shifts as household income increases; when household income increases by a certain amount (US$1000-1500) the consumption of electricity and natural gas increases significantly (Dougherty, 1993). Of course, changes in household energy mix need to take into account other factors such as utilities, resource endowments, cultural preferences, etc. However, because of possible economies of scale, demographic factors are important factors in examining structural changes in household energy consumption (Jingchao et al., 2012).

On the other hand, as urbanisation continues, the differences in energy consumption between urban and rural households have been widely discussed. Urban households have higher energy requirements per capita than other households due to differences in availability endowments (Lenzen et al., 2006). China's residential energy consumption has significant regional and stepwise characteristics: heating is the main component of winter energy consumption in northern cities; urban households have better energy consumption attributes than rural households in several aspects. Overall, total household energy consumption in China is low, dominated by coal consumption, and there is a large gap between urban and rural household energy consumption behaviour.

Micro-factors of Household Energy Consumption

Since the 1990s, the public has become aware of the fact that large emissions of greenhouse gases are the main cause of warming, and that greenhouse gases mainly originate from human consumption of energy, so that the relevant subjects concerning energy consumption and its carbon emissions have become the focus of academic research. Energy consumption in the household sector, on the other hand, has been a major area of study in recent years where the energy consumption sector is set to grow. Generally speaking, household energy consumption is contextually and morphologically diverse and can be divided into residential energy consumption generated in the internal space of the home and transport energy consumption generated outside the home through the use of private cars.

With regard to the micro perspective on the factors influencing household energy consumption, domestic studies have mainly conducted quantitative analysis from the perspective of household characteristics (e.g., housing type, household size, household type, etc.) and individual attributes (e.g. income level, education level, age stage, occupational category, etc.) (Saunders, 2013). However, the variability in habits and lifestyles of different households is one of the main reasons for intra-household differences in energy consumption. For example, changes in the lifestyles of household members, such as an increase in the number of dual-earner households and more time spent on leisure activities outside the home, can lead to a reduction in the amount of time people spend indoors and, consequently, a reduction in residential energy consumption. However, it has also been suggested that the relationship between residence time and energy consumption should be analysed more comprehensively in relation to the socio-economic characteristics of household members. Many foreign scholars have found that households with fewer members consume more residential energy. In the case of single-person households, for example, these households consume 17–30% more energy than households with two people living together.

This finding also demonstrates the importance of economies of scale in reducing energy consumption levels.

Energy Mix Characteristics of Household Energy Consumption

As China's economy develops and urbanisation progresses, the demand for household energy consumption has increased significantly. In the vast rural areas, it is still common for residents to rely on direct burning of firewood, straw and coal for their daily cooking. According to statistics, in 2014, China's per capita domestic energy use was 346 kg of standard coal, of which 364 kg of standard coal were used by urban residents and 325 kg of standard coal by rural residents. From the perspective of research, there are three perspectives on China's rural energy consumption: 1) Based on provincial-level rural energy statistics, the spatial pattern distribution and temporal distribution characteristics of China's rural energy consumption are studied. 2) A study of the structure and willingness to consume energy in rural areas, based on field survey data from rural households. 3) An analysis of the characteristics of rural individual energy consumption.

There are two different judgments on the transition of rural household energy consumption in academia: one view is that rural households in China still mainly use traditional biomass energy sources such as fuelwood, a representative survey report includes the household energy consumption survey by Renmin University of China (Baltruszewicz et al., 2021), and other studies using household surveys have also reached similar conclusions; the other view is that rural household energy consumption is dominated by coal, a representative study includes The other view is that rural household energy consumption is dominated by coal, and representative studies include a study by Tian Yishui of the Ministry of Agriculture and field research by other scholars. Therefore, it is difficult to judge the stage of rural household energy consumption in China through the results of existing studies alone, and to identify whether a fundamental shift from traditional non-commodity energy to commodity energy has been achieved. The lack of judgement on the current structure of rural household energy consumption will greatly influence the formulation of relevant public policies, such as whether to invest more in rural energy infrastructure, whether to promote policies aimed at eradicating rural energy poverty, and whether to increase efforts to transform and upgrade rural energy.

Regional Structural Characteristics of Household Energy Consumption

The Energy Ladder theory suggests that rural households with low incomes mostly use fuelwood or dung as cooking fuel, and as incomes increase, they gradually move up the 'energy ladder' to a new stage of using electrical lighting and fossil fuels for cooking activities. This shift to modern fuels is generally achieved when per capita incomes reach US$1,000 to US$1,500. This theory clarifies the link between income and the type of energy used and indicates the level of income required for the energy transition. According to the "energy ladder" the shift in the structure of energy consumption will be a gradual replacement of polluting energy by clean energy, inefficient energy by efficient energy, and less convenient energy by more convenient energy. Lu Hui et al. used hierarchical analysis to study the relationship between household income and energy consumption structure in Jiangsu and Anhui provinces, and showed that farmers with higher income levels placed more importance on comfort, convenience and hygiene when choosing energy sources (Lu and Lu, 2006).

Energy consumption in rural households in China is likely to go through a process of gradual substitution of high quality energy for low quality energy in a sequential manner (Wang et al., 2018; Zhong et al., 2020). **Table 1** presents the corresponding findings of representative literature based on household energy surveys of rural households in China in recent years. As can be seen, the point in time of the study, the area surveyed and the final conclusions vary greatly between the different literatures.

MODEL CONSTRUCTION AND DATA SOURCES

Household Energy Consumption Forecasting Model Construction

This paper applies the "linear regression-decision tree" method to analyse changes in energy consumption trends in urban households in China, based on a linear variable importance analysis. **Figure 1** illustrates the logic of the analysis in this paper. Linearity in this paper refers to linearity in a broad sense, i.e., the relationship between data and data.

1) Principle features of regression

Assuming that the data is x and the result is y, the model in the middle is actually an equation, which is a one-sided interpretation, but helps us to understand what a model really is. Mathematical modelling is about finding the relationship between data and data from the data given in the question, building a mathematical equation model, and getting the result to solve real world problems (Qi et al., 2021) and finding solutions to real-world problems based on data, especially the processing of random data in the context of big data era (Wang, et al., 2022). It is actually the same as the model in machine learning (Zhang et al., 2020). So what is the general model of linear regression? The general model expression for linear regression is.

$$h_\theta(x) = \sum_{i=0}^{n} \theta_i x_i = \theta^T x = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots \theta_n x_n \quad (1)$$

The mystery of the model has been unveiled to us in the form of this formula above. Don't be intimidated by the formula, just know what the model looks like. Suppose i = 0, which represents a quadratic equation, a straight line through the origin in the coordinate system, and so on. Loss function: This is used to estimate the extent to which the predicted value of your model, f(x), is inconsistent with the true value of YY. The smaller the loss function, the better the model will be.

TABLE 1 | Key findings on the structural characteristics of household energy consumption.

Author	Scope of the survey	Key findings
Tonooka et al. (2006)	Shaanxi Province	Biomass-based fuels
Li et al. (2013)	Jilin Province	Biomass-based fuels
Wang et al. (2007)	Shandong Province	Source coal-based
Zhang and Yang (2019)	Beijing, Tianjin, Hebei	Source coal-based
Xu et al. (2014)	Fujian, Shandong, Inner Mongolia, Guizhou, Hebei, Gansu and Qinghai Provinces	Similar proportion of biomass energy to coal use
Zhang et al. (2014)	Shanxi Province, Guizhou Province, Zhejiang Province	Similar proportion of biomass energy to coal use

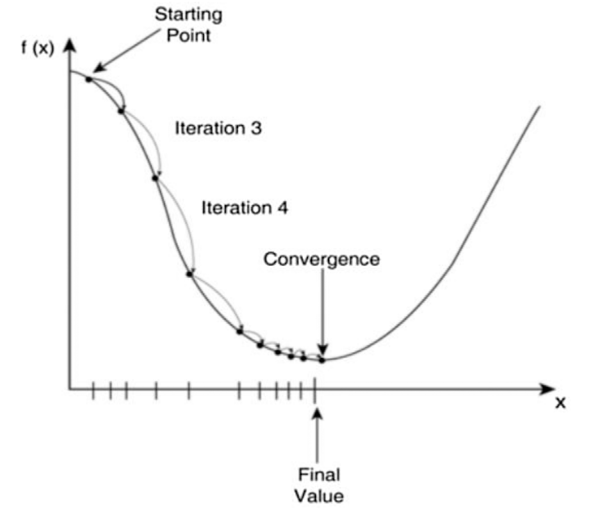


FIGURE 1 | Model training process simulation process.

$$f(\theta_0, \theta_1, \ldots, \theta_n) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 \tag{2}$$

At first the loss function is relatively large, but as the straight line keeps changing (the model keeps being trained), the loss function gets smaller and smaller, thus reaching the minima point, which is the final model we want to obtain. This method is collectively known as gradient descent. As the model continues to be trained, the gradient of the loss function becomes flatter and flatter until the point of minima, where the distance from the point to the line and the minimum, so that the line passes through all the points, which is the model (function) we require. By analogy, the same is true for a high-dimensional linear regression model. The model is optimised using gradient descent to find the extreme value points, which is the process of model training.

But there are two main issues in the process of model fitting. One is that in machine learning model training, the better the generalization ability of a model, the better the model performs. What is the generalisation capability of a model? A model's ability to generalise: how well a machine learning model learns concepts that the model has not encountered when it is in the process of learning. The generalisation ability of a model is a direct result of the over- and under-fitting of the model. Our goal is to minimise the sum of squares of the points to the line, so it is clear from the above illustration that the middle graph is a good fit, the leftmost case is an underfit, and the rightmost case is an overfit. Underfitting: The predicted value of the training set is quite wrong with the true value of the training set, which is called underfitting. Overfitting: The predicted value of the training set, which exactly fits the true value of the training set, is called overfitting. Underfitting is already well understood, that is, the error is relatively large, and overfitting is the training set on the performance is very good, a new batch of data for prediction results will be very unsatisfactory, generalization generalization is said to be a generalization. The solution uses a regularization term, which is a parameter to the gradient descent formula, i.e., Change in loss function from **Eq. 3** to **Eq. 4**

$$f(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 \tag{3}$$

$$f(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 + \lambda\sum_{j=1}^{n}\theta_j^2\right] \tag{4}$$

Adding this regularisation term has the advantage of controlling the magnitude of the parameters and not allowing the model to become "uncontrolled". Limiting the parameter search space solves the problem of underfitting and overfitting. As I said before, I'm explaining the principles and optimisation of linear regression models, but when it comes to actually using these methods, it's a simple matter of saying that someone else has already prepared these computational libraries, thanks to open source!

2) Algorithmic features of decision trees

A decision tree is a supervised learning algorithm. It applies to categories and continuous input (features) and output (predictor) variables (as shown **Figure 2**). The tree-based approach divides the feature space into a series of rectangles and then places a simple model (like a constant) for each rectangle. Conceptually, they are simple and effective. First we go through an example to understand decision trees. The process of creating a decision tree is then analysed using a formal analytical approach. Consider a simple data set of customers of a lending company. We are given all customers' checking account balances, credit history, length of tenure and previous loan status. The relevant task is to predict whether the customer's risk rating is credible. The problem can be solved using the following decision tree.

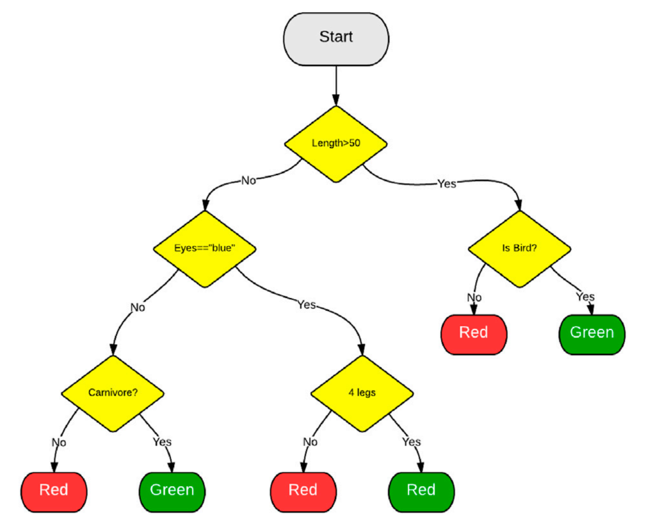


FIGURE 2 | General logical characteristics of a decision tree.

We now turn our attention to the details of the CART algorithm for regression trees. Briefly, the creation of a decision tree consists of two steps (Kupka et al., 2010).

1. divide the predictor space, i.e., the set of possible values X_1, X_2, . . ., X_p, into J distinct and non-overlapping regions R_1, R_2, . . ., R_J.
2. the same prediction is made for each sample observation entering region R_J, and that prediction is the mean of the training sample predictions in R_J.

In order to create J regions R_1, R_2,. . ., R_J, the predictor regions are divided into high-dimensional rectangles or boxes. The aim is to find the box region R_1, R_2, . . ., R_J that minimises the RSS by means of the following equation

$$\sum_{j=1}^{J}\sum_{i\in R_j}\left(y_i-\hat{y}_{R_j}\right)^2 \qquad (5)$$

where yhat_Rj is the average predicted value of the training observations in the jth box shape.

Since this spatial splitting is computationally infeasible, we often use a greedy approach to partition the region, called recursive binary splitting.

It is greedy because at each step of the tree creation process the best partition is selected at each particular step, rather than predicting the future and picking a partition that will appear in future steps and help create a better tree. Note that all partition regions R_j are rectangles. In order to perform a recursive binary partition, the predictor X_j is first selected and the cut point s

$$\sum_{i:\,x_i\in R_1(j,s)}\left(y_i-\hat{y}_{R_1}\right)^2+\sum_{i:\,x_i\in R_2(j,s)}\left(y_i-\hat{y}_{R2}\right)^2 \qquad (6)$$

where yhat_R1 is the average predicted value of the observed samples in region R_1(j,s) and yhat_R2 is the average predicted value of the observed samples in region R_2(j,s). This process is repeated to find the best predictors and cut points, and to further separate the data to minimise the RSS within each sub-region. However, we do not split the entire predictor space, we only split one or two of the previously identified regions. This process will continue until a stopping criterion is reached, for example we can set the stopping criterion to contain a maximum of m observations per region. Once we have created regions R_1, R_2, . . . R_J, given a test sample, we can use the average predicted value of all training samples in that region to predict the value of that test sample.

3) Least squares regression tree production algorithm

Input: training set—data set D.
Output: regression tree f(x).

In the input space where the training data set is located, recursively divide each region into two sub-regions and decide on the output value of each sub-region to construct a binomial decision tree.

Step 1: Choose the optimal cut variable j and cut point s. Solve that

$$\min_{j,s}\left[\min_{c1}\sum_{xi\in Ri(j,s)}\left(y_i-c_1\right)^2+\min_{c2}\sum_{xi\in R2(j,s)}\left(y_i-c_2\right)^2\right] \qquad (7)$$

Iterate over variable j, scanning the cut point s for a fixed cut variable j, choosing to use the above formula to bring it to a minimum value.

Step 2: Divide the region by the selected (j,s) and decide on the corresponding output value.

$$R_1(j,s)=\{x|x(j)\le s\} \qquad (8)$$

$$R_2(j,s)=\{x|x(j)>s\} \qquad (9)$$

$$\hat{c}_m=\frac{1}{N_m}\sum_{xi\in R_m(j,s)} yi,\quad x\in R_m,\quad m=1,2 \qquad (10)$$

Step 3: continue to call steps (1) and (2) for both subset regions until the stop condition is met.

Step 4: Divide the input space into M regions, R1, R2,. . ., Rm, to generate a decision tree.

$$f(x)=\sum_{m=1}^{M}\hat{c}_m I(x\in R_m) \qquad (11)$$

EMPIRICAL ANALYSIS

Analysis of Forecast Results for Urban Household Electricity Consumption

1) Linear forecast results for household electricity consumption

For the process of forecasting household energy consumption, both linear and non-linear forecasting techniques are used in this paper, and the results are

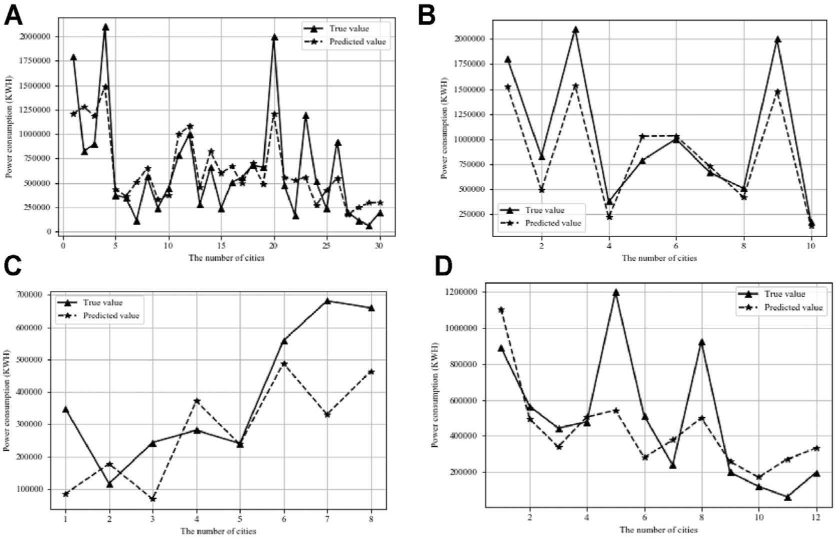


FIGURE 3 | Linear forecast results for national urban household electricity consumption.

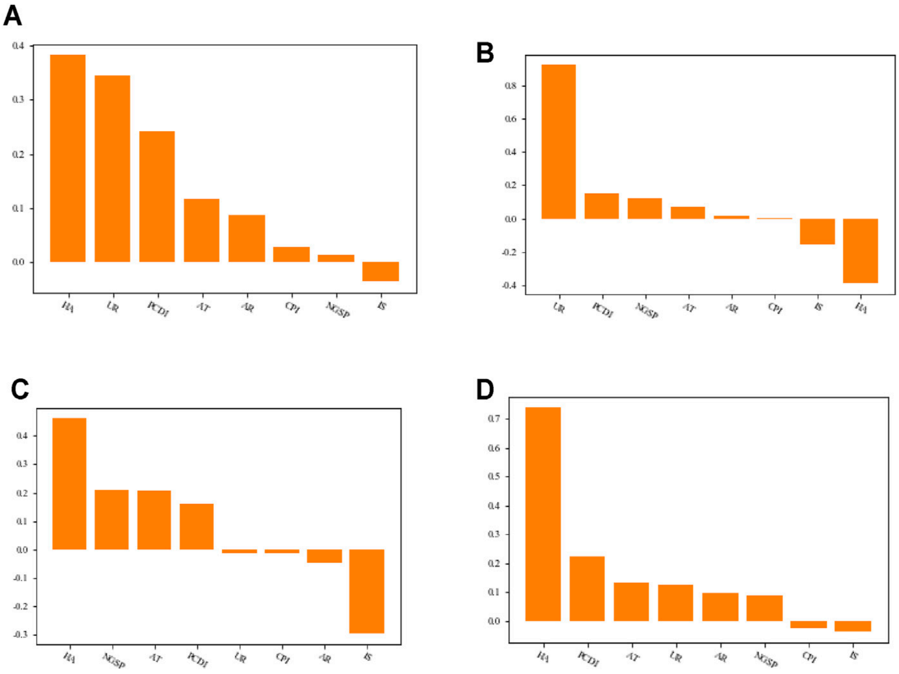


FIGURE 4 | Ranking the importance of linear predictor variables of electricity consumption in urban households across the country.

compared for two household energy varieties, electricity and natural gas consumption. **Figure 3** shows the sub-regional scenarios for linear forecasting of household electricity consumption. **Figure 3A** represents the national forecast of household electricity consumption levels; **Figure 3B** represents the forecast of household electricity consumption levels in the eastern region; **Figure 3C** represents the forecast of household electricity consumption levels in the central region; and **Figure 3D** represents the forecast of household electricity consumption levels in the western region.

The specific forecast results show a high degree of accuracy and trend consistency in the national forecast of household electricity consumption. The forecasted values of household electricity consumption at a national level, as expressed in the total sample forecast, are relatively close to the true value levels. By region, the forecast accuracy is better in the Eastern region,

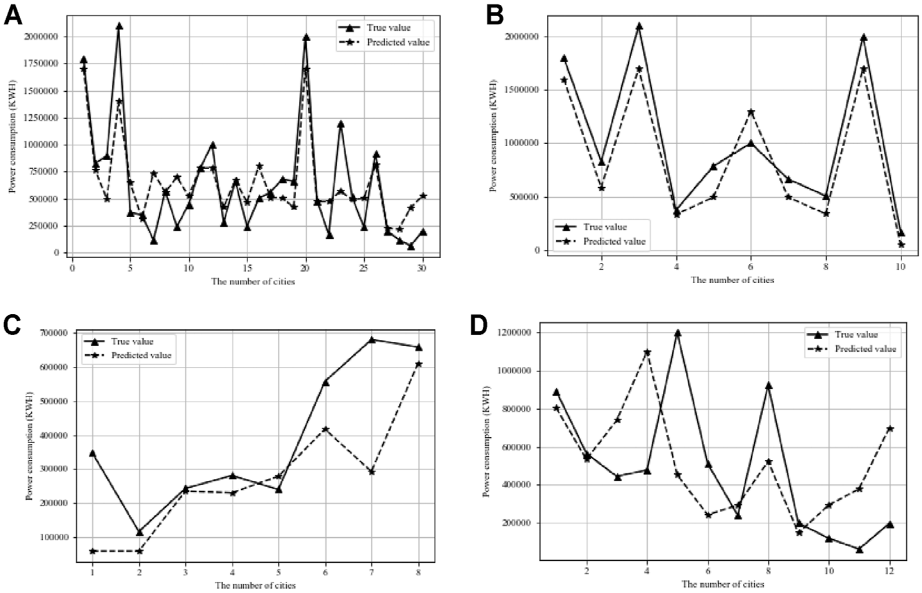


FIGURE 5 | Non-linear forecast results for national urban household electricity consumption.

followed by the Western region, and less accurate in the Central region. However, the trend in the distribution of forecast values shows that the forecast values for household electricity consumption in the East, Central and West regions show a better trend than the regional differences in electricity consumption levels.

Figure 4 shows the importance of the predictor variables for household electricity consumption at the national level. The national ranking shows that housing area, urbanisation rate and disposable income per capita are the top three factors in importance, indicating that living environment, urbanisation development and income play an important role in the growth of urban household electricity consumption; in terms of the directional characteristics of the impact, they all contribute positively to the growth of urban household electricity consumption.

2) Non-linear prediction results for household electricity consumption

Figure 5 shows the non-linear forecast of household electricity consumption by region. In particular, **Figure 5A** represents the national forecast of household electricity consumption levels; **Figure 5B** represents the forecast of household electricity consumption levels in the eastern region; **Figure 5C** represents the forecast of household electricity consumption levels in the central region; and **Figure 5D** represents the forecast of household electricity consumption levels in the western region. Compared to the linear forecast trend, the non-linear forecast results are more accurate.

Figure 5A shows the results of the non-linear forecasts of household electricity consumption at a national level. The non-linear forecasts capture the evolution of changes in household electricity consumption in terms of the trend between the forecasted and true values. Across the region, the predicted values of household electricity consumption are highly consistent with the true values and are more accurate than the linear forecasts. The Eastern region model forecasts accurately express the trends and levels of change in the true values, but the forecasts fluctuate more than the true values. At the same time, the predicted values accurately predict the direction of fluctuations in the true values. In the specific case of urban household electricity consumption in the central region, the forecast values accurately predict the trends in household electricity consumption, but are less accurate in terms of the magnitude and direction of fluctuations in the forecast values to the true values. In terms of the forecast for urban household electricity consumption in the western region, similar to the central region, the forecast values pounce well on the trends in total household electricity consumption in the western cities, but in individual cities the gap between the forecast values and the true values is large.

Figure 6 shows the degree of importance of factors influencing urban household electricity consumption in a non-linear forecasting scenario for urban household electricity consumption, from a national as well as a sub-regional perspective. From a national perspective, the level of urbanisation development is the number one factor influencing urban household electricity consumption and has a significant influence. In addition to the level of urbanisation, the top three influencing factors are the size of the housing stock and the price of electricity. The top three most important factors influencing urban household electricity consumption in the Eastern region are consistent with the national level, with the difference being in the order of importance. The first most important factor in the Eastern region is still the level of urbanisation, while the impact of electricity prices on household electricity consumption overtakes that of house

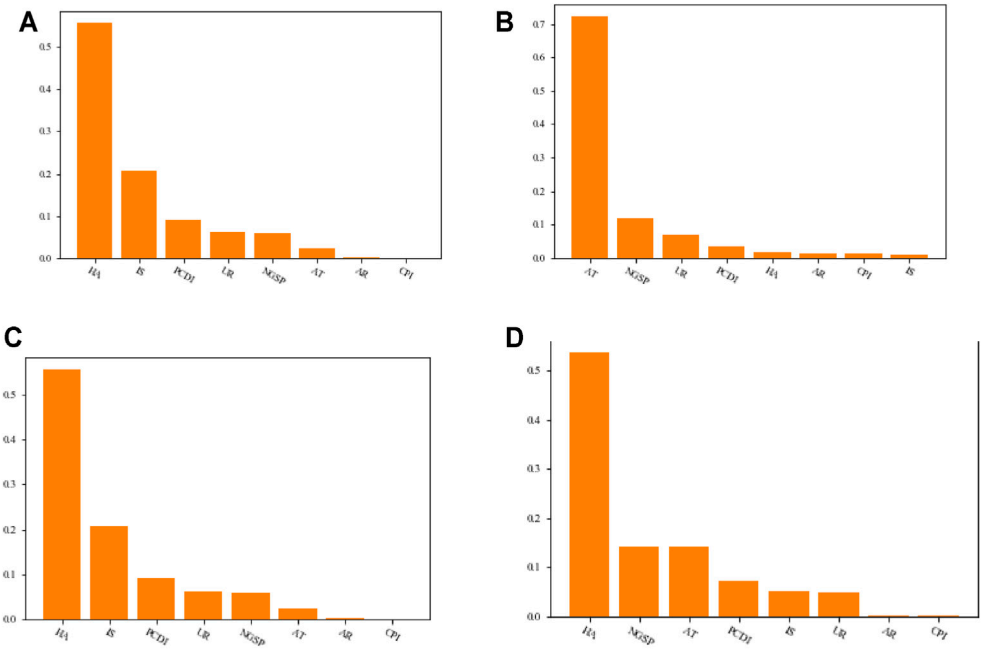


FIGURE 6 | Ranking the importance of non-linear predictor variables of electricity consumption in urban households across the country.

size in second place. In the central region, housing area, industrial structure and per capita disposable income are the three most important factors influencing urban household electricity consumption within the region. The area of the house is the first, while the industrial structure and disposable income per capita are the second and third. In the western region, the important factors influencing urban household electricity consumption are more distinctly different from those in the national, eastern and central regions, with house size, electricity price and average temperature being the three most important factors influencing urban household electricity consumption within the region. Further, in terms of economic development in the traditional sense, we find that the level of urbanisation and prices in the more economically developed regions, such as the coast, compared to the inland, are important factors influencing changes in household energy consumption. Factors affecting household electricity consumption in less economically developed regions in the traditional sense, such as the western region, are mainly manifested in factors such as climatic factors and the condition of residence. In the central region, in the context of the new economic development in recent years, the important factors affecting household electricity consumption in urban areas are economic factors such as industrial factors and household income factors.

Analysis of Forecast Results for Urban Household Gas Consumption

1) Linear prediction results for household gas consumption

Figure 7 shows the results of the linear forecast of urban household gas consumption. On a national scale, the overall forecast results are better. At the eastern scale, the trend in the forecast values is the same as the trend in the composition of the true values, with the forecast values largely reflecting the level of change in urban household gas consumption. At the central scale, the overall forecast is lower than the true value of household gas consumption in the central cities, but basically reflects the trend in the composition of household gas consumption, which is a good forecast. At the western end of the scale, the projections are generally closer to the level and composition of household gas consumption.

Figure 8 shows the ranking of the importance of the predictor variables in the context of linear consumption of natural gas by urban households. At a national level, the top three most important factors influencing household gas consumption are per capita disposable income, house size and industrial structure, which are positively influencing urban household gas consumption. At the same time, in terms of the negative relationship, we find that the level of urbanisation and natural conditions are negatively important factors influencing the change in urban gas consumption. At the eastern scale, disposable income per capita, industrial structure and average temperature are the main factors that positively influence the change in urban gas consumption, while the level of urbanisation and the volume of gas sales are important factors that negatively inhibit the change in household gas consumption. At the central level, the main factors influencing changes in household gas consumption are industrial structure, urbanisation and per capita disposable income, all of which contribute positively to the growth of household gas consumption. The CPI index, gas sales and weather conditions are the main factors inhibiting the growth of household gas consumption. On a western scale, the size of housing stock and disposable income per capita are the main drivers of change in urban

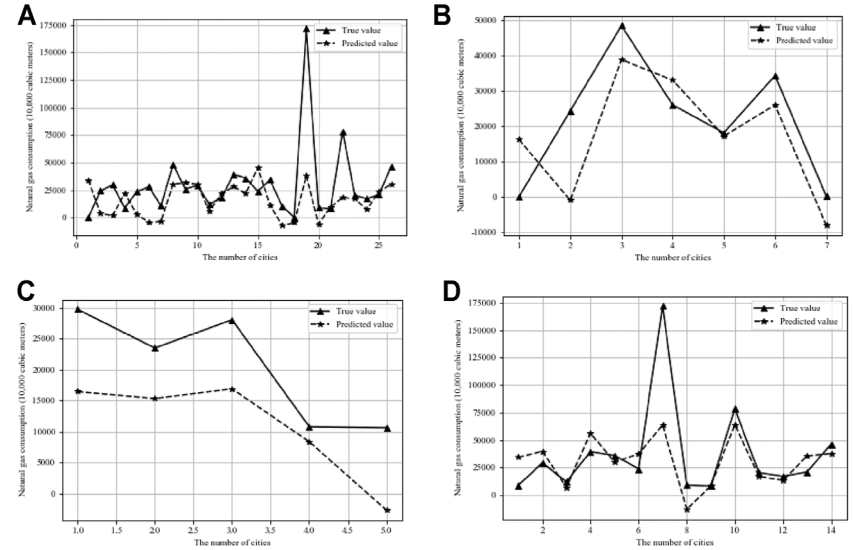


FIGURE 7 | Linear projection results for urban household gas consumption.

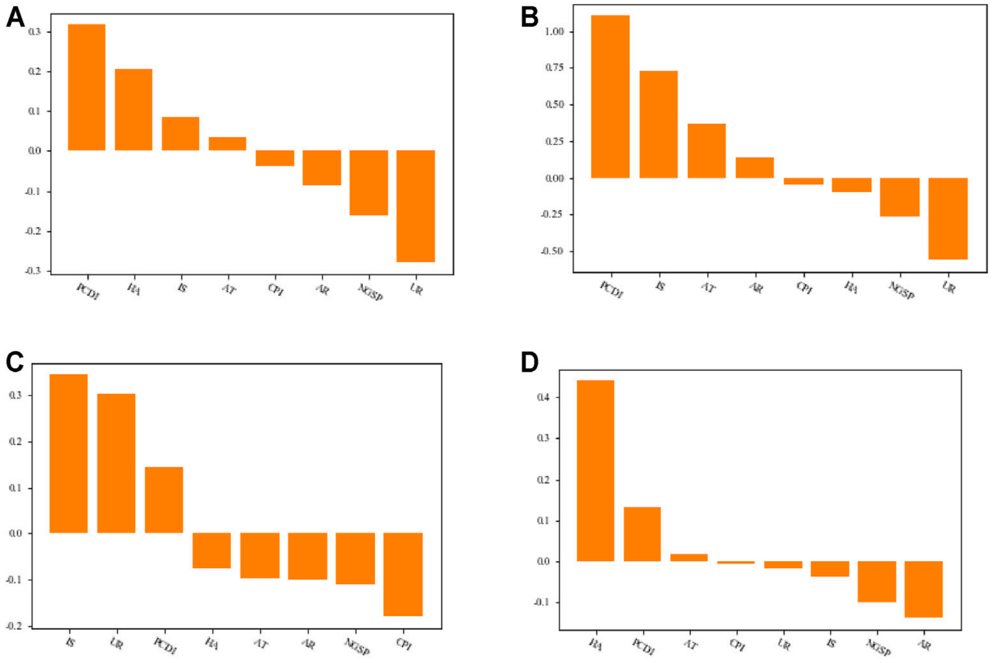


FIGURE 8 | Importance of linear predictor variables for urban household gas consumption.

household gas consumption, while annual rainfall, the price of gas sales and industrial structure are the main factors inhibiting the growth of urban household gas consumption.

2) Non-linear prediction results for household gas consumption

Figure 9 shows the results of the regional perspective on the forecasted changes in urban household natural gas consumption in China in the context of non-linear forecasting techniques. On a national scale, the non-linear forecasting results are closer to the true levels and composition trends of urban household natural gas consumption. On an eastern scale, the point projections for different cities are closer to the true value of urban household gas consumption, but the overall projections are smaller than the true value of gas consumption. At the central scale, the forecasts of urban

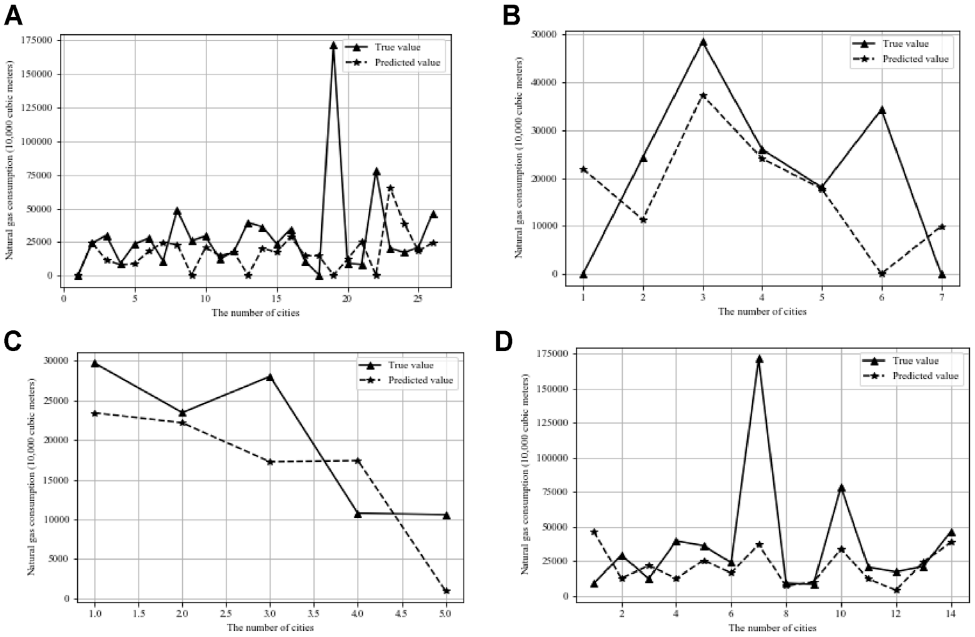


FIGURE 9 | Non-linear forecast results for urban household gas consumption.

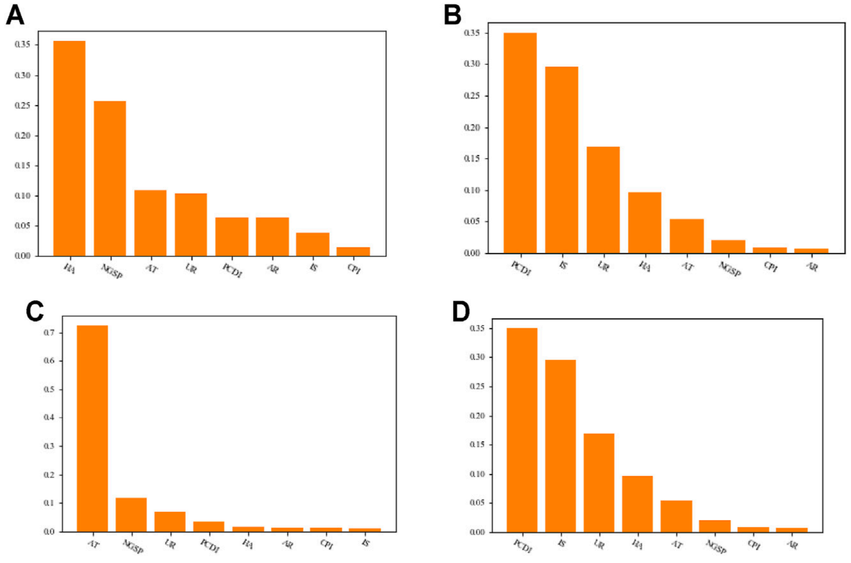


FIGURE 10 | Importance of non-linear predictor variables for urban household gas consumption.

household gas consumption are more accurate in predicting the trends and composition of the true values, but in the comparison of the forecasts of urban household gas consumption levels, the majority of the forecasts are lower than the true values. At the western end of the range, the forecasts for urban household gas consumption are good, with the forecasts for gas consumption levels in the sample cities generally close to the true levels and reflecting the composition of the true values, demonstrating the accuracy of the non-linear forecasts in the urban household gas consumption process.

Figure 10 shows the degree of importance of the influencing factors affecting changes in household gas consumption in the non-linear forecasting process for urban household gas consumption. At a national level, the factors influencing changes in urban household gas consumption remain the sales

TABLE 2 | Comparison of evaluation indexes of linear and non-linear results.

Region	Linear model (LR)			Decision trees (GBDT)			Support vector machines (SVR)		
	MAE	RMSE	DS	MAE	RMSE	DS	MAE	RMSE	DS
National Coal Power	240,946	9.94E+10	0.800	240,637	1.2E+11	0.830	233,234	9.81E+10	0.862
Eastern Coal Power	233,159	8.84E+10	0.890	222,139	7.3E+10	1.000	235,379	8.99E+10	1.000
Central Coal Power	150,975	3.48E+10	0.710	122,916	3E+10	0.720	150,163	3.57E+10	0.710
Western Coal Power	193,473	6.78E+10	0.820	219,857	8.4E+10	0.910	188,065	7.06E+10	0.900

price of gas, the size of the housing stock and disposable income per capita. At the eastern level, disposable income per capita, industrial structure and the level of urbanisation are the main factors that positively drive changes in urban household gas consumption. At the central scale, climatic conditions (average temperature) are the absolute drivers of growth in urban household gas consumption. At the western scale, the size of housing stock, the price of natural gas sales and the average temperature are important factors influencing the change in urban household gas consumption within the region. From a non-linear forecasting perspective, all influencing factors are positively associated with the level of change in urban household gas consumption.

The forecasting analysis of urban household electricity and natural gas consumption above shows that the overall non-linear forecasting technique is superior to the linear forecasting technique. From a sub-regional research perspective, the drivers affecting household energy consumption vary across regional scales, showing significant regional differences. This may be related to the differences in the scale and level of development of the more pronounced regional economies in China.

Analysis of Forecast Results for Urban Household Gas Consumption

1) Evaluation of urban household electricity consumption forecasts

This paper uses the DT-SVR non-linear technique for forecasting urban household energy consumption. Specific analyses were carried out to forecast and analyse urban household electricity consumption levels from four perspectives: national, eastern, central and western. In order to assess the effectiveness of the DT-SVR forecasting technique chosen in this paper, the forecasting results are evaluated in this paper, comparing the accuracy evaluation criteria of the linear and non-linear forecasting approaches respectively.

Table 2 indicates the accuracy evaluation indicators for the forecasts in this section. Firstly, the squared absolute error values of forecasts for national urban household electricity consumption show that the mean absolute error (MAE) values for the decision tree algorithm and the support vector machine approach are significantly smaller than the MAE values under the linear forecasting approach. At the same time, the root mean-square error (RMSE) values for the decision tree algorithm and the support vector machine approach are significantly smaller than the RMSE values under the linear forecasting approach. These types of evaluation indicators illustrate that at the national level, non-linear forecasting of household electricity consumption is more accurate. Secondly, the accuracy of the urban household electricity consumption forecasting results is seen in the eastern, central and western parts of the region, while the mean absolute error and root mean square error values of the non-linear forecasts are significantly smaller than the MAE and RMSE values of the linear forecasts. Further, this paper shows the results of linear and non-linear forecasting techniques by comparing the directional symmetry of the forecasts. The results show that the directional symmetry of directional symmetry (DS) under the national urban household electricity consumption level forecast is significantly higher than the DS value obtained under the non-linear forecasting technique.

These results further validate the role of non-linear forecasting techniques in influencing the level of urban household electricity consumption.

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(x_t - \hat{x}_t\right)^2} \tag{12}$$

$$MAE = \frac{1}{N}\sum_{t=1}^{N}\left|x_t - \hat{x}_t\right| \tag{13}$$

$$DS = \frac{1}{N}\sum_{t=1}^{N} d(t) \times 100\%,\ d(t) = \begin{cases} 1 & if\,(x_{t+1} - x_t)(x_{t+1} - x_t) \geq 0 \\ 0 & otherwise \end{cases} \tag{14}$$

4) Trends in non-linear forecasts of household electricity consumption

Table 3 shows the forecast trend of household electricity consumption in major cities in eastern China. The forecast results show that in mega cities, such as Beijing, Tianjin and Shanghai, household electricity consumption is forecast to show a year-on-year growth trend from 2019 to 2023. The DT-SVR non-linear forecast finds that household electricity consumption in cities such as Xiamen and Ningbo will show fluctuations in individual years from 2019 to 2023. In comparison, urban household electricity consumption in Hangzhou and Fuzhou will show individual year declines.

Table 4 shows the forecast trend of household electricity consumption in major cities in Central China. The forecast results indicate that urban household electricity consumption in the central region is expected to exhibit a continuous growth trend during 2019–2023. The forecast results show that in major coal

TABLE 3 | Forecast trends in urban household electricity consumption in the Eastern Region (billion kWh).

Year	Beijing	Tianjin	Shijiazhuang	Shenyang	Shanghai	Nanjing	Hangzhou	Xiamen	Qingdao	Shenzhen	Fuzhou	Ningbo
2018	89.41	56.57	28.45	47.87	120.00	51.31	23.88	92.42	19.89	11.96	6.26	19.70
2019	89.76	56.87	31.77	51.77	121.98	53.14	24.58	93.77	21.93	16.14	7.56	21.61
2020	91.52	57.89	32.45	53.86	123.84	54.19	26.87	94.18	22.78	22.98	8.65	22.68
2021	93.68	62.06	33.94	55.21	125.79	55.52	29.08	94.98	23.53	23.86	10.91	23.50
2022	95.79	63.33	35.61	57.09	127.36	57.90	32.55	95.39	24.76	25.15	11.06	24.78
2023	98.99	65.71	36.97	59.32	132.25	61.99	36.35	96.72	25.89	26.34	12.43	25.37

TABLE 4 | Forecast trends in urban household electricity consumption in the Central Region (in billion kWh).

	Taiyuan	Hohhot	Changchun	Harbin	Hefei	Nanchang	Zhengzhou	Wuhan	Changsha
2018	180.00	82.56	210.00	37.20	78.48	100.00	50.49	200.00	16.94
2019	183.46	83.63	212.01	44.41	77.78	105.61	52.23	201.23	18.30
2020	185.55	84.34	213.16	47.98	79.02	111.42	53.76	204.78	22.22
2021	188.98	85.23	215.16	51.74	80.32	113.36	55.87	211.65	24.17
2022	192.23	87.18	217.84	53.46	81.87	117.01	57.89	217.98	26.07
2023	195.52	90.22	220.59	56.66	83.56	123.67	61.13	222.31	28.07

TABLE 5 | Forecast trends in urban household electricity consumption in the Western Region (in billion kWh).

	Urumqi	Guiyang	Kunming	Xining	Lanzhou	Yinchuan	Xi'an	Chengdu
2018	34.76	11.64	24.34	28.12	24.03	55.77	68.07	65.93
2019	35.88	12.78	26.61	29.99	25.69	56.65	70.54	67.08
2020	37.32	14.67	28.15	31.33	26.99	57.77	73.44	69.55
2021	41.67	16.87	30.68	33.08	28.18	59.32	76.15	71.98
2022	42.98	18.36	32.33	35.01	30.39	60.46	77.28	74.44
2023	43.84	21.25	35.08	37.18	33.02	61.64	80.76	76.12

producing provinces, such as Shanxi, urban household electricity consumption shows a relatively large expected growth. Central cities, such as Wuhan, also show very strong growth in urban household electricity consumption. From a regional perspective, cities in the northeastern provinces, such as Changchun and Harbin, show greater potential for growth in urban household electricity consumption.

Table 5 shows the forecast trends in household electricity consumption levels for the major cities in the western region of China. Western cities' household electricity consumption will show a large fluctuating trend between 2019 and 2023. Chengdu, as one of the major representatives of cities in the western region, maintains a continuous growth momentum. While urban cities such as Lanzhou, Yinchuan and Xi'an have expected consumption fluctuations in urban household electricity consumption from 2019 to 2023. While cities in remote areas, such as Urumqi will continue to increase the level of urban household electricity consumption, maintaining a clear trend of growing household electricity consumption.

2) Evaluation of urban household gas consumption forecasts

This paper uses the DT-SVR non-linear technique for forecasting urban household energy natural gas. In order to assess the validity of the DT-SVR forecasting technique chosen in this paper, the results are evaluated by comparing the accuracy evaluation criteria of linear and non-linear forecasting methods. For the forecasting of urban household gas consumption levels, the paper also compares the model evaluation values of the above three forecasting techniques. The analysis shows that both at the national and sub-regional levels, the decision tree non-linear forecasting technique is more accurate than the non-linear analysis technique, in terms of the level of values taken for the three types of indicators. The mean absolute error (MAE) and root mean square error (RMSE) values for non-linear forecasting are smaller than the corresponding evaluation indicator values for the decision tree forecasting technique (Rasheed, 2021). Comparing the linear forecasting technique with the support vector machine analysis technique similarly expresses the high accuracy of the non-linear forecasting technique in the forecasting process of urban household gas consumption levels. Further, this paper compares the directional symmetry indicators of urban household natural gas consumption level forecasting and finds that none of the DS values of the non-linear forecasting techniques are smaller than those of the linear forecasting techniques, illustrating the good results of this paper's DS-SVR non-linear forecasting

TABLE 6 | Comparison of prediction results between linear and non-linear prediction methods.

Region	Linear model (LR)			Decision trees (GBDT)			Support vector machines (SVR)		
	MAE	RMSE	DS	MAE	RMSE	DS	MAE	RMSE	DS
National Gas	20,417	1.09E+09	0.600	20,186	1.7E+09	0.760	18,762	1.1E+09	0.760
Eastern Gas	10,806	1.69E+08	0.830	12,817	2.9E+08	0.830	9,162	1.28E+08	1.000
Central Gas	9,652	1.1E+08	0.750	8,933	1E+08	0.750	8,486	97922570	0.750
Western Gas	17,990	9.93E+08	0.850	15,231	1.9E+09	0.920	16,350	1.19E+09	1.000

TABLE 7 | Projected trends in urban household gas consumption in the Eastern Region.

Year	Beijing	Tianjin	Shijiazhuang	Shenyang	Shanghai	Nanjing	Hangzhou	Xiamen	Qingdao	Shenzhen	Fuzhou	Ningbo
2018	17.16	0.90	1.22	3.94	3.60	2.39	0.88	0.88	0.81	7.83	2.04	1.71
2019	18.91	1.03	1.99	4.66	3.88	3.98	1.08	0.97	1.21	8.12	2.32	1.88
2020	20.72	1.36	2.07	5.35	4.07	4.56	1.46	1.23	1.32	9.03	2.63	2.21
2021	22.24	1.69	2.34	6.13	4.33	4.79	1.99	1.45	1.45	10.17	2.89	2.42
2022	24.81	2.09	3.01	7.82	4.63	5.35	2.13	1.89	1.56	11.42	3.31	2.67
2023	25.97	2.56	3.44	8.74	4.98	6.02	2.45	2.02	2.12	12.18	3.47	3.01

TABLE 8 | Projected trends in urban household gas consumption in the Central Region.

	Taiyuan	Zhengzhou	Wuhan	Changsha	Zhengzhou	Nanchang
2018	2.97	2.35	2.80	1.08	1.06	1.47
2019	3.15	2.93	3.11	1.37	1.48	1.68
2020	3.46	3.75	3.47	1.62	1.93	2.43
2021	4.04	4.27	4.16	2.16	2.38	2.99
2022	4.47	4.99	4.64	2.55	2.97	3.32
2023	5.51	5.71	5.18	3.03	3.38	3.99

model in urban household natural gas consumption level forecasting (as shown in **Table 6**).

4) Trends in non-linear projections of household gas consumption

Table 7 shows the projected trends in the amount of change in household natural gas consumption in the major cities in the eastern region of China. Household natural gas consumption in the major cities all show an increasing trend of change. Natural gas, as the main source of energy consumption in major urban households, shows an almost consistent trend of growth in consumption in all large cities. In particular, cities such as Background, Tianjin and Qingdao. Relatively speaking, the total level of household natural gas consumption in the sub-developed cities on the eastern seaboard was largely flat, but still maintained its future growth trend.

Table 8 shows the projected trends in the level of household natural gas consumption in the central region of China. Looking at the forecast levels for the major cities, the incremental increases in household natural gas consumption in central cities are modest, but all have a trend of year-on-year growth. In comparison, cities along the Yangtze River, such as Wuhan and Changsha, have higher levels of total annual household gas consumption than most other cities. Taiyuan, as one of the major representatives of central cities, has a more pronounced growth trend in urban household gas consumption.

TABLE 9 | Projected trends in urban household gas consumption in the Western Region.

	Xi'an	Chengdu	Lanzhou	Xining	Yinchuan
2018	2.43	4.84	2.60	1.81	3.39
2019	2.86	5.35	2.95	2.22	3.88
2020	3.46	6.03	3.16	2.67	4.21
2021	3.97	6.88	3.42	3.02	4.53
2022	4.24	7.32	3.78	3.64	4.89
2023	4.87	7.87	4.31	4.17	5.03

Table 9 shows the projected trends in household natural gas consumption levels in the western region of China. As representatives of the cities in the western region, Xi'an and Chengdu cities have continued to increase their household energy consumption changes. In contrast, cities in remote areas such as Lanzhou and Xining cities show a smaller increase in household natural gas consumption. Yinchuan, on the other hand, is expected to show a more significant change in natural gas growth trends.

In summary, projections of natural gas consumption levels in urban households in China show significant regional variation.

There are regional differences not only in the annual aggregate characteristics of the change in natural gas consumption growth, but also in the comparative magnitude of growth. Projections of household gas consumption are generally higher in developed coastal cities than in less developed inland cities; and the growth trend in household gas consumption is generally higher in cities with faster economic growth inland than in cities with less developed economic development. These characteristics will play a positive role in the process of formulating and evaluating household energy consumption policies, and will help to objectively assess the effectiveness and regional synergistic effects of household energy policies.

DISCUSSION OF THE PREDICTED RESULTS OF URBAN HOUSEHOLD ENERGY CONSUMPTION

Discussion of Forecast Results for Household Electricity Consumption

Electricity consumption is the main type of energy consumed by urban households in China and has a diversity of uses. The level of urbanisation development becomes an important factor influencing urban household electricity consumption, both in non-linear and linear projections (Baltruszewicz et al., 2021). From an analytical perspective of household energy policy efficiency, urbanisation development is an important economic category in China, with two points of concern: firstly, the path of urbanisation development in China, whether it manifests itself in the form of population migration as urbanisation progresses, or in the form of local urbanisation. Secondly, the availability of household energy in the development of urbanisation. In terms of the first point, we argue that the development of urbanisation changes the quality and level of accessibility of household energy consumption. In particular, for the period of transition from rural to urbanisation, urbanisation has on the one hand reduced the structure of energy supply for rural households after urbanisation and on the other hand increased the cost of energy access for urban households (Shen et al., 2020). Electricity, as a clean, modern energy source, has two important apparent outcomes in the urbanisation process. One is the total growth of electricity consumption due to the influence of energy consumption rigidity; the other is the total growth of electricity consumption due to the substitution between energy consumer goods. This is particularly true for the movement of people in the development of urbanisation (rural-urban migration due to urbanisation).

The level of per capita income remains a major constraint on household energy consumption in central and western cities. In recent years, with China's Belt and Road Economic Belt initiative and a series of policies to vigorously develop content city clusters, public income levels have gradually increased under the operating mechanism of urbanisation development pushing back the economy (Mrówczyńska et al., 2020). In rural areas in particular, the phenomenon of energy poverty has gradually improved in individual areas. But the potential for the role of per capita income levels on household energy consumption is stronger in the central and western range. On the one hand, income factors remain the main constraint on the ability to spend on consumption in the traditionally less developed regions of the Midwest. On the other hand, the gradual increase in income levels has been accompanied by a gradual release of the potential for household energy consumption. In this sense, the potential for urban household energy consumption, particularly electricity consumption, in the central and western regions will continue to grow in the short term as income levels increase.

Discussion of Forecast Results for Household Gas Consumption

Natural gas consumption is one of the key categories of energy consumption in urban households (Wang et al., 2021). In this paper, both linear and non-linear forecasting techniques are applied to urban household gas consumption. The results show that the non-linear forecasts more accurately reflect the changes and trends in the composition of urban household natural gas consumption, but the important factors affecting urban household natural gas consumption vary slightly across regions. Overall, the main factor influencing urban household gas consumption is income level (Gassar et al., 2019). However, in China's urbanisation process, there are important structural factors in household gas consumption, namely the composition of the main consumers of gas in urban households. This is inextricably linked to the way in which China's urbanisation is progressing. As mentioned earlier, the transfer of rural to urban populations will contribute to the growth of urban household gas consumption in terms of aggregate uplift. The urbanisation of rural areas *in situ* will contribute to the change in the total volume of household gas consumption from the perspective of the structural composition of the consumer group. The process of urbanisation has, in some cases, reduced the ease and diversity of access to energy in rural areas and increased the consumption of key energy sources. This is also influenced to some extent by supply policies accordingly.

In terms of the impact of rising incomes on urban household gas consumption, the impact of disposable income on changes in urban household gas consumption within the Midwest is more pronounced (Jürisoo et al., 2019). From the indication of the forecast results, the sales price of natural gas is an important factor influencing natural gas consumption. On the one hand, urban residents are less price sensitive to household natural gas consumption due to the rigid nature of natural gas consumption in the context of urban life; on the other hand, the potential of income to stimulate consumption capacity is under-stimulated in the urban household consumption scenario. That is, with the further rise in income levels of urban residents in less developed regions and the further acceleration of urbanisation, the growth in total urban household natural gas consumption will increasingly manifest itself in the continued growth of total urban household natural gas consumption. At the same time, while taking into account economic factors, social factors, climate factors are also key elements affecting natural gas consumption. Natural gas consumption is a modern clean energy source and an important source of supporting urban household energy consumption in the urbanisation process. In the future, as the urbanisation process accelerates and develops, the total change in urban household natural gas consumption will continue to maintain a continuous growth process.

CONCLUSION

In this paper, we apply linear and non-linear forecasting techniques to forecast and analyse the trends in total urban household electricity and natural gas consumption and the factors affecting them. In the implementation of the forecasting analysis, we compare the forecasting accuracy of household electricity and natural gas consumption in the context of linear and non-linear analyses respectively. Also, the important factors influencing urban household electricity and natural gas consumption are examined and ranked separately in the context of the two forecasting techniques. Again on the basis of this the research implications and value of the forecasting results are analysed through discussion.

The findings show that non-linear forecasting techniques are highly effective in accurately portraying changes in urban household electricity consumption and changes in total natural gas consumption. When looking at the factors influencing urban household electricity consumption and natural gas consumption from four scoping perspectives - nationwide, eastern, central and western the degree to which the main influencing factors play a role varies and exhibits significant regional differences.

In general, the important influencing factor on household energy consumption in the eastern region is mainly manifested in the level of urbanisation development, the influencing factor on household energy consumption in the central region is mainly influenced by factors such as industrial structure, and the change in total urban household energy consumption in the western region is more influenced by natural conditions and income levels. From the traditional sense of the degree of economic development, within the less developed economic development regions, income level is still the main factor limiting the change in urban household energy consumption, income level on household energy consumption has not been stimulated in the process of economic development in these less developed regions, is not yet fully released, will be further manifested in the promotion of household energy consumption on the role. Urbanisation as an important factor in examining household energy consumption, its different development patterns and processes will gradually be reflected in the choice of urban household energy consumption and changes in total consumption and other scenarios. This is also an important consideration in the development of household energy policies.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

Conceptualization, LZ and XW; methodology, LZ; software, LZ; validation, LZ and XW; formal analysis, LZ; investigation, LZ; resources, XW; data curation, LZ; writing "original draft preparation, LZ; writing" review and editing, XW; visualization, LZ; funding acquisition, XW. All authors have read and agreed to the published version of the manuscript.

FUNDING

The paper is supported by the National Natural Science Foundation of China (No. 71874133).

ACKNOWLEDGMENTS

The authors would like to thank the referees as well as the editors.

REFERENCES

Baltruszewicz, M., Steinberger, J. K., Owen, A., Brand-Correa, L. I., and Paavola, J. (2021). Final Energy Footprints in Zambia: Investigating Links between Household Consumption, Collective Provision, and Well-Being. *Energ. Res. Soc. Sci.* 73 (6), 101960. doi:10.1016/j.erss.2021.101960

Ding, Y. X., and Peng, S. (2020). Study on the Spatial Distribution and Influencing Factors of Household Energy Consumption in China[J]. *Resource Development Market* 36 (04), 366–370. doi:10.3969/j.issn.1005-8141.2020.04.006

Dougherty, W. W. (1993). Statement and Process: Designing 'good' Arguments about the Rural Energy Problem in Developing Countries. *Environ. Plann. B* 20 (4), 379–390. doi:10.1068/b200379

Fan, J.-L., Zhang, Y.-J., and Wang, B. (2017). The Impact of Urbanization on Residential Energy Consumption in China: An Aggregated and Disaggregated Analysis. *Renew. Sustainable Energ. Rev.* 75 (10), 220–233. doi:10.1016/j.rser.2016.10.066

Gassar, A., Yun, G. Y., and Kim, S. (2019). Data-driven Approach to Prediction of Residential Energy Consumption at Urban Scales in London[J]. *Energy* 187 (Nov.15), 115973.1–115973.13. doi:10.1016/j.energy.2019.115973

Han, C., Huang, Y., Shen, H., Chen, Y., Ru, M., Chen, Y., et al. (2016). Modeling Temporal Variations in Global Residential Energy Consumption and Pollutant Emissions[J]. *Appl. Energ.* 184, 820–829. doi:10.1016/j.apenergy.2015.10.185

Jingchao, Z., and Kotani, K. (2012). The Determinants of Household Energy Demand in Rural Beijing: Can Environmentally Friendly Technologies Be Effective? *Energy Econ.* 34 (2), 381–388.

Jürisoo, M., Serenje, N., Mwila, F., Lambe, F., and Osborne, M. (2019). Old habits die hard: Using the energy cultures framework to understand drivers of household-level energy transitions in urban Zambia. *Energ. Res. Soc. Sci.* 53, 59–67. doi:10.1016/j.erss.2019.03.001

Kupka, J., Jirava, P., and Kaparová, M. (2010). Quality of Life Modelling Based on Decision Trees[J]. *E a M: Ekonomie a Management* 13 (3), 130–146.

Lenzen, M., Wier, M., Cohen, C., Hayami, H., Pachauri, S., Schaeffer, R. A., et al. (2006). A Comparative Multivariate Analysis of Household Energy Requirements in Australia, Brazil, Denmark, India and Japan[J]. *Energy* 31 (2/3), 181–207. doi:10.1016/j.energy.2005.01.009

Li, G., Ning, B., Abdu, M. A., Otsuka, Y., Yokoyama, T., Yamamoto, M., and Liu, L. (2013). Longitudinal Characteristics of Spread F Backscatter Plumes Observed

With the EAR and Sanya VHF Radar in Southeast Asia. *J. Geophys. Res.: Space Phys.* 118 (10), 6544–6557.

Lu, H., and Lu, L. (2006). An Empirical Analysis of the Impact of Farmers' Income Level on the Energy Consumption Structure of Rural Households [J]. *Finance Trade Res.* 2006 (03), 28–34. doi:10.19337/j.cnki.34-1093/f.2006.03.005

Mrówczyńska, M., Skiba, M., Bazan-Krzywoszańska, A., and Sztubecka, M. (2020). Household Standards and Socio-Economic Aspects as a Factor Determining Energy Consumption in the City[J]. *Appl. Energ.* 264, 114680. doi:10.1016/j.apenergy.2020.114680

Qi, L., Fengde, W., Jindong, L., and Wensheng, X. (2021). A Hybrid Support Vector Regression with Multi-Domain Features for Low-Velocity Impact Localization on Composite Plate Structure[J]. *Mech. Syst. Signal Process.* 154, 107547. doi:10.1016/j.ymssp.2020.107547

Rasheed, A. A. (2021). Improving Prediction Efficiency by Revolutionary Machine Learning Models[J]. *Mater. Today Proc.* 2021 (1). doi:10.1016/j.matpr.2021.04.014

Saunders, H. (2013). Is what We Think of as "rebound" Really Just Income Effects in Disguise? *Energy Policy* 57 (jun), 308–317. doi:10.1016/j.enpol.2013.01.056

Shen, M., Lu, Y., Wei, K. H., and Cui, Q. (2020). Prediction of Household Electricity Consumption and Effectiveness of Concerted Intervention Strategies Based on Occupant Behaviour and Personality Traits[J]. *Renew. Sustainable Energ. Rev.* 127, 109839. doi:10.1016/j.rser.2020.109839

Tonooka, Y., Liu, J., Kondou, Y., Ning, Y., and Fukasawa, O. (2006). A Survey on Energy Consumption in Rural Households in the Fringes of Xian City. *J. Energy Build.* 38 (11), 1335–1342.

Wang, F., Zhou, Z., Dai, Z., Gong, X., Yu, G., Liu, H., and Yu, Z. l. (2007). Development and Demonstration Plant Operation of an Opposed Multi-Burner Coal-Water Slurry Gasification Technology. *J. Front. Energy Power Eng. China* 1 (3), 251–258.

Wang, S., Liu, X., Jiang, S., and Zhan, Y. (2018). Reducing Energy Bill of Data center via Flexible Partial Execution. *J. Ambient Intell. Hum. Comput.* doi:10.1007/s12652-018-1157-9

Wang, S., Sun, S., Zhao, E., and Wang, S. (2021). Urban and Rural Differences with Regional Assessment of Household Energy Consumption in China. *Energy* 232, 121091. doi:10.1016/j.energy.2021.121091

Wang, S., Xu, Z., and Ha, J. (2022). Secure and Decentralized Framework for Energy Management of Hybrid AC/DC Microgrids Using Blockchain for Randomized Data. *Sustainable Cities Soc.* 76, 103419. doi:10.1016/j.scs.2021.103419

Xu, K., Chen, C., Liu, H., Tian, Y., Li, X., and Yao, H. (2014). Effect of Coal Based Pyrolysis Gases on the Performance of Solid Oxide Direct Carbon Fuel Cells. *Internat. J. Hydrogen Energy* 39 (31), 17845–17851.

Yuan, B., Ren, S., and Chen, X. (2015). The Effects of Urbanization, Consumption Ratio and Consumption Structure on Residential Indirect CO2 Emissions in China: A Regional Comparative Analysis. *Appl. Energ.* 140 (feb.15), 94–106. doi:10.1016/j.apenergy.2014.11.047

Zhang, R., Wei, T., Glomsrød, S., and Shi, Q. (2014). Bioenergy Consumption in Rural China: Evidence From a Survey in Three Provinces. *J. Energy Policy* 75, 136–145.

Zhang, C., and Yang, J. (2019). Economic Benefits Assessments of "Coal-to-Electricity" Project in Rural Residents Heating Based on Life Cycle Cost. *J. Cleaner Prod.* 213, 217–224.

Zhang, K., Zhu, D., Li, J., Gao, X., and Lu, J. (2020). Learning Stacking Regression for No-Reference Super-resolution Image Quality Assessment[J]. *Signal. Process.* 2020, 107771. doi:10.1016/j.sigpro.2020.107771

Zheng, X., Wei, C., Qin, P., Guo, J., Yu, Y., Song, F., et al. (2014). Characteristics of Residential Energy Consumption in China: Findings from a Household Survey. *Energy Policy* 75, 126–135. doi:10.1016/j.enpol.2014.07.016

Zhong, W., Song, J., Yang, W., Fang, K., and Liu, X. (2020). Evolving Household Consumption-Driven Industrial Energy Consumption under Urbanization: A Dynamic Input-Output Analysis[J]. *J. Clean. Prod.* 289 (7), 125732. doi:10.1016/j.jclepro.2020.125732

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Energy Research

ORIGINAL RESEARCH
published: 20 December 2021
doi: 10.3389/fenrg.2021.764635

Ensemble Forecasting Frame Based on Deep Learning and Multi-Objective Optimization for Planning Solar Energy Management: A Case Study

Yongjiu Liu[1], *Li Li*[1]* *and Shenglin Zhou*[2]

[1]*School of Statistics, Shandong Technology and Business University, Yantai, China,* [2]*School of Political Science and Public Administration, Shandong University, Qingdao, China*

There are many prediction models that have been adopted to predict uncertain and non-linear photovoltaic power time series. Nonetheless, most models neglected the validity of data preprocessing and ensemble learning strategies, which leads to low forecasting precision and low stability of photovoltaic power. To effectively enhance photovoltaic power forecasting accuracy and stability, an ensemble forecasting frame based on the data pretreatment technology, multi-objective optimization algorithm, statistical method, and deep learning methods is developed. The proposed forecasting frame successfully integrates the advantages of multiple algorithms and validly depict the linear and nonlinear characteristic of photovoltaic power time series, which is conductive to achieving accurate and stable photovoltaic power forecasting results. Three datasets of 15-min photovoltaic power output data obtained from different time periods in Belgium were employed to verify the validity of the proposed system. The simulation results prove that the proposed forecasting frame positively surpasses all comparative hybrid models, ensemble models, and classical models in terms of prediction accuracy and stabilization. For one-, two-, and three-step predictions, the MAPE values obtained from the proposed frame were less than 2, 3, and 5%, respectively. Discussion results also verify that the proposed forecasting frame is obviously different from other comparative models, and is more stable and high-efficiency. Thus, the proposed frame is highly serviceable in elevating photovoltaic power forecasting performance and can be used as an efficient instrument for intelligent grid programming.

Keywords: artificial intelligence, ensemble forecasting system, photovoltaic power forecasting, renewable energy management, smart grid management

OPEN ACCESS

Edited by:
Lean Yu,
Beijing University of Chemical Technology, China

Reviewed by:
Narottam Das,
Central Queensland University, Australia
Akin Tascikaraoglu,
Muğla University, Turkey

***Correspondence:**
Li Li
lili_sdtbu@163.com

Specialty section:
This article was submitted to Smart Grids, a section of the journal Frontiers in Energy Research

Received: *25 August 2021*
Accepted: *17 November 2021*
Published: *20 December 2021*

Citation:
Liu Y, Li L and Zhou S (2021) Ensemble Forecasting Frame Based on Deep Learning and Multi-Objective Optimization for Planning Solar Energy Management: A Case Study. Front. Energy Res. 9:764635. doi: 10.3389/fenrg.2021.764635

1 INTRODUCTION

The exhaustion of fossil energy and global warming have been inescapable events for humans (Das et al., 2015; Takilalte et al., 2019; Irfan et al., 2021). To work out these events, exploring and exploiting renewable energy worldwide should be the ultimate focus of attention (Islam, 2017; Shezan et al., 2017; Liu et al., 2020; Elavarasan et al., 2021). Photovoltaic (PV) power, which is unlimited, green, and available, has become a key point in new energy resource research (Jithin and Roykumar, 2018; Shelat et al., 2019; Zhu and Pi, 2020; Tan et al., 2021). The International Energy Agency announced that, until 2019, the global accumulative installed capacity of PV power exceeded 627 GW.[1] Nevertheless, PV power is labile and fluctuates at high frequency, which unpredictably impacts the facility wastage and grid stability of the intelligent electric system. Therefore, enhancing the forecasting accuracy and stability of PV power must be considered to help solve the aforementioned tasks and optimize the intelligent electric system operation.

By reviewing past studies, we can see that several forecasting models have been proposed and developed to enhance prediction precision and effectiveness (Yildiz and Acikgoz, 2021). With respect to the calculative mechanism, the forecasting model can be summarized as the following rough categories (Abdel-Nasser and Mahmoud, 2019; Liu et al., 2022): physical, statistical, and artificial intelligent models. Physical models rely on sky cameras and satellite data to forecast PV power (Dong et al., 2020). PV power prediction with satellite imaging or sky cameras has been developed as a key theme based on data capture and cloud movement (Elsinga and van Sark, 2017). Physical methods exhibit satisfactory performance when the state of the weather is stabilized (Li et al., 2020). In contrast to physical models, statistical models use sufficient actual data to conduct short-term PV power forecasting, and these models with regard to short-term forecasting surpass physical models in terms of performance (Zhang et al., 2019). The prediction performance of statistical models is impacted when the input variables have a nonlinear relationship. Autoregressive moving average (ARMA) (David et al., 2016), autoregressive integrated moving average (ARIMA) (Pedro and Coimbra, 2012), Kalman filter (Soubdhan et al., 2016), and other statistical models have been adopted and gained significant prediction results. In addition, artificial intelligent models, which incorporate artificial neural networks (ANNs) (Yacef et al., 2014), fuzzy logic methods (Tanaka et al., 2011), and deep learning methods (DLMs) (Jiang et al., 2020), are widely adopted tools for short-term PV power forecasting (Yagli et al., 2019; Devaraj et al., 2021). Based on their outstanding capabilities, DLMs can deal with the fuzzy relationship between the actual data and forecasting data. As a booming branch of artificial intelligence methods, DLM has attracted wide attention in numerous fields (Zhou et al., 2020). Compared with the two models, DLMs depend on historical data and have high fault tolerance, which means that DLMs can robustly and adaptively predict PV power. In addition, DLMs can dispose of nonlinear data, conduct adaptive forecasting, and judge fuzzy relationships (Li, 2020). Nonetheless, DLMs have instinctive shortcomings, including over-fitting, easy to local optimum, and low convergence speed (Jiang and Liu, 2019). Apart from the abovementioned forecasting models, hybrid models have also received great attention. Hybrid approaches can overcome the limitation of individual model by combining predictor with other algorithms (Kushwaha and Pindoriya, 2019). For example, Qu et al. (2021b) established a hybrid gated recurrent unit (GRU) to forecast day-ahead PV generation and proved hybrid GRU is superior to individual GRU in terms of forecasting accuracy. Korkmaz (2021) used variational mode decomposition approach and convolutional neural network (CNN) to improve PV power forecasting ability. Relative to benchmark deep learning models, the proposed hybrid model can provide better forecasting results. Eseye et al. (2018) developed a novel hybrid short-term forecasting method, which integrated wavelet transform (WT), particle swarm optimization (PSO) with support vector machine (SVM) to enhance PV power forecasting precision. By comparing with various prediction approaches, the proposed model showed excellent prediction performance, which is helpful to integrate PV into power grid. However, forecasting performance of a definite forecasting model is different with respect to different datasets and observation sites. Thus, one forecasting approach cannot be applied to all forecasting situations.

The drawbacks of the aforementioned methods can be concluded as follows:

(1) Physical models cannot obtain satisfactory results pertaining to short-term PV power prediction based on several disadvantages: running efficiency is lower, consumed computing resources are expensive, and forecasting results are unsatisfactory. Hence, physical models cannot offer a satisfactory service for short-term PV power forecasting.
(2) Statistical models are poor in predicting data with high fluctuation and nonlinear characteristics. It cannot effectively forecast PV power based on the linear hypothesis (Niu and Wang, 2019).
(3) Compared with the aforementioned models, the artificial intelligence model, such as DLMs, can detect the nonlinear relationship between the historical and forecasted values. It has attracted several researchers over the past several years for the validity to forecast complicated relationships (Feng et al., 2017). Nonetheless, DLMs have instinctive shortcomings, such as over-fitting, easy to local optimum, and low convergence speed (Iversen et al., 2016).
(4) Because of the instinctive drawbacks of each model, the individual model cannot forecast time-series data that vary under the changing environment, resulting in poor forecasting performance in some situations.

To overcome the above disadvantages, the ensemble learning strategy based on multiple forecasting models that proposed by Bates and Granger (1969) has been widely used by researchers. Ensemble strategy employs multiple forecasting models to achieve an aggregated result that is superior to every base forecasting model (Opitz and Maclin, 1999). The main principle of this strategy is to obtain optimal weights to ensure the minimum sum of squared errors of the training set (Hao and Tian, 2019). By combing multiple predictors, we can better utilize more useful information and remove particular deviations brought by individual predictor. Moreover, the ensemble strategy can successfully integrate the merit of all involved sub-predictors, such as their good ability to grasp different data characteristic and the good property to overcome

[1] https://news.solarbe.com/202004/29/324368.html.

TABLE 1 | Relevant studies.

Literature	Methods of construction	Year
Literature 1 (Yin et al., 2020)	Extreme learning machine, non-iterative correction theory, seasonal model	2020
Literature 2 (Niu et al., 2020)	Random forest feature selection, complete ensemble empirical mode decomposition, backpropagation, particle swarm optimization	2020
Literature 3 (Zhang et al., 2020a)	Dendritic neural network, wavelet transform algorithm	2020
Literature 4 (Li et al., 2020)	Wavelet packet decomposition, LSTM	2020
Literature 5 (Agga et al., 2021)	CNN, LSTM, ConvLSTM	2021
Literature 6 (Mellit et al., 2021)	LSTM, Bidirectional LSTM, GRU, Bidirectional GRU, CNN, CNN-LSTM, CNN-GRU	2021
Literature 7 (Luo et al., 2021)	Pearson correlation coefficient, LSTM, physical constraints	2021
Literature 8 (Zhen et al., 2021)	Genetic algorithm, Bidirectional LSTM	2021
Literature 9 (Qu et al., 2021a)	CNN, LSTM, CNN-LSTM	2021
etc.		

negative effect (e.g., overfitting), which is proved to be effective to improve forecasting performance in many forecasting fields (Xiao et al., 2015; Liu et al., 2019; Wang et al., 2021). Yang and Dong (2018) proposed a seasonal time series ensemble model that used six component models from different families and 8 ensemble methods to conduct PV power output forecasting. A simple remedy was added to the ensemble model, which was proved to be effective to improve forecasting ability. Li et al. (2020) used wavelet packet decomposition (WPD) to decompose PV power data and used long short-term memory (LSTM) to forecast the decomposed series. The predicted sub-series are ultimately integrated based on linear weighting strategy to obtain the final forecasting values. Simulation results verified its high-quality forecasting ability. Sharma et al. (2021) proposed a novel forecasting frame, where the maximal overlap discrete wavelet transform technique was used for decomposition and the LSTM was used for sub-series forecasting. By integrating the sub-series forecasting results, the final PV power forecasting results were finally obtained. More studies about ensemble forecasting models are listed in **Table 1**. From the above review, we can find that most existing ensemble models are more likely to use a single forecasting model. However, PV power series is fluctuant and uncertain with both intricate linear and nonlinear characteristics, which must be captured by different class of forecasting models. To this end, in this paper, both statistical model and DLMs are combined together to better grasp the linear and nonlinear characteristics of PV power series. Moreover, some existing ensemble models use linear weighting method to calculate the final ensemble forecasting results. Considering linear weighting method may not reflect the importance of the prediction results of each component, a multi-objective optimization algorithm (MOOA) is used to optimize the combining weights, which can effectively improve PV power forecasting performance. Besides, data preprocessing is an important process in PV power forecasting because it can filter the high-frequency noise in original time series and retain the useful information. Nevertheless, most studies may ignore the importance of data preprocessing or adopt poor preprocessing methods. In this paper, an effective data preprocessing method, namely singular spectrum analysis (SSA), is used to preprocess the historical PV power output forecasting, which can better grasp the data characteristic of PV power series and effectively improve forecasting ability.

In our study, proposed ensemble forecasting frame (PEFF) is built, which incorporates SSA, multi-objective grasshopper algorithm (MOGOA), ARIMA, and DLMs. Specifically, SSA was selected to eliminate irregular fluctuations of observed values in a complex environment. SSA can effectively process the original time series to enhance the forecasting performance. ARIMA and three DLMs (i.e., deep belief network (DBN), GRU, LSTM) were adopted to conduct PV power forecasting, and the ensemble coefficient of each model was obtained using MOGOA. ARIMA can effectively predict the linear trend of PV power generation, whereas DLMs can effectively predict nonlinear trends. The PEFF fills the gap between the statistical and artificial intelligence models. MOGOA can effectively combine forecasting results based on an effective style. The PEFF that integrates the benefits of individual models with data pretreatment techniques and intelligent optimization algorithms can validly improve the PV power prediction ability (Tian and Hao, 2018).

The leading course of our study relative to other studies in the domain of PV power forecasting is summarized below:

(1) A data pretreatment technique was adopted to relieve the random fluctuation of PV power sequences in real time. The observed PV power output time series will be disintegrated into several subseries; then, the subseries with the highest frequency fluctuation is abnegated, and the residuals are structured to conduct PV power forecasting. Considering this disposal, the essential character of PV power could be better extracted, and hence, the forecasting performance can be greatly improved.
(2) The statistical model is beneficial to grasp linear characteristics, while DLMs make for nonlinear characteristics. For the sake of comprehensive control of the linear and nonlinear characteristics of PV power, ARIMA (the statistical model) is used to forecast the linear trend, and three DLMs are used for the nonlinear trends.
(3) MOGOA, as an effective parameter optimization technology, can determine the optimal coefficient of each sub-model. MOGOA with an archive to determine approximative values

of the Pareto optimal solution can prompt prediction precision and prediction stability. MOGOA can help deal with an intricate optimization problem.

(4) The developed ensemble frame (EF) can assist in the operation and optimization of smart grids. Based on the realistic PV power data and comprehensive prediction result analyses, the PEFF is verified as an effective forecasting frame and can be applied to other forecasting fields in future.

At present, an accurate and stable forecasting system is urgently needed for renewable energy generation. However, in the current study, the developed prediction models have defects. Therefore, we propose a PEFF for PV power generation prediction to compensate for the defects of the current prediction model and provide a new scheme for PV power generation prediction.

2 METHODS

In this section, SSA and MOGOA are presented in detail, and a particular process of the PEFF is introduced.

2.1 Data Preprocessing Strategy

SSA, as an instrumental data preprocessing technique to process the observed PV power values, has been continually adopted in various fields, such as biology (Hassani and Ghodsi, 2015), physics (Krishnannair et al., 2016), climatology (Unnikrishnan and Jothiprakash, 2018), and economics (de Carvalho and Rua, 2017). The flow of the SSA is listed as follows:

Step 1. Embedding

Conversing original time series $\mathrm{X} = (x_1, x_2, \cdots x_N)$ into $\mathrm{Z} = (z_1, z_2, \cdots, z_K)$ as **Eq. 1**.

$$\mathrm{X} = (x_1, x_2, \cdots, x_N) \rightarrow \mathrm{Z} = (z_1, z_2, \cdots, z_K), \tag{1}$$

where $z_i = (x_\mathrm{i}, x_{i+1}, \cdots, x_{i+L-1})^\mathrm{T} \in R^L$, $K = N - L + 1$, $L \in [2, N]$. The consequence of this mapping is embodied as a trajectory matrix with the mathematical expression of

$$\mathbf{Z} = [\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_K] = \left(\mathbf{z}_{ij}\right)_{i,j=1}^{L,K} = \begin{pmatrix} x_1 & x_2 & \cdots & x_K \\ x_2 & x_3 & \cdots & x_{K+1} \\ \cdots & \cdots & \cdots & \cdots \\ x_L & x_{L+1} & \cdots & x_N \end{pmatrix}, \tag{2}$$

Step 2. Singular values decomposition

Given a covariance matrix ($\mathbf{S} = XX^\mathrm{T}$), this step is employed to obtain L eigenvalues $(\lambda_1, \lambda_2, \cdots, \lambda_L)$ and eigenvectors $(\mathbf{U}_1, \mathbf{U}_2, \cdots, \mathbf{U}_L)$. Suppose $t = \boldsymbol{max}(i, \textit{ such that } \lambda_\mathrm{t} > 0)$ and $\mathbf{V}_i = \mathbf{X}^\mathbf{T}\mathbf{U}_i\sqrt{\lambda_i}$ $(i = 1, 2, ..., \mathrm{t})$, then, $S = XX^\mathrm{T}$ in this step can be indicated by

$$\mathbf{Z} = \mathbf{E}_1 + \mathbf{E}_2 + \cdots + \mathbf{E}_t, \tag{3}$$

where $\mathbf{E}_i = \sqrt{\lambda_i}\mathbf{U}_i\mathbf{V}_i$ and the rank of Z_i is 1. Therefore, $\mathbf{V}_1, \mathbf{V}_2, \cdots, \mathbf{V}_\mathrm{t}$ are the principal components, and $(\sqrt{\lambda_i}, \mathbf{U}_i, \mathbf{V}_i)$ denotes the characteristic root of the trajectory matrix ($\mathbf{Z}$).

Step 3. Grouping

The interval $(\mathrm{i} = 1, 2, ..., \mathrm{t})$ is disintegrated into several components $(S_1, S_2, \cdots, S_\mathrm{m})$ without a connection between them. Suppose that $\mathbf{S} = (\mathrm{s}_1, \mathrm{s}_2, \cdots, \mathrm{s}_\mathrm{p})$, then $\mathbf{Z}_S$ is defined as $\mathbf{Z}_S = \mathbf{Z}_{s_1} + \mathbf{Z}_{s_2} + \cdots + \mathbf{Z}_{s_p}$, and $\mathbf{Z}$ can be disintegrated into $\mathbf{Z} = \mathbf{Z}_{S_1} + \mathbf{Z}_{S_2} + \cdots + \mathbf{Z}_{S_m}$.

Step 4. Diagonal averaging

In this step, the grouping result is converted into a sequence with N points. Assume that $\mathbf{Z}$ is an $L * K$ matrix, $L^* = \min(\mathrm{L}, \mathrm{K})$ and $K^* = \max(\mathrm{L}, \mathrm{K})$. If $L < K$, then $\mathrm{z}_\mathrm{ij}^* = \mathrm{z}_\mathrm{ij}$, or else, $\mathrm{z}_\mathrm{ij}^* = \mathrm{z}_\mathrm{ji}$. Finally, $\mathbf{Z}$ is turned into a sequence $(r_1, r_2, \cdots, r_N)$ based on the following formula:

$$r_k = \begin{cases} \frac{1}{k+1}\sum_{q=1}^{k+1} z_{q,k-q+1}^*, \ \mathbf{1} \le k \le L^* \\ \frac{1}{L^*}\sum_{q=1}^{L^*} Z_{q,k-q+1}^* \ L^* \le k \le K^* \\ \frac{1}{N-K+1}\sum_{q=1}^{N-K^*+1} Z_{q,k-q+1}^* \ K^* \le k \le N \end{cases}, \tag{4}$$

2.2 Intelligent Optimization Algorithm

MOGOA simulates the location of the grasshopper population, which is used to search for the optimal solution to a definite problem. Based on related articles (Mirjalili et al., 2018), the operating mechanism of the MOGOA can be summarized as follows:

The motion of each grasshopper is principally influenced by individual interactions, weight, and wind strength. In addition, X_i represents the location of the ith grasshopper, as shown in **Eq. 5**.

$$X_i = S_i + G_i + A_i, \tag{5}$$

where S_i, G_i, and A_i denote the individual interaction, weight, and wind strength of each grasshopper, respectively.

S_i can be quantized by subsequent equations:

$$S_i = \sum_{\substack{j=1 \\ j \neq i}}^{N} s\left(d_{ij}\right)\hat{d}_{ij}, \tag{6}$$

$$d_{ij} = \left|X_j - X_i\right|, \tag{7}$$

$$\hat{d}_{ij} = \left(X_j - X_i\right)/d_{ij}, \text{and} \tag{8}$$

$$s(r) = f\mathbf{e}^{-r/l} - \mathbf{e}^{-r}, \tag{9}$$

where d_{ij} denotes the space between the ith and jth grasshopper and $\hat{d}_{ij}$ denotes a normalized vector from the ith grasshopper to the jth grasshopper. $s(r)$ quantizes individual interactions based on f and l.

The weight is computed *via* **Eq. 10**:

$$G_i = -g\hat{\mathbf{e}}_g, \tag{10}$$

Here, g denotes the gravitational coefficient, and $\hat{\mathbf{e}}_g$ defines a normalized vector to the earth's core. In addition, the wind strength of each grasshopper can be calculated using **Eq. 11**.

$$A_i = u\hat{\mathbf{e}}_w, \tag{11}$$

Here, u defines a constant parameter, and $\hat{\mathbf{e}}_w$ denotes a vector normalized to wind direction. Moreover, **Eq. 5** can be expressed in detail using **Eq. 12**.

$$X_i = \sum_{\substack{j=1 \\ j \neq i}}^{N} s\left(\left|X_j - X_i\right|\right) \frac{X_j - X_i}{d_{ij}} - g\hat{\mathbf{e}}_g + u\hat{\mathbf{e}}_w, \tag{12}$$

Here, N denotes population size. Moreover, the aforementioned formulas simulate the motion of *the i*th grasshopper under hypothetical status.

The force applied by gravitation is insignificant. The wind strength is related to the orientation ($\hat{T}_d$). Therefore, X_i can be extended as follows:

$$X_i^d = \mathbf{c}\left(\sum_{\substack{j=1 \\ j \neq i}}^{N} \mathbf{c}\frac{ub_d - lb_d}{2} s\left(\left|X_j^d - X_i^d\right|\right) \frac{X_j - X_i}{d_{ij}}\right) + \hat{T}_d, \tag{13}$$

Here, ub_d and lb_d represent the upper and lower boundaries of *the d*th variable, respectively. $\hat{T}_d$ denotes the dth variable value of the optimal solution. In addition, c determined using **Eq. 14** can reduce exploration and improve exploitation such that the operation speed can be correspondingly decreased based on the iteration number.

$$\mathbf{c} = c_{max} - l\frac{c_{max} - c_{min}}{L}, \tag{14}$$

Here, c_{max} and c_{min} denote the maximum and minimum values, respectively, *and* l and L represent the present iteration and max iteration, respectively.

To conduct multi-objective optimization *via* GOA, a Pareto optimal solution is adopted to modify the solution distribution. The distance between each solution and neighboring solutions is quantized. Then, the neighboring solution number is adopted to measure the density of the Pareto optimal solutions. The probability of selecting the search objective of the archive of the current iteration is defined in **Eq. 15**.

$$\mathbf{P}_i = \frac{1}{N_i}, \tag{15}$$

Here, N_i represents the neighboring solution number of the ith solution.

2.3 Flow of the PEFF

Bates et al. proved that the effective ensemble prediction accuracy of different forecasting models far surpasses that of the individual models (Bates and Granger, 1969). 1,450 values were collected from three periods: the 1st–1160th values were selected as the training set, the 1161st–1392nd values were considered as the validation set, and the 1393rd–1450th values were selected as the testing set. In prediction process, rolling forecasting mechanism is used, and the principle of rolling forecasting is that updating the input data by discarding the old data for each loop to perform the forecasting. In our study, the input set for each loop is 5 samples $\{\boldsymbol{yy^{PV}}(t-4), \boldsymbol{yy}^{PV}(t-3), \boldsymbol{yy}^{PV}(t-2), \boldsymbol{yy}^{PV}(t-1), \boldsymbol{yy}^{PV}(t)\}, (t=$ 5, 6, ..., 1,449), and the outputs of forecasting models are $\{\hat{\boldsymbol{y}}^{PV}(t+1)\}$, $\{\hat{\boldsymbol{y}}^{PV}(t+2)\}$, and $\{\hat{\boldsymbol{y}}^{PV}(t+3)\}$ from one-step to three-step forecasting, respectively. In this study, PEFF forecasts the linear and nonlinear trends of the PV power output sequence, and the flow is listed in this subsection and exhibited in **Figure 1**.

2.3.1 Operating Mechanism 1: Data Preprocessing

SSA is adopted to conduct the real-time treatment of the initial PV power series, so that the dominating feature of the PV power sequence will be mastered, and effective forecasting will be conducted subsequently.

2.3.2 Operating Mechanism 2: Prediction of Hybrid Predictors

Based on the linear and nonlinear characteristics of the PV power sequence, ARIMA and DLMs were selected to build the PEFF. By combining SSA and these models, hybrid models were employed as sub-models to predict PV power. The PV power output values corresponding to the validation set were forecasted based on the rolling forecasting mechanism. Based on real data, hybrid models perform single-step and multi-step predictions. The linear model (SSA–ARIMA) in sub-models can predict the linear trend of the PV power sequence, and nonlinear models (SSA-DLMs) can predict the nonlinear trend.

The input vector of DLMs in the training set is as follows:

$$input\_train_{DLMs} = \begin{bmatrix} \mathbf{yy}(1th) & \mathbf{yy}(2th) & \cdots & \mathbf{yy}(5th) \\ \mathbf{yy}(2th) & \mathbf{yy}(3th) & \cdots & \mathbf{yy}(6th) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{yy}((1156-k)th) & \mathbf{yy}((1157-k)th) & \cdots & \mathbf{yy}((1160-k)th) \end{bmatrix} \tag{16}$$

$$output\_train_{DLMs} = \begin{bmatrix} \mathbf{y}((5+k)th) & \mathbf{y}((6+k)th) & \cdots \\ & \mathbf{y}(1160th) & \end{bmatrix}^{\top} \tag{17}$$

The input vector of ARIMA in the training set is as follows:

$$input\_train_{ARIMA} = \begin{bmatrix} \mathbf{yy}(1th) & \mathbf{yy}(2th) & \cdots & \mathbf{yy}(295th) \\ \mathbf{yy}(2th) & \mathbf{yy}(3th) & \cdots & \mathbf{yy}(296th) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{yy}((866-k)th) & \mathbf{yy}((867-k)th) & \cdots & \mathbf{yy}((1160-k)th) \end{bmatrix} \tag{18}$$

$$output\_train_{ARIMA} = \begin{bmatrix} \mathbf{y}((295+k)th) & \mathbf{y}((296+k)th) \\ \cdots & \mathbf{y}(1160th) \end{bmatrix}^{\top} \tag{19}$$

where k denotes the forecasting step, and $\mathbf{y}$ denotes the actual PV values, and $\mathbf{yy}$ denotes the processed PV values.

The input vector of DLMs in the validation set is as follows:

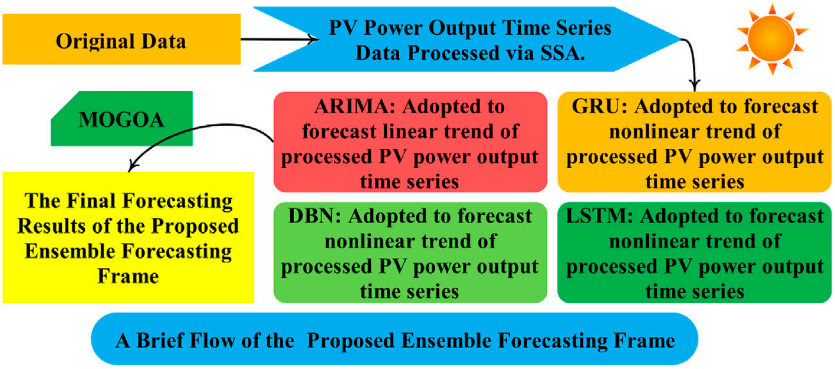

Original Data
PV Power Output Time Series Data Processed via SSA.
MOGOA
ARIMA: Adopted to forecast linear trend of processed PV power output time series
GRU: Adopted to forecast nonlinear trend of processed PV power output time series
The Final Forecasting Results of the Proposed Ensemble Forecasting Frame
DBN: Adopted to forecast nonlinear trend of processed PV power output time series
LSTM: Adopted to forecast nonlinear trend of processed PV power output time series
A Brief Flow of the Proposed Ensemble Forecasting Frame

FIGURE 1 | Flowchart of the proposed ensemble forecasting system (including data preprocessing, sub-model forecasting, and ensemble forecasting based on MOGOA).

$$input\_validation_{DLMs} = \begin{bmatrix} \mathbf{yy}(1157th) & \mathbf{yy}(1158th) & \cdots & \mathbf{yy}(1161th) \\ \mathbf{yy}(1158th) & \mathbf{yy}(1159th) & \cdots & \mathbf{yy}(1162th) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{yy}((1388-k)th) & \mathbf{yy}((1389-k)th) & \cdots & \mathbf{yy}((1392-k)th) \end{bmatrix} \tag{20}$$

$$output\_validation_{DLMs} = \begin{bmatrix} \mathbf{y}(1161th) & \mathbf{y}(1162th) & \cdots \\ & \mathbf{y}(1392th) & \end{bmatrix}^{\mathrm{T}} \tag{21}$$

The input vector of ARIMA in the validation set is as follows:

$$input\_validation_{ARIMA} = \begin{bmatrix} \mathbf{yy}((867-k)th) & \mathbf{yy}((868-k)th) & \cdots & \mathbf{yy}((1161-k)th) \\ \mathbf{yy}((868-k)th) & \mathbf{yy}((869-k)th) & \cdots & \mathbf{yy}((1162-k)th) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{yy}((1098-k)th) & \mathbf{yy}((1099-k)th) & \cdots & \mathbf{yy}((1392-k)th) \end{bmatrix} \tag{22}$$

$$output\_validation_{ARIMA} = \begin{bmatrix} \mathbf{y}(1161th) & \mathbf{y}(1162th) & \cdots \\ & \mathbf{y}(1392th) & \end{bmatrix}^{\mathrm{T}} \tag{23}$$

where k denotes the forecasting step, and $\mathbf{y}$ denotes the actual PV values, and $\mathbf{yy}$ denotes the processed PV values.

The input vector of DLMs in the testing set is as follows:

$$input\_test_{DLMs} = \begin{bmatrix} \mathbf{yy}(1389th) & \mathbf{yy}(1390th) & \cdots & \mathbf{yy}(1393th) \\ \mathbf{yy}(1390th) & \mathbf{yy}(1391th) & \cdots & \mathbf{yy}(1394th) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{yy}((1435-k)th) & \mathbf{yy}((1436-k)th) & \cdots & \mathbf{yy}((1450-k)th) \end{bmatrix} \tag{24}$$

$$output\_test_{DLMs} = \begin{bmatrix} \mathbf{y}(1393th) & \mathbf{y}(1394th) & \cdots & \mathbf{y}(1450th) \end{bmatrix}^{\mathrm{T}} \tag{25}$$

The input vector of ARIMA in the testing set is as follows:

$$input\_test_{ARIMA} = \begin{bmatrix} \mathbf{yy}((1099-k)th) & \mathbf{yy}((1100-k)th) & \cdots & \mathbf{yy}((1393-k)th) \\ \mathbf{yy}((1100-k)th) & \mathbf{yy}((1101-k)th) & \cdots & \mathbf{yy}((1394-k)th) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{yy}((1156-k)th) & \mathbf{yy}((1157-k)th) & \cdots & \mathbf{yy}((1450-k)th) \end{bmatrix} \tag{26}$$

$$output\_validation_{ARIMA} = \begin{bmatrix} \mathbf{y}(1393th) & \mathbf{y}(1394th) & \cdots \\ & \mathbf{y}(1450th) & \end{bmatrix}^{\mathrm{T}} \tag{27}$$

where k denotes the forecasting step, and $\mathbf{y}$ denotes the actual PV values, and $\mathbf{yy}$ denotes the processed PV values.

2.3.3 Operating Mechanism 3: Ensemble Forecasting

In this stage, MOGOA is applied to determine the best weight coefficient of the forecasting values of each sub-model. Based on MOGOA, prediction values matching the validation set of four prediction sub-models obtained from Process 2 are used to search for the best weight coefficient of each sub-model, and real values matching the testing set are used to test the forecasting performance of the PEFF. Finally, the final PV power prediction result is aggregated *via* the prediction values matching the testing set of each sub-model and the optimal weight coefficients corresponding to each sub-model. The objective functions of MOGOA are prediction accuracy and stability in PEFF, and its fitness function is provided:

$$min\begin{cases} Ob_1 = mean\left(abs\left(\mathbf{y}-\hat{\mathbf{y}}\right)/\mathbf{y}\right) \times 100\% \\ Ob_2 = std\left(\mathbf{y}-\hat{\mathbf{y}}\right) \end{cases} \tag{28}$$

where $\mathbf{y}$ denotes the actual PV values, and $\hat{y}$ denotes the forecasting PV values.

The fitness function can be rewritten as:

$$\begin{cases} \{\mathbf{w}\} = arg\ \underset{\{\mathbf{w}\}}{min}\begin{cases} Ob_1 = mean\left(abs\left(\mathbf{y}-\hat{\mathbf{y}}\right)/\mathbf{y}\right) \times 100\% \\ Ob_2 = std\left(\mathbf{y}-\hat{\mathbf{y}}\right) \end{cases} \\ s.t.\ -2 \le w \le 2 \\ \hat{\mathbf{y}} = sub\_\hat{y}_{ARIMA} \times \mathbf{w}_{ARIMA} + sub\_\hat{y}_{DBN} \times \mathbf{w}_{DBN} \\ \quad + sub\_\hat{y}_{GRU} \times \mathbf{w}_{GRU} + sub\_\hat{y}_{LSTM} \times \mathbf{w}_{LSTM}, \\ \mathbf{w} = \{\mathbf{w}_{ARIMA}, \mathbf{w}_{DBN}, \mathbf{w}_{GRU}, \mathbf{w}_{LSTM}\} \end{cases} \tag{29}$$

The weights are optimized to achieve good forecasting performance in validation set by MOGOA. Ultimately, the final forecasting results are calculated as $\hat{\mathbf{y}}(1393th - 1450th)$.

TABLE 2 | Four performance indicators.

Metric	Definition	Equation
MAE (Aygül et al., 2019)	Average absolute error	$\mathbf{MAE} = \sum_{i=1}^{M} \lvert \hat{\mathbf{e}}_i - \mathbf{e}_i \rvert / M$
MAPE (Zhang et al., 2020b)	Mean absolute percentage error	$\mathbf{MAPE} = (\sum_{i=1}^{M} \lvert (\mathbf{e}_i - \hat{\mathbf{e}}_i)/\mathbf{e}_i \rvert / M) \times 100\%$
RMSE (Nie et al., 2020)	Root mean square error	$\mathbf{RMSE} = \sqrt{\sum_{i=1}^{M} (\hat{\mathbf{e}}_i - \mathbf{e}_i)^2 / M}$
SDE (Liu et al., 2021)	Standard deviation of error	$\mathbf{SDE} = \sqrt{\sum_{i=1}^{M} (\mathbf{e}_i - \hat{\mathbf{e}}_i)^2 / M}$

Note: $\mathbf{e}_i$ *denotes the actual PV power output at point* i, *and* $\hat{\mathbf{e}}_i$ *denotes the forecasting PV power output at point* i. *MAE, MAPE, and RMSE are used to measure prediction accuracy, and standard deviation is used to measure prediction stability.*

2.3.4 Operating Mechanism 4: Forecasting Performance Assessment

The forecasting accuracy and stability were assessed using four indicators (see **Table 2** for details) based on three experiments, and five discussions were held to further analyze the prediction effect of PEFF.

3 EXPERIMENTAL SETUP AND RESULT ANALYSES

In this section, the experimental setup and forecasting result analyses based on three PV power datasets are presented to verify the forecasting ability of our PEFF.

3.1 Datasets

Initial PV power data were acquired from three datasets in Belgium with a time interval of 15 min. When the light intensity reaches a certain level, the PV power generation has sufficient output; therefore, this study considers PV power generation data from 9:00 to 16:00 every day as the verification dataset. Specifically, 1,450 data points for continuous 50 days from different time period were adopted as the reference dataset. The detailed data characteristics of the PEFF are shown in **Figure 2**.

There is no official or specific procedure to select the optimal training-to-test ratio. In actual application, with the improvement of training-to-test, the forecasting accuracy can be obviously improved, while too many training data may result in overfitting issue. In this paper, based on previous experiences and researches, the ratio of training, validation, and test set is set to 20:4:1. Specifically, the 1st–1160th values were selected as the training set, the 1161st-1392nd values were considered as the validation set, and the 1393th-1450th values were selected as the testing set. The relevant data characteristics are listed in **Table 3**.

3.2 Assessment Indicators of Forecasting Performance

There must be a scientific evaluation system to determine whether the prediction performance is satisfactory. In this section, four indicators, including the mean absolute error (MAE), mean absolute percent error (MAPE), root mean square error (RMSE), and standard deviation of error (SDE), are introduced to verify the forecasting effort of our PEFF. The concepts and equations of the four indicators are listed in **Table 2**.

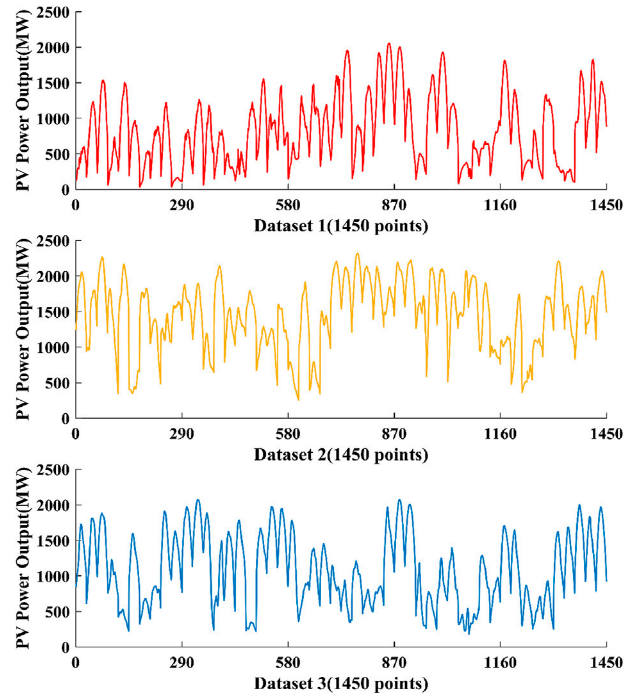


FIGURE 2 | Original PV power output time series in these studied datasets.

3.3 Experimental Setup

Based on the PV power dataset, three experiments were designed to compare the PEFF and reference models. In these experiments, Experiment I contrasted the prediction ability of the PEFF and hybrid models. Experiment II compared the PEFF with the EFs, employing different data pretreatment strategies and MOOAs in terms of forecasting effect. Experiment III compared the prediction capacity of the PEFF and classical models. The prediction ability of 1-step to 3-step prediction is testified *via* four indicators, and experimental result analyses are described.

Experiment I was conducted to verify the advantages of PEFF compared with hybrid models. The parameter setting of the SSA is the same as that of the PEFF, and the rolling number of the models was set to 5.

TABLE 3 | Relevant data characteristics of three datasets.

Datasets	Datasets	Number	Mean	Std	Min	Max	Kurtosis	Skewness
Dataset 1	Training Set	1,160	830.33	476.36	32.97	2055.73	−0.41	0.54
	Validation Set	232	768.81	506.05	101.23	1818.68	−1.19	0.39
	Testing Set	58	1,208.61	343.51	519.97	1825.68	−0.73	−0.06
	All Samples	1,450	835.62	482.98	32.97	2055.73	−0.63	0.45
Dataset 2	Training Set	1,160	1,486.61	473.34	253.56	2,320.85	−0.38	−0.56
	Validation Set	232	1,281.88	453.91	363.65	2,211.87	−0.67	−0.02
	Testing Set	58	1,648.02	272.66	1,101.78	2069.63	−1.10	−0.05
	All Samples	1,450	1,460.31	471.21	253.56	2,320.85	−0.48	−0.48
Dataset 3	Training Set	1,160	1,083.22	497.29	183.47	2073.17	−1.06	0.15
	Validation Set	232	1,067.73	473.69	252.92	2001.07	−1.21	0.07
	Testing Set	58	1,507.73	322.29	786.71	1971.44	−0.87	-0.40
	All Samples	1,450	1,097.72	494.67	183.47	2073.17	−1.09	0.09

TABLE 4 | Comparison of the prediction performance of the PEFF and hybrid models.

Datasets	Models	1-Step				2-Step				3-Step			
		MAE	MAPE	RMSE	SDE	MAE	MAPE	RMSE	SDE	MAE	MAPE	RMSE	SDE
Dataset 1	SSA-ARIMA	64.5753	6.4809	80.5735	80.6501	87.5710	6.2232	85.3681	85.3949	98.9754	7.4362	107.1025	106.0687
	SSA-DBN	20.8679	2.0754	28.8039	29.0033	58.0964	5.7688	75.5247	74.8270	81.7015	7.1736	101.2765	100.4727
	SSA-GRU	37.2952	3.7035	48.5440	39.1455	65.7595	5.6530	72.4347	43.8018	73.7984	7.3753	97.2019	96.0304
	SSA-LSTM	41.5859	3.6510	46.4083	28.2302	57.0653	5.0747	64.8244	41.1532	71.4108	6.3817	89.8253	80.8380
	PEFF	16.0007	1.7722	28.9642	28.5218	24.1139	2.5428	36.5499	36.7695	40.6881	4.3568	57.1474	57.5807
Dataset 2	SSA-ARIMA	45.1211	2.7987	49.3694	49.4504	47.5442	3.0058	65.7468	65.8141	57.7359	3.8265	86.1445	86.2185
	SSA-DBN	15.1359	1.0225	22.2644	22.3720	35.4919	2.3595	51.7319	48.7248	56.1531	3.7176	81.2890	77.9517
	SSA-GRU	27.8566	1.7620	32.2005	27.8333	35.5072	2.2519	45.2732	35.8688	58.4671	3.7642	94.6169	94.9409
	SSA-LSTM	42.0471	2.6084	45.8709	24.1338	31.1521	2.0103	40.4655	34.0953	52.4383	3.3406	68.4523	68.8433
	PEFF	13.5285	0.9495	21.3817	21.5418	18.3995	1.2623	29.2054	29.0429	31.3365	2.0810	44.2778	44.3717
Dataset 3	SSA-ARIMA	41.5486	2.9235	49.0275	49.0931	40.1157	2.9690	48.7420	48.7538	45.6637	3.5356	73.9258	73.8198
	SSA-DBN	12.8713	1.0139	19.7119	19.6991	38.4924	2.8621	55.3783	55.6886	43.8516	3.2763	64.9805	62.5742
	SSA-GRU	32.1875	2.2983	35.5315	18.6984	37.6604	2.7164	43.5752	29.5510	73.7979	4.8642	86.3725	61.1768
	SSA-LSTM	37.6564	2.6550	40.6847	19.9823	39.9714	2.8341	46.1460	32.2411	46.0131	3.2804	58.1740	51.5115
	PEFF	11.0289	0.9060	16.4936	16.6610	18.0331	1.4887	26.4301	24.4284	31.9175	2.4017	41.5250	34.7977

Experiment II was conducted to prove that the ensemble learning strategy of PEFF surpasses the EFs structured *via* other data pretreatment techniques (complete ensemble empirical mode decomposition (CEEMD)) and MOOAs (multi-objective dragonfly algorithm (MODA) and multi-objective grey wolf optimizer (MOGWO)). For each EF, the ensemble learning strategy changes, and the input and output settings remain unchanged.

Experiment III was employed to reveal the forecasting superiority of the PEFF with classical models (backpropagation (BP) neural network, extreme learning machine (ELM), Elman neural network (ENN), echo state network (ESN), least squares support vector machine (LSSVM) and radical basis function (RBF)).

3.4 Experiment I: Comparison With Hybrid Predictors

The experimental results are listed in **Table 4**. For Dataset 1, PEFF has an unrivaled characteristic in one-step and multi-step predictions. In particular, the MAPE value is 1.7722% in one-step, 2.5428% in two-step, and 4.3568% in three-step predictions, which are minimum compared with the involved models. For Dataset 2, the lowest MAE, MAPE, RMSE, and SDE were obtained from the PEFF in one step, with values of 13.5285, 0.9495%, 21.3817, and 21.5428, respectively. In multi-step prediction, the most satisfactory results are achieved by the PEFF, confirming the forecasting effect of our PEFF. For Dataset 3, the forecasting accuracy and stability of PEFF signally precede that of the reference models. This implies that although hybrid models can improve the prediction precision weakly, the PEFF is better.

Remark. The PEFF obtains a more satisfactory prediction ability with the smallest error indicator values among all of the involved models, proving the short-term prediction availability of the proposed PEFF in PV power output.

3.5 Experiment II: Comparison With EFs Adopting Diverse Ensemble Strategies

Experiment II compares the EFs with different data pretreatment techniques (CEEMD) and MOOAs (MODA and MOGWO). The forecasting results are listed in **Table 5**. For Dataset 1, the PEFF is precise and stabilized

TABLE 5 | Comparison of the forecasting performance of the PEFF and EFs employing other ensemble strategies.

Datasets	Models	1-Step				2-Step				3-Step			
		MAE	MAPE	RMSE	SDE	MAE	MAPE	RMSE	SDE	MAE	MAPE	RMSE	SDE
Dataset 1	CEEMD-MOGOA-EF	31.8597	2.4202	40.9902	40.8996	37.9407	3.8876	51.3962	49.2323	61.3274	6.2315	78.4916	73.9839
	SSA-MODA-EF	18.8108	2.0113	28.5070	28.7373	37.1732	3.5273	45.4530	40.5328	60.3734	6.0215	71.5867	63.0439
	SSA-MOGWO-EF	18.6543	1.9854	27.7369	27.9771	30.2334	3.0368	40.4392	39.0266	58.6560	5.7436	69.5111	57.5834
	PEFF	16.0007	1.7722	28.9642	28.5218	24.1139	2.5428	36.5499	36.7695	40.6881	4.3568	57.1474	57.5807
Dataset 2	CEEMD-MOGOA-EF	15.7684	1.0647	22.4003	24.5794	31.6600	2.0732	39.0680	35.1800	56.1855	3.4025	72.8122	70.6055
	SSA-MODA-EF	14.5178	1.0158	23.1657	22.5536	27.3173	1.8540	37.8721	33.7889	47.2092	3.1136	67.6010	67.5988
	SSA-MOGWO-EF	13.7641	0.9688	22.8846	22.2488	23.2763	1.5750	33.9391	33.5204	41.2373	2.7012	53.5360	44.8950
	PEFF	13.5285	0.9495	21.3817	21.5418	18.3995	1.2623	29.2054	29.0429	31.3365	2.0810	44.2778	44.3717
Dataset 3	CEEMD-MOGOA-EF	15.9487	1.2555	22.8784	23.0553	40.9837	3.1094	63.2146	63.0417	58.7871	4.8862	84.1269	78.6112
	SSA-MODA-EF	12.4153	0.9827	18.6229	18.6425	31.6632	2.3163	40.8688	26.8852	41.0942	3.0999	55.7911	47.0912
	SSA-MOGWO-EF	11.8032	0.9550	18.3029	18.4288	23.8026	1.8575	32.7021	25.9686	40.6278	3.0528	54.5274	44.3540
	PEFF	11.0289	0.9060	16.4936	16.6610	18.0331	1.4887	26.4301	24.4284	31.9175	2.4017	41.5250	34.7977

TABLE 6 | Comparison of the prediction performance of the PEFF and reference models.

Datasets	Models	1-Step				2-Step				3-Step			
		MAE	MAPE	RMSE	SDE	MAE	MAPE	RMSE	SDE	MAE	MAPE	RMSE	SDE
Dataset 1	BP	36.1011	6.5977	81.6468	81.7840	83.3276	18.1566	132.0443	131.5471	105.8290	20.7587	164.0412	164.1236
	ELM	42.5785	7.5592	91.4609	91.6177	81.1682	15.0630	142.4280	142.5972	122.4549	21.9137	205.6969	206.0437
	ENN	39.7156	7.4723	84.4419	84.5207	82.3793	14.7120	143.5108	143.7415	119.3047	21.7489	203.0020	203.3021
	ESN	46.5811	8.7681	98.2482	98.3220	90.3716	17.6039	143.7276	143.9530	133.0377	26.0459	189.5262	189.8139
	LSSVM	42.4525	7.4831	92.1394	91.9860	78.2665	14.5620	141.7317	141.6096	116.6006	21.9949	190.9299	190.8851
	RBF	44.9547	8.4727	102.9092	102.9386	83.7915	16.1391	159.0064	159.1626	120.7738	23.7272	196.3467	196.4828
	PEFF	16.0007	1.7722	28.9642	28.5218	24.1139	2.5428	36.5499	36.7695	40.6881	4.3568	57.1474	57.5807
Dataset 2	BP	32.5137	3.7464	75.4105	75.5294	64.2328	7.6081	117.2039	117.1351	96.1471	11.1675	156.8650	156.7773
	ELM	36.3188	4.3381	77.0213	77.0773	65.9400	7.5464	116.5061	116.6311	98.3678	11.1553	155.4301	155.2397
	ENN	35.6443	4.1737	77.1195	77.1545	68.5001	7.8467	119.2438	119.2861	96.3027	10.9769	154.5617	153.9480
	ESN	46.7721	4.9832	108.2536	108.4385	86.3844	9.1621	145.8431	145.9740	122.5856	12.9829	183.3423	183.1392
	LSSVM	45.1141	5.6019	97.4178	97.5851	70.9837	7.9522	128.6216	128.7701	93.4934	9.9382	155.3779	155.4104
	RBF	35.4258	4.2141	78.9399	79.0719	63.1031	7.1286	116.0497	116.2490	89.6376	9.9341	149.9513	150.1800
	PEFF	13.5285	0.9495	21.3817	21.5418	18.3995	1.2623	29.2054	29.0429	31.3365	2.0810	44.2778	44.3717
Dataset 3	BP	30.2430	4.0074	64.9896	64.9369	61.3442	7.6853	109.0110	109.1889	89.7525	11.4595	146.1884	146.3795
	ELM	29.6854	3.9093	65.1172	65.1585	63.7582	7.9744	111.1659	111.2437	96.6048	11.8075	151.4255	151.1732
	ENN	31.3175	4.0124	66.2060	66.2620	62.3144	8.0096	108.9393	108.8660	95.4399	11.9277	150.2073	150.0207
	ESN	36.2387	4.2851	80.6081	80.5945	71.8532	8.2486	123.1829	122.9609	110.2528	12.5923	166.9108	166.8145
	LSSVM	29.0768	3.7328	64.2934	64.3394	59.7767	7.5425	105.9502	106.0854	92.8919	11.5781	145.6523	145.8945
	RBF	30.0137	3.8859	64.9002	64.9930	62.5905	7.9905	108.4614	108.6286	95.1872	11.9129	148.0287	148.2561
	PEFF	11.0289	0.9060	16.4936	16.6610	18.0331	1.4887	26.4301	24.4284	31.9175	2.4017	41.5250	34.7977

in PV power prediction, which can be concluded based on the MAPE values (1.7722, 2.5428, and 4.3568%) in each forecasting step. For Dataset 2, the four assessment indicator values in each forecasting step obtained from PEFF are the most satisfactory. The MAPE in one-step forecasting obtained from the PEFF is 0.9495%, which is 0.2132% higher than that of SSA-MOGWO-EF, which is second in the prediction effect. As for Dataset 3, regardless of the prediction step, the PEFF obtains the optimal forecasting result proved by obviously lower error indicator values. For instance, in three steps, the PEFF provides the lowest MAE, MAPE, RMSE, and SDE of 31.9175, 2.4017%, 41.5250, and 34.7977, respectively, while the highest MAPE was obtained from CEEMD-MOGOA-EF at 4.8862%.

Remark. The assessment indicator values in Experiment II show that the PEFF precedes the EFs based on other ensemble strategies in terms of forecasting precision and stability, regardless of the prediction step and dataset.

3.6 Experiment III: Comparison With Classic Models

The experimental results reveal the forecasting ability of the PV power sequence by comparing the PEFF with classic models (BP, ELM, ENN, ESN, LSSVM, and RBF). The prediction results are listed in **Table 6**. For Dataset 1, with regard to the one-step prediction, PEFF exhibits the optimal forecasting performance. With regard to two- and three-step

TABLE 7 | DM results of the models included in this study.

Models	Dataset 1			Dataset 2			Dataset 3		
	1-step	2-step	3-step	1-step	2-step	3-step	1-step	2-step	3-step
BP	8.0975[a]	8.2785[a]	8.5469[a]	8.9575[a]	8.9649[a]	8.1576[a]	8.9706[a]	8.9572[a]	8.4854[a]
ELM	8.3003[a]	7.6419[a]	7.9218[a]	8.4157[a]	8.2922[a]	8.4595[a]	8.1557[a]	7.5357[a]	8.3491[a]
ENN	9.4340[a]	9.1787[a]	9.2577[a]	9.2431[a]	8.8922[a]	9.1555[a]	8.6712[a]	9.2060[a]	8.5318[a]
ESN	8.7769[a]	8.5462[a]	8.5971[a]	9.3235[a]	9.1948[a]	8.8171[a]	9.4502[a]	8.5344[a]	8.9387[a]
LSSVM	8.0816[a]	8.4655[a]	8.4952[a]	7.8869[a]	8.1898[a]	8.1456[a]	8.3463[a]	8.4094[a]	8.4547[a]
RBF	6.7760[a]	7.1797[a]	7.1551[a]	6.6626[a]	6.6190[a]	6.9984[a]	7.4597[a]	6.8404[a]	7.0853[a]
SSA-ARIMA	2.0238[b]	2.5513[b]	2.0551[b]	2.3060[b]	2.4991[b]	2.6909[a]	2.7593[a]	2.3472[b]	1.9986[b]
SSA-DBN	1.9993[b]	2.0575[b]	2.6407[a]	2.0543[b]	2.6143[a]	2.0435[b]	2.7293[a]	2.1500[b]	1.9966[b]
SSA-GRU	2.0511[b]	2.4160[b]	2.2733[b]	2.1517[b]	2.6308[a]	2.3853[b]	2.3497[b]	2.7172[a]	2.0858[b]
SSA-LSTM	2.5572[b]	2.5537[b]	2.1804[b]	2.3678[b]	1.9759[b]	1.9740[b]	2.3308[b]	2.5792[a]	2.7340[a]
CEEMD-MOGOA-EF	2.0899[b]	2.5288[b]	2.4294[b]	1.9719[b]	2.2971[b]	2.1222[b]	2.7543[a]	2.2712[b]	2.4885[b]
SSA-MODA-EF	2.1256[b]	2.5620[b]	2.2230[b]	2.6141[a]	2.6492[a]	2.7082[a]	2.4105[b]	2.0438[b]	2.1890[b]
SSA-MOGWO-EF	2.8733[b]	2.1124[b]	2.7858[a]	2.4983[b]	2.9561[a]	2.0382[b]	2.4027[b]	2.0667[b]	2.9219[a]

Note:
[a] *99% significance level (critical value = 2.576).*
[b] *95% significance level (critical value = 1.960).*

forecasting processes, the assessment indicators of the PEFF are minimally compared with the classical models, which indicate that the PEFF is more valid in PV power prediction. For Dataset 2, classical models achieved unsatisfactory prediction effects with higher values of MAE, MAPE, RMSE, and SDE. Specifically, in 2-step forecasting, the MAPE values of BP, ELM, ENN, ESN, LSSVM, and RBF are 7.6081, 7.5464, 7.8467, 9.1621, 7.9522, and 7.1286%, respectively, and the MAPE values of PEFF were 1.2623, 6.3458, 6.2841, 6.5844, 7.8999, 6.6899, and 5.8664%. As for Dataset 3, the PEFF precedes other involved models with average values of the evaluation criteria of 20.3265, 1.5988%, 28.1495, and 25.2957, respectively, in three steps.

Remark. Based on the results of this experiment, we can conclude that the PEFF has a stronger effect than the classical models in short-term PV power prediction.

4 DISCUSSION

In this section, the PEFF is discussed in detail, including the significance, sensitivity analysis, operational efficiency, practical applications, defects, and future directions of the PEFF.

4.1 Forecasting Significance of the PEFF

To investigate whether there is a prominent difference in the prediction ability between the PEFF and reference models, the Diebold–Mariano (DM) test (Jiang et al., 2021) was conducted. The concrete theory of this test can be found in (Zhang et al., 2021).

As for our study, **Table 7** lists the DM values from 1-step to 3-step prediction based on the three datasets. First, the PEFF is different from classical models (BP, ELM, ENN, ESN, LSSVM, and RBF) at a significance level of 99%. Moreover, although the DM values computed based on the difference between the PEFF and hybrid models are lower than that computed based on the difference between the PEFF and each classical model, the PEFF has a distinguishing prediction capacity compared with each hybrid model at a significance level of 95%. Then, when comparing the PEFF with the EFs adopting disparate ensemble strategies, the DM statistical magnitude pertaining to one-step to three-step prediction based on each dataset exceeds the critical value at a significance level of 95%, which illustrates that there is a 95% possibility that we will not reject H_1.

TABLE 8 | Four designed indicators of sensitivity analysis.

Metrics	Definition	Equations
$\mathbf{S}_{MAE}$	STD of MAE of *n* times prediction	$\mathbf{S}_{MAE} = Std\,(MAE_1, MAE_2, ..., MAE_n)$
$\mathbf{S}_{MAPE}$	STD of MAPE of *n* time prediction	$\mathbf{S}_{MAPE} = Std\,(MAPE_1, MAPE_2, ..., MAPE_n)$
$\mathbf{S}_{RMSE}$	STD of RMSE of *n* time prediction	$\mathbf{S}_{RMSE} = Std\,(RMSE_1, RMSE_2, ..., RMSE_n)$
$\mathbf{S}_{SDE}$	STD of SDE of *n* time prediction	$\mathbf{S}_{SDE} = Std\,(SDE_1, SDE_2, ..., SDE_n)$

Based on the DM statistical magnitude, the forecasting results of PEFF are significantly different from those of classical models (BP, ELM, ENN, ESN, LSSVM, and RBF), hybrid models (SSA-ARIMA, SSA-DBN, SSA-GRU, and SSA-LSTM), and EFs using diverse ensemble strategies (CEEMD-MOGOA-EF, SSA-MODA-EF, and SSA-MOGWO-EF). Thus, it is valuable to exploit PEFF and employ it in practical PV power forecasting.

4.2 Sensitivity Analysis of the PEFF

To explore the prediction ability of the PEFF when a certain parameter changes, sensitivity analysis was performed to measure the output result sensitivity of PEFF based on the parameter settings of SSA and MOGOA. The standard deviation (STD) of error indicators, as shown in **Table 8**, was adopted to assess the level at which the parameter setting impacted the properties of PEFF (Liu et al., 2021). The results of the sensitivity analysis are listed in **Table 9**, where the window length and principal

TABLE 9 | STD values of the results acquired by changing parameters.

Datasets	Algorithms	Parameters	1-Step				2-Step				3-Step			
			$\mathbf{S}_{MAE}$	$\mathbf{S}_{MAPE}$	$\mathbf{S}_{RMSE}$	$\mathbf{S}_{SDE}$	$\mathbf{S}_{MAE}$	$\mathbf{S}_{MAPE}$	$\mathbf{S}_{RMSE}$	$\mathbf{S}_{SDE}$	$\mathbf{S}_{MAE}$	$\mathbf{S}_{MAPE}$	$\mathbf{S}_{RMSE}$	$\mathbf{S}_{SDE}$
Dataset 1	SSA	Window Length	2.4813	0.2350	2.3475	2.3814	2.4059	0.3147	2.6224	2.9407	2.3837	0.2948	3.4107	2.7419
		PCDN	2.3360	0.2054	2.3083	2.3926	2.2226	0.2817	2.8869	2.8968	2.7934	0.2097	2.4408	2.7966
	MOGOA	Population Size	1.1213	0.0917	1.1665	0.8483	1.7892	0.1428	1.2537	1.1912	0.4999	0.1025	1.2675	1.5429
		Iteration Number	1.2564	0.0336	1.1484	1.2494	1.9742	0.1214	1.0036	1.8143	0.4316	0.0543	0.6807	0.8152
		Archive Size	0.3559	0.1024	0.6964	0.4870	0.4273	0.0452	0.7552	0.8221	0.7834	0.1017	1.3152	1.3772
Dataset 2	SSA	Window Length	2.3800	0.2428	1.8090	2.0273	4.4983	0.3094	4.5220	0.9046	6.5667	0.4377	5.1595	0.4256
		PCDN	2.7747	0.2052	2.5614	1.7477	4.5408	0.2985	4.4322	0.9292	4.6348	0.4184	5.0983	0.4358
	MOGOA	Population Size	1.1489	0.0616	0.5079	0.3770	2.0190	0.1531	2.0844	0.2835	2.8884	0.2013	2.4218	0.2682
		Iteration Number	1.1993	0.0717	0.6826	0.6831	2.6139	0.1805	2.4364	0.2175	1.9861	0.1461	1.7761	0.2883
		Archive Size	1.3714	0.0815	1.1076	0.7743	0.9389	0.0742	0.9619	0.5533	2.0549	0.1344	1.9448	0.2166
Dataset 3	SSA	Window Length	2.0610	0.1698	1.9280	2.6148	2.0962	0.1984	3.2311	3.8987	3.7061	0.2162	4.0595	4.8314
		PCDN	2.2278	0.1803	2.2001	1.9136	2.8146	0.2428	2.3597	2.6772	5.3466	0.3361	6.0052	6.1229
	MOGOA	Population Size	0.3832	0.0237	0.3812	0.4114	1.2668	0.1207	1.8670	1.9202	0.7750	0.0640	0.4889	0.7453
		Iteration Number	0.9925	0.0607	0.8133	0.7673	1.7904	0.0701	0.9975	1.8172	0.9138	0.0791	0.6670	0.5771
		Archive Size	0.7819	0.0500	0.3134	0.3099	0.6165	0.0530	0.5860	1.7270	0.7736	0.0549	0.6798	0.8035

TABLE 10 | Run time of each model.

Models	Dataset 1			Dataset 2			Dataset 3			Average
	1-step	2-step	3-step	1-step	2-step	3-step	1-step	2-step	3-step	
BP	1.6324	2.0975	1.2785	1.9572	2.4854	1.8003	2.1419	1.4218	1.9157	1.8590
ELM	1.0469	0.9575	0.9649	0.8922	0.9595	0.9557	1.0357	0.8491	0.9340	0.9551
ENN	5.1576	4.9706	4.9572	4.6787	4.7577	4.7431	5.3922	4.6555	5.1712	4.9427
ESN	3.9595	3.6557	3.0357	2.7655	2.7952	3.1869	3.4898	3.4456	2.6463	3.2200
LSSVM	2.4854	2.8003	3.1419	2.7060	3.0318	3.2769	3.0462	3.0971	2.8235	2.9343
RBF	7.8491	6.9340	6.6787	6.7094	6.7547	7.2760	6.6797	6.6551	7.1626	6.9666
SSA-ARIMA	15.4218	16.9157	15.7922	15.6948	16.3171	15.9502	17.0344	16.4387	16.3816	16.2163
SSA-DBN	129.6555	110.1712	112.7060	121.7513	113.2551	109.5060	113.6991	124.8909	119.9593	117.2883
SSA-GRU	144.0318	153.2769	144.0462	135.5472	126.1386	136.1493	125.2575	123.8407	124.2543	134.7270
SSA-LSTM	152.7577	161.7431	141.3922	143.1190	152.4984	145.9597	133.3404	122.5853	154.2238	145.2911
CEEMD-MOGOA-EF	209.0971	198.8235	199.6948	190.8143	191.2435	189.9293	198.3500	203.1966	211.2511	199.1556
SSA-MODA-EF	218.8147	216.9058	217.1270	218.6160	219.4733	220.3517	218.8308	217.5853	217.5497	218.3616
SSA-MOGWO-EF	223.9134	225.6324	226.0975	225.9172	226.2858	225.7572	224.7537	223.3804	225.5678	225.2562
PEFF	163.3171	160.9502	161.0344	164.0759	165.0540	163.5308	167.7792	166.9340	168.1299	164.5339

Note: The running time is measured in seconds (s).

component decomposition number (PCDN) belong to SSA and the population size, iteration number, and archive size belong to MOGOA.

Sensitivity analyses were conducted by changing one parameter, and the remaining parameters remained unchanged. It must be known that each parameter value is assigned as 40, 45, 50, 55, and 60 in terms of window length, and 10, 15, 20, 25, and 30 in terms of PCDN, respectively. Meanwhile, the parameter is set as 10, 30, 50, 70, and 90 in terms of population size; 300, 400, 500, 600, and 700 in terms of iteration number; and 100, 150, 200, 250, and 300 in terms of archive size.

(1) As the parameters of SSA change, the $\mathbf{S}_{MAE}$, $\mathbf{S}_{MAPE}$, $\mathbf{S}_{RMSE}$, and $\mathbf{S}_{SDE}$ values of the two parameters become higher. For instance, in the three-step prediction based on Dataset 1, the $\mathbf{S}_{MAPE}$ value is 0.2948 for window length and 0.2097 for PCDN, which are lower than the $\mathbf{S}_{SDE}$ values but still higher than the $\mathbf{S}_{MAPE}$ values of MOGOA parameters. The above results indicate that the SSA significantly impacts the forecasting performance of the PEFF.

(2) When the parameter in MOGOA is altered, compared with the sensitivity analysis results obtained from SSA, the measured indicators of MOGOA are lower than those of SSA, indicating that the fluctuation of forecasting performance generated by parameter alteration in MOGOA is slight.

4.3 Operational Efficiency of the PEFF

To further explore the operational efficiency of PEFF, the run time of each model based on three datasets, regardless of the forecasting step, is listed in **Table 10**. In particular, the mean value of the operational time of the PEFF is 164.5339 s, which is shorter than the EFs based on different ensemble strategies. The computing time of PEFF is shorter than that of SSA-MODA-EF, which confirms the superiority of the MOGOA adopted in PEFF. Moreover, in contrast to hybrid and classical models, the average operational time of PEFF is longer, which is reasonable owing to its excellent prediction ability. The operational efficiency of the PEFF can be improved by adopting a high-powered computer.

4.4 Practical Applications of the PEFF

In practical scenarios, real-time missions considering PV power generation planning and grid security safeguards require effective forecasting. In particular, precise and stable PV power prediction can solve the challenge caused by the irregular undulations of PV power, which is the key point for the businesslike running of PV power generation systems and can improve the stability and efficiency of the energy market and energy industry. Accurate and stable PV power forecasting can also effectively boost the PV penetration degree, reduce the use of fossil fuels, and enhance economic and environmental benefits, which is conductive to the sustainable development of the society.

Moreover, the forecasting results of the PV power output support decision-makers in maintaining the power system stability, installing large PV power stations, and monitoring the security of power systems. When the predicted PV power output result is inconsistent with the real data, energy producers can assess efficiency degradation caused by motor aging or motor faults and deal with it in time to reduce economic loss. In other words, accurate PV output forecasting provides valuable assistance for monitoring the running status of equipment, which saves maintenance costs and reduces the risk of power grid breakdown.

Furthermore, accurate forecasting is essential for grid operators to help them determine balancing power that can satisfy unnecessary demand for fossil fuels. By referring to PV power forecasting results, decision makers can determine reasonable power supply volumes of PV power and fossil fuel power plants so as to satisfy the country's power demand. Meanwhile, accurate PV power forecasting is conductive to setting reasonable rotating reserve capacity so as to enhance energy economy and reduce the risk of PV abandonment.

4.5 Defects and Future Directions

The main limitation of PEFF is that the applied area is limited to power systems containing PV power stations, instead of finance, such as future price predictions.

After PV power prediction, adaptable improvements for future are as follows:

(1) Finding more effective data preprocessing methods to process PV power data and process the irregular characteristics of the initial PV power data more effectively.
(2) Enhancing sub-models to provide satisfactory forecasting results for the subsequent forecasting of EF.
(3) The operation efficiency of the proposed PEFF should be improved by GPU acceleration.
(4) More underlying external factors, such as weather and solar irradiation, must be taken into consideration to obtain better forecasting results for longer forecasting horizons.

5 CONCLUSION

We developed an ensemble forecasting frame that capitalizes the data preprocessing technique and optimization algorithm to forecast PV power. The proposed system has been proved to be effective and efficient to improve the prediction accuracy and stability of short-term PV power. Specifically, a data preprocessing technique is employed to disintegrate the original PV power sequence and integrate a processed sequence to decrease prediction errors created by the irregular undulations of the PV power series. ARIMA and three DLMs were adopted as sub-models to forecast PV power sequences. Further, MOGOA was adopted to compute the weight of each sub-model of the PEFF and obtain the final prediction result. Simulation results prove that the proposed system (SSA–MOGOA–EF) surpasses the comparative models. Specifically, in Experiment I, the lowest average MAPE based on each dataset was obtained from PEFF with values of 2.89, 1.43, and 1.60%, which were reduced by 3.82, 1.78, and 1.54%, respectively, compared with the maximum values obtained from SSA–ARIMA. This revels that the proposed ensemble forecasting scheme is obviously superior to the comparative hybrid models in terms of accuracy and stability. The ensemble strategy can successfully improve short-term PV power forecasting performance. In Experiments II, the MAPE values of PEFF based on all datasets are the most satisfactory, which implies that the PEFF based on SSA and MOGOA technologies exceeds the comparative ensemble models based on other data preprocessing technologies and optimization algorithms. Thus, it is a wise choice to use SSA–MOGOA–EF for PV power forecasting. Similarly, in Experiment III, the improvement of the proposed forecasting system over the classical individual models is more significant, further testifies the effectiveness of the proposed ensemble system. Five discussions are further conducted to testify the performance of the proposed frame. Based on the discussions, we testify that there is an observable difference between the prediction results of the PEFF and the benchmark models, and the proposed forecasting frame is less sensitive to the parameter change of MOGOA than that of SSA. Furthermore, the proposed forecasting frame incurs a lower cost compared with EFs adopting other ensemble strategies. Thus, we can conclude that the PEFF successfully improves the forecasting accuracy and stability of PV power and can achieve more efficient and time-saving forecasting results, which can provide useful support for smart grid planning.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

Software, YL; Supervision, LL; Writing—original draft, SZ; Writing—review and editing, YL.

REFERENCES

Abdel-Nasser, M., and Mahmoud, K. (2019). Accurate Photovoltaic Power Forecasting Models Using Deep LSTM-RNN. *Neural Comput. Applic* 31, 2727–2740. doi:10.1007/s00521-017-3225-z

Agga, A., Abbou, A., Labbadi, M., and El Houm, Y. (2021). Short-Term Self Consumption PV Plant Power Production Forecasts Based on Hybrid CNN-LSTM, ConvLSTM Models. *Renew. Energ.* 177, 101–112. doi:10.1016/j.renene.2021.05.095

Aygül, K., Cikan, M., Demirdelen, T., and Tumay, M. (2019). Butterfly Optimization Algorithm Based Maximum Power point Tracking of Photovoltaic Systems under Partial Shading Condition. *Energy Sourc. A: Recovery, Utilization, Environ. Effects*, 1–19. doi:10.1080/15567036.2019.1677818

Bates, J. M., and Granger, C. W. J. (1969). Combination of Forecasts. *Oper. Res. Q.*

Das, N., Wongsodihardjo, H., and Islam, S. (2015). Modeling of Multi-junction Photovoltaic Cell Using MATLAB/Simulink to Improve the Conversion Efficiency. *Renew. Energ.* 74, 917–924. doi:10.1016/j.renene.2014.09.017

David, M., Ramahatana, F., Trombe, P. J., and Lauret, P. (2016). Probabilistic Forecasting of the Solar Irradiance with Recursive ARMA and GARCH Models. *Solar Energy* 133, 55–72. doi:10.1016/j.solener.2016.03.064

de Carvalho, M., and Rua, A. (2017). Real-time Nowcasting the US Output gap: Singular Spectrum Analysis at Work. *Int. J. Forecast.* 33, 185–198. doi:10.1016/j.ijforecast.2015.09.004

Devaraj, J., Madurai Elavarasan, R., Shafiullah, G., Jamal, T., and Khan, I. (2021). A Holistic Review on Energy Forecasting Using Big Data and Deep Learning Models. *Int. J. Energ. Res.* 45, 13489–13530. doi:10.1002/er.6679

Dong, J., Olama, M. M., Kuruganti, T., Melin, A. M., Djouadi, S. M., Zhang, Y., et al. (2020). Novel Stochastic Methods to Predict Short-Term Solar Radiation and Photovoltaic Power. *Renew. Energ.* 145, 333–346. doi:10.1016/j.renene.2019.05.073

Elavarasan, R. M., Leoponraj, S., Vishnupriyan, J., Dheeraj, A., and Gangaram Sundar, G. (2021). Multi-Criteria Decision Analysis for User Satisfaction-Induced Demand-Side Load Management for an Institutional Building. *Renew. Energ.* 170, 1396–1426. doi:10.1016/j.renene.2021.01.134

Elsinga, B., and van Sark, W. G. J. H. M. (2017). Short-term Peer-To-Peer Solar Forecasting in a Network of Photovoltaic Systems. *Appl. Energ.* 206, 1464–1483. doi:10.1016/j.apenergy.2017.09.115

Eseye, A. T., Zhang, J., and Zheng, D. (2018). Short-term Photovoltaic Solar Power Forecasting Using a Hybrid Wavelet-PSO-SVM Model Based on SCADA and Meteorological Information. *Renew. Energ.* 118, 357–367. doi:10.1016/j.renene.2017.11.011

Feng, C., Cui, M., Hodge, B.-M., and Zhang, J. (2017). A Data-Driven Multi-Model Methodology with Deep Feature Selection for Short-Term Wind Forecasting. *Appl. Energ.* 190, 1245–1257. doi:10.1016/j.apenergy.2017.01.043

Hao, Y., and Tian, C. (2019). The Study and Application of a Novel Hybrid System for Air Quality Early-Warning. *Appl. Soft Comput.* 74, 729–746. doi:10.1016/j.asoc.2018.09.005

Hassani, H., and Ghodsi, Z. (2015). A Glance at the Applications of Singular Spectrum Analysis in Gene Expression Data. *Biomol. Detect. Quantification* 4, 17–21. doi:10.1016/j.bdq.2015.04.001

Irfan, M., Elavarasan, R. M., Hao, Y., Feng, M., and Sailan, D. (2021). An Assessment of Consumers' Willingness to Utilize Solar Energy in China: End-Users' Perspective. *J. Clean. Prod.* 292, 126008. doi:10.1016/j.jclepro.2021.126008

Islam, S. (2017). "Challenges and Opportunities in Grid Connected Commercial Scale PV and Wind Farms," in Proc. 9th Int. Conf. Electr. Comput. Eng. ICECE 2016, 1–7. doi:10.1109/ICECE.2016.7853843

Iversen, E. B., Morales, J. M., Møller, J. K., and Madsen, H. (2016). Short-term Probabilistic Forecasting of Wind Speed Using Stochastic Differential Equations. *Int. J. Forecast.* 32, 981–990. doi:10.1016/j.ijforecast.2015.03.001

Jiang, P., Liu, Z., Niu, X., and Zhang, L. (2021). A Combined Forecasting System Based on Statistical Method, Artificial Neural Networks, and Deep Learning Methods for Short-Term Wind Speed Forecasting. *Energy* 217, 119361-11936. doi:10.1016/j.energy.2020.119361

Jiang, P., and Liu, Z. (2019). Variable Weights Combined Model Based on Multi-Objective Optimization for Short-Term Wind Speed Forecasting. *Appl. Soft Comput.* 82, 105587. doi:10.1016/j.asoc.2019.105587

Jiang, P., Liu, Z., Wang, J., and Zhang, L. (2021). Decomposition-selection-ensemble Forecasting System for Energy Futures price Forecasting Based on Multi-Objective Version of Chaos Game Optimization Algorithm. *Resour. Pol.* 73, 102234. doi:10.1016/j.resourpol.2021.102234

Korkmaz, D. (2021). SolarNet: A Hybrid Reliable Model Based on Convolutional Neural Network and Variational Mode Decomposition for Hourly Photovoltaic Power Forecasting. *Appl. Energ.* 300, 117410. doi:10.1016/j.apenergy.2021.117410

Krishnannair, S., Aldrich, C., and Jemwa, G. T. (2016). Detecting Faults in Process Systems with Singular Spectrum Analysis. *Chem. Eng. Res. Des.* 113, 151–168. doi:10.1016/j.cherd.2016.07.014

Kushwaha, V., and Pindoriya, N. M. (2019). A SARIMA-RVFL Hybrid Model Assisted by Wavelet Decomposition for Very Short-Term Solar PV Power Generation Forecast. *Renew. Energ.* 140, 124–139. doi:10.1016/j.renene.2019.03.020

Li, C. (2020). Designing a Short-Term Load Forecasting Model in the Urban Smart Grid System. *Appl. Energ.* 266, 114850. doi:10.1016/j.apenergy.2020.114850

Li, P., Zhou, K., Lu, X., and Yang, S. (2020). A Hybrid Deep Learning Model for Short-Term PV Power Forecasting. *Appl. Energ.* 259, 114216. doi:10.1016/j.apenergy.2019.114216

Liu, L., Zhao, Y., Wang, Y., Sun, Q., and Wennersten, R. (2019). "A Weight-Varying Ensemble Method for Short-Term Forecasting PV Power Output," in *Energy Procedia*, 158, 661–668. doi:10.1016/j.egypro.2019.01.180*Energ. Proced.*

Liu, Z., Jiang, P., Wang, J., and Zhang, L. (2021). Ensemble Forecasting System for Short-Term Wind Speed Forecasting Based on Optimal Sub-model Selection and Multi-Objective Version of Mayfly Optimization Algorithm. *Expert Syst. Appl.* 177, 114974-11497. doi:10.1016/j.eswa.2021.114974

Liu, Z., Jiang, P., Wang, J., and Zhang, L. (2022). Ensemble System for Short Term Carbon Dioxide Emissions Forecasting Based on Multi-Objective tangent Search Algorithm. *J. Environ. Manage.* 302, 113951. doi:10.1016/j.jenvman.2021.113951

Liu, Z., Jiang, P., Zhang, L., and Niu, X. (2020). A Combined Forecasting Model for Time Series: Application to Short-Term Wind Speed Forecasting. *Appl. Energ.* 259, 114137. doi:10.1016/j.apenergy.2019.114137

Luo, X., Zhang, D., and Zhu, X. (2021). Deep Learning Based Forecasting of Photovoltaic Power Generation by Incorporating Domain Knowledge. *Energy* 225, 120240. doi:10.1016/j.energy.2021.120240

Mellit, A., Pavan, A. M., and Lughi, V. (2021). Deep Learning Neural Networks for Short-Term Photovoltaic Power Forecasting. *Renew. Energ.* 172, 276–288. doi:10.1016/j.renene.2021.02.166

Mirjalili, S. Z., Mirjalili, S., Saremi, S., Faris, H., and Aljarah, I. (2018). Grasshopper Optimization Algorithm for Multi-Objective Optimization Problems. *Appl. Intell.* 48, 805–820. doi:10.1007/s10489-017-1019-8

Nie, Y., Jiang, P., and Zhang, H. (2020). A Novel Hybrid Model Based on Combined Preprocessing Method and Advanced Optimization Algorithm for Power Load Forecasting. *Appl. Soft Comput.* 97, 106809. doi:10.1016/j.asoc.2020.106809

Niu, D., Wang, K., Sun, L., Wu, J., and Xu, X. (2020). Short-term Photovoltaic Power Generation Forecasting Based on Random forest Feature Selection and CEEMD: A Case Study. *Appl. Soft Comput.* 93, 106389. doi:10.1016/j.asoc.2020.106389

Niu, X., and Wang, J. (2019). A Combined Model Based on Data Preprocessing Strategy and Multi-Objective Optimization Algorithm for Short-Term Wind Speed Forecasting. *Appl. Energ.* 241, 519–539. doi:10.1016/j.apenergy.2019.03.097

Opitz, D., and Maclin, R. (1999). Popular Ensemble Methods: An Empirical Study. *jair* 11, 169–198. doi:10.1613/jair.614

Pedro, H. T. C., and Coimbra, C. F. M. (2012). Assessment of Forecasting Techniques for Solar Power Production with No Exogenous Inputs. *Solar Energy* 86, 2017–2028. doi:10.1016/j.solener.2012.04.004

Qu, J., Qian, Z., and Pei, Y. (2021a). Day-ahead Hourly Photovoltaic Power Forecasting Using Attention-Based CNN-LSTM Neural Network Embedded with Multiple Relevant and Target Variables Prediction Pattern. *Energy* 232, 120996-12099. doi:10.1016/j.energy.2021.120996

Qu, Y., Xu, J., Sun, Y., and Liu, D. (2021b). A Temporal Distributed Hybrid Deep Learning Model for Day-Ahead Distributed PV Power Forecasting. *Appl. Energ.* 304, 117704. doi:10.1016/j.apenergy.2021.117704

Sharma, N., Mangla, M., Yadav, S., Goyal, N., Singh, A., Verma, S., et al. (2021). A Sequential Ensemble Model for Photovoltaic Power Forecasting. *Comput. Electr. Eng.* 96, 107484. doi:10.1016/j.compeleceng.2021.107484

Shelat, N., Das, N., Khan, M. M. K., and Islam, S. (2019). "Nano-structured Photovoltaic Cell Design for High Conversion Efficiency by Optimizing Various Parameters," in 2019 29th Australas. Univ. Power Eng. Conf. AUPEC 2019, 26–29. doi:10.1109/AUPEC48547.2019.211859

Shezan, S. K. A., Das, N., and Mahmudul, H. (2017). Techno-economic Analysis of a Smart-Grid Hybrid Renewable Energy System for Brisbane of Australia. *Energ. Proced.* 110, 340–345. doi:10.1016/j.egypro.2017.03.150

S., J., and M., R. (2018). "Reconfigurable Solar Converter with Inverter, Chopper and Rectifier Operation for Residential PV Applications," in Proc. 2018 IEEE Int. Conf. Power Electron. Drives Energy Syst. PEDES 2018, 1–4. doi:10.1109/PEDES.2018.8707571

Soubdhan, T., Ndong, J., Ould-Baba, H., and Do, M.-T. (2016). A Robust Forecasting Framework Based on the Kalman Filtering Approach with a Twofold Parameter Tuning Procedure: Application to Solar and Photovoltaic Prediction. *Solar Energy* 131, 246–259. doi:10.1016/j.solener.2016.02.036

Takilalte, A., Harrouni, S., and Mora, J. (2019). Forecasting Global Solar Irradiance for Various Resolutions Using Time Series Models - Case Study: Algeria. *Energ. Sourc. Part A: Recovery, Utilization, Environ. Effects* 0, 1–20. doi:10.1080/15567036.2019.1649756

Tan, B., Raga, S. R., Rietwyk, K. J., Lu, J., Fürer, S. O., Griffith, J. C., et al. (2021). The Impact of spiro-OMeTAD Photodoping on the Reversible Light-Induced Transients of Perovskite Solar Cells. *Nano Energy* 82, 105658. doi:10.1016/j.nanoen.2020.105658

Tanaka, K., Uchida, K., Ogimi, K., Goya, T., Yona, A., Senjyu, T., et al. (2011). Optimal Operation by Controllable Loads Based on Smart Grid Topology Considering Insolation Forecasted Error. *IEEE Trans. Smart Grid* 2, 438–444. doi:10.1109/TSG.2011.2158563

Tian, C., and Hao, Y. (2018). A Novel Nonlinear Combined Forecasting System for Short-Term Load Forecasting. *Energies* 11, 712. doi:10.3390/en11040712

Unnikrishnan, P., and Jothiprakash, V. (2018). Daily Rainfall Forecasting for One Year in a Single Run Using Singular Spectrum Analysis. *J. Hydrol.* 561, 609–621. doi:10.1016/j.jhydrol.2018.04.032

Wang, J., Li, Q., and Zeng, B. (2021). Multi-layer Cooperative Combined Forecasting System for Short-Term Wind Speed Forecasting. *Sustainable Energ. Tech. Assessments* 43, 100946. doi:10.1016/j.seta.2020.100946

Xiao, L., Wang, J., Dong, Y., and Wu, J. (2015). Combined Forecasting Models for Wind Energy Forecasting: A Case Study in China. *Renew. Sustain. Energ. Rev.* 44, 271–288. doi:10.1016/j.rser.2014.12.012

Yacef, R., Mellit, A., Belaid, S., and Şen, Z. (2014). New Combined Models for Estimating Daily Global Solar Radiation from Measured Air Temperature in Semi-arid Climates: Application in Ghardaïa, Algeria. *Energ. Convers. Manage.* 79, 606–615. doi:10.1016/j.enconman.2013.12.057

Yagli, G. M., Yang, D., and Srinivasan, D. (2019). Automatic Hourly Solar Forecasting Using Machine Learning Models. *Renew. Sustain. Energ. Rev.* 105, 487–498. doi:10.1016/j.rser.2019.02.006

Yang, D., and Dong, Z. (2018). Operational Photovoltaics Power Forecasting Using Seasonal Time Series Ensemble. *Solar Energy* 166, 529–541. doi:10.1016/j.solener.2018.02.011

Yildiz, C., and Acikgoz, H. (2021). A Kernel Extreme Learning Machine-Based Neural Network to Forecast Very Short-Term Power Output of an On-Grid Photovoltaic Power Plant. *Energ. Sourc. Part A: Recovery, Utilization, Environ. Effects* 43, 395–412. doi:10.1080/15567036.2020.1801899

Yin, W., Han, Y., Zhou, H., Ma, M., Li, L., and Zhu, H. (2020). A Novel Non-iterative Correction Method for Short-Term Photovoltaic Power Forecasting. *Renew. Energ.* 159, 23–32. doi:10.1016/j.renene.2020.05.134

Zhang, L., Dong, Y., and Wang, J. (2019). Wind Speed Forecasting Using a Two-Stage Forecasting System with an Error Correcting and Nonlinear Ensemble Strategy. *IEEE Access* 7, 176000–176023. doi:10.1109/ACCESS.2019.2957174

Zhang, L., Wang, J., Niu, X., and Liu, Z. (2021). Ensemble Wind Speed Forecasting with Multi-Objective Archimedes Optimization Algorithm and Sub-model Selection. *Appl. Energ.* 301, 117449. doi:10.1016/j.apenergy.2021.117449

Zhang, T., Lv, C., Ma, F., Zhao, K., Wang, H., and O'Hare, G. M. P. (2020a). A Photovoltaic Power Forecasting Model Based on Dendritic Neuron Networks with the Aid of Wavelet Transform. *Neurocomputing* 397, 438–446. doi:10.1016/j.neucom.2019.08.105

Zhang, W., Zhang, L., Wang, J., and Niu, X. (2020b). Hybrid System Based on a Multi-Objective Optimization and Kernel Approximation for Multi-Scale Wind Speed Forecasting. *Appl. Energ.* 277, 115561. doi:10.1016/j.apenergy.2020.115561

Zhen, H., Niu, D., Wang, K., Shi, Y., Ji, Z., and Xu, X. (2021). Photovoltaic Power Forecasting Based on GA Improved Bi-LSTM in Microgrid without Meteorological Information. *Energy* 231, 120908. doi:10.1016/j.energy.2021.120908

Zhou, Q., Wang, C., and Zhang, G. (2020). A Combined Forecasting System Based on Modified Multi-Objective Optimization and Sub-model Selection Strategy for Short-Term Wind Speed. *Appl. Soft Comput.* 94, 106463. doi:10.1016/j.asoc.2020.106463

Zhu, E., and Pi, D. (2020). Photovoltaic Generation Prediction of CCIPCA Combined with LSTM. *Complexity* 2020, 1–11. doi:10.1155/2020/1929372

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Energy Research

ORIGINAL RESEARCH
published: 03 January 2022
doi: 10.3389/fenrg.2021.739653

Identifying the Asymmetric Channel of Crude Oil Risk Pass-Through to Macro Economy: Based on Crude Oil Attributes

*Shuaishuai Jia[1], Hao Dong[2] and Zhenzhen Wang[3]**

[1]Guangzhou Institute of International Finance, Guangzhou University, Guangzhou, China, [2]Lingnan (University) College, Sun Yat-Sen University, Guangzhou, China, [3]School of Mathematics and Statistics, Guangdong University of Foreign Studies, Guangzhou, China

The impact channel of crude oil market risk on the macroeconomy is highly related to oil attributes. This paper uses a stepwise test method with dummy variables to identify the channel effect of commodity market risk as well as financial market risk and explore the characteristics of the channel effect in different periods dominated by different oil attributes. Furthermore, this paper investigates the asymmetric characteristics of the channel effect under the condition of crude oil returns heterogeneity. The empirical results show that: First, commodity market risk, as well as financial market risk plays a channel role in the impact of crude oil market risk on the macroeconomic operation. Second, there is a significant difference in the ability of the commodity market and financial market to cope with shocks of crude oil market risk in periods dominated by different attributes. During the period dominated by the commodity attribute of oil, both commodity market and financial market play the role of "risk buffer"; during the period dominated by dual attributes of oil, the commodity market risk plays the role of "risk buffer", while the financial market risk plays the role of "magnifier" of the crude oil market risk. Third, the channel effect pattern and degree of commodity market risk and financial market risk are significantly asymmetric.

Keywords: channel identification, upward and downward return risk, macro economy, Commodity market, financial market

OPEN ACCESS

Edited by:
Xun Zhang,
Academy of Mathematics and Systems Science (CAS), China

Reviewed by:
Muhammad Mohsin,
Jiangsu University, China
Xuerong Li,
Academy of Mathematics and Systems Science (CAS), China

****Correspondence:***
Zhenzhen Wang
wangzhzh@gdufs.edu.cn

Specialty section:
This article was submitted to Sustainable Energy Systems and Policies, a section of the journal Frontiers in Energy Research

Received: *11 July 2021*
Accepted: *12 November 2021*
Published: *03 January 2022*

Citation:
Jia S, Dong H and Wang Z (2022) Identifying the Asymmetric Channel of Crude Oil Risk Pass-Through to Macro Economy: Based on Crude Oil Attributes.
Front. Energy Res. 9:739653.
doi: 10.3389/fenrg.2021.739653

1 INTRODUCTION

The impact of crude oil market risk on the smooth operation of the macroeconomy is related to the dual attributes of the oil. As a key input factor, the price fluctuation of crude oil will directly or indirectly affect macroeconomic factors such as inflation, interest rate and price level (Bloch et al., 2015; Ratti and Vespignani, 2016; Sodeyfifi and Katircioglu, 2016; Kang et al., 2017; Shi and Sun, 2017; Ji et al., 2019a; Saeed and Ridoy, 2020). The dependence of economic development on crude oil is the main factor that determines the direct shock of the crude oil market. Since 1993, China has become a net importer of crude oil, and the import volume shows an upward trend. In 2014, China became the world's largest importer of crude oil. China's consumption of crude oil is also on the rise, making it the second-largest consumer of crude after the United States. In addition, the commodity and financial attributes of oil reflect the indirect impact of the crude oil market on the smooth operation of the macroeconomy. Crude oil, on the one hand as the main commodity, its price fluctuation will cause the price of other commodities to change in the same direction; on the other hand, as the necessity of production and life, the change of its price as well as the change of other commodities prices will inevitably lead to the change of many macroeconomic indicators, such as the price level. In addition, due to the influence of demand, the international crude oil market price has a great uncertainty, which in turn changes the macroeconomic operation (Guo et al., 2016; Choi et al., 2018; Gong and Lin, 2018; Humbatova et al., 2019). Given this, violent oil price fluctuations must have an impact on the smooth operation of the economy, and the impact of the crude oil market on the economy is heterogeneous under different oil attributes.

The focus of investigating the ability of the commodity market and financial market to cope with the oil market risk impact is mainly on the channel effect. In terms of commodity markets, on the one hand, China, as the largest importer, could be changed by the fluctuation of crude oil price, thus affecting the supply of crude oil. Moreover, changes in crude oil market price would lead to changes in commodity market price, which indirectly affect economic development (Ji et al., 2018). On the other hand, China focuses on strengthening the energy reserves and infrastructures construction to buffer shocks of the external environment. The decline of crude oil market price reduces the constraints of the cost factor on China and provides conditions for expanding energy reserves and increasing energy infrastructure construction. From the perspective of the financial market, the negative correlation between crude oil price and the financial market creates more profit space for investors, especially fund investors (Wen et al., 2019a; Zheng and Du 2019). Investors take crude oil as a hedge asset, grasp the downward trend of crude oil prices, and obtain excess profits in a short time. This kind of capital flow challenges the effectiveness of the financial market and then affects economic development (Ji et al., 2019a; Meng et al., 2020).

The main purpose of this paper is to identify the channels through which crude oil market risk affects macroeconomic operation and to analyze the asymmetric characteristics of channel effect under the condition of oil returns heterogeneity. The formation of dual attributes of oil has enhanced the channel effect of the commodity market as well as a financial market in the impact of the crude oil market on economic development. In addition, the rise and fall of returns in the crude oil market will affect the speed of obtaining market information and investors' expectations, thus the correlation between risks in the crude oil market and commodity market is different from that between risks in the crude oil market and financial market (Cheng et al., 2016; Ji et al., 2019a; Wu et al., 2021). The financial attribute of crude oil has attracted more and more investors to consider oil as a financial commodity while providing opportunities for speculators. Due to the heterogeneity of information access between investors and speculators, investors and speculators have different responses to market price fluctuations under different trends (Ahmed et al., 2017; Ji et al., 2019b). This paper investigates the asymmetric impact of the international crude oil market risks on the stable economic operation from the perspective of the rising and falling returns of the crude oil market, which is conducive to improving the cross-market information sharing mechanism and preventing the shock from the price fluctuation in the international crude oil market to the stable economic operation.

The existing literature mainly identifies the impact channels of the crude oil market and price level. Alvarez et al. (2011) analyzed the indirect effect of the crude oil market on price levels in Spain and Europe. Razmi et al. (2016) investigated the mediating mechanism of currency in the impact of crude oil price on the price level. The results show that before the financial crisis, the mediating effect of currency is not obvious, but after the financial crisis, the price of crude oil will not only directly affect the price level, but also have an indirect effect on the price level through currency. In addition, the interaction mechanism between crude oil price and price level can also be realized through other channels, such as the channel of interest rate level (Smets and Peersman, 2001; Tillmann, 2008; Kose et al., 2012), the channel of credit level (Wulandari, 2012), the mediating effect of exchange rate market (Takhtamanova, 2010), and the mediating effect of stock price (Gregoriou and Kontonikas, 2010; Nistico, 2012; Chen et al., 2020). Sek (2017) analyzed the channel role of crude oil export price and production cost in the relationship between crude oil price and price level. The results show that crude oil export price and production cost play a channel role in the relationship between crude oil price and price level, but in the long run, the crude oil price has no indirect effect on the price level. Sek (2019) further analyzed the heterogeneity of the indirect effect of crude oil price on price level between crude oil exporting and importing countries. Chen et al. (2020) studied the impact mechanism of Brent (WTI) crude oil price on CPI/PPI. The results show that Brent has a significant negative indirect impact on CPI, PPI and MPPI, while the financial market has no channel effect in the impact of WTI crude oil price on the price level.

However, despite the channel effect behind the relationships between crude oil markets and macroeconomics, there are still several gaps in most of the existing studies. Firstly, most of the existing literature focuses on the channel effect of the financial market, and analysis of the channel effect of the commodity market focuses on a single commodity such as gold or natural gas. Secondly, the channel effect needs to be updated with the oil dual attributes. the existing related literature ignores the perspective of different oil attributes. Finally, according to the review of the relevant literature concerning the influence mechanism of the international crude oil market, existing researches ignore the asymmetry of the channel effect caused by investors' expectation when the returns of the international crude oil market rise and fall.

To overcome the deficiencies, the main contributions are as follows. The first is to identify the channel effect of commodity market risk. As a core part of the commodity, the oil risk evolution not only exhibits the main risk resource for natural gas or gold, but also snapshots the risk evolution of the commodity market index. Additionally, the commodity market plays an important role in macroeconomic. In this sense, this paper identifies the channel role of the commodity market risk as well as financial market risk in the impact of crude oil market on the macroeconomic operation. Secondly, this paper illustrates the channel effects of commodity market risk or financial market risk taking accounting into the dual attributes of the oil. On the one hand, there are differences in the risk evolution characteristics of the crude oil market in periods dominated by oil's financial attributes and commodity attributes. On the other hand, the formation mechanism of different oil attributes leads to a different correlation between the crude oil market and the commodity market as well as the financial market. Based on this, this paper analyzes the correlation between the influence channels of the crude oil market and oil attributes. Finally, we analyze the asymmetric channel effect commodity market risk or financial market risk existing in the oil pass-through to macroeconomic under different oil return trends.

The overall framework of this paper is as follows: **Section 2** proposes the research hypotheses of this paper; **Section 3** elaborates the research design; **Section 4** identifies the channel effect of commodity market and financial market and analyzes its asymmetry. The main conclusions of this paper are summarized in the fifth Section.

2 HYPOTHESES

The identification of the impact channels of crude oil market risks includes two aspects: the direct impact of international crude oil market risks on the smooth operation of the macroeconomy and the indirect impact of commodity market risks as well as financial market risks. Changes in the international crude oil market risks will affect investors' expectations, enterprise operations and external environment, and then directly affect the smooth operation of the macroeconomy (Hamilton, 1983; Killian, 2009; Sek, 2017, 2019; Gong and Lin, 2018; He and Lin, 2019; Huang et al., 2019). From the perspective of investors' expectation, the change of the crude oil market risk means the increase of the crude oil price uncertainty and the change of investors' expectation, which in turn changes investors' capital allocation, leading to the fluctuation of output (Gonzalez-Concepcion et al., 2018; Wang et al., 2021). From the perspective of enterprise operation, changes in crude oil market risks affect the operating cost of enterprises. To control costs and maximize their profits, manufacturers will adjust their oil input. If the input of other production factors remains unchanged, the change of oil input will affect the total production amount, that is, the actual output level of the enterprise will change, and then affect the smooth operation of the macroeconomy (Long and Liang, 2018). The external environment mainly includes two aspects: industrial structure adjustment and monetary policy adjustment. The effect of crude oil price fluctuations on industrial structure adjustment is mainly due to the difference in oil dependence, and the adjustment of resource allocation is cyclical to a certain extent, which eventually increases the burden of economic self-regulation (Chen et al., 2020). The effect of monetary policy adjustment is also affected by the risk of the crude oil market (Wen et al., 2019b). On the one hand, crude oil market risk causes the transfer of wealth between oil importing and exporting countries (Wei, 2019); on the other hand, monetary authorities will formulate corresponding policies to mitigate the impact of changes in crude oil prices. Wealth transfer and policy change will lead to the change of money supply in the domestic market. In the case of constant money demand, the change of money supply will affect the domestic investment environment, change the investment strategy of enterprises, and ultimately affect the smooth operation of the economy. To sum up, the international crude oil market risk will change the stability of macroeconomic operations.

The commodity market and financial market are the key channels through which the risks of the crude oil market affect the smooth operation of the economy (Fan et al., 2021; Xiao et al., 2021). Commodity market risks are directly related to macroeconomic operation such as domestic output and price level (Shi and Sun, 2017). Fundamentally, fluctuation in commodity prices changes the transportation costs of related products, thus affecting the price changes of final products. Therefore, the channel effect of the commodity market is highly correlated with such factors as manufacturer cost, wealth transfer, monetary policy, industrial structure adjustment and consumer expectation (Gong and Lin, 2018; He and Lin, 2019). So commodity market is one channel. In addition, the development of crude oil market financialization provides more new assets for investors or enterprises in pursuit of profits. The price fluctuation of the crude oil market will increase the uncertainty of the financial market, and then affect the behavior of investors, and finally affect the macroeconomy (Cong and Shen, 2013; Coronado et al., 2018). Specifically, the price fluctuation in the crude oil market will change the allocation proportion of investors or enterprises' funds between the real economy and the financial market through the expected effect; the cost effect will change the oil demand of manufacturers; the monetary policy effect will change the investment strategy. Changes in these factors ultimately affect the smooth operation of the macro-economy. Therefore, the financial market plays a channel role in the influence of the crude oil market on the macroeconomic operation.

Based on the above analysis, this paper puts forward the basic hypothesis of the channel role of the commodity market as well as the financial market.

Hypothesis 1: *Commodity market risks, as well as financial market risks play a channel role in the impact of the crude oil market on the smooth operation of the macroeconomy.*

The channel effect of the commodity market is significantly related to the dual attributes of the oil. Commodity prices are directly related to the macro-economic operation, such as domestic output and price level (Shi and Sun, 2017; Song et al., 2019). Fundamentally, the price fluctuation of the crude oil market will lead to the change of commodity price, and then lead to the change of commodity import price, which will be transferred to the change of product price. Due to the differences in the formation mechanism of different oil attributes, there are differences in the ability of the commodity market to cope with the risk impact of the crude oil market (**Figure 1**). From the perspective of oil's commodity attribute, fluctuations in crude oil price will lead to changes in the price of raw materials, and then change the production costs of enterprises (Hewitt et al., 2019). The change of enterprise production cost will lead to the change of enterprise investment strategy, and ultimately affect the smooth operation of the macro-economy. Consumers are recipients of commodity prices. As a consumer necessity, the price fluctuation of agricultural products such as grain will affect the operation of the macro-economy. Monetary policy has a moderating effect on price stability. Price fluctuations in the crude oil market will affect the prices of other commodities. Monetary policies such as quantitative easing can increase the money supply by ensuring low interest rates, and then stimulate investment to slow down the impact of changes in the prices of crude oil and other commodities, and finally affect the economic situation (Yang and Zhou, 2017).

From the perspective of the financial attribute of oil, the information effect of commodity price on macroeconomic operation increases with the financialization of oil and other commodities (Zhang et al., 2017). Changes in commodity prices triggered by fundamentals (changes in the relationship between supply and demand of commodities) will transmit different signals to other markets about the functioning of the global economy, changing market confidence, and thus affecting the smooth operation of the macroeconomy. If commodity financialization improves the information content of commodity prices, it will help the commodity market to play a signal function, and lead to a stronger response of the macroeconomy to commodity price shocks. If commodity financialization brings more speculative noise, it will increase the distortion of market price signals, which will interfere with market players and further amplify the change of output.

Investor expectations have a significant impact on the mediating effect of commodity markets (Dong et al., 2019; Chen et al., 2020). When returns of the crude oil market rise, the market investors' expectations are high and the market sentiment is positive. Enterprises adjust their investment strategies and focus on the demand for a single commodity. When returns fall, the diversity of the commodity market is the main way for market participants to cushion the blow of a major event. Through the purchase of diversified commodities, investors have reduced oil demand, driving down international oil prices and pushing up the prices of other commodities. Due to the difference in the impact of investors' expectations on the demand changes of different commodities, the channel effect of commodity market risk is greater when returns fall (Ji et al., 2019a).

Based on the above analysis, this paper puts forward the basic hypothesis on channel asymmetry in the commodity market.

Hypothesis 2a: *Under different attributes of oil, the channel effect of commodity market risk is different.*

Hypothesis 2b: *Under the condition of returns heterogeneity, the mediating effect of commodity market risk is asymmetric.*

Financial market risk is one of the key influence factors for the smooth operation of the macroeconomy. The mediating effect of the financial market is mainly related to the economic effect of financial market risks, specifically including the following aspects (**Figure 2**): first, enterprises solve their financing constraints through the financial market to improve economic activities; second, financial market influences the efficiency of capital allocation and industrial structure; third, price fluctuation in the financial market has monetary policy effect; fourth, financial market has wealth effect (Dong et al., 2019; Sek, 2019; Chen et al., 2020). Based on the commodity attribute of oil, risks of the crude oil market change operating costs of enterprises. The financial market is one of the main financing channels for enterprises. Enterprises finance by issuing financial assets such as stocks and bonds for their development. When the capital market is prosperous, the economic activities of enterprises are active, and the cost of financing through the financial market is reduced. The high liquidity of the capital market attracts a large number of funds to enter the market. Risks in the financial market will affect the expectations of investors, and then change their investment strategies to seek investment benefits.

The mediating effect of the financial market is also related to investor behavior (Song et al., 2019; Chen et al., 2020; Li and Zhong, 2020). The financial market can curb inflation by attracting idle funds in the financial field. Fast financial market development has attracted a large number of risk preference investors to enter the market for trading, which has a storage function for the idle funds of the real economy. This function changes the inflation level through price fluctuation of the financial market. In addition, the price fluctuation of the financial market will make the capital flow from the real economy into the virtual economy (Li et al., 2020). Specifically, financial market risks make asset prices falsely high, and the profits of investing in the virtual economy far outweigh those in the real economy, causing capital flow from the real economy to the financial markets, resulting in money supply shortage in the real economy, which leads to the decline of output, the reduction of efficiency, the lack of motivation for technological innovation, and finally reduced allocation of social resources (Gong and Lin, 2018; He and Lin, 2019). Compared with the decline of returns, the rise of crude oil returns makes it easier for investors to obtain expected profits, reducing financing constraints of enterprises as well as the cost of obtaining capital. In addition, market uncertainty increases when returns fall, and market investors are risk-preference ones, reducing the funds' storage function of the financial market. Therefore, the crude oil market risk has a greater impact on the macroeconomic operation through the financial market when returns rise.

Based on this analysis, this paper puts forward the basic hypothesis of asymmetric channel effect in the financial market.

Hypothesis 3a: *The channel effect of financial market risk is different with different attributes of crude oil.*

Hypothesis 3b: *There is an asymmetry in the mediating role of financial market risk under different trends.*

3 STUDY DESIGN

3.1 Selection for Channel Identification Model

This paper uses a stepwise test approach to identify the impact channels of crude oil market risk. If crude oil market risk has an impact on the smooth operation of the macroeconomy by influencing commodity market risk or financial market risk, it

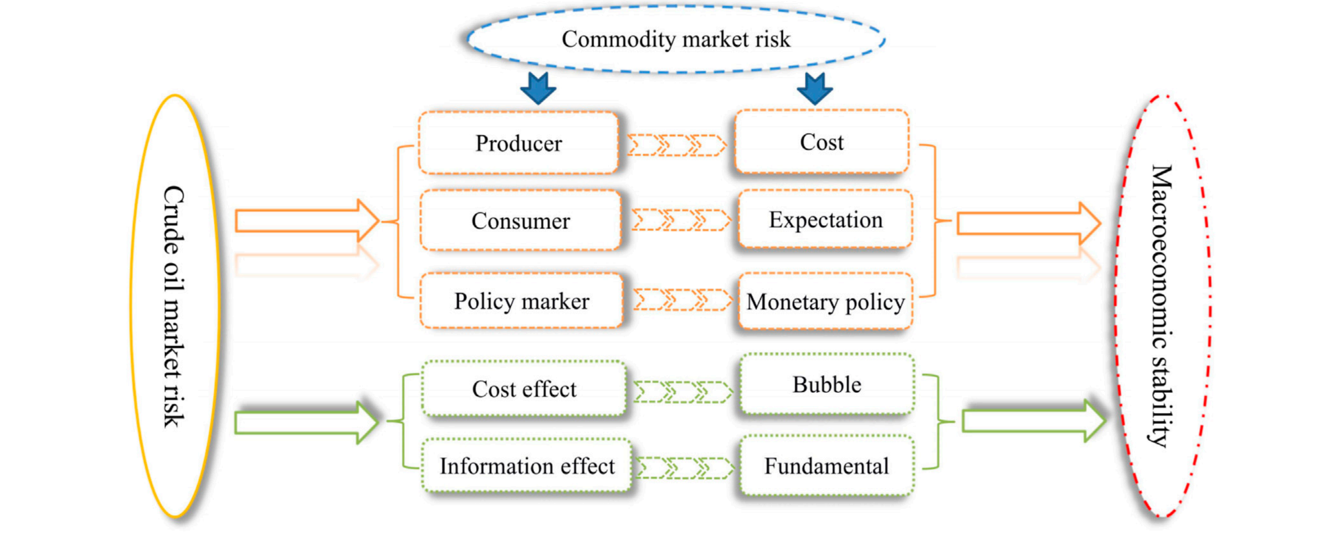


FIGURE 1 | Formation mechanism of the channel effect of commodity market risk.

is said that commodity market risk or financial market risk has a mediating effect. There are two methods to test the mediating effect: structural equation model and stepwise test. Bentler (1980) first proposed the structural equation model and analyzed the path of attitude influencing behavior. Since then, the structural equation model has been widely used in psychological research. From the perspective of engineering project operation, Qureshi and Kang (2015) analyzed the influence of project size on project complexity and analyzed its influence path. Dong et al. (2020), based on the structural equation model, analyzed mediating roles of economic conditions and financial conditions in the business cycle affecting the health system financing process. By setting the relationship between latent variables and explicit variables, the structural equation model can be used to further analyze the interaction between latent variables and latent variables. But this model requires high data quantity, so it is frequently used in questionnaire analysis. The stepwise test model verifies the mediating effect of variables by sequentially testing coefficients significance. This model was first proposed by Baron and Kenny (1986). Chen et al. (2020) analyzed the channel roles of stock prices and local government debt in the influence of crude oil prices on price levels. Compared with the structural equation model, the stepwise test method requires less data quantity. Therefore, this paper selects the stepwise test method.

The stepwise test consists of three steps. The first step is to fit the regression model as 1) to test the significance of the influence coefficient c of international crude oil market risk on the price level.

$$pl_t = \beta_{00} + cBRISK_t + \boldsymbol{\beta X_t} + \varepsilon_t, \tag{1}$$

where pl_t represents stable operation of China's macro economy at time t; $BRISK_t$ is international crude oil market risk; X_t is the control variable.

In the second step, regression models such as 2) and 3) are fitted to test the significance of coefficients a and b in turn.

$$cha_t = \beta_{01} + aBRISK_t + \boldsymbol{\beta X_t} + \varepsilon_t, \tag{2}$$

$$pl_t = \beta_{03} + c'BRISK_t + bcha_t + \boldsymbol{\beta X_t} + \varepsilon_t, \tag{3}$$

where cha_t represents the channel variable, specifically refers to commodity market risk.

In the third step, if both coefficients a and b are significant, their significance is tested; If at least one of them is not significant, the Sobel test is used to further analyze the mediating effect of commodity market risk or financial market risk.

Accordingly, the mediating effect test can be divided into three steps. The first step is to test the significance of c. The significance of coefficient c, which describes the significance of crude oil market risk and macroeconomic operation. If the effect is not significant, it is not necessary to conduct a subsequent mediating effect test. If the effect is significant, the second step test is carried out. The second step is to test whether coefficients a and b are significant. The significance of coefficient a reflects the significance of the impact of crude oil market risk on the mediating variables (commodity market risk or financial market risk), and the significance of coefficient b represents the significance of the impact of mediating variables on the smooth operation of the macroeconomy. $a \times b$ reflects the indirect effect of crude oil market risk on the smooth operation of the macroeconomy. Therefore, checking whether $a \times b$ is 0 is the key of the third step. In Step 3, $a \times b$ is tested according to different situations. 1) If both coefficients a and b are significant, it indicates that $a \times b$ is not 0, then the significance of the direct effect of crude oil market risk on the smooth operation of the macroeconomy (c') is tested. 2) If at least one of the coefficients a and b is not significant, the Sobel test is used to analyze whether $a \times b$ is 0. If c' is significant, it indicates that the risk of the crude oil market has both direct and indirect effects on

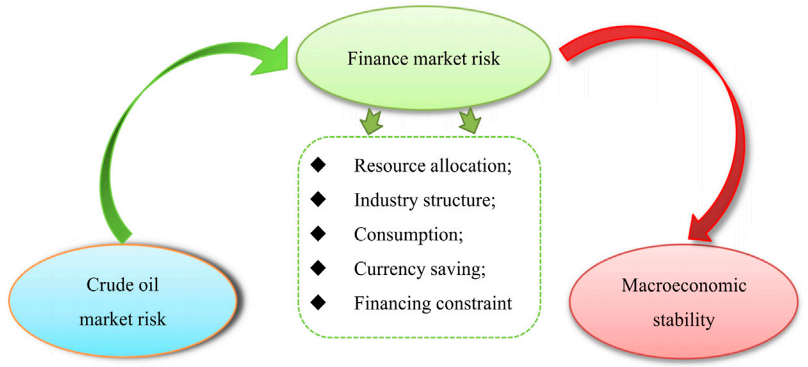


FIGURE 2 | Formation mechanism of the channel effect of financial market risk.

macroeconomic operation; if c' is not significant, it indicates that the risk of the crude oil market only has an indirect effect on macroeconomic operation. In addition, if the Sobel test is passed, it means that $a \times b$ is not significant, that is, the risk of the crude oil market has an indirect effect on macroeconomic operation, so we continue to test the significance of the coefficient c'. On the contrary, it indicates that the risk of the crude oil market has no indirect effect on macroeconomic operation, so the test should be stopped.

The Sobel test statistic (S.TE) is proposed by Sobel (1982), and the null hypothesis of the test is $a \times b = 0$. The statistical calculation method is shown in.

$$S.TE = \frac{a \times b}{\left(b^2 \times S_a^2 + a^2 \times S_b^2\right)^{1/2}}. \tag{4}$$

where S_a^2 and S_b^2 represent the standard error of the estimation of coefficients a and b, respectively.

3.2 Variable Selection and Measurement

According to the model selection part, variables selected in this paper include macro-economic operation, international crude oil market risk, commodity market risk, financial market risk and related control variables. The description and measurement methods of the variables selected in this paper are summarized in **Table 1**.

This paper employs four type variables, that is Explained variable, Explanatory variable, Mediating variable and Control variable (shown in **Table 1**). Specifically, we use consumer price index (CPI) and producer price index (PPI) to measure the Explained variable. And this paper employs Conditional Autoregression quantile Value-at-Risk (CAViaR) to model the Explanatory variable (crude oil market risk) and Mediating variable (commodity market risk or financial market risk). Finally, the Control variables include money supply and lagged CPI/PPI. Monetary policy is the key variable affecting CPI and PPI, and money supply is the direct embodiment of monetary policy. In this way, to eliminate the seasonal effect in monetary supply, we use the year-on-year ratio of M2 to measure the monetary supply. Moreover, since the influence of insufficient selection of control variables on the empirical results, this paper also selects the first-order lag of the explained variable as control variables (Li et al., 2021b).

This paper employs the CAViaR model to measure the key explanatory variable as well as channel variables, crude oil market risk, commodity market risk and financial market risk. Existing risk measurement methods are mainly based on different natures of crude oil market returns. On the one hand, market risks are measured from the perspective of the heteroscedasticity nature of asset returns. Relevant literature uses static and dynamic VaR based on the GARCH model to predict financial market risks such as stock market, crude oil market and virtual financial asset market (Bernardi and Catania, 2016; Ferraty and Quintela-Del-Río, 2016; Gkillas and Katsiampa, 2018; Li et al., 2018; Saculsan and Kanamura, 2020). On the other hand, market risks are measured from the perspective of asset returns with an agglomeration nature. Most of the relevant literature uses the expected shortfall (ES) method to measure risks. ES mainly forecasts financial market risks from the perspective of extreme events to make up for the characteristics that ordinary VaR cannot capture. The above methods have two common features. One is based on a specific distribution of crude oil market returns. The other is based on parameter estimation. For the former feature, returns of the crude oil market are usually limited to some specific distributions, such as normal distribution, t distribution and GED distribution. According to historical experience, a parametric model is used to measure the risks of the crude oil market. In addition, for parametric models, the accuracy of parameter estimation and the degree of model fitting are two aspects that need to be considered in model construction. According to the definition of risk, the measurement of crude oil market risk is forecasting quantiles. Therefore, a conditional autoregressive value at risk model (CAViaR) proposed by (Engle and Manganeli, 2004) is adopted, considering the agglomeration effect of international crude oil market returns and the application of quantile regression in risk measurement. The CAViaR model does not need to presuppose the distribution of the international crude oil market returns, and it uses the quantile regression to calculate quantiles; meanwhile, considering the agglomeration nature of international crude

TABLE 1 | Selection of mediating effect test variables.

Variable type	Variable	Indicator	Measure method
Explained variable	Macroeconomic Operation	CPI PPI	CPI PPI
Explanatory variable	Returns upward risk in international oil market Returns downward risk in international oil market	BURISK BDRISK	CAViaR
Mediating variable	Commodity market risk Financial market risk	CRISK FRISK	CAViaR
Control variable	Monetary policy Lag term	M2 CPI/PPI	M2 chain index First-order lag of the explained variables

oil market risks, the model adds the lag term of risks. By using the four model forms of CAViaR, existing literature predicts the dynamic risks of upward and downward asset returns (Meng and Taylor, 2018; Li et al., 2020, 2021a).

The basic form of the CAViaR model is as (5):

$$Risk_t(\beta) = \beta_1 + \sum_{i=1}^{q} \beta_i Risk_{t-i}(\beta) + \sum_{j=1}^{r} \beta_j l(\boldsymbol{R_{t-j}}), \tag{5}$$

where $Risk_t$ represents international crude oil market risk in month t; l $(\boldsymbol{R_{t-j}})$ is a function of exogenous variables, mainly describing the impact of different forms of international crude oil market returns on risks; the lag term $Risk_{t-i}(\boldsymbol{\beta})$ describes the agglomeration of international crude oil market risk. Based on different forms of international crude oil market returns and different model variants, (Engle and Manganeli, 2004) further put forward four forms of CAViaR model: absolutely symmetric model, asymmetric model, indirect GARCH model and adaptive model, with specific forms as (6)–(9).

Absolute symmetry model:

$$Risk_t(\boldsymbol{\beta}) = \beta_1 + \beta_2 Risk_{t-1}(\boldsymbol{\beta}) + \beta_3|R_{t-1}|. \tag{6}$$

Asymmetry model:

$$Risk_t(\boldsymbol{\beta}) = \beta_1 + \beta_2 Risk_{t-1}(\boldsymbol{\beta}) + \beta_3 (R_{t-1})^{+} + \beta_4 (R_{t-1})^{-}, \tag{7}$$

where $(R_{t-1})^{+} = max(R_{(t-1)}, 0)$, $(R_{t-1})^{-} = min(R_{(t-1)}, 0)$, depicting positive or negative monthly returns of international crude oil market in the previous period.

Indirect GARCH(1,1) model:

$$Risk_t(\boldsymbol{\beta}) = (\beta_1 + \beta_2 Risk^2_{t-1}(\boldsymbol{\beta}) + \beta_3 R^2_{t-1})^{1/2}. \tag{8}$$

Adaptive model:

$$\begin{aligned} Risk_t(\beta_1) &= Risk_{t-1}(\beta_1) \\ &\quad + \beta_1\{[1 + \exp(G[R_{t-1} - Risk_{t-1}(\beta_1)])]^{-1} - \alpha\}, \end{aligned} \tag{9}$$

where G is a finite positive integer. When returns exceed the measured value of risk, the value of G should be increased appropriately; on the contrary, the value of G should be

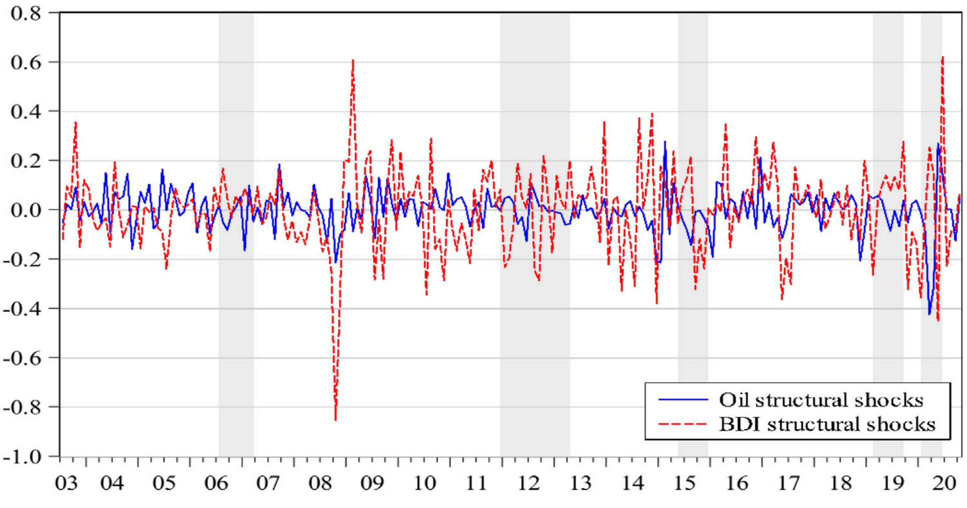


FIGURE 3 | Dynamic feature recognition of oil commodity attribute. Note: Shadow parts show negative relationship between the structural shocks of oil demand and oil price.

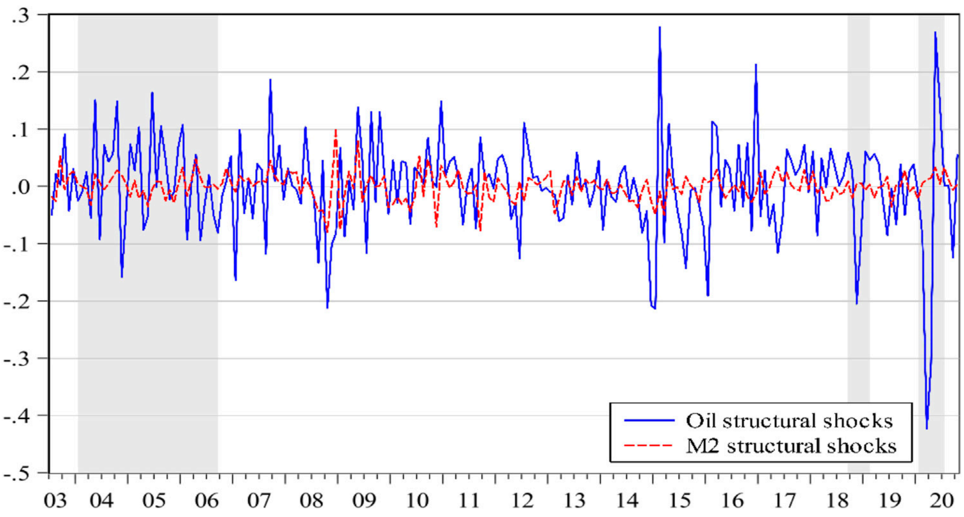


FIGURE 4 | Dynamic feature recognition of oil financial attribute. Note: Shadow parts show negative relationship between the structural shocks of oil demand and monetary policy.

reduced appropriately. Only in this way can the goodness of fit of the adaptive model be increased. This paper mainly constructs the appropriate CAViaR model from the former three models to measure the international crude oil market risk[1]

The test of the model fitting effect is to compare the fitting situation of different risk measurement methods to the international crude oil market risk. This paper uses the values of HIT and DQ statistics proposed by Engle and manganeli (2004), based on the properties of VaR and dynamic quantile test. The HIT mainly examines the difference between the risk measurement results and the returns of the international crude oil market. The HIT test statistic is expressed as (10),

$$Hit_t(\boldsymbol{\beta}) = I(R_t < Risk_t(\boldsymbol{\beta})) - \alpha, \tag{10}$$

where $I(\cdot)$ is the indicative function, when $R_t < Risk_t(\boldsymbol{\beta})$, the value of $Hit_t(\boldsymbol{\beta})$ is $1-\alpha$; otherwise, it is $-\alpha$. In addition, according to the quantile function definition, the value of the statistic $Hit_t(\boldsymbol{\beta})$ is 0 when data of period $t-1$ are given. In other words, the value of $Hit_t(\boldsymbol{\beta})$ is not related to international crude oil market risk and its lag term, so the HIT test may not be sufficient to test the goodness of fit of the model. Further, (Engle and Manganeli, 2004) proposed the DQ test, including fitting sample test and test sample test. The statistics of DQ test are expressed as (11) and (12),

$$DQ_{IS} = \frac{Hit'(\hat{\beta})X(\hat{\beta})\left(\widehat{M_T}\widehat{M_T'}\right) - X'(\hat{\beta})Hit'(\hat{\beta})}{\alpha(1-\alpha)} \sim \chi_q^2, \tag{11}$$

$$\begin{aligned} DQ_{OS} = {} & N_R^{-1}Hit'\left(\widehat{\beta_{T_R}}\right)X\left(\widehat{\beta_{T_R}}\right)\left[X'\left(\widehat{\beta_{T_R}}\right)X\left(\widehat{\beta_{T_R}}\right)\right]^{-1} \\ & \times X'\left(\widehat{\beta_{T_R}}\right)Hit'\left(\widehat{\beta_{T_R}}\right)/\alpha(1-\alpha) \sim \chi_q^2, as\ R \to \infty, \end{aligned} \tag{12}$$

where DQ_{IS} and DQ_{OS} refer to DQ test statistics of fitting samples and test samples, respectively; $X(\hat{\beta})$ is related to $\hat{\beta}$, for measuring the returns information of the international crude oil market of the fitting sample, i.e., $Hit(\hat{\beta}) = [Hit_1(\hat{\beta}), Hit_2(\hat{\beta}), \ldots, Hit_T(\hat{\beta})]'$. Similarly, assume T_R represents the size of fitting sample data and N_R is the size of test sample data, $X(\hat{\beta}_{T_R})$ is related to $\hat{\beta}_{T_R}$, $n = T_R + 1, T_R + 2, \ldots, T_R + N_R$, for measuring the returns information of the international crude oil market of the test sample, i.e., $Hit(\hat{\beta}_{T_R}) = [Hit_{T_R+1}(\hat{\beta}_{T_R}), Hit_{T_R+2}(\hat{\beta}_{T_R}), \ldots, Hit_{T_R+N_R}(\hat{\beta}_{T_R})]'$, and

$$\widehat{M_T} = X'\left(\hat{\beta}\right) - \left\{\left(2T\widehat{C_T}\right)^{-1}\sum_{t=1}^{T} I\left(|R_t < Risk_t\left(\hat{\beta}\right)| < \widehat{C_T}\right)\right.$$

$$\left. \times X'\left(\hat{\beta}\right)\nabla Risk_t\left(\hat{\beta}\right)\right\}\widehat{D_T^{-1}}\nabla' Risk_t\left(\hat{\beta}\right).$$

3.3 Sampling Scheme

In this paper, the structural vector autoregressive model (SVAR) is used to identify the dominant period of different oil attributes. The financial and commodity attributes of oil determine that the price of the crude oil market is positively affected by monetary policy and crude oil demand respectively. The financial attribute of oil means that the formation and fluctuation of crude oil market price have the basic characteristics of financial products and can play a role in the financial market (Chen et al., 2016; Zhang et al., 2017; Raheem et al., 2020). As a demand regulation policy, expansionary monetary policy will increase oil demand, reduce the uncertainty of the crude oil market, and release good news for investors. Investors' knowledge of the news enhances their optimistic expectations, which in turn changes the allocation of their funds in real and financial investments, increasing speculative demand (Tang and Xiong, 2010; Oleg and Ekaterina, 2020).

Oil supply is inelastic in the short term, and OPEC's regulation of oil supply has lagged effect, so the price of the crude oil market is affected by the demand in the short term. The demand for crude oil is usually correlated with the total demand of the national economy (Ghassan and Alhajhoj, 2016). In the long

[1] We refer to Section 4 of Engle and Manganeli, (2004) for model parameter estimation.

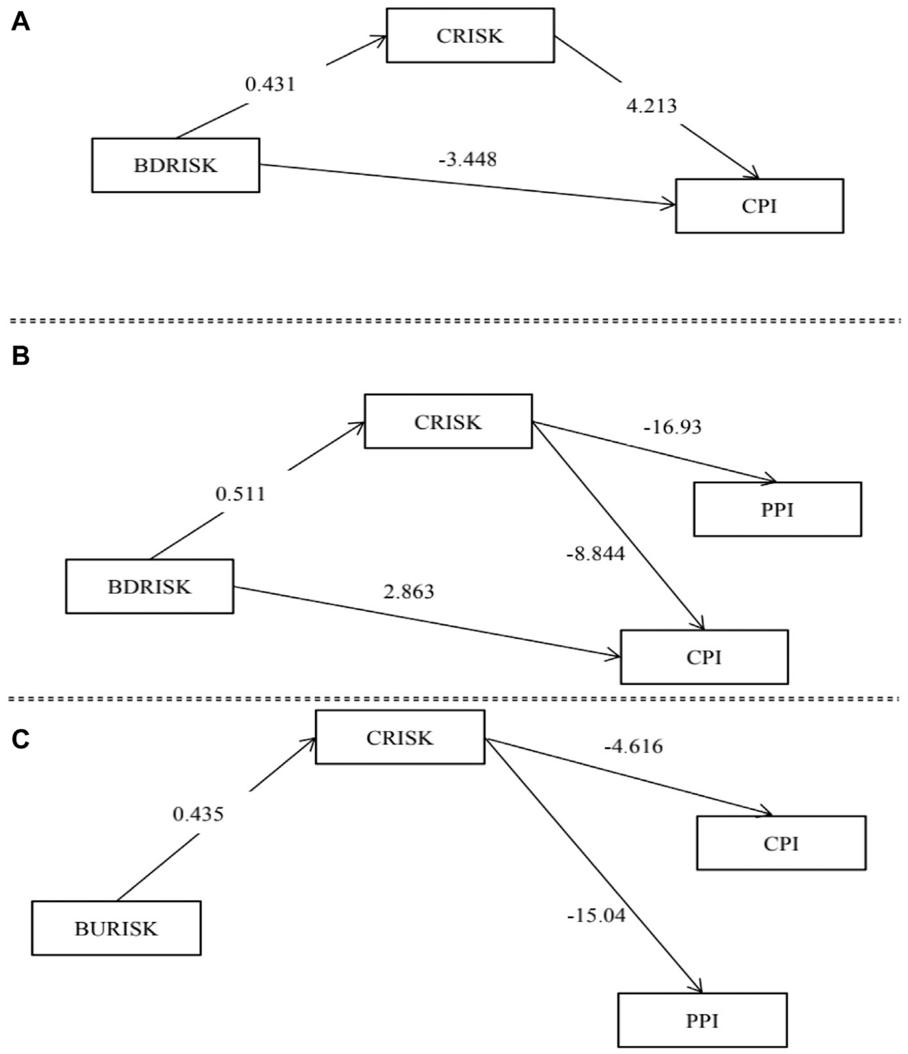


FIGURE 5 | The channel effect of commodity market risk under different returns trends. Note:**(A)–(B)** show the channel effect of commodity market risk in the impact of international crude oil market downward returns risk when oil commodity attribute and dual attribute dominate, respectively; **(C)** represents the channel effect of the commodity market risk when the dual attribute of oil dominates in the international crude oil market upward returns risk. The number on the arrow represents the impact coefficient.

run, the supply of crude oil is more elastic. However, as oil is a non-renewable resource, its reserves, resource endowment, production cost, production capacity and OPEC resolutions all limit the supply of crude oil (Loutia et al., 2016).

Changes in crude oil price will alter the production cost of enterprises as well as the oil demand, and then affect a country's inflation level. As one of the objectives of monetary policy, to maintain price stability, countries formulate corresponding monetary policy to adjust the inflation level, so the crude oil market price will also affect monetary policies. However, traditional linear models cannot fully describe the correlation between the crude oil market, monetary policy and crude oil demand. Considering the immediate impact of international crude oil price, oil demand and monetary policy, as well as the characterization of structural shocks on the correlation, this paper refers to Kilian (2009) and constructs an SVAR model to identify the dual attributes of the oil.

The basic form of SVAR(p) model constructed in this paper is shown in **Formula (13)**,

$$B_0\boldsymbol{X_t} = \sum_{i=1}^{p} B_i\boldsymbol{X_{t-i}} + \varepsilon_t. \tag{13}$$

where $\boldsymbol{X_t} = (opi_t, dem_t, mpo_t)'$ is a 3 × 3 vector; opi_t represents the international crude oil price at time t; dem_t refers to the oil demand at time t; mpo_t represents the monetary policy at time t; p is the lag order which is identified with the SC criterion; B_0 describes the immediate effect of international crude oil market price, oil demand and monetary policy, and similarly, B_i describes the marginal impact lagged i order.

Since B_0 is reversible, **Formula (13)** can be simplified as (14).

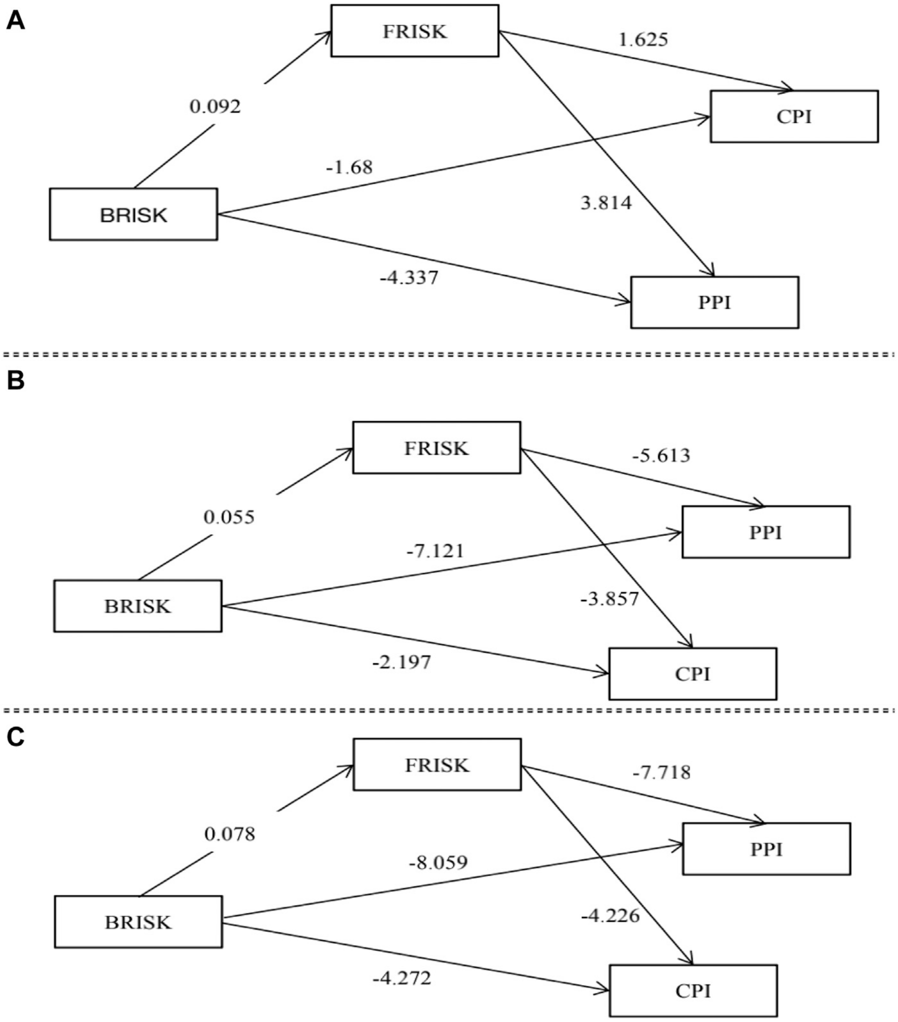


FIGURE 6 | The channel effect of financial market risk under different returns trends. Note:**(A,B)** show the channel effect of financial market risk in the impact of international crude oil market downward returns risk when oil commodity attribute and dual attribute dominate, respectively; **(C)** represents the channel effect of the financial market risk when the dual attribute of oil dominates in the international crude oil market upward returns risk. The number on the arrow represents the impact coefficient.

$$\boldsymbol{X_t} = \sum_{i=1}^{p} B_0^{-1} B_i \boldsymbol{X_{t-i}} + B_0^{-1} \varepsilon_t, \tag{14}$$

where $\varepsilon_t = (\varepsilon_t^{price-shock}, \varepsilon_t^{demand-shock}, \varepsilon_t^{policy-shock})'$ is the structural vector of international oil price shocks, including specific oil price shocks, international oil price demand shocks and international oil price monetary policy shocks.

Combined with the research purpose of this paper and existing literature, this paper imposes short-term zero constraints on the immediate impact matrix, and constructs the SVAR model. The specific constraint matrix is shown in,

$$B_0 \boldsymbol{X_t} = \begin{bmatrix} 1 & b_{12} & b_{13} \\ 0 & 1 & b_{23} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} opi_t \\ dem_t \\ mpo_t \end{bmatrix} \tag{15}$$

The corresponding position in the matrix represents the immediate impact between variables. Specifically, the crude oil price will respond to the shock of oil demand and monetary policy, so the first element of the first row of the constraint matrix is 1, and the other elements are not 0. Although global economic activities take oil as the main raw material and fuel, when oil price changes, the oil demand will have a lag effect on the crude oil price due to the development of enterprise investment plans and oil reserves, that is, the oil price shock will not affect the current oil demand, $b_{21} = 0$. In addition, changes in crude oil prices and oil demand will not cause changes in monetary policy, i.e. $b_{31} = b_{32} = 0$. But changes in monetary policy will cause changes in oil demand in the current period.

3.4 Summary

In different periods dominated by different attributes of oil, the mediating effect of commodity market risk and financial market risk is different. To distinguish different oil attribute dominant periods, based on the identification of oil attributes

TABLE 2 | Stage characteristics dominated by dual attributes of oil.

Attribute Name	Specific dominant period	Maximum duration	Minimum duration	Period proportion
Commodity attribute	2003.7–2006.8; 2007.3–2007.6 2013.4–2015.5; 2016.1–2019.2	38 months	4 months	51%
Financial attribute	2006.9–2007.2; 2007.7–2007.12 2012.1–2013.3; 2015.6–2015.12 2019.3–2019.9; 2020.7-	15 months	5 months	23%
Dual attributes	2008.1–2011.12; 2019.10–2020.7	48 months	10 months	26%

TABLE 3 | Channel effect test of commodity market risk and financial market risk for the full sample.

D.V.	Step one: (1)		Step two: (2)		Step three: (3)			
	CPI	PPI	CRISK	FRISK	CPI	CPI	PPI	PPI
BRISK	−2.597*	−7.155*	0.495*	0.067*	−2.137*	−0.728	−6.657*	−1.409
	(0.732)	(1.231)	(0.027)	(0.030)	(0.747)	(1.321)	(1.264)	(2.171)
CRISK	—	—	—	—	—	−3.252*		−10.01*
	—	—	—	—	—	(1.919)		(3.152)
FRISK	—	—	—	—	−2.285*	—	−2.493	—
	—	—	—	—	(0.958)	—	(1.546)	—
C.V.	Y	Y	Y	Y	Y	Y	Y	Y
R2	0.914	0.954	0.764	0.840	0.916	0.915	0.954	0.956
P(F)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
P(S.TE)	—	—	—	—	—	—	0.191	

*Note: Considering the length of the paper, the table only reports the key variables and omit the control variables; * indicates that the coefficient is significant at the confidence level of 0.05; D.V. represents the explained variable of each step; C.V. reports whether the control variable is included; R^2 is the adjusted R^2; P(F) refers to the P value of statistic F of model goodness of fit; and P(S.TE) reports the P value of the Sobel test.*

in 3.3, three dummy variables are added in this paper, representing oil commodity attribute dominance (DC), oil financial attribute dominance (DF) and oil dual-attribute dominance (DB), and the difference of mediating effect is analyzed with the stepwise test method. The analysis process is shown in **(16)–(18)**,

$$pl_t * D_t = \beta_{00} + cBRISK_t * D_t + \boldsymbol{\beta X_t} * \boldsymbol{D_T} + \varepsilon_t, \tag{16}$$

$$cha_t * D_t = \beta_{01} + aBRISK_t * D_t + \boldsymbol{\beta X_t} * \boldsymbol{D_T} + \varepsilon_t, \tag{17}$$

$$pl_t * D_t = \beta_{03} + c'BRISK_t * D_t + bcha_t * D_t + \boldsymbol{\beta X_t} * \boldsymbol{D_T} + \varepsilon_t. \tag{18}$$

where D_t refers to the dummy variable, representing different dominant periods of different oil attributes.

4 ASYMMETRIC EFFECTS OF INTERNATIONAL CRUDE OIL MARKET RISK WITH DIFFERENT RETURNS TRENDS

4.1 Inferred Oil Attributes

According to **Section 3.3**, the variables of oil attribute identification include international crude oil market price, oil demand and monetary policy. In this paper, Brent oil spot price is used as the proxy variable of international crude oil market price. Oil demand is measured by the growth rate of the Baltic dry bulk index (BDI). Imitating the practice of Kilian (2009), considering the close relationship between shipping index and oil demand, this paper selects dry bulk freight index as the proxy variable of oil demand. In addition, this paper selects the global money supply to measure monetary policy. After obtaining the money supply of the United States, Japan and the European Union, we use historical bilateral exchange rate data to convert the money supply into US dollars and then aggregate them to get the value of the monetary policy proxy variables. Considering the seasonal effect of variables, this paper uses X12 to adjust the international crude oil price, BDI and money supply (GM2) seasonally. On this basis, to eliminate heteroscedasticity, this paper further implements logarithmic processing on the data. The above-mentioned data are from the Wind database[2].

The dominant position of oil commodity attribute and financial attribute has dynamic characteristics. **Figure 3** and **Figure 4** respectively show the dynamic feature recognition of oil commodity attribute and financial attribute. On the one hand, from the correlation between demand, monetary policy and structural shock of crude oil price, it can be seen that

[2] The results of SVAR model stability test and lag order test can be obtained from the author

TABLE 4 | Test on the mediating effect of commodity market risk and financial market risk when oil commodity attribute dominates.

D.V.	Step one: (16)		Step two: (17)		Step three: (18)			
	CPI	PPI	CRISK	FRISK	CPI	CPI	PPI	PPI
BRISK	−1.498*	-3.876*	0.431*	0.092*	−1.68*	−3.448*	−4.337*	−1.561
	(0.548)	(1.231)	(0.023)	(0.024)	(0.555)	(1.018)	(1.249)	(2.305)
CRISK	—	—	—	—	—	4.213*	—	-4.980
	—	—	—	—	—	(1.863)	—	(4.194)
FRISK	—	—	—	—	1.625*		3.814*	
	—	—	—	—	(0.928)		(2.089)	
C.V.	Y	Y	Y	Y	Y	Y	Y	Y
R2	0.945	0.946	0.881	0.945	0.946	0.947	0.954	0.946
P(F)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
P(S.TE)	—	—	—	—	—	—	—	0.236

*Note: Considering the length of the paper, the table only reports the key variables and omit the control variables; * indicates that the coefficient is significant at the confidence level of 0.05; D.V. represents the explained variable of each step; C.V. reports whether the control variable is included; R^2 is the adjusted R^2; P(F) refers to the P value of statistic F of model goodness of fit; and P(S.TE) reports the P value of the Sobel test.*

TABLE 5 | Test on the mediating effect of commodity market risk and financial market risk when oil financial attribute dominates.

D.V.	Step one: (16)		Step two: (17)		Step three: (18)			
	CPI	PPI	CRISK	FRISK	CPI	CPI	PPI	PPI
BRISK	−2.666*	−3.876*	0.333*	0.021	−2.633*	−2.630*	−4.365*	−4.941
	(0.855)	(1.231)	(0.031)	(0.031)	(0.891)	(1.131)	(1.053)	(1.343)
CRISK	—	—	—	—	—	0.099	—	2.172
	—	—	—	—	—	(2.070)	—	(2.688)
FRISK	—	—	—	—	−0.161	—	1.297*	—
	—	—	—	—	(1.218)	—	(1.526)	—
C.V.	Y	Y	Y	Y	Y	Y	Y	Y
R2	0.964	0.946	0.806	0.965	0.964	0.964	0.960	0.960
P(F)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
S.TEST					0.896	0.961	0.590	0.420

*Note: Considering the length of the paper, the table only reports the key variables and omit the control variables; * indicates that the coefficient is significant at the confidence level of 0.05; D.V. represents the explained variable of each step; C.V. reports whether the control variable is included; R^2 is the adjusted R^2; P(F) refers to the P value of statistic F of model goodness of fit; and P(S.TE) reports the P value of the Sobel test.*

crude oil demand and structural shock of crude oil price change in the same direction, except for 2006M07-2007M03, 2011M12-2013M04, 2015M05-2015M12, 2019M02-2019M09, 2020M01-2020M06. Monetary policy and crude oil price structure shocks also basically change in the same direction, except for 2004M01-2006M09, 2018M09-2019M02, 2020M01-2020M07. On the other hand, to further compare and analyze the dynamic characteristics of the dual attributes of oil in the sample period except for the above-mentioned periods, this paper compares the symbols of the structural shocks of crude oil demand and monetary policy. For example, in the second half of 2003, there is a positive relationship between crude oil demand shock and monetary policy shock as well as crude oil price shock, but the shock of monetary policy on international crude oil price has a certain lag effect. This indicates that the fluctuation of international crude oil price is mainly affected by the oil demand, while the influence of monetary policy lags, that is, the fluctuation of crude oil price is mainly regulated by the relationship between oil supply and oil demand (Jia et al., 2021). Therefore, at this time, the oil commodity attribute is dominant. For example, from the second half of 2007 to the beginning of 2008, although the oil demand shock is positively related to the shock of the crude oil market price, the shock is still negative; while the monetary policy shock is positively related to the shock of crude oil market price, and the shock is positive. This shows that the financialization of the commodity market has gradually taken shape, and a large amount of oil has entered the reserve field as an investment or even speculation, but not into the production field. Therefore, the financial attribute of oil is dominant at this time. The same has happened since July 2020. To sum up, this paper obtains the stage characteristics dominated by dual attributes of oil as shown in **Table 2**.

Different attributes of oil alter dynamically (Zhao et al., 2020a). On the one hand, oil has the attributes of commodity and finance. On the other hand, the dominance of different attributes of oil is dynamic, and there is a situation of dual

TABLE 6 | Test on the mediating effect of commodity market risk and financial market risk when dual attribute of oil dominates.

D.V.	Step one: (16)		Step two: (17)		Step three: (18)			
	CPI	PPI	CRISK	FRISK	CPI	CPI	PPI	PPI
BRISK	−2.928*	−8.360*	0.511*	0.055*	−2.197*	2.863*	−7.121*	2.324
	(0.616)	(0.948)	(0.022)	(0.025)	(0.603)	(1.458)	(0.961)	(2.161)
CRISK	—	—	—	—	—	−8.844*		−16.93*
	—	—	—	—	—	(2.039)		(3.130)
FRISK	—	—	—	—	-3.857*	—	-5.613*	—
	—	—	—	—	(0.838)	—	(1.415)	—
C.V.	Y	Y	Y	Y	Y	Y	Y	Y
R2	0.968	0.960	0.914	0.955	0.971	0.971	0.963	0.965
P(F)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

*Note: Considering the length of the paper, the table only reports the key variables and omit the control variables; * indicates that the coefficient is significant at the confidence level of 0.05; D.V. represents the explained variable of each step; C.V. reports whether the control variable is included; R^2 is the adjusted R^2; P(F) refers to the P value of statistic F of model goodness of fit; and P(S.TE) reports the P value of the Sobel test.*

TABLE 7 | Test results of the spillover mechanism of international crude oil market upward returns risk under full sample.

D.V.	Step one: (16)		Step two: (17)		Step three: (18)			
	CPI	PPI	CRISK	FRISK	CPI	CPI	PPI	PPI
BRISK	−3.334*	−5.285*	0.401*	0.085	−2.407*	−0.724	−3.876	3.753
	(1.375)	(2.482)	(0.090)	(0.056)	(1.396)	(1.588)	(1.656)	(2.624)
CRISK	—	—	—	—	—	−3.823*	—	−13.27*
	—	—	—	—	—	(1.251)	—	(2.045)
FRISK	—	—	—	—	−2.572*	—	−3.879*	—
	—	—	—	—	(0.971)	—	(1.656)	—
C.V.	Y	Y	Y	Y	Y	Y	Y	Y
R2	0.911	0.946	0.394	0.837	0.914	0.915	0.947	0.957
P(F)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
P(S.TE)	—	—	—	—	0.187	—	0.202	

*Note: Considering the length of the paper, the table only reports the key variables and omit the control variables; * indicates that the coefficient is significant at the confidence level of 0.05; D.V. represents the explained variable of each step; C.V. reports whether the control variable is included; R^2 is the adjusted R^2; P(F) refers to the P value of statistic F of model goodness of fit; and P(S.TE) reports the P value of the Sobel test.*

attributes co-domination. Most of the dominant periods of oil financial attribute are before and after the occurrence of special events (Hu et al., 2021; Jia et al., 2021), mainly including six periods: 2006.9–2007.3, 2007.7–2007.12, 2011.12–2013.4, 2015.5–2015.12, 2019.2–2019.9 and 2020.7-. The dominant period of oil commodity attribute is in the stable period of the crude oil market, which includes four periods: 2003.7–2006.9, 2007.3–2007.6, 2013.4–2015.5 and 2016.1–2019.2. In addition, the period in which oil dual attributes jointly dominate the international crude oil market price is related to special events (Zhao et al., 2020b, 2020c; Xie et al., 2020), which are 2008.1–2011.12 and 2019.10–2020.7.

4.2 Channel Test of Crude Oil Market Risk

According to the dynamic feature identification results of different oil attributes dominant periods in **Section 4.1**, based on the availability of data, this paper extracts the oil attributes dominant period 2006.7–2020.10 to test the channel effects of commodity market risk and financial market risk. According to the research design, EVIEWS 8.0 software is used for the stepwise test, and online test tools (quantpsy.org/sobel/sobel.htm) are used for the Sobel test. **Tables 3–6** show the channel effect of commodity market risk, as well as financial market risk in the impact of crude oil market downward returns risk on macroeconomic operation; **Tables 7–9** report the channel test results of the impact of international crude oil market upward returns risk on the stable operation of the macroeconomy.

In the full sample, the impact of international crude oil market risk on CPI/PPI is different through commodity market risk or financial market risk. **Table 3** reports the stepwise results of the channel effect of commodity market risk and financial market risk in the full sample. According to the results of the first step in **Table 3**, the impact of international crude oil market risk on CPI/PPI is significantly negative. The second step is to test the significance of the impact of international crude oil market risk on commodity market risk as well as financial market risk. Similarly, international crude oil market risk has a significant positive impact on commodity market risk as well as financial market risk. Finally, the significance of the coefficients is analyzed by **Formula (3)**. As shown in **Table 3**, both international crude oil market risk and financial market risk have significant impacts on CPI, while financial market risk has

TABLE 8 | Test results on the spillover mechanism of international crude oil market upward returns risks when single oil attribute dominates.

D.V.	(a)Commodity attribute dominates	Step one: (16)	(b)Financial attribute dominates	Step one: (16)
	CPI	PPI	CPI	CPI
BURISK	−0.175 (0.826)	0.124 (1.917)	−0.982 (1.034)	−0.675 (1.305)
CRISK	—	—	—	—
FRISK	—	—	—	—
C.V.	Y	Y	Y	Y
R2	0.943	0.942	0.963	0.956
P(F)	0.000	0.000	0.000	0.000

*Note: Considering the length of the paper, the table only reports the key variables and omit the control variables; * indicates that the coefficient is significant at the confidence level of 0.05; D.V. represents the explained variable of each step; C.V. reports whether the control variable is included; R^2 is the adjusted R^2; (a) shows the test results of the spillover mechanism when oil commodity attribute dominates; (b) shows the test results of the spillover mechanism when oil financial attribute dominates.*

TABLE 9 | Test results on the spillover mechanism of international crude oil market upward returns risks when dual attribute of oil dominates.

D.V.	Step one: (16)		Step two: (17)		Step three: (18)			
	CPI	PPI	CRISK	FRISK	CPI	CPI	PPI	PPI
BURISK	−4.932*	−10.08*	0.435*	0.078*	0.078*	−0.995	-8.059*	1.890
	(1.063)	(1.705)	(0.079)	(0.041)	(0.995)	(1.412)	(1.656)	(2.023)
CRISK	—	—	—	—	—	−4.616*	—	−15.04*
	—	—	—	—	—	(1.148)	—	(1.794)
FRISK	—	—	—	—	−4.226*	—	−7.178*	—
	—	—	—	—	(0.804)	—	(1.493)	—
C.V.	Y	Y	Y	Y	Y	Y	Y	Y
R2	0.968	0.951	0.705	0.955	0.972	0.971	0.957	0.965
P(F)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

*Note: Considering the length of the paper, the table only reports the key variables and omit the control variables; * indicates that the coefficient is significant at the confidence level of 0.05; D.V. represents the explained variable of each step; C.V. reports whether the control variable is included; R^2 is the adjusted R^2; P(F) refers to the P value of statistic F of model goodness of fit; and P(S.TE) reports the P value of the Sobel test.*

an insignificant impact on PPI. Based on this, this paper uses the Sobel test to further identify the mediating effect of financial market risk in the impact of international crude oil market risk on PPI, and the test results fall into the acceptance domain, that is, there is no mediating effect. In addition, the impact on CPI/PPI is significant, but after controlling the commodity market risk, the impact of international crude oil market risk on CPI/PPI is not significant, that is, the direct effect does not exist. According to the comprehensive results of the stepwise test, under the full sample, international crude oil market risk will affect CPI through financial market risk, but will not affect PPI, that is, the mediating effect of financial market risk only exists in the relationship between international crude oil market risk and CPI (Chen et al., 2020). In addition, international oil market risks indirectly affect CPI/PPI through commodity market risk, while direct effects do not exist. So, Hypothesis 1 is true.

Under the dominance of oil commodity attribute, crude oil market risk affects CPI/PPI through financial market risk, while through commodity market risk, it only affects CPI. **Table 4** reports the stepwise test results of the channel effect of commodity market risk and financial market risk when oil commodity attribute dominates. The first step fits **Formula (16)** to test the significance of the impact of crude oil market risk on CPI/PPI. As can be seen from the table, the impact of crude oil market risk on CPI/PPI is significantly negative. The second step tests the significance of the impact of crude oil market risk on commodity market risks or financial market risks through fitting **Formula (17)**. Similarly, the crude oil market risk has a significant positive impact on commodity market risk as well as financial market risk. Finally, **Formula (18)** is fitted to analyze the significance of coefficients. As shown in **Table 4**, both crude oil market risk and financial market risk have significant impacts on CPI/PPI. Crude oil market risk and financial market risk have a significant impact on CPI, but not on PPI. Based on this, this paper uses the Sobel test to further identify the mediating effect of financial market risk in the impact of crude oil market risk on PPI, and the test results fall into the acceptance domain, that is, there is no mediating effect. For the mediating effect test of commodity market risk, the test results fall into the acceptance domain, that is, crude oil market risk will not influence PPI through commodity market risk. According to the comprehensive results of the stepwise test, under the dominance of oil commodity attribute, there is a mediating effect of financial market risk in the impact of crude oil

TABLE 10 | Comparison of the spillover mechanism of international crude oil market risk under different returns trends.

	Commodity market	PPI	Financial market	PPI
	CPI		CPI	
	(a) International crude oil market downward returns risk			
Full sample	*(–)	*(–)	√(–)	
Commodity attribute dominates	√(+)		√(+)	√(+)
Financial attribute dominates				
Dual attribute dominates	√(–)	*(–)	√(–)	√(–)
	(b) International crude oil market upward returns risk			
Full sample	*(–)	*(–)		
Dual attribute dominates	*(–)	*(–)	√(–)	√(–)

*Note: √ represents channel effect exists and * represents indirect effect exists. The pluses and minuses in parentheses indicate the direction of the effect. (a) shows the test results of the spillover mechanism of downward returns risk in international crude oil market; (b) shows the test results of the spillover mechanism of upward returns risk in international crude oil market.*

market risk on CPI and PPI, while the mediating effect of the commodity market only exists in the impact of crude oil market risk on CPI (Ji et al., 2018; Jia et al., 2021).

In the dominant period of oil financial attribute, crude oil market risk will not affect CPI/PPI through financial market risk and commodity market risk. **Table 5** reports the stepwise test results of the channel effect of commodity market risk as well as a financial market risk when the financial attribute of oil is dominant. The first step is fitting 16) to test the impact of international crude oil market risk on CPI/PPI. It can be seen from the table that the impact of international crude oil market risk on CPI/PPI is significantly negative, and the impact on PPI is greater than that on CPI, which indicates that the increase of crude oil market risk can reduce the level of CPI/PPI. The second step is to test the significance of the impact of crude oil market risk on commodity market risk or financial market risk by fitting (17). Similarly, the impact of crude oil market risk on commodity market risk is significantly positive, while the impact on financial market risk is not significant. Finally, the significance of the coefficients is analyzed by fitting **Formula (18)**. It can be seen from the table that the impact of commodity market risk on both CPI and PPI is not significant; the impact of financial market risk on CPI is not significant, and the impact is significant on PPI. Since the impact of international crude oil market risk and financial market risk is not significant in Step two, we need to do the Sobel test for all the mediating effects in the third step. All the test results fall into the acceptance domain, indicating that there is no mediating effect of commodity market risk and financial market risk (Gregorious and Kontonikas 2010; Meng et al., 2020; Jia et al., 2021). According to the stepwise test results, when the financial attribute of oil is dominant, the risk of the crude oil market directly affects CPI/PPI.

During the period dominated by the dual attribute of oil, the impact of crude oil market risk on CPI and PPI is different through financial market risk and commodity market risk (Chen et al., 2020; Meng et al., 2020). **Table 6** reports the stepwise test results of the channel effect of commodity market risk and financial market risk when the dual attribute of oil dominates. The first step fits **Formula (16)** to test the significance of the impact of crude oil market risk on CPI and PPI. As can be seen from the table, the impact of crude oil market risk on CPI and PPI is significantly negative, and the impact on PPI is greater than that on CPI, which indicates that the increase of crude oil market risk can reduce the level of both CPI and PPI. The second step is to test the significance of the impact of crude oil market risks on commodity market risks as well as financial market risks through fitting (17). Similarly, the risk of the crude oil market has a significant positive impact on the commodity market risk as well as the financial market risk, but the impact on the commodity market risk is greater than that on the financial market risk. Finally, **Formula (18)** is fitted to analyze the significance of coefficients. As can be seen from the table, both crude oil market risk and financial market risk have significant impacts on CPI and PPI. In addition, the impact of commodity market risk on CPI/PPI is also significant. After controlling the risk of the commodity market, the impact of crude oil market risk on CPI is significant, but the impact on PPI is not significant. The comprehensive stepwise test results show that when the dual attribute of oil dominates, the financial market risk has a mediating effect, while the mediating effect of the commodity market risk only exists in the impact of the international crude oil market risk on CPI, and the crude oil market risk indirectly affects PPI through the commodity market risk.

Under the full sample, the impacts of crude oil market upward returns risk on CPI/PPI are different through commodity market risk and financial market risk (Ji et al., 2019b). **Table 7** reports the stepwise test results of the risk of the upward returns in the crude oil market for the full sample. The first step is to examine the significance of the impact of the crude oil market risk on CPI/PPI. It can be seen from the table that the impact of the crude oil market upward returns risk on CPI/PPI is significantly negative, and the impact on PPI is greater than that on CPI, which indicates that the increase of international oil market risk can reduce the levels of both CPI and PPI. The second step is to examine the significance of the impact of crude oil market risk on commodity market risk as well as financial market risk. Similarly, the risk of the crude oil market has a significant positive impact on the commodity market risk, but has no significant impact on the financial market risk. Finally, **Formula (3)** is fitted to analyze the significance of coefficients. It can be seen from **Table 7** that financial market risk has a significant impact on both CPI and PPI, while the impact of crude oil market upward returns risk on PPI is not significant. Based on the results of the second step, this paper further uses the Sobel test for the mediating effect of financial market risk in the impact of the crude oil market on CPI/PPI, and the test results fall into the acceptance domain, that is, there is no mediating effect. In addition, the impact of the commodity market risk on CPI/PPI is significant, but after controlling the commodity market risk, the impact of the crude oil market risk on CPI/PPI is not significant, that is, the direct effect does not exist. According to the stepwise test results, there is no mediating effect of financial market risk under the full sample. Crude oil market risks indirectly affect both CPI and PPI through commodity market risks, while direct effects do not exist. Further, this section examines the spillover mechanism of the crude oil market upward returns risk in different oil attribute periods.

When dominated by a single attribute of oil, neither commodity market risk nor financial market risk has a mediating effect. **Table 8** reports the stepwise test results of the spillover mechanism of the international crude oil market upward returns risk during the period dominated by a single attribute of oil. The first step fit **Formula (16)** to test the significance of the impact of the international crude oil market upward returns risk on CPI/PPI. It can be seen from the table that regardless of oil commodity attribute or financial attribute, the impact on CPI/PPI is not significant. Comprehensive stepwise test results show that the risk of upward returns in the international crude oil market will not affect CPI/PPI through commodity market risk and financial market risk, and the mediating effect is not tenable.

During the period dominated by the dual attribute of oil, the impacts of crude oil market risk on both CPI and PPI are different through financial market risk and commodity market risk. **Table 9** reports the stepwise test results of the spillover mechanism of the return rise risk in the crude oil market when the dual attribute of oil dominates. The first step fits **Formula (16)** to test the significance of the impact of crude oil market risk on CPI/PPI. As can be seen from the table, the impact of crude oil market risk on both CPI and PPI is significantly negative, and the impact on PPI is greater than that on CPI, which indicates that the increase of crude oil market risk can reduce the levels of CPI and PPI. The second step is to test the significance of the impact of crude oil market risk on commodity market risk as well as the financial market risk through fitting (17). Similarly, the risk of the crude oil market has a significant positive impact on commodity market risk as well as financial market risk. Finally, **Formula (18)** is fitted to analyze the significance of coefficients. As can be seen from the table, both international oil market risk and financial market risk have significant impacts on CPI and PPI. In addition, the impact of commodity market risk on CPI/PPI is also significant. After controlling the commodity market risk, the impact of crude oil market risk on CPI/PPI is not significant. The comprehensive stepwise test results show that when the dual attribute of oil dominates, the financial market risk has a mediating effect, while the crude oil market risk will indirectly affect CPI/PPI through the commodity market risk.

Under the condition of heterogeneous comprehensive returns, the channel effects are asymmetric for the impact of crude oil market risk on the macro-economic operation. This paper further summarizes the stepwise test results (**Table 10**), compares and analyzes the numerical differences of the channel effects of commodity market risk and financial market risk during different periods dominated by different oil attributes.

The channel effect of commodity market risk is related to the returns trend of the crude oil market. **Table 10** (a) reports the summary of the channel effect test of commodity market risk under different returns trends. When the returns rise, the crude oil market risk indirectly affects CPI/PPI through the commodity market risk, and the direction is negative. When the oil commodity attribute is dominant, the commodity market risk has a positive mediating effect on the crude oil market risk and CPI. When the financial attribute of oil is dominant, there is no mediating effect of commodity market risk. When the dual attribute of oil dominates, crude oil market risk harms CPI through commodity market risk, but a significant negative indirect impact on PPI. When the returns fall, the crude oil market risk indirectly affects CPI/PPI through the commodity market risk when the dual attribute of oil dominates for the full sample. This is mainly due to asset diversification under different returns trends. When the returns rise, investors in the crude oil market have higher expectations, and the market sentiment is better. Enterprises adjust their investment strategies and focus on the demand for a single commodity. When returns fall, the diversity of commodity markets is the main way for market participants to mitigate the impact of major events (Li et al., 2021b). Through purchasing diversified commodities, investors reduce the demand for crude oil, which makes the international crude oil price fall, but increases the prices of other commodities.

Different trends of crude oil market returns have a significant influence on the channel effect of the financial market. As can be seen from **Table 10** (b), in the case of the full sample, the risk of the downward returns of the crude oil market significantly reduces the CPI level through financial market risk, but does not affect the PPI level. Under the domination of oil commodity attributes, the risk of the downward returns of the crude oil market has a significantly positive impact on CPI and PPI through the financial market risk. On the contrary, the financial market risk has a significant negative mediating effect when dominated by the dual attribute of oil. When returns rise, the significant negative mediating effect of financial market risk only plays a role when the dual attribute of oil dominates. This is mainly due to the difference in investor behavior in different returns trends. Compared with the downward returns, the rise in crude oil returns makes it easier for investors to obtain expected profits, enterprises have fewer financing constraints, and the cost of obtaining funds is also lower (Song et al., 2019; Chen et al., 2020). In addition, market uncertainty increases when returns fall, and most investors in the market are risk-prone, which reduces the storage function of financial markets for funds. To sum up, Hypotheses 2a and 3a are true.

4.3 Asymmetry Analysis of the Impact Channel of International Crude Oil Market Risk

This section further analyzes the numerical characteristics of the impact channels of upward and downward return risk of the international crude oil market under different oil attributes. The numerical characteristics of the channel effect of commodity market risk and financial market risk are shown in **Figure 5** and **Figure 6** respectively.

Different attributes of oil have significant impacts on the channel effect of commodity market risk under different returns trends. **Figure 5A,B**) and (b) show the numerical characteristics of the levels of CPI and PPI influenced by the risk of downward return of the crude oil market during the period dominated by different oil attributes through the commodity market risk. **Figure 5C** depicts the numerical characteristics of the levels of CPI and PPI influenced by the risk of the upward return of the crude oil market during the period dominated by dual attribute of oil through the commodity market risk. When the returns fall, under the domination of the oil commodity attribute, the direct effect of international crude oil market risk on CPI is −3.488, and the indirect effect of international crude oil market risk through the channel of commodity market risk is 1.815 (= 0.431 × 4.213). Under the domination of the dual attribute of oil, the direct effect of international crude oil market risk on CPI is 2.863, and the indirect effect through the channel of commodity market risk is −4.519 (= 0.511 × (−8.844)). In addition, the indirect effect of international oil market risk on PPI through commodity market risk is −8.651 (= 0.511 × (−16.93)). When returns rise, the indirect effect of commodity markets risk occurs only when the dual attribute of oil dominates. The indirect effect of international oil market risk on CPI through commodity market risk is −2.007 (= 0.435 × (−4.616)). Besides, the indirect effect of international oil market risk on PPI through commodity market risk is −6.542 (= 0.435 × (−15.04)). The channel effect of commodity market risk is greater when returns fall because investors' expectations lead to different influences on demand changes of different commodities. So Hypothesis 2b is true.

When returns rise, the impact of crude oil market risk on the macroeconomic operation is greater through financial market risk compared with commodity market risk. **Figure 6A,B** show the numerical characteristics of the impact of the risk of downward return of the crude oil market on the levels of CPI and PPI through the financial market in the period dominated by different oil attributes. **Figure 6C** depicts the numerical characteristics of the impact of the risk of upward return of the crude oil market on CPI/PPI through the financial market in the period dominated by the dual attribute of oil. Under the oil commodity attribute, the direct effect of international crude oil market risk on CPI is −1.68, and the indirect effect on CPI through financial market risk is 0.1495 (= 0.092 × 1.625). The direct impact of international oil market risk on PPI is −4.337, and the indirect impact on PPI through financial market risk is 0.351 (= 0.092 × 3.814). Under the domination of dual attribute of oil, the direct impact of international crude oil market risk on CPI is −2.197, and the indirect impact on CPI through financial market risk is −0.212 (= 0.055 × (−3.857)). The direct impact of international crude oil market risk on PPI is −7.121, and the indirect impact on PPI through financial market risk is −0.308 (= 0.055 × (−5.613)). When returns rise, the direct impact of international crude oil market risk on CPI is −4.272, and the indirect effect on CPI through financial market risk is −0.329 (= 0.078 × (−4.226)). The direct impact of international crude oil market risk on PPI is −8.059, and the indirect impact on PPI through financial market risk is −0.602 (= 0.078 × (−7.718)). In conclusion, Hypothesis 3b holds.

5 CONCLUSION

The channels through which crude oil market risk impacts macroeconomic operation are affected by the double effects of returns trend and oil attributes. This paper first uses SVAR to identify the dynamic features of oil attributes domination from July 2003 to October 2020. Secondly, this paper uses a stepwise regression test to analyze the channel effects of commodity market risk and financial market risk. Finally, the paper analyzes the asymmetric characteristics of the channels through which crude oil market risk impacts the smooth operation of the macroeconomy under the condition of returns heterogeneity. The conclusions are as follows:

The commodity market risk and financial market risk snapshot a "buffer" or "magnifier" role on crude oil risk pass-through to macroeconomic during oil dual attributes dominance. When the oil commodity attribute is dominant, the commodity market risk and the financial market risk show the role of "buffer", that is, the direct and indirect effects of the crude oil market risk on the smooth operation of the macro-economy through the commodity market and the financial market risk are opposite. Specifically, the direct effect of crude oil risk on CPI/PPI is negative, while the channel effect of commodity (financial) market risk on oil pass-through to CPI/PPI is positive during oil commodity attribute dominating. When oil financial attribute is dominant, commodity market risk and financial market risk have no channel effect. When dominated by the dual attribute of oil, commodity market risk is an effective channel for the macroeconomic operation to cope with the impact of crude oil market risk, while financial market risk acts as a "magnifier" of crude oil market risk. In other words, when crude oil risk increases, the degree of CPI/PPI could direct decrease. Considering the channel role of commodity (financial) market risk, the increase of crude oil risk would lead to an increase in CPI/PPI.

There are significant asymmetric channel effects of commodity market risk or financial market risk on the relationship between crude oil risk and macroeconomic under different oil return trends. On the one hand, when the returns of crude oil market rise, the channel effect of commodity market risk as well as financial market risk is mainly reflected in the dominant period of oil commodity attribute and dual attribute. When the international crude oil market returns fall, the commodity market risk and the financial market risk play the channel effect in the oil dual attribute dominant period. On the other hand, the risk of falling returns in the international crude oil market has a greater impact on CPI(PPI) through the commodity market risk than when the returns rise, which is 4.519 (8.651) and 2.007 (6.542), respectively. In contrast, the risk of rising returns in the international oil market has a greater impact on CPI/PPI through financial market risk, which is 0.329 (0.602) and 0.212 (0.308), respectively.

Based on the above conclusions, this paper puts forward the following policy suggestions. First, we should pay attention to the mediating role of the commodity market and financial market in preventing the impact of crude oil market risk in the dominant period of different oil attributes. Counter-cyclical regulation and other policies can be adopted to prevent and defuse the impact of crude oil market risks on the macroeconomy. The second is to pay attention to both the impact of the risk of the downward return and the upward returns risk of the crude oil market. The downward returns risk in the international crude oil market is a key factor affecting the operation of the commodity market, the financial market and the macroeconomy. However, the risk of the upward return provides more possibilities for investors to obtain expectations and also causes information transmission between markets.

This paper bears several limitations. Despite we presented the channel effect of commodity market risk or financial market risk on crude oil risk pass-through to macroeconomic, this paper neglects shocks of major events to the crude oil market. Thus, we could further explore the effect of structural breaks in the crude oil market on macroeconomic stability. Moreover, further studies about the moderating effect of commodity market risk or financial market risk and the mixture of mediating with moderating effect on crude oil risk pass-through to macroeconomic could be regarded as a valuable area.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

SJ, HD, and ZW contributed to conception and design of the study. HD and ZW organized the database. SJ, HD, and ZW performed the statistical analysis. SJ, HD, and ZW wrote the first draft of the manuscript. SJ, HD, and ZW wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This work was supported by National Office for Philosophy and Social Sciences (No. 21CTJ013) and Humanities and Social Sciences Research Ministry of Education (No.18YJC910008).

REFERENCES

Alvarez, L. J., Hurtado, S., Sanchez, I., and Thomas, C. (2011). The Impact of Oil Price Changes on Spanish and Euro Area Consumer Price Inflation. *Econ. Model.* 28 (1), 422–431. doi:10.1016/j.econmod.2010.08.006

Baron, R. M., and Kenny, D. A. (1986). The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations. *J. Personal. Soc. Psychol.* 51, 1173–1182. doi:10.1037/0022-3514.51.6.1173

Bentler, P. M. (1980). Multivariate Analysis with Latent Variables: Causal Modeling. *Annu. Rev. Psychol.* 31 (1), 419–456. doi:10.1146/annurev.ps.31.020180.002223

Bernardi, M., and Catania, L. (2016). Comparison of Value-At-Risk Models Using the MCS Approach. *Comput. Stat.* 31 (2), 579–608. doi:10.1007/s00180-016-0646-6

Bloch, H., Rafiq, S., and Salim, R. (2015). Economic Growth with Coal, Oil and Renewable Energy Consumption in China: Prospects for Fuel Substitution. *Econ. Model.* 44, 104–115. doi:10.1016/j.econmod.2014.09.017

Chen, H., Liu, L., Wang, Y., and Zhu, Y. (2016). Oil price Shocks and U.S. Dollar Exchange Rates. *Energy* 112, 1036–1048. doi:10.1016/j.energy.2016.07.012

Chen, S., Ouyang, S., and Dong, H. (2020). Oil price Pass-Through into Consumer and Producer Prices with Monetary Policy in China: Are There Non-linear and Mediating Effects. *Front. Energ. Res.* 8, 35. doi:10.3389/fenrg.2020.00035

Cheng, X., He, L., Lu, H., Chen, Y., and Ren, L. (2016). Optimal Water Resources Management and System Benefit for the Marcellus Shale-Gas Reservoir in Pennsylvania and West Virginia. *J. Hydrol.* 540, 412–422. doi:10.1016/j.jhydrol.2016.06.041

Choi, S., Furceri, D., Loungani, P., Mishra, S., and Poplawski-Ribeiro, M. (2018). Oil Prices and Inflation Dynamics: Evidence from Advanced and Developing Economies. *J. Int. Money Finance* 82, 71–96. doi:10.1016/j.jimonfin.2017.12.004

Cong, R. G., and Shen, S. (2013). Relationships Among Energy price Shocks, Stock Market, and the Macroeconomy: Evidence from China. *ScientificWorldJournal* 2013, 171868. doi:10.1155/2013/171868

Coronado, S., Jimenez-Rodriguez, R., and Rojas, O. (2018). An Empirical Analysis of the Relationships between Crude Oil, Gold and Stock Markets. *Energ. J.* 39, 193–207. doi:10.5547/01956574.39.si1.scor

Dong, H., Li, Z., and Failler, P. (2020). The Impact of Business Cycle on Health Financing: Subsidized, Voluntary and Out-Of-Pocket Health Spending. *Ijerph* 17 (6), 1928. doi:10.3390/ijerph17061928

Dong, H., Liu, Y., Liu, Y., and Chang, J. (2019). The Heterogeneous Linkage of Economic Policy Uncertainty and Oil Return Risks. *Green. Finance* 1 (1), 46–66. doi:10.3934/gf.2019.1.46

Engle, R. F., and Manganelli, S. (2004). CAViaR. *J. Business Econ. Stat.* 22 (4), 367–381. doi:10.1198/073500104000000370

Fan, P., Deng, R., QiuZhao, J. Z., Zhao, Z., and Wu, S. (2021). Well Logging Curve Reconstruction Based on Kernel ridge Regression. *Arab J. Geosci.* 14, 1559. doi:10.1007/s12517-021-07792-y

Ferraty, F., and Quintela-Del-Río, A. (2016). Conditional VaR and Expected Shortfall: a New Functional Approach. *Econometric Rev.* 35 (2), 263–292. doi:10.1080/07474938.2013.807107

Ghassan, H. B., and AlHajhoj, H. R. (2016). Long Run Dynamic Volatilities between OPEC and Non-OPEC Crude Oil Prices. *Appl. Energ.* 169, 384–394. doi:10.1016/j.apenergy.2016.02.057

Gkillas, K., and Katsiampa, P. (2018). An Application of Extreme Value Theory to Cryptocurrencies. *Econ. Lett.* 164, 109–111. doi:10.1016/j.econlet.2018.01.020

Gong, X., and Lin, B. (2018). Time-varying Effects of Oil Supply and Demand Shocks on China's Macro-Economy. *Energy* 149, 424–437. doi:10.1016/j.energy.2018.02.035

González-Concepción, C., Gil-Fariña, M. C., and Pestano-Gabino, C. (2018). Wavelet Power Spectrum and Cross-Coherency of Spanish Economic Variables. *Empir Econ.* 55, 855–882. doi:10.1007/s00181-017-1295-5

Grace Saculsan, P., Kanamura, T., and Kanamura, T. (2020). Examining Risk and Return Profiles of Renewable Energy Investment in Developing Countries: the Case of the Philippines. *Green. Finance* 2 (2), 135–150. doi:10.3934/gf.2020008

Gregoriou, A., and Kontonikas, A. (2010). The Long-Run Relationship between Stock Prices and Goods Prices: New Evidence from Panel Cointegration. *J. Int. Financial Markets, Institutions Money* 20 (2), 166–176. doi:10.1016/j.intfin.2009.12.002

Guo, J., Zheng, X., and Chen, Z.-M. (2016). How Does Coal price Drive up Inflation? Reexamining the Relationship between Coal price and General price Level in China. *Energ. Econ.* 57, 265–276. doi:10.1016/j.eneco.2016.06.001

Hamilton, J. D. (1983). Oil and the Macroeconomy since World War II. *J. Polit. Economy* 91, 228–248. doi:10.1086/261140

He, Y., and Lin, B. (2019). Regime Differences and Industry Heterogeneity of the Volatility Transmission from the Energy price to the PPI. *Energy* 176, 900–916. doi:10.1016/j.energy.2019.04.025

Hewitt, R. J., Bradley, N., Baggio Compagnucci, A., Barlagne, C., Ceglarz, A., Cremades, R., et al. (2019). Social Innovation in Community Energy in Europe: a Review of the Evidence. *Front. Energ. Res.* 7, 31. doi:10.3389/fenrg.2019.00031

Hu, Q., Li, T., Li, X., and Dong, H. (2021). Dynamic Characteristics of Oil Attributes and Their Market Effects. *Energies* 14, 3927. doi:10.3390/en14133927

Huang, Z., Liao, G., and Li, Z. (2019). Loaning Scale and Government Subsidy for Promoting green Innovation. *Technol. Forecast. Soc. Change* 144, 148–156. doi:10.1016/j.techfore.2019.04.023

Humbatova, S. İ. Q., Garayev, A. I. O., Tanriverdiev, S. M. O., and Hajiyev, N. Q.-O. (2019). Analysis of the Oil, price and Currency Factor of Economic Growth in Azerbaijan. *Jesi* 6, 1335–1353. doi:10.9770/jesi.2019.6.3(20)

Ji, Q., Bouri, E., Roubaud, D., and Kristoufek, L. (2019b). Information Interdependence Among Energy, Cryptocurrency and Major Commodity Markets. *Energ. Econ.* 81, 1042–1055. doi:10.1016/j.eneco.2019.06.005

Ji, Q., Bouri, E., Roubaud, D., and Shahzad, S. J. H. (2018). Risk Spillover between Energy and Agricultural Commodity Markets: a Dependence-Switching Covar-Copula Model. *Energ. Econ.* 75, 14–27. doi:10.1016/j.eneco.2018.08.015

Ji, Q., Liu, B.-Y., and Fan, Y. (2019a). Risk Dependence of CoVaR and Structural Change between Oil Prices and Exchange Rates: A Time-Varying Copula Model. *Energ. Econ.* 77, 80–92. doi:10.1016/j.eneco.2018.07.012

Jia, S., Dong, H., and Yang, H. (2021). Asymmetric Risk Spillover of the International Crude Oil Market in the Perspective of Crude Oil Dual Attributes. *Front. Environ. Sci.* 9, 720278. doi:10.3389/fenvs.2021.720278

Kang, W., Ratti, R. A., and Vespignani, J. L. (2017). Oil price Shocks and Policy Uncertainty: New Evidence on the Effects of Us and Non-us Oil Production. *Energ. Econ.* 66, 536–546. doi:10.1016/j.eneco.2017.01.027

Kilian, L. (2009). Not All Oil Price Shocks Are Alike: Disentangling Demand and Supply Shocks in the Crude Oil Market. *Am. Econ. Rev.* 99 (3), 1053–1069. doi:10.1257/aer.99.3.1053

Kose, N., Emirmahmutoglu, F., and Aksoy, S. (2012). The Interest Rate-Inflation Relationship under an Inflation Targeting Regime: The Case of Turkey. *J. Asian Econ.* 23 (4), 476–485. doi:10.1016/j.asieco.2012.03.001

Li, Z., Chen, L., and Dong, H. (2021b). What Are Bitcoin Market Reactions to Its-Related Events? *Int. Rev. Econ. Finance.*

Li, Z., Dong, H., Dong, H., Huang, Z., and Failler, P. (2018). Asymmetric Effects on Risks of Virtual Financial Assets (VFAs) in Different Regimes: A Case of Bitcoin. *Quantitative Finance Econ.* 2 (4), 860–883. doi:10.3934/qfe.2018.4.860

Li, Z., Dong, H., Floros, C., Charemis, A., and Failler, P. (2021a). Re-examining Bitcoin Volatility: A CAViaR-Based Approach. *Emerging Markets Finance and Trade.* doi:10.1080/1540496x.2021.1873127

Li, Z., Wang, Y., and Huang, Z. (2020). Risk Connectedness Heterogeneity in the Cryptocurrency Markets. *Front. Phys.* 8, 243. doi:10.3389/fphy.2020.00243

Li, Z., and Zhong, J. (2020). Impact of Economic Policy Uncertainty Shocks on China's Financial Conditions. *Finance Res. Lett.* 35, 101303. doi:10.1016/j.frl.2019.101303

Long, S., and Liang, J. (2018). Asymmetric and Nonlinear Pass-Through of Global Crude Oil price to China's PPI and CPI Inflation. *Econ. Research-Ekonomska Istraživanja* 31, 240–251. doi:10.1080/1331677x.2018.1429292

Loutia, A., Mellios, C., and Andriosopoulos, K. (2016). Do OPEC Announcements Influence Oil Prices? *Energy Policy* 90, 262–272. doi:10.1016/j.enpol.2015.11.025

Meng, J., Nie, H., Mo, B., and Jiang, Y. (2020). Risk Spillover Effects from Global Crude Oil Market to China's Commodity Sectors. *Energy* 202, 117208. doi:10.1016/j.energy.2020.117208

Meng, X., and Taylor, J. W. (2018). An Approximate Long-Memory Range-Based Approach for Value at Risk Estimation. *Int. J. Forecast.* 34 (3), 377–388. doi:10.1016/j.ijforecast.2017.11.007

Nisticò, S. (2012). Monetary Policy and Stock-price Dynamics in a DSGE Framework. *J. Macroeconomics* 34, 126–146. doi:10.1016/j.jmacro.2011.09.008

Oleg, S., and Ekaterina, V. (2020). Financial and Non-financial Investments: Comparative Econometric Analysis of the Impact on Economic Dynamics. *Quantitative Finance Econ.* 4 (3), 382–411.

Qureshi, S. M., and Kang, C. (2015). Analysing the Organizational Factors of Project Complexity Using Structural Equation Modelling. *Int. J. Project Manag.* 33 (1), 165–176. doi:10.1016/j.ijproman.2014.04.006

Raheem Ahmed, R., Vveinhardt, J., Štreimikienė, D., Ghauri, S. P., and Ahmad, N. (2017). Estimation of Long-Run Relationship of Inflation (Cpi & Wpi), and Oil Prices with Kse-100 Index: Evidence from Johansen Multivariate Cointegration Approach. *Technol. Econ. Dev. Economy* 23, 567–588. doi:10.3846/20294913.2017.1289422

Raheem, I. D., Bello, A. K., and Agboola, Y. H. (2020). A New Insight into Oil price-inflation Nexus. *Resour. Pol.* 68, 101804. doi:10.1016/j.resourpol.2020.101804

Ratti, R. A., and Vespignani, J. L. (2016). Oil Prices and Global Factor Macroeconomic Variables. *Energ. Econ.* 59, 198–212. doi:10.1016/j.eneco.2016.06.002

Razmi, F., Azali, M., Chin, L., and Shah Habibullah, M. (2016). The Role of Monetary Transmission Channels in Transmitting Oil price Shocks to Prices in ASEAN-4 Countries during Pre- and post-global Financial Crisis. *Energy* 101, 581–591. doi:10.1016/j.energy.2016.02.036

Saeed, S. J., and Ridoy, D. N. (2020). Covid-19, Oil price and UK Economic Policy Uncertainty: Evidence from the ARDL Approach. *Quantitative Finance Econ.* 4 (3), 503–514.

Sek, K. (2019). Effect of Oil Price Pass-Through on Domestic Price Inflation: Evidence from Nonlinear ARDL Models. *Panoeconomicus* 66 (1), 69–91. doi:10.2298/pan160511021s

Sek, S. K. (2017). Impact of Oil price Changes on Domestic price Inflation at Disaggregated Levels: Evidence from Linear and Nonlinear ARDL Modeling. *Energy* 130, 204–217. doi:10.1016/j.energy.2017.03.152

Shi, X., and Sun, S. (2017). Energy price, Regulatory price Distortion and Economic Growth: a Case Study of China. *Energ. Econ.* 63, 261–271. doi:10.1016/j.eneco.2017.02.006

Smets, F., and Peersman, G. (2001). The Monetary Transmission Mechanism in the Euro Area: More Evidence from Var Analysis. Working Paper No. 91.

Sobel, M. E. (1982). "Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models," in *Sociological Methodology*. Editor S. Leinhardt (Washington, DC: American Sociological Association), 13, 290–312. doi:10.2307/270723

Sodeyfi, S., and Katircioglu, S. (2016). Interactions between Business Conditions, Economic Growth and Crude Oil Prices. *Econ. Research-Ekonomska Istraživanja* 29, 980–990. doi:10.1080/1331677x.2016.1235504

Song, Y., Ji, Q., Du, Y.-J., and Geng, J.-B. (2019). The Dynamic Dependence of Fossil Energy, Investor Sentiment and Renewable Energy Stock Markets. *Energ. Econ.* 84, 104564. doi:10.1016/j.eneco.2019.104564

Takhtamanova, Y. F. (2010). Understanding Changes in Exchange Rate Pass-Through. *J. Macroeconomics* 32, 1118–1130. doi:10.1016/j.jmacro.2010.04.004

Tang, K., and Xiong, W. (2010). Index Investment and Financialization of Commodities. NBER Working Paper No. w16385.

Tillmann, P. (2008). Do interest Rates Drive Inflation Dynamics? an Analysis of the Cost Channel of Monetary Transmission. *J. Econ. Dyn. Control.* 32 (9), 2723–2744. doi:10.1016/j.jedc.2007.10.005

Wang, H., Han, Y., Fidrmuc, J., and Wei, D. (2021). Confucius Institute, Belt and Road Initiative, and Internationalization. *Int. Rev. Econ. Finance* 71, 237–256. doi:10.1016/j.iref.2020.09.011

Wei, Y. (2019). The Relationship between Oil and Non-oil Commodity Prices and China's PPI and CPI: an Empirical Analysis. *Energ. Sourc. B: Econ. Plann. Pol.* 14, 125–146. doi:10.1080/15567249.2019.1630032

Wen, F., Min, F., Zhang, Y. J., and Yang, C. (2019b). Crude Oil price Shocks, Monetary Policy, and China's Economy. *Int. J. Fin Econ.* 24, 812–827. doi:10.1002/ijfe.1692

Wen, F., Zhao, Y., Zhang, M., and Hu, C. (2019a). Forecasting Realized Volatility of Crude Oil Futures with Equity Market Uncertainty. *Appl. Econ.* 51, 6411–6427. doi:10.1080/00036846.2019.1619023

Wu, B., Fang, H., Jacoby, G., Li, G., and Wu, Z. (2021). Environmental Regulations and Innovation for Sustainability? Moderating Effect of Political Connections. *Emerging Markets Rev.*

Wulandari, R. (2012). Do Credit Channel and Interest Rate Channel Play Important Role in Monetary Transmission Mechanism in Indonesia? A Structural Vector Autoregression Model. *Proced. - Soc. Behav. Sci.* 65, 557–563. doi:10.1016/j.sbspro.2012.11.165

Xiao, N., Xinyi, R., Xiong, Z., Xu, F., Zhang, X., Xu, Q., et al. (2021). A Diversity-Based Selfish Node Detection Algorithm for Socially Aware Networking. *J. Sign Process. Syst.* 93 (7), 811–825. doi:10.1007/s11265-021-01666-y

Xie, W., Zhang, R., Zeng, D., Shi, K., and Zhong, S. (2020). Strictly Dissipative Stabilization of Multiple-Memory Markov Jump Systems with General Transition Rates: A Novel Event-Triggered Control Strategy. *Int. J. Robust Nonlinear Control.*

Yang, Z., and Zhou, Y. (2017). Quantitative Easing and Volatility Spillovers across Countries and Asset Classes. *Manag. Sci.* 63, 333–354. doi:10.1287/mnsc.2015.2305

Zhang, Y.-J., Chevallier, J., and Guesmi, K. (2017). "De-financialization" of Commodities? Evidence from Stock, Crude Oil and Natural Gas Markets. *Energ. Econ.* 68, 228–239. doi:10.1016/j.eneco.2017.09.024

Zhao, C., Liu, X., Zhong, S., Shi, K., Liao, D., and Zhong, Q. (2020b). Secure Consensus of Multi-Agent Systems with Redundant Signal and Communication Interference via Distributed Dynamic Event-Triggered Control. *ISA Trans.*

Zhao, C., Zhong, S., Zhang, X., Zhong, Q., and Shi, K. (2020a). Novel Results on Nonfragile Sampled-Data Exponential Synchronization for Delayed Complex Dynamical Networks. *Int. J. Robust Nonlinear Control.*

Zhao, C., Zhong, S., Zhong, Q., and Shi, K. (2020c). Synchronization of Markovian Complex Networks with Input Mode Delay and Markovian Directed Communication via Distributed Dynamic Event-Triggered Control. *Nonlinear Anal. Hybrid Syst.* 36, 100883. doi:10.1016/j.nahs.2020.100883

Zheng, Y., and Du, Z. (2019). A Systematic Review in Crude Oil Markets: Embarking on the Oil price. *Green. Finance* 1, 328–345. doi:10.3934/gf.2019.3.328

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Energy Research

ORIGINAL RESEARCH
published: 07 January 2022
doi: 10.3389/fenrg.2021.772508

A Hybrid Model for Power Consumption Forecasting Using VMD-Based the Long Short-Term Memory Neural Network

*Yingjun Ruan, Gang Wang, Hua Meng and Fanyue Qian**

School of Mechanical Engineering, Tongji University, Shanghai, China

Energy consumption prediction is a popular research field in computational intelligence. However, it is difficult for general machine learning models to handle complex time series data such as building energy consumption data, and the results are often unsatisfactory. To address this difficulty, a hybrid prediction model based on modal decomposition was proposed in this paper. For data preprocessing, the variational mode decomposition (VMD) technique was used to used to decompose the original sequence into more robust subsequences. In the feature selection, the maximum relevance minimum redundancy (mRMR) algorithm was chosen to analyse the correlation between each component and the individual features while eliminating the redundancy between individual features. In the forecasting module, the long short-term memory (LSTM) neural network model was used to predict power consumption. In order to verify the performance of the proposed model, three categories of contrast methods were applied: 1) Comparing the hybrid model to a single predictive model, 2) Comparing the hybrid model with the backpropagation neural network (BPNN) to the hybrid model with the LSTM and 3) Comparing the hybrid model using mRMR and the hybrid model using mutual information maximization (MIM). The experimental results on the measured data of an office building in Qingdao show that the proposed hybrid model can improve the prediction accuracy and has better robustness compared to VMD-MIM-LSTM. In the three control groups mentioned above, the R^2 value of the hybrid model improved by 10, 3 and 3%, respectively, the values of the mean absolute error (MAE) decreased by 48.9, 41.4 and 35.6%, respectively, and the root mean square error (RMSE) decreased by 54.7, 35.5 and 34.1%, respectively.

Keywords: load forecasting, variational mode decomposition, feature selection, machine learning, deep learning

OPEN ACCESS

Edited by:
Jian Chai,
Xidian University, China

Reviewed by:
Zheng Qian,
Beihang University, China
Gabriel Mendonça Paiva,
Universidade Federal de Goiás, Brazil

****Correspondence:***
Fanyue Qian
qianfanyue91@163.com

Specialty section:
This article was submitted to Smart Grids, a section of the journal Frontiers in Energy Research

Received: *08 September 2021*
Accepted: *06 December 2021*
Published: *07 January 2022*

Citation:
Ruan Y, Wang G, Meng H and Qian F (2022) A Hybrid Model for Power Consumption Forecasting Using VMD-Based the Long Short-Term Memory Neural Network. Front. Energy Res. 9:772508. doi: 10.3389/fenrg.2021.772508

1 INTRODUCTION

Energy is critical in modern society, and energy consumption is a major issue that has long plagued humanity. Increasing demand for energy is gradually drawing attention to energy conservation issues around the world. Among energy sources, building electricity consumption accounts for a large proportion of total social energy consumption. From a global perspective, building energy consumption accounts for about 40% of the global energy consumption, and this proportion is likely to increase in the future.

Scientists have explored various methods for predicting building electricity consumption, aiming to achieve intelligent energy management and energy-saving building reconstruction based on predicted energy consumption. However, building electricity forecasting continues to be a challenging effort due to the variety of factors that affect energy consumption, such as building structure, equipment, weather conditions, and energy-use behaviours of the building occupants.

Building electricity consumption predictions can be divided into three methods according to the type of data input and processing method used: White-box physics-based models, grey-box reduced-order models and black-box data-driven models.

White-box physics-based models rely on thermodynamic rules for detailed energy modelling and analysis. The construction of the physical model requires a large number of physical parameters related to the building and a detailed setting of the system operation. Its accuracy depends on the input parameters and the selected simulation software. Zhu et al. (Zhu et al., 2012; Said, 2016) compared the Dest, Energy Plus and DOE-2 simulation software calculation methods, and their research results showed that the difference of load between the simulation results of Dest and Energy Plus was less than 10%. However, some detailed architectural data may not be readily available to researchers, resulting in an inability to provide accurate inputs and thus leading to poor predictive performance.

Grey-box modelling approaches offer a combination of physical and data-driven prediction models, leveraging the advantages and minimizing the disadvantages of both approaches. In grey-box models, some internal parameters and equations are physically interpretable (Eom et al., 2012; Amasyali and El-Gohary, 2018). Grey-box models may also show better performance compared to black-box and white-box models. For example, Dong et al. (Dong et al., 2016) developed a hybrid model which coupled a data-driven model and a thermal network model for predicting the total energy consumption of residential areas and compared its prediction performance to artificial neural networks (ANN), support vector machines (SVM) and least square support vector machine (LSSVM)-based models.

Unlike physical models, black-box data-driven models do not require detailed building data, but rather they learn from the available historical data to make predictions. Common machine learning algorithms include SVM and ANN. These algorithms have a wide range of applications in the field of energy consumption prediction. Currently, about 47% of studies use ANN to predict energy consumption (Liu et al., 2019). For example, Mansoor et al. (Muhammad et al., 2020) compared two different neural network models, feed-forward neural networks (FFNN) and echo state networks (ESN) for electrical load forecasting in real commercial buildings; their results indicated that the ESN model generally performed slightly better than the FFNN model Katarina. Liu et al. (Liu et al., 2020) proposed a hybrid forecasting model that combined the Jaya algorithm and SVM. In this model, the representative features of the input data were selected and the hyper-parameters of SVM were optimized by using the Jaya optimization algorithm to efficiently improve the forecasting accuracy of wind speed. Mendonça et al. (de Paiva et al., 2020) investigated the application of machine learning models for solar radiation intensity prediction. They evaluated multigene genetic programming (MGGP) and the multilayer perceptron (MLP) ANN. The results showed that MGGP produced better results in the case of a single prediction, while ANN presented more accurate results for ensemble forecasting. Anderso et al. (Marcello Anderson et al., 2017) applied portfolio theory to solar and wind energy forecasting to improve resource forecasting for specific solar and wind energy conditions in the Brazilian region. Their study showed that the optimal combination of 30% solar and 70% wind resources generated the smallest calculated standard deviation.

However, the original time series were often unstable due to the disturbance of uncertainty. For this type of data, a single model did not produce excellent results (He et al., 2018). To improve the prediction accuracy, the segregation of these series with different frequencies from the energy data was considered as a possible solution.

Empirical mode decomposition (EMD) was proposed by Dr Norden E. Huang in 1998 (Huang Norden et al., 1998) as a method for processing nonstationary signals; it is an adaptive time-frequency localization analysis method, the number of decomposed IMFs depends on the data itself. Liu et al. (Liu et al., 2012) proposed a standard hybridization of EMD with the backpropagation neural network (BPNN) method. In this study, all intrinsic mode functions (IMFs) and the residue were forecasted with BPNN models. Similarly, Guo et al. (Guo et al., 2011) proposed a modified EMD–FFNN model in the form of an EMD-based FFNN ensemble learning paradigm. This study showed that the first IMF containing high-frequency components was mostly unsymmetrical and disordered, which led to the generation of large forecasting disturbances. The simplest combinations of hybrid EMD–SVM models are presented in the literature (Lin and Peng, 2011; Zhang et al., 2015). These models decomposed wind data into a series of components (IMFs) using EMD, and then different models were built with various kernel functions and parameters for each component using the SVM model.

However, the IMF components obtained by EMD often exhibit mode mixing, resulting in inaccurate IMF components. To solve this problem, many scholars have proposed improved algorithms. Wu and Huang (Wu and Huang, 2009) suggested the ensemble empirical mode decomposition (EEMD) method. Numerous articles in distinct research areas have claimed the superior performance of the EEMD method over hybrid EMD models. The hybrid EEMD–SVM model has been used in the literature and has achieved better prediction accuracy than other models (Hu et al., 2013; Wu et al., 2018). In one work (Wu et al., 2018), wind speed data was decomposed into seven IMF components with EEMD and then the IMFs were predicted using the appropriate SVM models. Elsewhere, a similar approach was used in which the first IMF (IMF1) was removed from the prediction analysis and all remaining IMFs were forecasted with SVM models (Hu et al., 2013). Yu et al. (Wu et al., 2018) proposed a novel model based on EEMD and LSTM for crude oil price forecasting. In this study, a method to select the same number of proper inputs in various decomposition scenarios was developed. To extract features from the selected components more adequately, LSTM was introduced as a forecasting method to predict price movement directly. Dragomiretskiy and Zosso (Dragomiretskiy and Zosso, 2014) introduced the variational mode decomposition (VMD) method in 2014. The VMD algorithm is more robust in that it inherits the advantages of the EMD algorithm while solving the mode mixing problem of the EMD algorithm. In recent years, the VMD algorithm has been successfully applied in many fields, such as fault diagnosis research (Zhang et al., 2017) and forecast research (Liu et al., 2018; Niu et al., 2020). The studies of He (He et al., 2019) and Li (Li et al., 2018) have shown that the combination model based on "decomposition-prediction" can achieve high prediction accuracy in heating and cooling seasons. He et al. developed a VMD-LSTM forecasting model for electricity load forecasting in Hubei province. They divided the 1-year data into four parts, corresponding to four seasons. The results show that the proposed forecasting model has high forecasting accuracy on all four data sets. The lowest prediction accuracy is found in summer, attributed to the higher fluctuation and uncertainty of load in summer.

Studies using signal decomposition methods have some shortcomings. Firstly, some literature uses different prediction methods for different IMF frequencies while ignoring the feature selection variability of IMFs. Secondly, it is difficult to provide a reasonable explanation for the physical meaning of each component using signal decomposition methods. To address these inadequacies, a hybrid system was developed that comprises three modules to predict the electricity load of public buildings in Qingdao. Compared with existing studies on short-term load forecasting, the main contributions of this paper are as follows:

1) A novel deep learning-based method for predicting building electricity consumption is proposed. The idea of "decomposition–reconstruction–integration" results in a feasible and efficient method to model and forecast nonlinear, non-stationary, complex time series.
2) Due to the volatility and uncertainty of the load data, VMD is used to decompose the raw load into more stable series. Most of the literature does not detail the determination of the number of VMD components (Sun et al., 2019b). In this paper, the mean value of the instantaneous frequency of each component is used to determine the number of K.
3) Most of the literature does not provide a reasonable interpretation of the components decomposed by the modal decomposition algorithm. In this paper, the highly volatile load is decomposed into several subsequences by VMD. The redundancy between features is removed by the mRMR algorithm so that each subsequence has a suitable feature. With the features selected by mRMR, this paper attempts to analyse the physical meaning of each subsequence.

This study is organized as follows. **Section 2** outlines the principles of the methods related to the proposed hybrid system. In addition, a case study is presented in **Section 3**. Finally, the study's conclusions and avenues for future work are presented in **Section 4**. Note that the data decomposition and feature selection were performed on a laptop with an Intel(R) i5-7400 CPU with MATLAB 2020a installed, and the deep learning model was performed on a laptop with an Intel(R) i5-7400 CPU with Python 3.8 installed.

2 METHODOLOGY

The main contents of this section introduce the algorithm used in this paper: Variational mode decomposition (VMD), Max-Relevance and Min-Redundancy (mRMR) and Long Short-Term Memory Neural Network (LSTM). The flow chart of the hybrid model is shown in **Figure 1**.

2.1 Principle of Variational Mode Decomposition

2.1.1 VMD Principle

In order to solve the modal mixing problem existing in EMD, Dragomiretskiy et al. (Dragomiretskiy and Zosso, 2014) proposed the VMD algorithm, which is essentially a set of adaptive Wiener filter sets. The decomposition number K of VMD is determined artificially. Theoretically, if the value of K is more reasonable, it can effectively suppress the modal mixing phenomenon. the main process of VMD is divided into five steps:

1) Suppose u_k is the Kth order mode of the original signal f and $\delta(t)$ is a Dirac distribution. The analytic signal of the mode u_k is calculated by Hilbert transform, then its unilateral frequency spectrum can be expressed as:

$$\left(\delta(t)+\frac{j}{\pi t}\right)^{\star} u_k(t) \tag{1}$$

2) Adding a pre-estimated center frequency to the resolved signal of the mode, the frequency of the mode can be modulated to the corresponding baseband:

$$\left[\left(\delta(t)+\frac{j}{\pi t}\right)^{\star} u_k(t)\right] e^{-jw_k t} \tag{2}$$

3) Calculating the bandwidth of each modal signal, the constrained optimization problem is expressed as:

$$\min_{\{u_k\},\{w_k\},}\left\{\sum_{k=1}^{k}\left\|\sigma_t\left[\left(\delta(t)+\frac{j}{\pi t}\right)^{\star} u_k(t)\right]e^{-jw_k t}\right\|_2^2\right\} \tag{3}$$

where the constraint of **Eq. 3** is: $\sum_{k=1}^{k} u_k(t) = f(t)$.

4) The Lagrangian function $\lambda(t)$ and quadratic penalty factor α are introduced to solve the optimal solution of the constrained problem and transform the constrained optimization problem into an unconstrained optimization problem.

$$\begin{aligned}L[\{u_k\},\{w_k\},\lambda] = \alpha\sum_{k=1}^{k}\left\|\sigma_t\left[\left(\delta(t)+\frac{j}{\pi t}\right)^{\star} u_k(t)\right]e^{-jw_k t}\right\|_2^2 \\ +\left\|f(t)-\sum_{k=1}^{k}u_k(t)\right\|_2^2 + \left\langle\lambda(t), f(t)-\sum_{k=1}^{k}u_k(t)\right\rangle\end{aligned} \tag{4}$$

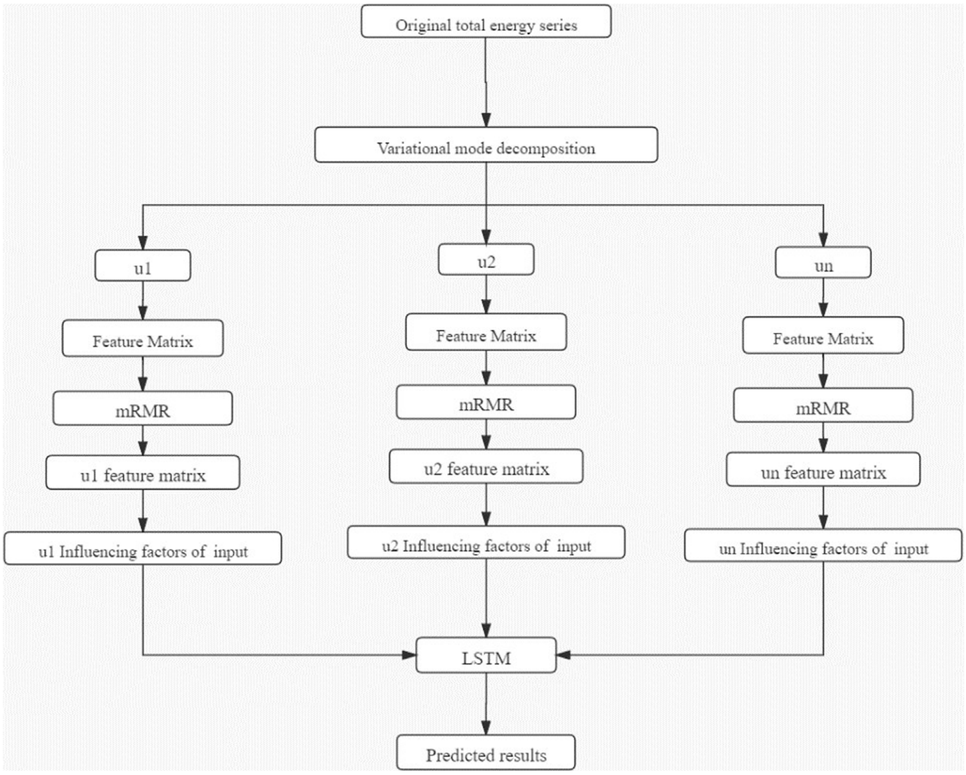

Original total energy series
Variational mode decomposition
u1
u2
un
Feature Matrix
Feature Matrix
Feature Matrix
mRMR
mRMR
mRMR
u1 feature matrix
u2 feature matrix
un feature matrix
u1 Influencing factors of input
u2 Influencing factors of input
un Influencing factors of input
LSTM
Predicted results

FIGURE 1 | Hybrid model flow chart.

5) Use the multiplicative operator alternating direction method to update u_k^{n+1}, w_k^{n+1}, λ^{n+1} alternately in both directions until the following iteration conditions are satisfied:

$$\sum_k \left\| \hat{u}_k^{n+1} - \hat{u}_k^{n} \right\|_2^2 \Big/ \left\| \hat{u}_k^{n} \right\|_2^2 < \varepsilon \tag{5}$$

Where $\varepsilon > 0$, u_k^{n+1}, w_k^{n+1}, λ^{n+1} are denoted as:

$$\hat{u}_k^{n+1}(w) = \frac{\hat{f}(\omega) - \sum_{i<k} \hat{u}_k^{n+1}(\omega) - \sum_{i>k} \hat{u}_k^{n}(\omega) + \frac{\hat{\lambda}^n(w)}{2}}{1 + 2\alpha\left(\omega - \omega_k^n\right)^2} \tag{6}$$

$$\omega_k^{n+1} = \frac{\int_0^\infty \omega \left| \hat{u}_k^{n+1}(w) \right|^2 d\omega}{\int_0^\infty \left| \hat{u}_k^{n+1}(w) \right|^2 d\omega} \tag{7}$$

$$\hat{\lambda}^{n+1}(\omega) = \hat{\lambda}^{n}(\omega) + \tau\left(\hat{f}(w) - \hat{u}_k^{n+1}(\omega)\right) \tag{8}$$

where τ is the updated noise parameter.

2.1.2 VMD parameter determination

1) Modal Number

The number of modalities K should be determined before the VMD is used to decompose. Too large or too small a value of K will affect the accuracy of the model. In this paper, the mean value of instantaneous frequency of each component is used to determine the number of K. When the value of k is too large, and the high-frequency component will be broken. It means that the instantaneous frequency at the break of the high-frequency component is 0. As a result, the high-frequency component breaks lead to a decrease in the average instantaneous frequency. **Figure 2** shows the mean values of instantaneous frequencies for the nine cases of VMD components. It can be seen from the figure that the number of VMD components increases to a certain number, and the curve has an obvious bending phenomenon. To sum up, the value of K is chosen as 4.

2) Penalty Factor

The penalty factor changes the constrained variational problem into a non-constrained variational problem. According to Ref. (Wu, 2016), when the value of the penalty factor is set to 2000 has strong adaptability and can ensure a certain convergence speed.

2.2 Principle of Max-Relevance and Min-Redundancy

Peng et al.(Peng et al., 2005) proposed a feature selection method based on Mutual Information, which uses Mutual Information to measure the dependency between two variables while taking into account the redundancy between features.

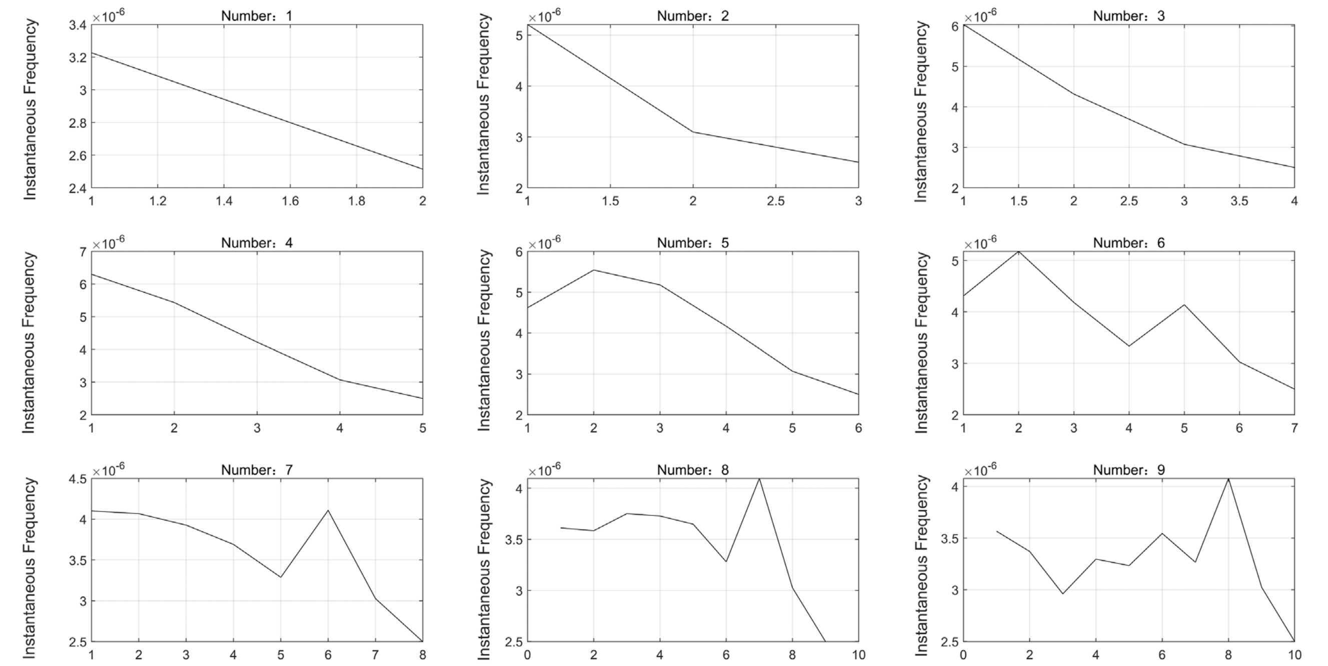


FIGURE 2 | Average instantaneous frequency when K from 1 to 9

2.2.1 Max-Relevance

The maximum correlation criterion solution can be expressed as the average of the mutual information between the feature x_i and the target variable y:

$$\max D(J, y) = \frac{1}{|J|}\sum_{x_i \in J} I(x_i, y) \tag{9}$$

where x_i is the characteristic; y represents the target variable; J is the set containing the x_i; $I(x_i, y)$ represents the mutual information between the feature x_i and the target variable y. The expression is as follows:

$$I(x_i, y) = \int\int p(x_i, y)\log\frac{p(x_i, y)}{p(x_i)p(y)}dx_i dy \tag{10}$$

where $p(x_i)$, $p(y)$ are the edge probability density functions of x_i and y, respectively , $p(x_i, y)$ is the joint probability density function of of x_i and y.

2.2.2 Minimum Redundancy

The overlapping information between any two feature variables is called redundancy information. The features selected according to **Eq. 9** only consider the degree of correlation and do not consider the existence of redundancy between features. The input of redundant features increases the number of input features, and decreases the accuracy of the prediction model. The Minimum Redundancy expression is as follows:

$$\min R(J) = \frac{1}{|J|^2}\sum_{x_i \in J, x_j \in J} I(x_i, x_j) \tag{11}$$

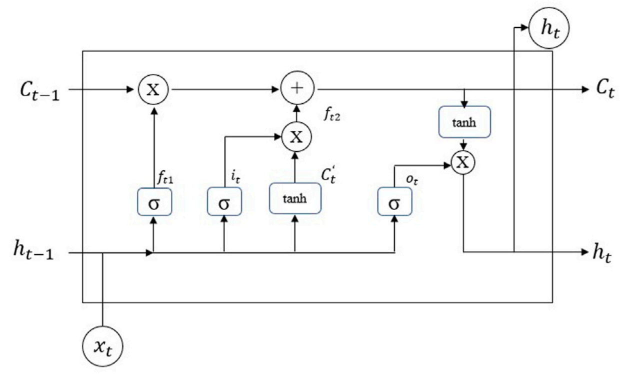


FIGURE 3 | LSTM structure.

mRMR can be expressed by **Eqs 9**, **11** as:

$$\begin{gathered}\max \psi(D, R),\\ \psi = D - R.\end{gathered} \tag{12}$$

2.3 Prediction Model

The prediction part uses Long Short-Term Memory Neural Network (LSTM) model, which was proposed by Hochreiter and Schmidhuber (Hochreiter and Schmidhuber, 1997) to learn long-term dependence information. It can handle more complex problems, and has more mature applications in the field of load prediction (Sun et al., 2019a). Long short-term memory neural network is a special form of the recurrent neural network.

LSTM is composed of cells with the same structure (**Figure 3**). In this model, the data of the next moment is predicted each time by the previous data and historical data, which is processed by the cells. Each cell has three input parameters: Historically stored information C_{t-1}, historical data X_t and h_{t-1}, which represent the prediction results of the last cell and input parameter to the cell. Each cell contains four parts, including the forgotten gate, the input gate, the update gate and the output gate.

The data h_{t-1} that processed by the previous cell, and the input data of the current time X_t are linked by a matrix and obtain $X_t^{'}$,

$$X_{t}^{'} = [h_{t-1}, X_t] \tag{13}$$

In the forgotten gate, LSTM can decide what information to discard from the cell. After the sigmoid function processing, $X_t^{'}$ can get f_{t1}. LSTM can remember large amounts of historical data by f_{t1} filtered data.

$$f_{t1} = \sigma\left(W_f \cdot X^{'} + b_f\right) \tag{14}$$

In the input gate, LSTM acquires the new data, After the sigmoid function processing, $X_t^{'}$ can get i_t. The i_t decides the useful data in $X_t^{'}$. Moreover, $X_t^{'}$ is processed by the tanh function to calculate $C_t^{'}$,

$$i_t = \sigma\left(W_i \cdot X^{'} + b_t\right) \tag{15}$$

$$C_{t}^{'} = \tan\text{h}\left(W_c \cdot X^{'} + b_c\right) \tag{16}$$

$$f_{t2} = i_t \times C_{t}^{'} \tag{17}$$

C_t is updated in the update gate. To obtain the historical data, C_{t-1} and f_{t1} are multiplied by the matrix, In order to keep more accurate rules in the cell for accurate prediction, f_{t2} is added to the equation to get output C_t,

$$C_t = f_{t1} \times C_{t-1} + f_{t2} \tag{18}$$

LSTM outputs the result in the output gate. After the sigmoid function processing, $X_t^{'}$ can get O_t. O_t decides which C_t needs to be retained as the result. In addition, C_t is processed by the tanh function to get $h_t^{'}$, $h_t^{'}$ and O_t are multiplied to obtain the final data h_t,

$$O_t = \sigma\left(W_o \cdot X_{t}^{'} + b_o\right) \tag{19}$$

$$h_t = O_t \times tanh\left(C_t\right) \tag{20}$$

$$\tan\text{h}\left(x\right) = \frac{e^{2x} - 1}{e^{2x} + 1} \tag{21}$$

$$\sigma\left(x\right) = \frac{e^x}{e^x + 1} \tag{22}$$

In order to compare the prediction results of different models, three evaluation metrics will be used in this paper: Decision factor: R-square (R^2), mean absolute error (MAE), and root mean square error (RMSE). The specific calculation of these metrics is described as follows:

$$R_2 = 1 - \frac{\sum_i \left(P_i - Q_i\right)^2}{\sum_i \left(\bar{Q}_i - Q_i\right)^2} \tag{23}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Q_i - P_i| \tag{24}$$

$$RMSE = \sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}\left(Q_i - P_i\right)^2\right)} \tag{25}$$

Where Q_i is the recorded value of building power consumption at time i, and P_i is the predicted value of building power consumption at time i. These three criteria describe the closeness of the predicted data to the actual data in three different ways, The value of R^2 is between 0 and 1, with 0 indicating wo$_{r}$se than the mean and 1 indicating perfect prediction, And for MAE and RMSE, the smaller the value, the better the prediction result of the model.

3 CASE STUDY

3.1 Data Introduction

The building electricity consumption data obtained in this article was obtained from the Qingdao civil building energy consumption monitoring platform. Raw data was selected from three summer cooling months (June, July and August) with a time granularity of 1 hour. The maximum and minimum values of the original data are 803.5 KW and 65 KW; the difference between the maximum and minimum values is 738.5 KW, which demonstrates the volatility of the data. The mean and standard deviation of this data are 333.48 KW and 199.97 KW, respectively, which shows the large dispersion of the data. In **Figure 4**, which illustrates the sequence of the original data, it can be seen that the raw load fluctuates considerably.

3.2 Comparison of Decomposition by EEMD and VMD

Figure 5A shows how EEMD decomposes the original load into 11 intrinsic mode functions (IMFs) and a residual, and **Figure 5B** shows the spectrum after passing the Fourier transform. Even with the improved EMD algorithm, the phenomenon of modal mixing is still evident. Modal mixing occurs when one modal component is decomposed into multiple components. In the figure, the frequency band of IMF4 overlaps with the frequency bands of IMF3 and IMF5. Modal mixing is a defect of the EEMD algorithm and leads to degradation of the model accuracy, so it is important to avoid this phenomenon.

The VMD algorithm solves the modal mixing problem inherent in the EEMD algorithm. In **Figure 6A**, the VMD decomposition results are shown for $K = 4$ and the penalty parameter $a = 2000$, which were determined in **Section 2.1.2**. u_1 is the lowest frequency component, u_2 is the medium frequency component and u_3 and u_4 are the highest frequency components. According to the additional analysis supplied by the spectrogram in **Figure 6B**, there is no overlap in the frequencies of the components, which indicates that the VMD algorithm solves the problem of modal mixing.

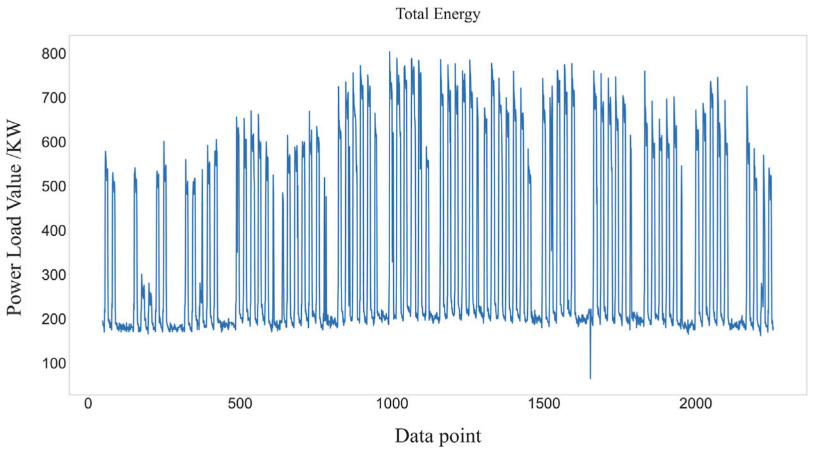


FIGURE 4 | Original total energy series.

In summary, both the EEMD and VMD algorithms are capable of handling volatile raw data, and both algorithms decompose the load into several stable components. The EEMD algorithm is an improved algorithm based on EMD, but it is limited due to the phenomenon of modal confusion. The VMD algorithm overcomes this shortcoming. The VMD algorithm can sufficiently decompose the raw load data to obtain more physically meaningful components and improve the accuracy of the model prediction.

3.3 Feature Selection

Building electricity consumption is influenced by climate and historical load. However, the raw data contains only meteorological factors. To fully consider the independence of each component and research the physical significance of each component, a set of feature matrices are established in this paper. The appropriate feature set is selected by the mRMR algorithm for input into the prediction model. The established feature matrices and their representations are shown in **Table 1**.

The time interval of the load data collected in this paper is 1 h. In **Table 1**, D_t represents the load point for the previous 48 h at time t. Similarly, T_t, H_t and Dp_t represent the temperature, humidity and dew point temperature, respectively, for the previous 48 h at time t. The wind speed, temperature, humidity and dew point are the meteorological characteristics of the dataset.

After establishing the feature matrix, each component of the decomposition is used as the target variable y, and x_i is the data point in the feature matrix. The mRMR values are calculated according to **Eq. 12** and the calculation results are sorted in descending order. The top 15 influencing factors are selected as the feature matrix of the input model. The final selection results are shown in **Table 2**.

To illustrate the superiority of the mRMR algorithm, mutual information maximization (MIM) (Novakovic et al., 2011) is used as a comparison in this paper. The MIM algorithm is based on the theory of mutual information, but unlike mRMR, the MIM algorithm only considers the correlation between features and target variables and does not consider the redundancy between features. The results of the MIM feature selection are shown in **Table 2**.

Consider u_1 and u_4 in **Table 2** as an example. The low-frequency components u_1 and u_2 are mainly influenced by D_t, which indicates that the u_1 and u_2 components are influenced more heavily by the historical load of the past 48 h. It is further seen through **Figure 6A** that although both u_1 and u_2 are strongly influenced by historical loads, u_1 presents a load variation trend with a week as a period, while u_2 presents a load variation trend with a 24-h period. In contrast, the high-frequency components u_3 and u_4 are not only influenced by the historical load but also by the weather factor. Weather factors are usually seen as uncertainty factors. The influence of humidity on u_4 is ranked second among all the features. This explains why u_4 is more volatile than u_1: u_1 is mainly influenced by historical load and has a certain regularity, while u_4 is influenced by uncertainties such as humidity. Thus, u_4 is more irregular.

Table 3 shows that the results of the MIM feature selection method are similarly ranked, with the higher-ranked features all being historical loads at a given moment. This is especially apparent for the u_3 and u_4 components. The top five features selected using MIM have a high degree of overlap because the MIM algorithm only considers the maximum correlation between features and variables while ignoring the degree of redundancy between features. This is improved by using the mRMR algorithm. For u_3 and u_4, the feature overlap selected using the mRMR algorithm is not high, and features that are not considered by MIM, such as wind speed and humidity, are taken into account by the mRMR algorithm.

In summary, the mRMR algorithm considers not only the correlation between features and target variables but also the degree of redundancy between features. The selected features can better reflect some characteristics of the modal components and reduce the dimensionality of the feature matrix.

3.3 Model Predictions

In this paper, the LSTM model is used for predictions. The training and test sets are divided for a total of 2,208 data

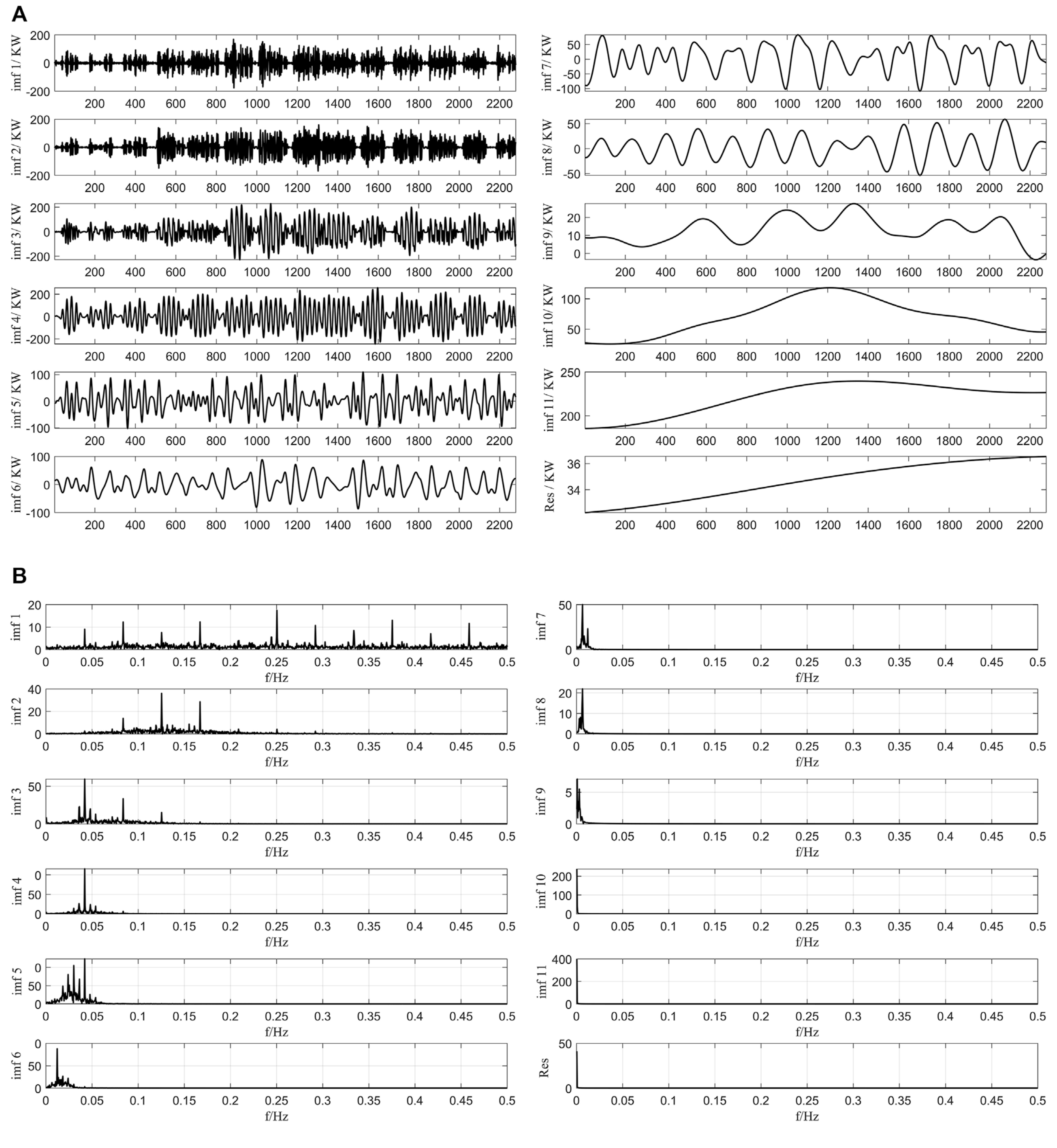


FIGURE 5 | (A) The decomposition results of EEMD. **(B)** The decomposition spectrogram of EEMD.

points from June 1, 2017 to August 31, 2017. Of this data, 80% is used to build the model and 20% is used to check the validity of the established models.

The number of layers of the LSTM model serves to remember important information, and theoretically, more hidden layers give the model an improved nonlinear fitting ability and a better learning effect. However, increasing the number of layers consumes a considerable amount of computation time. According to the literature (Pan, 2018; Li et al., 2019), the number of implied layers generally does not exceed 3, so the number of implied layers in this paper has been determined to be 1.

The number of nodes in the hidden layer affects the performance of the model. If the number of nodes in the implicit layer is too small, less effective information is

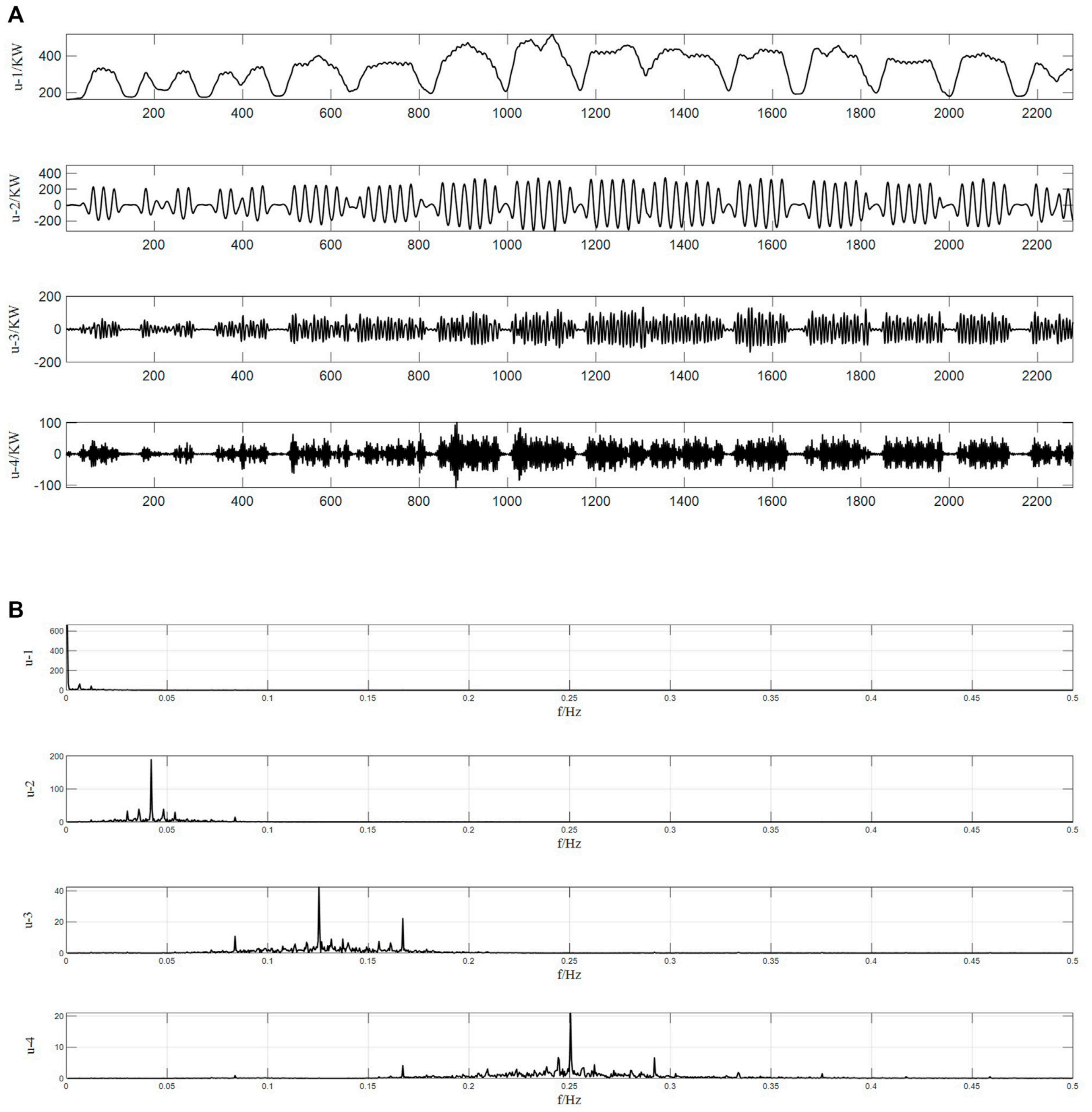


FIGURE 6 | (A) The decomposition results of VMD. **(B)** The decomposition spectrogram of VMD.

obtained in the prediction process. If the number of nodes in the implicit layer is too large, it may lead to a longer training time and overfitting problems. According to the literature (Xu et al., 2020), the number of nodes in the hidden layer can be determined by **Eq. (26)**:

$$l = \sqrt{m+n} + a \tag{26}$$

where l is the number of nodes in the hidden layer, m is the number of input nodes, n is the number of output nodes, and a is a constant from 1 to 10. By calculation, the number of nodes l in the hidden layer is determined to be 8. The LSTM network is trained using the Adam optimization algorithm (Wang et al., 2019). By referring to relevant literature (Kong et al., 2019; Pei et al., 2020) and experimental measurements, the remaining parameters are set: The number of iterations of the neural network is set to 1,000, the learning rate is set to 0.01, and the expected error is set to 0.0004. The prediction results are shown in **Figure 7** and **Table 4**.

TABLE 1 | Construction of feature matrix.

Feature name	Representation
D_t	D_1,D_2,D_3. . .D_48
T_t	T_1,T_2,T_3. . .T_48
H_t	H_1,H_2,H_3. . .H_48
Dp_t	Dp_1,Dp_2,Dp_3. . .Dp_48
Wind speed	Wind speed
Temperature	Temperature
Humidity	Humidity
Dew point	Dew point

Figure 7 and **Table 3** demonstrate that the integrated model proposed in this paper achieves better results for the prediction of each component. In general, the prediction results for the low and medium frequency components (u_1 and u_2, respectively) are better, with R^2 values of 0.997 and 0.994, respectively. The u_3 component also achieved a better prediction, having an R^2 value of 0.992. In contrast, the prediction results for the high-frequency component u_4 are slightly worse, with an R^2 value of only 0.982.

Table 3 also shows the prediction results for each component obtained using the MIM feature selection method. The R^2 values of the prediction results for all four components are lower than those obtained by the mRMR method, especially for the u_2 and u_3 components. The main reason for this result is because the MIM feature selection algorithm does not consider the redundancy among the features, which leads to a certain degree of repetitiveness of the selected features.

3.4 Model Comparison

The proposed model is compared and analysed alongside other models to verify its reliability. The other models are singular and include a model using EEMD decomposition (EEMD–mRMR–BPNN), a model using the MIM algorithm (VMD–MIM–LSTM) and a model using the BPNN algorithm (VMD–mRMR–BPNN). The prediction results and evaluation metrics of all models are shown in **Figure 8** and **Table 4**.

The predictions of the integrated model with the addition of the modal decomposition algorithm are more accurate compared to the single prediction model (LSTM), as shown in **Figure 8**. This indicates that the modal decomposition algorithm can indeed handle more complex data and improve the accuracy of the model. In addition, the prediction model proposed in this paper has the highest prediction accuracy among the four models.

According to the evaluation metrics analysis in **Table 4**, the prediction error of the single LSTM model is larger than the prediction error of the integrated model. This is mainly due to the instability of the load data and the limitations of the input

TABLE 2 | Feature selection results.

Number	u_1		u_2		u_3		u_4	
	mRMR	MIM	mRMR	MIM	mRMR	MIM	mRMR	MIM
1	D_8	D_8	D_1	D-1	D_2	D-2	D_1	D-1
2	D_46	D-15	D_13	D-2	D_15	D-3	Humidity	D-2
3	D_15	D-16	D_43	D-6	Humidity	D-4	D_12	D-3
4	D_1	D-5	D_7	D-13	D_8	D-1	D_21	D-4
5	D_36	D-4	D_27	D-5	D_23	D-7	1D_8	D-5
6	D_26	D-46	D_36	D-10	D_45	D-6	D_2	D-6
7	D_17	D-30	D_47	D-7	D_4	D-5	D_41	D-8
8	D_30	D-26	D_17	D-3	D_13	D-10	D_47	D-23
9	D_44	D-10	D_2	D-9	D_33	D-9	D_14	D-24
10	D_4	D-47	D_10	D-4	D_1	D-8	D_3	D-12
11	D_10	D-48	D_16	D-12	D_17	D-13	Wind speed	D-25
12	D_48	D-14	D_6	D-8	D_7	D-23	D_10	D-7
13	D_39	D-28	D_48	D-27	D_46	D-12	D_23	D-9
14	D_16	D-22	D_38	D-11	D_3	D-14	D_4	D-10
15	D_3	D-39	D_3	D-28	D_14	D-11	D_35	D-22

TABLE 3 | VMD component prediction result.

Model	Subsequences	MAE (kWh)	RMSE (kWh)	R_2
VMD + mRMR + LSTM	u_1	2.61	3.34	0.997
	u_2	8.61	11.23	0.994
	u_3	2.90	4.02	0.992
	u_4	2.18	3.07	0.982
VMD + MIM + LSTM	u_1	8.83	10.44	0.977
	u_2	25.63	35.95	0.943
	u_3	7.08	11.14	0.960
	u_4	2.11	3.09	0.980

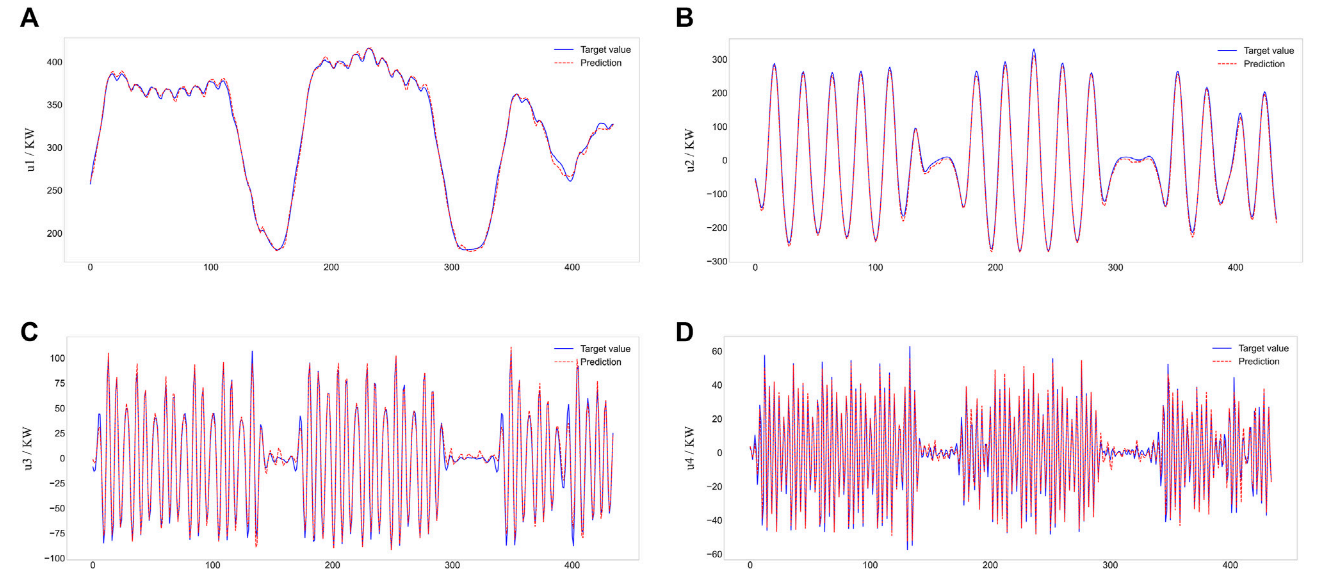


FIGURE 7 | VMD component prediction figure.**(A)** u1, **(B)** u2, **(C)** u3, **(D)** u4.

TABLE 4 | Evaluation metrics of each model.

Model	MAE (kWh)	RMSE (kWh)	R_2
LSTM	41.77	67.70	0.87
VMD + mRMR + BPNN	36.42	47.54	0.94
EEMD + mRMR + BPNN	50.39	62.21	0.89
VMD + MIM + LSTM	33.15	46.42	0.94
VMD + mRMR + LSTM	21.36	30.64	0.97

features. The modal decomposition algorithm can decompose the fluctuating data into several stable IMFs, and the mRMR algorithm can select suitable features for the model. Thus, the prediction results of the integrated model are better than those of the single LSTM model. In addition, the integrated model using the VMD modal decomposition method (VMD–mRMR–BPNN) predicts better results than the integrated model using EEMD (EEMD–mRMR–BPNN). The R^2 is improved by 5.6% and the

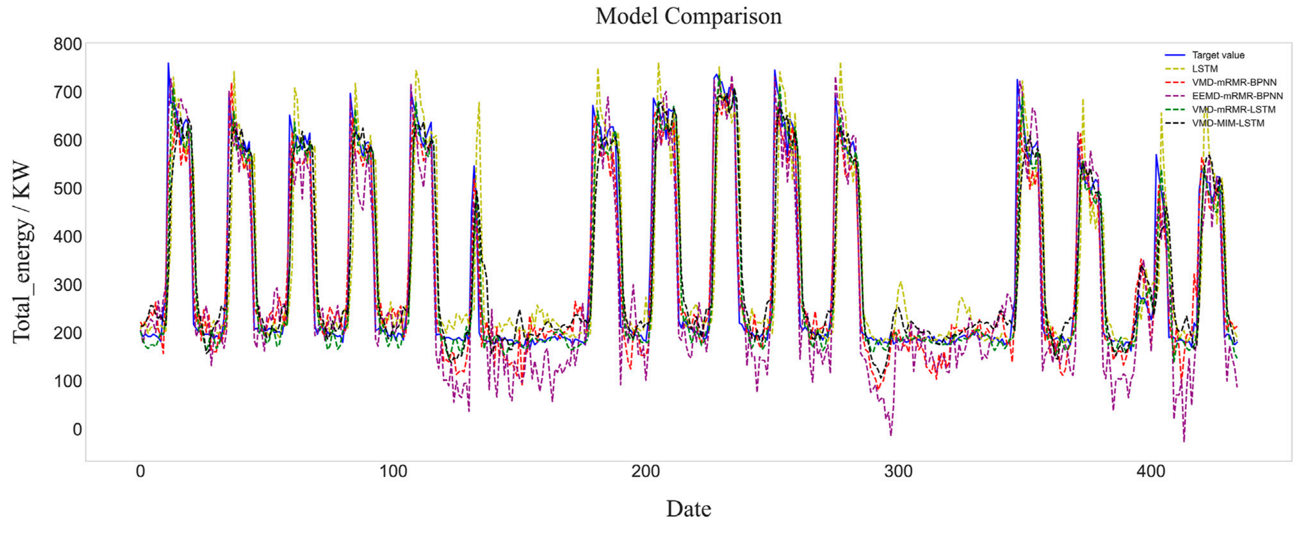


FIGURE 8 | Prediction results of each model.

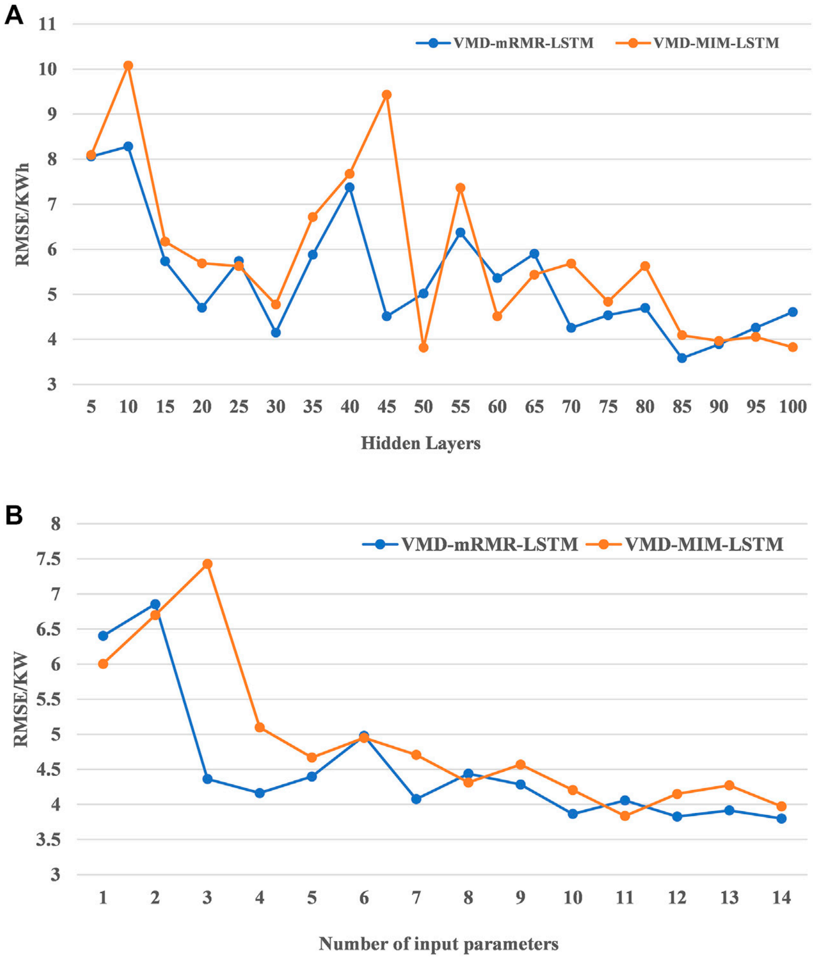


FIGURE 9 | (A) Effect of the number of hidden layers on RMSE of u4. **(B)** Effect of the number of input parameters on RMSE of u4.

MAE and RMSE are reduced by 27.7 and 23.6%, respectively, because the VMD algorithm solves the problems of modal aliasing and elusive components.

Comparing the VMD–MIM–LSTM and VMD–mRMR–LSTM integrated models, the mRMR algorithm, which takes into account the redundancy between features, achieves better prediction accuracy for the feature selection algorithm. The R^2 is improved by 3.0%, the value of the MAE is reduced by 35.6% and the value of RMSE is reduced by 34.1%. This is because the mRMR algorithm takes into account the redundancy between features and can select the appropriate feature matrix for each IMF.

Comparing the VMD–mRMR–LSTM and VMD–mRMR–BPNN prediction models, the integrated model using LSTM outperforms the integrated model using BPNN. The R^2 of the LSTM integrated model is improved by 3.0%, the value of MAE is reduced by 41.4% and the value of RMSE is reduced by 35.5%. The power load series is a sample of power load variation over time, and the BPNN model has shortcomings in analysing these types of time series. For the time series, the LSTM model better mines the relationship between the data points. In brief, the model proposed in this paper has the highest prediction accuracy.

3.5 Model Robustness

The experimental results show that the hybrid model proposed in this paper has high prediction accuracy. In this section, the robustness of the hybrid model is analysed by varying the number of input feature parameters and the number of neurons in the hidden layer. For simplicity, the VMD-decomposed u_4 has been selected as the target dataset.

3.5.1 Number of Neurons in The Hidden Layers

In theory, with the increase of the number of neurons in the hidden layer and the more abstract features extracted by deep learning, the more accurate a time series will be, which is favourable for predictions (Zhang et al., 2020). **Figure 9A** shows the RMSE of the VMD–mRMR–LSTM and VMD–MIM–LSTM models when the number of neurons in the hidden layer is changed. When the number of hidden layer nodes is between 5 and 25, the RMSE shows a gradual decrease; when the number of hidden layer nodes is between 25

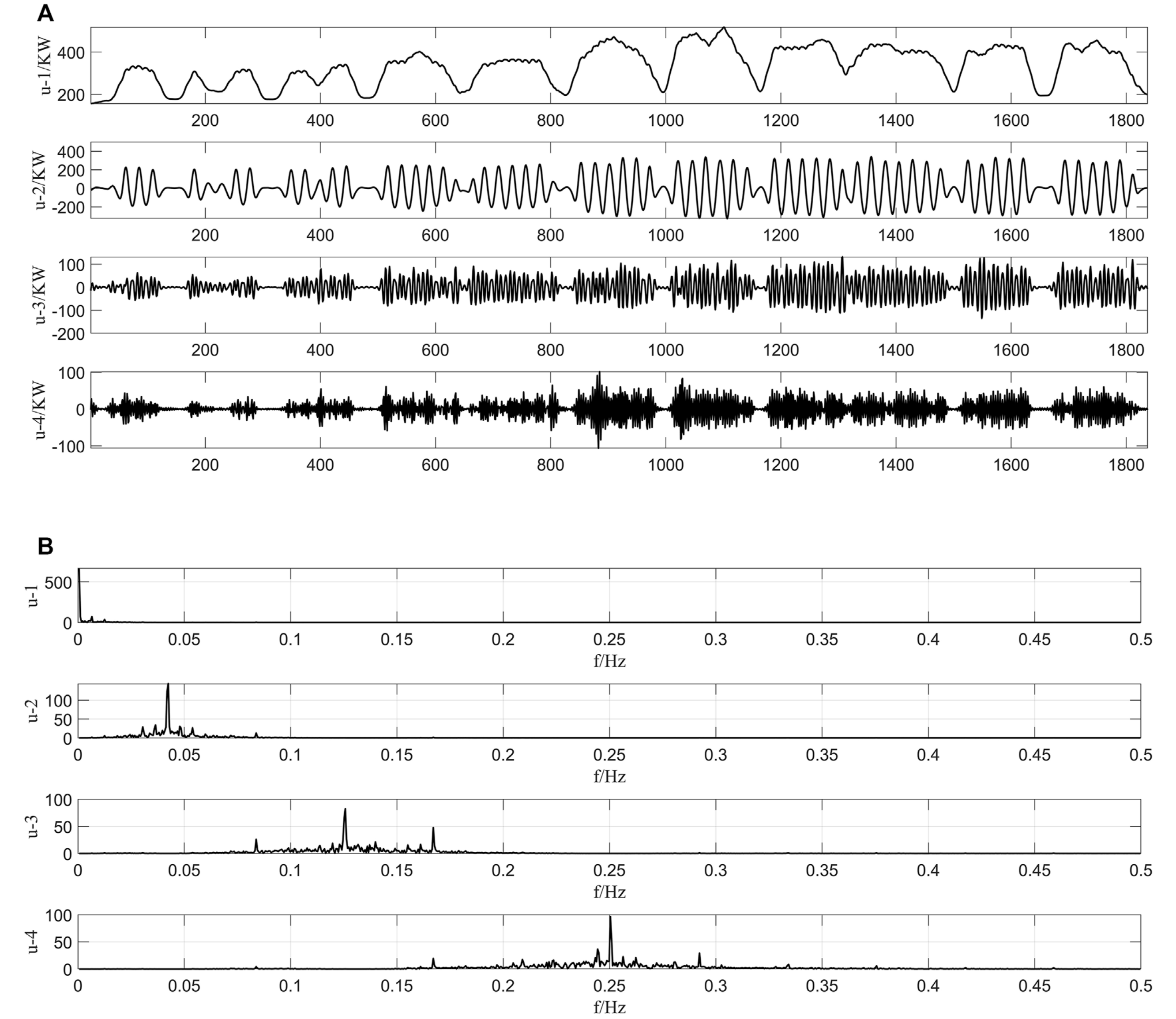


FIGURE 10 | (A) The decomposition results of training set. **(B)** The decomposition spectrogram of training set.

and 65, the RMSE has a large fluctuation; when the number of hidden layer nodes is between 65 and 100, the RMSE tends to be smooth, its value is mostly between 3 and 5 and the prediction error is relatively stable, which indicates that these two models are highly robust. In addition, the RMSE of VMD–mRMR–LSTM is lower than that of VMD–MIM–LSTM in most situations, which indicates that VMD–mRMR–LSTM has better prediction performance and more stable robustness.

3.5.2 Number of Input Parameters

The redundancy between features is theoretically taken into account by the mRMR algorithm so that more input parameters lead to a better prediction performance of the model. However, too many feature parameters can increase the complexity of the model and increase the computing cost. **Figure 9B** illustrates the effect of the number of feature parameters on the accuracy of the model. When the number of input features is between 1 and 6, the RMSE of the models is decreasing and fluctuates. When the number of input features is between 6 and 15, the RMSE of both models decreases smoothly with values in the range of 3.5–4.5. The prediction error of the VMD–mRMR–LSTM model is smaller than that of the VMD–MIM–LSTM model. Therefore, the VMD–mRMR–LSTM model is more robust than the VMD–MIM–LSTM model when the number of input features is changed.

3.5.3 Effect of Input Data

To investigate the effect of different input data on the accuracy of the model, the training set in the 3.1 section is used as the

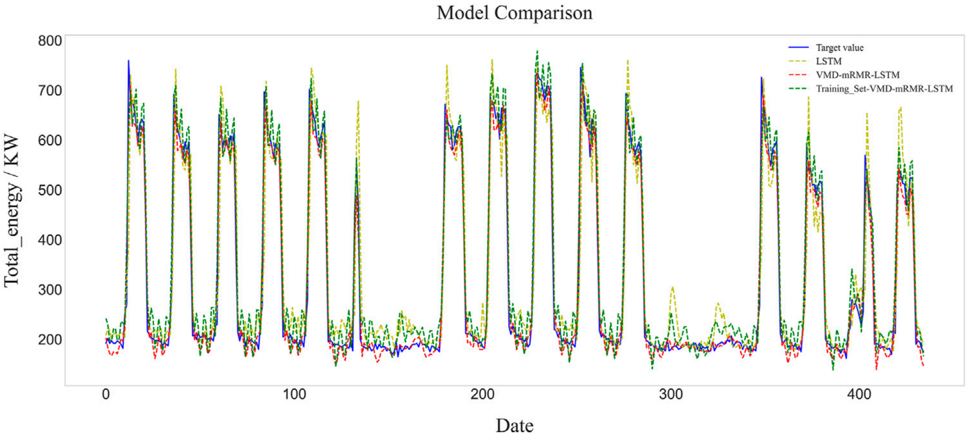


FIGURE 11 | Prediction results of model with training set.

raw data input to the hybrid model. **Figure 10A** shows the results of the training set decomposed by VMD. Comparing the decomposition results in **Figure 6A**, **10A**, it can be seen that the trend of the training set is similar to the original data. Further comparing the spectrograms in **Figure 6B**, **10B**, although the peak frequency of the training data set and the original data set is different, they appear at the same locations. The decomposition results of the training set are input into the hybrid model, and the prediction result is shown in **Figure 11**. The predicted values of MAE, RMSE, and R^2 are 30.96 kWh, 38.96 kWh, and 0.95, respectively. Compared with the results of decomposing the original data (VMD–mRMR–LSTM), the R^2 of the decomposed training dataset model (Training_Set-VMD–mRMR–LSTM) decreased by 2%, and the RMSE and MAE increased by 27 and 44.9%, respectively, indicating that the selection of the input data can have an impact on the accuracy of the model. Training_Set- VMD–mRMR–LSTM still has higher accuracy than the single LSTM model, and the R^2 improved by 7%, RMSE and MAE reduced by 42.4 and 25.9%, respectively. In conclusion, the use of training data as model input reduces the accuracy of the model, but the impact is small in general. Compared with a single model, the proposed hybrid model still has a greater superiority.

4 CONCLUSION

A hybrid short-term load forecasting model, namely VMD–mRMR–LSTM, was proposed in this paper. To solve the modal mixing problem presented by the EMD algorithm, the VMD algorithm was used, and the value of its decomposition number K was determined by the average instantaneous frequency. For feature selection, the mRMR algorithm was used to select the related feature by analysing the correlation between each component and feature as well as the redundancy between features. Finally, the LSTM model was used for the prediction model. The case study in this paper demonstrated the following:

1) Compared to single prediction models, hybrid models have higher accuracy and are more robust in the field of energy consumption prediction and have a broad application prospect for the short-term prediction of building energy consumption.
2) Using VMD to decompose the original sequence can have a better decomposition effect than when EEMD is used. Decomposition by VMD solves the problem of modal confusion so that the decomposed sequence is stable. The prediction results of the hybrid model using VMD are higher than those of the hybrid model using EEMD.
3) The mRMR algorithm can eliminate the redundancy between features and show the influencing factors of the modal components. The experimental results prove that the features selected by the mRMR algorithm have a higher prediction accuracy and better interpretability than those selected by MIM, which is supported by the value of R^2 increasing by 3%, the value of MAE decreasing by 35.6% and the value of RMSE decreasing by 34.1%.
4) The hybrid model proposed in this paper can achieve an R^2 value of 0.97, and its prediction results are higher than those of the single model (LSTM) and the general integrated model (VMD–MIM–LSTM). Therefore, the proposed VMD–mRMR–LSTM approach has a high potential for practical applications in energy systems, such as forecasting building energy consumption.
5) By varying the number of input feature parameters and the number of neurons in the hidden layer, the model is proven to have good robustness.

In this paper, all decomposed components were predicted using the LSTM model. However, since the frequency of each component varied, the LSTM may not have produced ideal results for each component. For example, in **Table 3**, the difference between the R^2 values of the u_1 and u_4 components for the VMD–mRMR–LSTM model was not negligible. Choosing appropriate prediction models for the different frequency components may lead to better results. We will conduct more research in this direction in the future.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: This study is applicable to the air conditioning cooling and heating load data of various building users. Requests to access these datasets should be directed to FQ, qianfanyue91@163.com.

AUTHOR CONTRIBUTIONS

YR gave guidance on the framework and ideas of the paper. GW built the load forecasting model and compared it with conventional method. HM gave guidance on the framework and ideas of the paper. FQ gave guidance on the framework and ideas of the paper.

FUNDING

This research has been supported by the national key R&D project (No.2020YFD1100504-05).

REFERENCES

Amasyali, K., and El-Gohary, N. M. (2018). A Review of Data-Driven Building Energy Consumption Prediction Studies[J]. *Renew. Sust. Energ. Rev.* 81. doi:10.1016/j.rser.2017.04.095

de Paiva, G. M., Pires Pimentel, Sergio., Pinheiro Alvarenga, Bernardo., Marra, E. G., Mussetta, Marco., and Leva, S. (2020). Multiple Site Intraday Solar Irradiance Forecasting by Machine Learning Algorithms: MGGP and MLP Neural Networks[J]. *Energies* 13 (11).

Dong, B., Li, Z., Rahman, S. M. M., and Vega, R. (2016). A Hybrid Model Approach for Forecasting Future Residential Electricity Consumption. *Energy and Buildings* 117, 341–351. doi:10.1016/j.enbuild.2015.09.033

Dragomiretskiy, K., and Zosso, D. (2014). Variational Mode Decomposition. *IEEE Trans. Signal. Process.* 62 (3), 531–544. doi:10.1109/TSP.2013.2288675

Eom, J., Clarke, L., Kim, S. H., Kyle, P., and Patel, P. (2012). China's Building Energy Demand: Long-Term Implications from a Detailed Assessment. *Energy* 46 (1), 405–419. doi:10.1016/j.energy.2012.08.009

Guo, Z., Zhao, W., Lu, H., and Wang, J. (2011). Multi-step Forecasting for Wind Speed Using a Modified EMD-Based Artificial Neural Network Model[J]. *Renew. Energ.* 37 (1).

He, F., Zhou, J., Feng, Z-k., Liu, G., and Yang, Y. (2019). A Hybrid Short-Term Load Forecasting Model Based on Variational Mode Decomposition and Long Short-Term Memory Networks Considering Relevant Factors with Bayesian Optimization Algorithm[J]. *Appl. Energ.* 237. doi:10.1016/j.apenergy.2019.01.055

He, Q., Wang, J., and Lu, H. (2018). *A Hybrid System for Short-Term Wind Speed forecasting[J]*. Kidlington, Oxford: Elsevier BV, 226.

Hochreiter, S., and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Comput.* 9 (8), 1735–1780. doi:10.1162/neco.1997.9.8.1735

Hu, J., Wang, J., and Zeng, G. (2013). A Hybrid Forecasting Approach Applied to Wind Speed Time Series[J]. *Renew. Energ.* 60. doi:10.1016/j.renene.2013.05.012

Huang Norden, E., Zheng, S., Long Steven, R., Wu, M. C., Shih Hsing, H., Zheng, Q., et al. (1998). The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Non-stationary Time Series Analysis[J]. *Proc. R. Soc. A: Math. Phys. Eng. Sci.*, 454.

Kong, W., Dong, Z. Y., Jia, Y., David, J., Hill, Y. X., and Zhang, Y. (2019). Short-Term Residential Load Forecasting Based on LSTM Recurrent Neural Network.[J]. *IEEE Trans. Smart Grid* 10 (1).

Li, C., Chen, Z. Y., and Liu, J. B. (2019). "Power Load Forecasting Based on the Combined Model of LSTM and XGBoost[C]//PRAI '19," in *Proceedings of the 2019 the International Conference on Pattern Recognition and Artificial Intelligence* (Wenzhou, China: ACM), 46–51.

Li, C., Tao, Y., Ao, W., Yang, S., and Bai, Y. (2018). Improving Forecasting Accuracy of Daily enterprise Electricity Consumption Using a Random forest Based on Ensemble Empirical Mode Decomposition[J]. *Energy* 165. doi:10.1016/j.energy.2018.10.113

Lin, Y ., and Peng, L. (2011). Combined Model Based on EMD-SVM for Short-Term Wind Power Prediction. *Proc. CSEE* 31, 102–108.

Liu, H., Chen, C., Tian, H-Q., and Li, Y-F. (2012). A Hybrid Model for Wind Speed Prediction Using Empirical Mode Decomposition and Artificial Neural Networks[J]. *Renew. Energ.* 48. doi:10.1016/j.renene.2012.06.012

Liu, H., Mi, X., and Li, Y. (2018). Smart Multi-step Deep Learning Model for Wind Speed Forecasting Based on Variational Mode Decomposition, Singular Spectrum Analysis, LSTM Network and ELM[J]. *Energ. Convers. Manag.* 159. doi:10.1016/j.enconman.2018.01.010

Liu, M., Cao, Z., Zhang, J., Wang, L., Huang, C., and Luo, X. (2020). Short-term Wind Speed Forecasting Based on the Jaya-SVM Model[J]. *Int. J. Electr. Power Energ. Syst.* 121. doi:10.1016/j.ijepes.2020.106056

Liu, Z., Wu, D., Liu, Y., Han, Z., Lun, L., Gao, J., et al. (2019). Accuracy Analyses and Model Comparison of Machine Learning Adopted in Building Energy Consumption Prediction[J]. *Energy Exploration & Exploitation* 37 (4). doi:10.1177/0144598718822400

Marcello Anderson, F. B., Lima, P. C. M. C., Carneiro, T. C., Leite, J. R., Luiz, J., Neto, D. B., et al. (2017). Portfolio Theory Applied to Solar and Wind Resources Forecast[J]. *IET Renew. Power Generation* 11 (7). doi:10.1049/iet-rpg.2017.0006

Muhammad, M., Francesco, G., Sonia, L., and Marco, M. (2020). Comparison of echo State Network and Feed-Forward Neural Networks in Electrical Load Forecasting for Demand Response Programs[J]. *Mathematics Comput. Simulation*, 184.

Niu, H., Xu, K., and Wang, W. (2020). *A Hybrid Stock price index Forecasting Model Based on Variational Mode Decomposition and LSTM network[J]*. Dordrecht, Netherlands: Springer US.

Novakovic, J., Strbac, P., and Bulatovic, D. (2011). Toward Optimal Feature Selection Using Ranking Methods and Classification Algorithms. *Yugoslav J. Operation* 21 (1), 119–135. doi:10.2298/YJOR1101119N

Pan, B. (2018). Application of XGBoost Algorithm in Hourly PM2.5 Concentration Prediction[J]. *IOP Conf. Series:Earth Environ. Sci.*, 113, 1–7.

Pei, S., Qin, H., Yao, L., Liu, Y., Wang, C., and Zhou, J. (2020). Multi-Step Ahead Short-Term Load Forecasting Using Hybrid Feature Selection and Improved Long Short-Term Memory Network[J]. *Energies* 13 (16). doi:10.3390/en13164121

Peng, H., Long, F., and Ding, C. (2005). Feature Selection Based on Mutual Information: Criteria of max-dependency, max-relevance, and Min-Redundancy. *IEEE Trans. Pattern Anal. Mach Intell.* 27 (8), 1226–1238. doi:10.1109/TPAMI.2005.159

Said, G. (2016). A New Ensemble Empirical Mode Decomposition (EEMD) Denoising Method for Seismic Signals[J]. *Energ. Proced.* 97.

Sun, Z., Zhao, S., and Zhang, J. (2019). Short-Term Wind Power Forecasting on Multiple Scales Using VMD Decomposition, K-Means Clustering and LSTM Principal Computing[J]. *IEEE Access* 7. doi:10.1109/access.2019.2942040

Sun, Z., Zhao, S., and Zhang, J. (2019). Short-Term Wind Power Forecasting on Multiple Scales Using VMD Decomposition, K-Means Clustering and LSTM Principal Computing. *IEEE Access* 7, 166917–166929. doi:10.1109/ACCESS.2019.2942040

Wang, S., Sun, J., and Xu, Z. (2019). HyperAdam: A Learnable Task-Adaptive Adam for Network Training[J]. *Proc. AAAI Conf. Artif. Intelligence* 33. doi:10.1609/aaai.v33i01.33015297

Wu, Y-X., Wu, Q-B., and Zhu, J-Q. (2018). Improved EEMD-Based Crude Oil price Forecasting Using LSTM Networks[J]. *Physica A: Stat. Mech. its Appl.*, 516.

Wu, Y. J. (2016). *Research on Fault Diagnosis of Wind Turbine Transmission System Based on Variational Mode Decomposition. Dissertation*. Beijing: North China Electric Power University.

Wu, Z., and Huang, N. E. (2009). Ensemble Empirical Mode Decomposition: A Noise-Assisted Data Analysis Method[J]. *Adv. Adaptive Data Anal.* 1 (1).

Xu, L., Zhou, C., Hu, Y., Li, G., and Xi, F. (2020). Energy Consumption Prediction of Chiller Based on Long Short-Term Memory [J]. *Refrigeration & Air Conditioning* 34 (06), 664–669.

Zhang, M., Jiang, Z., and Kun, F. (2017). Research on Variational Mode Decomposition in Rolling Bearings Fault Diagnosis of the Multistage Centrifugal Pump[J]. *Mech. Syst. Signal Process.* 93. doi:10.1016/j.ymssp.2017.02.013

Zhang, W., Liu, F., Zheng, X., and Li, Y . (2015). "A Hybrid EMD-SVM Based Short-Term Wind Power Forecasting Model," in Proceedings of the 2015 IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC) (BrisbaneAustralia: QLD), 151–185. doi:10.1109/appeec.2015.7380872

Zhang, Y., Yan, B., and Aasma, M. (2020). A Novel Deep Learning Framework: Prediction and Analysis of Financial Time Series Using CEEMD and LSTM[J]. *Expert Syst. Appl.* 159. doi:10.1016/j.eswa.2020.113609

Zhu, D., Yan, D., and Wang, C. (2012). Comparison of Building Energy Simulation Software: DeST, EnergyPlus and DOE-2[J]. *Building Sci.* 28 (S2), 213–222.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Energy Research

ORIGINAL RESEARCH
published: 10 February 2022
doi: 10.3389/fenrg.2021.793413

Green Bond Index Prediction Based on CEEMDAN-LSTM

Jiaqi Wang[1], *Jiulin Tang*[1] *and Kun Guo*[1,2]*

[1]School of Economics and Management, University of Chinese Academy of Sciences, Beijing, China, [2]Research Center on Fictitious Economy & Data Science, Chinese Academy of Science, Beijing, China

Green bonds, which are designed to finance for environment-friendly or sustainable projects, have attracted more and more investors' attention. However, the study in this field is still relatively limited, especially in forecasting the market's future trends. In this paper, a hybrid model combining CEEMDAN and LSTM is introduced to predict green bond market in China (represented by CUFE-CNI High Grade Green Bond Index). In order to evaluate the performance of our model, we also use EMD to decompose the green bond index. Our empirical result suggests that, compared with EMD-LSTM and LSTM models, CEEMDAN-LSTM is the most accurate model in green bond index forecasting. Meanwhile, we find that indices from the crude oil market and green stock market are both effective predictors, which also provides ground on the correlations between the green bond market and other financial markets.

Keywords: green bonds, CEEMDAN, LSTM, green finance, machine learning

Edited by:
Xun Zhang,
Academy of Mathematics and Systems Science (CAS), China

Reviewed by:
Vidya C. T.,
Centre for Economic and Social Studies (CESS), India
Jian Li Jane Hao,
Xi'an Jiaotong-Liverpool University, China

****Correspondence:***
Kun Guo
guokun@ucas.ac.cn

Specialty section:
This article was submitted to Sustainable Energy Systems and Policies,
a section of the journal Frontiers in Energy Research

Received: *12 October 2021*
Accepted: *28 December 2021*
Published: *10 February 2022*

Citation:
Wang J, Tang J and Guo K (2022) Green Bond Index Prediction Based on CEEMDAN-LSTM.
Front. Energy Res. 9:793413.
doi: 10.3389/fenrg.2021.793413

INTRODUCTION

In order to achieve the goal of peak carbon dioxide emission and carbon neutrality, China is now attempting to transit to low-carbon economy, leading to the urgent financing demand of many green projects. Therefore, the development of various green finance instruments has enjoyed a rapidly growing attention. Compared with traditional financial alternatives, these green instruments are specially designed to support the environment-friendly or sustainable programs. Among them, green bonds, also called climate bonds at times, are viewed as promising green securities to meet the immense capital needs for low-economy projects (Kochetygova and Jauhari, 2014). Different from other bonds, green bonds have some unique features: firstly, the purpose of issuing green bonds is to support environmental companies or programs; secondly, there is a strict procedure of evaluating and choosing green projects; thirdly, the funds raised by green bonds can only be used in the environmental programs and the use of funds will be tracked transparently; finally, annual reports about funds are disclosed every year, which enables the investors to supervise the use of funds (World Bank Group, 2015).

The first green bond emerged in 2007, when the European Investment Bank announced the raising of money for environment-protecting programs by issuing bonds. Shortly after that, in 2008, the first worldwide green bond was issued by the World Bank. Based on statistics from the Climate Bonds Initiative (CBI, 2021), from 2013 to 2020, the international green bond market has developed dramatically, with the amount of green bonds issued each year growing 26 times from approximately 11 billion dollars to over 290 billion dollars. Until the first half of 2021, the cumulative amount of bonds issued has reached 1.3 trillion and the growth rate is still growing. At the same time, the geographic features of green bonds issuance have also changed remarkably. The emerging market began to participate in the green bond market in 2014, which only occupied 2% of the global market at first. However, at the end of 2020, the proportion reached 16%, demonstrating the rapid growth of the emerging market. In China, green bonds appeared in April 2015, and the People's Bank of China together with the Ministry of Finance promulgated *Guidance on building a green financial system* in August 2016, which marks China as the first country in the world to provide explicit government support in the establishment of green financial systems (Chen and Zhao, 2021).

Although the green bonds did not become prevalent in China until the recent years, they have experienced fast development. According to CBI (2020), China merely issued 1.3 billion dollars in green bonds in 2015, which occupied no more than 3% of the global green bond issued. However, the amount of issued green bonds was roaring by nearly 20 times in the next year, exceeding 20 billion dollars, and accounted for 25% of global green bonds issued. Since then, the figure has been growing year by year, except for 2020 due to COVID-19. In addition, it is estimated that the green bond market in China is bigger than statistics suggests because the evaluation and information disclosing standards of green bonds in China are not consistent with international standards, which may decrease the attractiveness of Chinese green bonds in the global markets (Zhang, 2020). Statistics indicates that almost half of the green bonds issued by China in 2020 failed to meet the standards of CBI. Fortunately, China is now attempting to revise its classification system to make it in closer alignment with global taxonomy by publishing the *Green Bonds Endorsed Projects Catalogue (2021 edition)*. It is foreseeable that China is sure to be one of the most profound green bond markets in the future once the standards are more harmonized.

While the green bond market is booming these days, the researches in this field are relatively inadequate. Particularly, there are few studies relevant to the prediction of green bond markets. As green bond indices are beneficial in improving the market efficiency and transparency, investors are eager to anticipate the future trend of the green bond market by predicting green bond indices (Kochetygova and Jauhari, 2014). Therefore, the purpose of this study is to forecast the green bond index through various machine learning models.

However, compared with stock prices, it is much more complex to predict the bond prices because of the dearth of trading information (Ganguli and Dunnmon, 2017). Owing to the asymmetric information needed and offered, the price of bonds cannot reflect the fair value accurately at times. In this paper, we attempt to forecast the green bond index based on two frameworks. The first one is to predict the bond prices and returns depending on indicators from the bond market, stock market, and commodity market (Lin et al., 2018; Choi and Kim, 2018; Chordia et al., 2014; Nazlioglu et al., 2020; Gormus et al., 2018). As the work related to bond returns prediction is limited, some studies have already chosen several widely used stock price predictors to forecast bond returns on the basis of co-movement between the stock and bond markets (Connolly et al., 2005). For instance, motivated by stock price forecasting, Devpura et al. (2021) choose 12 predictor variables to predict bond returns. Fong and Wu (2020) utilize the typical technical rules in the stock market to testify the predictability of 48 sovereign bond markets, and the result suggests that technical indicators are suitable predictors especially when machine learning method is used. The second topic that is closely affiliated to our work is to apply the machine learning method in the prediction of financial markets (Henrique et al., 2019; Gu et al., 2020; Jiang, 2021). Relevant literature has shown that the nonlinear algorithms perform better in forecasting bond returns (Bauer and Rudebusch, 2017; Huang et al., 2020; Giacoletti et al., 2021). As a result, the machine learning models are proven to achieve highest accuracy in the financial time series prediction (Ghoddusi et al., 2019; Bianchi et al., 2021; Sadorsky, 2021).

Based on previous literature, in this work we utilize machine learning methods to forecast the closing price of the green bond index. The index predicted in the empirical study is the CUFE-CNI High Grade Green Bond Index, which appears to be one of the most representative green bond indices in China. In terms of choosing predictor variables, we are inspired by the literature of bond returns and stock indices prediction. Historical prices and other trading indicators are widely acknowledged predictors to forecast financial markets (Jiang, 2021). Owing to the limited transaction information about green bonds, we select several historical trading indicators, including the closing price, the opening price, the trading volume, the turnover of trading volumes, and the daily return rate. Moreover, many studies have proven that there are significant relations between the green bond market and other financial markets (e.g., stock market, crude oil market, carbon emission market), implying that indices from other markets can be effective predictors (Reboredo, 2018; Reboredo and Ugolini, 2020; Dutta et al., 2021). The co-movements, however, do not necessarily lead to the predictability of green bond market unless the leading roles of other markets are confirmed. Thus, the Grey relational analysis is applied to examine whether our predictor variables are leading indicators of the closing price of the green bond index. In addition, as the machine learning method is largely used in predicting financial series, we use Long Short-Term Memory Networks (LSTM), an effective model in stock indices prediction for its advantages of combining long-term and short-term information, to forecast the green bond index (Cao et al., 2019; Sanboon et al., 2019; Sethia and Raut, 2019). Since the green bond index is unstable and nonlinear, the Empirical Mode Decomposition (EMD) and Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) are introduced to decompose the green bond index into several intrinsic mode functions (IMFs) and a residue. After decomposing, the original index is separated into several stable series, and thus, the prediction accuracy could be enhanced.

Given the results of our empirical study, it is suggested that CEEMDAN-LSTM model is the most effective tool in analyzing the future trends of the green bond market. We also apply four loss functions in this study to measure the out-of-sample prediction errors of CEEMDAN-LSTM, EMD-LSTM, and LSTM. The results of loss functions also indicate CEEMDAN-LSTM model is optimal. Meanwhile, our paper demonstrates that indices from the crude oil market and green stock market are both suitable predictor variables for the green bond market, as the prediction accuracy is significantly improved after the two indices are involved in our model.

The contribution of our work is twofold. On the one hand, our paper is the first to predict the future trend of the green bond market. As indices are normally used to assess the overall performance of assets, investors in the green bond market will greatly benefit from the prediction of the green bond index (Partridge and Medda, 2018). On the other hand, by examining the predictability of the stock index and crude oil index to the green bond index, we reinforce the finding that there are co-movements between the green bond market and other financial markets.

The rest of the paper is organized as follows: *Literature Review* highlights the previous work about the green bond market and predicting methods; *Data and Methodology* introduces the data and methods we have employed in this paper, including EMD, CEEMDAN, Grey relational analysis, LSTM, and some loss functions; *Empirical Results* presents the process of training and the out-of-sample prediction results; and *Conclusion* summarizes the whole study and puts forward the conclusion.

LITERATURE REVIEW

In this section, we first summarize literature about two topics: green bond market and financial market prediction models, which are both closely related to our work. After presenting the existing literature, we also illustrate how our work is related to and different from these studies.

Green bond market

Recently, there has been a wide concern about some environment-friendly financial products. Broadly speaking, these green financial instruments play important roles in both environmental protection and financing. *Copenhagen Accord* introduced in 2009 points that financial innovation is a powerful way to defeat against global warming. Many economies are also in agreement that it is urgent to transform the economic development mode by attracting investors to green their portfolios, and one of the most effective ways is to provide more appealing green financial instruments (Piñeiro-Chousa et al., 2021). On the contrary, a few works also raise the different voices. For example, Bracking (2015) argues whether the virtual green financial assets could boost the development of real asset markets, after studying the case of the Clean Development Mechanism in South Africa. Christophers (2019) also questions the correlations between green financial derivative market and energy commodity market.

At the same time, green financial tools, serving as a kind of financial innovation product, can exert positive influence in two aspects: the companies' perspective and the investors' perspective. Going green has been one of the managing aims in many companies. Some studies have shown that companies are able to get extra green premium by providing socially responsible products or investing in green projects (Besley and Ghatak, 2007; Orlitzky et al., 2011). Friede et al. (2015) find a positive relationship between environmental, social, and governance (ESG) investing and corporate financial performance by studying 2,200 empirical cases. Tang and Zhang (2019) believe the institutional ownership will increase after a company announces the issuance of green stocks. Other studies also show that companies who pay much more attention on sustainability and environment protection normally behave better in financial performance (Khan et al., 2016; Trinks et al., 2018). Zerbid (2019), on the other hand, finds there is a small negative premium on green bonds. From investors' perspective, the issuance of green bonds is likely to promote the disclosure of corporate ESG information (Piñeiro-Chousa et al., 2021). Flammer (2021) suggests that companies' long-term value and environmental performance can be enhanced after issuing green bonds, which will benefit long-term green investors.

However, while environmental stocks have been on heated discussion, the study in green bonds is relatively limited. One reason is that compared with other kinds of financial tools, the green bond market still occupies a relatively small share (less than 1%) of the whole bond market (CBI, 2018), and since it is a newly-emerged financial product, the data related in this field are also inadequate. Due to the insufficient materials, there are still some research gaps in this field.

Previous works on green bonds mainly focus on the green premium of green bonds, as well as the dynamic relationship between the green bond market, and other types of financial markets (Hachenberg and Schiereck, 2018). Febi et al. (2018) use the LOT liquidity model raised by Lesmond et al. (1999) to explain the yield spread of green bonds and suggest that the liquid risk of green bonds is so minor that it can be negligible. Sheng et al. (2021) examine the green bonds issuance in China and propose that there is a negative premium in green bonds issuing, which is more significant in state-owned enterprises. Liaw (2020) reports an opposite conclusion after surveying and believes that compared with traditional bonds, the yield of green bonds is lower.

As for the relations between the green bond market and other markets, most studies are concentrated on their spillover effects. Reboredo and Ugolini (2020) find the green bond market is closely connected with the fixed-income and currency markets, while the correlations between the green bond market and the stock and energy markets are weak. Dutta and Noor (2021) examine the correlations between the climate bond market and other markets during COVID-19, and the empirical result suggests that there are bidirectional spillover effects between the climate bond market and the stock, gold, and oil markets. Hammoudeh et al. (2020) apply a novel time-varying Granger causality test in the study and find the significant relationships between green bond index and the US 10-year Treasury bond index and the carbon dioxide emissions index. Meanwhile, some scholars also raise other factors that will affect the price of green bonds. Pham et al. (2020) find that the attention of investors can influence the returns and volatility of green bonds. They use Google Search Volume Index (GSVI) to represent investor attention and several green bond indices to represent the performance of green bond market. The vector auto-regression (VAR) model is employed to explore the correlation between these two variables. The result has shown that the relationship between investor attention and green bond index varies over time, but in the short term, the relation is stronger. Similarly, Piñeiro-Chousa et al. (2021) analyze how the social network will influence the green bond market and argue that investor sentiment plays an important role in green bond market fluctuation.

Although the contemporaneous correlations and causality between green bond market and other financial markets are on heated discussion, few studies illustrate the lagged relationships among these markets. Our work contributes to the lagged correlations between the green bond market and other financial markets (crude oil and stock markets) by testify the ability of crude oil and stock prices (represented by the Crude Oil Price Index and the CNI EP Index) to predict the green bond price (represented by the CUFE-CNI High Grade Green Bond Index). The empirical result demonstrates that in China, the crude oil index and stock index are both effective predictor variables in forecasting the green bond market, implying the transmission of lagged prices between the green bond market and the other two markets.

Prediction models

The predictability of financial markets has been a classic research topic in financial fields. For this problem, Fama (1965) gives a discouraging answer by proposing an efficient market hypothesis. As markets are efficient, prices vary following random paths, suggesting no analysis can be utilized to predict the markets accurately. However, there are some opposite voices. Many empirical studies have shown that future trends of various financial markets can be predicted, which may ascribe to some psychological factors and the immature markets (Henrique et al., 2019). Traditional prediction of financial instrument prices is based on technical analysis and fundamental analysis (Jiang, 2021). A series of classic time series prediction techniques, like moving average and auto-regressive, are also employed to forecast financial markets (Kumar and Thenmozhi, 2014). Thanks to the fast development of artificial intelligence, more and more advanced methods such as machine learning are used in the field of predictions. Given the nonlinear, unstable, noisy, dynamic nature of financial time series, it is reasonable to apply machine learning into financial market prediction (Hsu et al., 2016; Bezerra and Albuquerque, 2017; Zhang et al., 2017; Shah et al., 2019).

Among various machine learning approaches, the neutral network algorithm appears to be the most accurate method in financial market prediction (Li and Ma, 2010). Kim et al. (2021) have employed several approaches to forecast the corporate bond yield spreads, including linear regression, nonlinear regression, support vector machine (SVM), random forest, and neutral network. Among them, neutral network is reported to outperform any other technique. Similarly, after comparing the prediction result, Gao and Chai (2018) find the recurrent neural network (RNN) works most accurately when it comes to the prediction of stock indices. Based on different types of neutral networks, some studies have already created new hybrid forecasting models. Sun et al. (2019) have put forward a new model combining the auto-regressive and moving average model (ARMA), generalized auto-regressive conditional heteroskedasticity (GARCH), and neutral network to detect the shock hitting the US stock market by using the high-frequency data of the US stock market. This model proves to forecast the high-frequency market accurately in their study. Huynh et al. (2017) propose a new method based on bidirectional gated recurrent unit (BGRU) in order to investigate the potential relation between investor sentiment and stock price. Consequence shows that the prediction accuracy can be enhanced to nearly 60% when BGRU is used in forecasting S&P 500 index and it stills perform well in corporate stock prediction.

More and more researchers pay attention to the powerful predicting ability of LSTM in financial markets forecasting. Compared with other machine learning techniques, LSTM is considered to be the advanced RNN, capable of combining short-time memory and long-time memory (Zhang et al., 2018; Kamal et al., 2020). Because of its advantages, it is frequently used to predict financial data. Akita et al. (2016) employ the LSTM method to forecast 10 listed companies' stock prices, based on textual information collected from newspapers articles. The experiment demonstrates that the effectiveness of LSTM is higher than multi-layer perceptron (MLP) and support vector regression (SVR) and RNN. Gite et al. (2021) collect the information from a famous Indian financial news website and create a sentiment indicator to predict stock price with LSTM. They find that financial news has a great influence on the volatility of stock prices, and the predicting accuracy of LSTM can go up to 96.2%. Lin et al. (2021) use several models to testify whether S&P500 and CSI300 can be forecasted. The result shows that all the models are capable of predicting these stock indices, while the forecasting error rate of CEEMDAN-LSTM model is the lowest.

Nevertheless, since financial product prices are results of a series of combined factors, it is still difficult to estimate the future movements at times. Zhou et al. (2018) point that some factors influence the prices in the short term, while others exert a longer impact. Thus, the financial data would be predicted more accurately once it is decomposed into several parts according to the frequency. In the related literature, EMD and CEEMDAN are commonly utilized to deal with the unstable, volatile financial time series before the sequence is applied into prediction (Xian et al., 2020). Some studies have confirmed the noise reducing ability of EMD and CEEMDAN in market prediction. For example, Vlasenko et al. (2020) has proposed a hybrid model based on EMD and multi-dimensional Gaussian neuro-fuzzy analysis. The prediction result suggests that after the financial time series is decomposed, the prediction accuracy can be enhanced significantly. Cao et al. (2019) also find that the combined model CEEMDAN-LSTM outperforms single LSTM, MLP, and SVM. Lin et al. (2021) draw the similar conclusion, suggesting CEEMDAN is an effective tool in stock indices prediction. Moreover, this paper also compares the performance of EMD-LSTM and CEEMDAN-LSTM. The latter novel method achieves higher accuracy, probably owing to the mode mixing effect of EMD.

All in all, forecasting financial time series is attached with a substantial consequence and significant challenge. Compared to traditional forecasting techniques, the machine learning approach performs better in dealing with unstable and nonlinear financial data. Among various machine learning methods, LSTM is reported to be the most suitable tool to predict financial markets. Since stable financial series with lower volatility will be predicted more accurately, the decomposing approaches EMD and CEEMDAN are often utilized to reduce noise in the forecasting processes. Particularly, as CEEMDAN avoids the problem of mode mixing, it is regarded as a useful tool to predict the prices of various assets (Colominas et al., 2014).

After summarizing the literature about financial data prediction, we are surprised to find there are limited works about bond indices prediction, partly because of the insufficient trading data in the bond market. In particular, no work has predicted the green bond market. According to Ganguli and Dunnmon (2017), though the bond prices are more challenging to forecast for a lack of trading information, the machine learning technique can be used to settle down this problem to some extent. Therefore, our work attempts to fill the study gap by applying hybrid model CEEMDAN-LSTM to predict the green bond index. The result is consistent with the previous work, demonstrating the powerful forecasting ability of LSTM in the prediction of the green bond market.

As green bonds are becoming prevalent these days, transactions in the green bond markets are becoming more frequent as well. As a result, the volatility of the green bond market can change dramatically in short terms, resulting in unexpected market risks. At the same time, serving as a type of promising environmental financial products, green bonds not only help investors to diversify their portfolios, but also play vital roles in mitigating the negative impact of economy development. In that case, they are sure to become important financial tools in the future. Therefore, it is of great significance to study the returns and volatility of the green bond market. The green bond index, consisting of several representative green bonds, is an effective instrument for us to analyze the market. Nonetheless, there is no previous work about the forecasting of green bond indices. Considering the long time-span and the high volatility characteristics of financial time series, we choose the widely used method, LSTM, to predict the green bond index. Also, instability has been one of the major difficulties in dealing with financial data. In this paper, we attempt to use EMD and CEEMDAN to stabilize the green bond index, which improve the prediction accuracy greatly.

DATA AND METHODOLOGY

Data

With the fast development of green bonds in China, several green bond indices have emerged to explain the financial performance of green bonds, including the ChinaBond China Green Bond Index, the CUFE-CNI High Grade Green Bond Index, the FTSE Chinese (Onshore CNY) Green Bond Index, and so on. In this paper, we use CUFE-CNI High Grade Green Bond Index, one of the most representative green bond indices in China, as the benchmark of the green bond market. The CUFE-CNI High Grade Green Bond Index, which consists of labeled and non-labeled green bonds in the China onshore bond market, was launched by the International Institute of Green Finance (IIGF) in the Central University of Finance and Economics (CUFE) and Shenzhen Security Information Co., Ltd. (SSI) in March 2017. Compared with other types of green bond indices, the CUFE-CNI High Grade Green Bond Index mainly focuses on high quality green bonds, including bonds issued by government-related organization or AAA-rated corporations. As the index is designed to present the financial performance of green bonds whose proceeds are used exclusively for environmental projects, the weights are determined by the green asset amount of constituent bonds.

As for predictors, some trading indicators such as opening price and turnover rate are widely used to predict the financial indices. Many studies use the historical price to predict the trend of the stock market (Assis et al., 2018; Chen et al., 2019). Al-Thelaya et al. (2019) augment the predictors into some technical indicators. Dingli and Fournier (2017) also select some technical indicators including momentum, volume, and volatility rate. Given the previous literature, in this paper, we also choose some trading indicators as our predictor variables, including the closing price, the opening price, the trading volume, the turnover of trading volumes, and daily return rate. Daily return rate (*DRR*) describes the increment percentage of today's closing price (P_t) relative to yesterday's closing price (P_{t-1}), and it is calculated as follows:

$$DRR = \frac{P_t - P_{t-1}}{P_t} \tag{1}$$

At the same time, some macroeconomic indicators can also be used to predict the green bond market. Dingli and Fournier (2017) argue that since the financial markets are closely connected, the movement of other markets can result in the changes of the stock market. As a result, they utilize the price of commodities and the currency exchange rate to forecast the stock price. Similarly, Zhong and Enke (2017) employ the factors from bond market and currency exchange market into the forecasting model. Considering the possible spillover effects between the green bond market and other financial markets (Reboredo, 2018), we also take the price changes of other markets into consideration.

In this paper, two price indices from the commodity market and stock market are used to forecast the CUFE-CNI High Grade Green Bond Index as well. The crude oil market is represented by the Crude Oil Price Index, while the environment-friendly stock market is represented by the CNI EP Index. The Crude Oil Price Index is designed to reflect the daily price of crude oil based on the closing prices of WTI and Brent crude oil future contracts. The CNI EP Index is the one of the benchmarks presenting the environment-friendly stocks in China, which comprises 40 representative company stocks related to the environmental protection, accounting for the overall performance of the listed environmental companies in the China A-share market.

For the three indices, we collect daily data from January 4, 2013, to December 31, 2020, to reflect the prices of the green bond market, the crude oil market, and the environment-friendly stock market in China. In order to testify whether the crude oil price and the stock price can be employed to predict the green bond index effectively, we later use the Grey relational analysis to

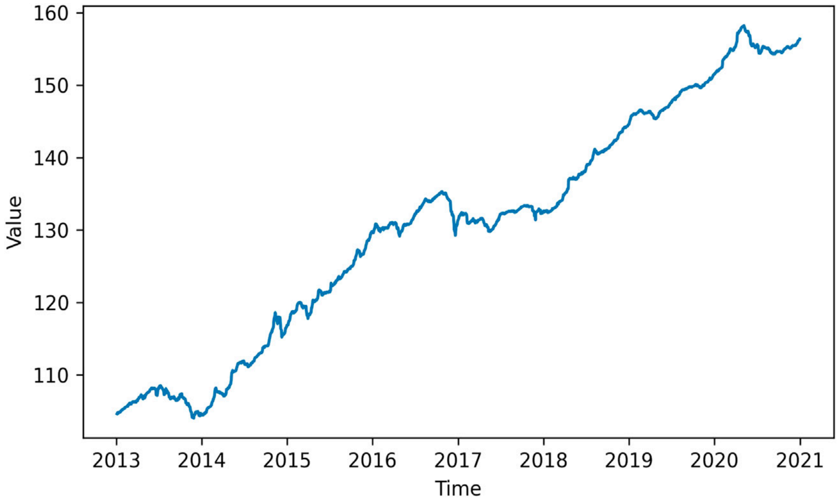


FIGURE 1 | The closing price of the CUFE-CNI High Grade Green Bond Index from 2013 to 2020.

examine the correlations between the lagged two indices and the green bond index. The data of the CUFE-CNI High Grade Green Bond Index and the Crude Oil Price Index are available on the China Stock Market & Accounting Research Database (CSMAR), while the data of the CNI EP Index is obtained from the Wind database.

The data set is separated into two parts: the training set and the test set. Since the data are time series, we use 80% data sorted by chronological order to train the model and the remaining 20% are used for out-of-sample prediction. **Figure 1** presents the closing price of the CUFE-CNI High Grade Green Bond Index from 2013 to 2020. It is clearly shown that it has maintained an upward tendency since 2013. By the end of 2020, the closing price had increased by more than 50% compared with that index in 2013, reaching 156.37. It is obvious that the green bond market has been booming during the past few years, which also indicates investors' growing preferences of green bonds. Meanwhile, we can see there are some small fluctuations in short terms, reflecting the volatility of the market. Therefore, it is essential to stabilize the series by decomposing it into several parts before prediction.

Table 1 gives the statistical description of the CUFE-CNI High Grade Green Bond Index, Crude Oil Price Index, and CNI EP Index. The total number of observations is 1944. Compared with the Crude Oil Price Index and CNI EP Index, the price of the CUFE-CNI High Grade Green Bond Index changed in a considerable small range from 2013 to 2020, indicating it had less volatility. Therefore, the green bond index could be a useful fixed income instrument to diversify the risk of investment portfolios. The Crude Oil Price Index and CNI EP Index, on the other hand, changed dramatically in the 7 years, suggesting the extremely high market risks. **Table 1** also shows that the skewness of the CUFE-CNI High Grade Green Bond Index is −0.0839, suggesting the index is skewed to the left. The kurtosis of the CUFE-CNI High Grade Green Bond Index also indicates that this index does not accord with normal distribution, so it is reasonable to standardize it before predicting. Meanwhile, the distribution of the CUFE-CNI High Grade Green Bond Index is more closed to the normal distribution than the other two indices, which means it can be predicted more accurately by means of machine learning.

Methodology

In this paper, three machine learning models are used for forecasting the green bond index, which are CEEMDAN-LSTM, EMD-LSTM, and LSTM. When choosing our predictor variables, our paper takes the possible lagged correlations between the green bond market and other financial markets into consideration based on previous literature. In order to testify whether the crude oil index and the stock index are suitable predictors, following (Hou et al., 2018), we employ the Grey relational analysis to examine the correlations between the predictors and the next day's closing price. Meanwhile, as the green bond index is unstable, the CEEMDAN and EMD are utilized to decompose the index into several sequences according to their frequency. Finally, after they are normalized, these time sequences are used to predict the future green bond trend with LSTM.

EMD

EMD is an adaptive signal time-frequency processing method (Huang et al., 1998). It decomposes the time series into a number of IMFs, according to the time scale feature of data. EMD is widely used in predicting stock price (Wang and Wang, 2017, Rezaei et al., 2021), sovereign bond yield (Wang et al., 2017), and crude oil price (Yu et al., 2008; Zhang et al., 2009).

The specific decomposition process of EMD is as follows. First, for an original data sequence $s(t)$, find all its maximum points as the upper envelope and all its minimum points as the lower envelope by using cubic spline interpolation, and work out $m(t)$ as the mean value of the upper envelope and the lower envelope. Then calculate the intermediate signal $h_1(t) = s(t) - m(t)$ and judge whether it is an IMF. If so, define $h_1(t) = IMF_1(t)$. Next,

TABLE 1 | The statistical description of three indexes

Index	Count	Min	Max	Mean	Standard deviation	Skewness	Kurtosis
CUFE-CNI High Grade Green Bond Index	1944	104.03	158.23	130.7882	15.9270	-0.0839	1.9325
Crude Oil Price Index	1944	192.64	1446.87	826.4793	295.0567	0.6810	2.3139
CNI EP Index	1944	2068.17	5593.01	3152.3180	624.4271	0.7550	3.7546

calculate $m_1(t)$ as the mean value of the upper and lower envelope of $IMF_1(t)$ and determine the second component $IMF_2(t)$ using the same method. Then, repeat this process for n times until the residue $r_n(t)$ is a constant or a monotone function. Finally, the decomposition process is terminated. $s(t)$ is decomposed into several IMFs and a residue, as shown in the following:

$$s(t) = \sum_{i=1}^{n} IMF_i(t) + r_n(t), i = 1, 2, \ldots, n \tag{2}$$

where t is time, $s(t)$ is the original sequence, $IMF_i(t)$ is the i_{th} IMF decomposed from $s(t)$, and $r_n(t)$ is the residue.

CEEMDAN

Although EMD does well in decomposing time series and has the adaptability to process data with complex structure, this method was criticized for its mode mixing effect. In order to solve this problem, Ensemble Empirical Mode Decomposition (EEMD) was proposed when the normally distributed white noise is added to the original sequence, largely eliminating mode mixing in EMD (Wu and Huang, 2009); but at the same time, this method has a new problem, that is, white noise cannot be completely cancelled after lumped average, resulting in reconstruction errors. Therefore, CEEMDAN was put forward by Torres et al. (2011), which adds adaptive noise to each component decomposed by EEMD, settling down the mode mixing and reconstruction errors simultaneously. Similarly, the CEEMDAN method has many applications in the studying stock market (Jothimani and Yadav, 2019), commodity market (Li et al., 2019; Zhou et al., 2019), and sovereign credit default swap market (Li et al., 2021).

Different from EMD, the Gaussian white noise sequence with standard normal distribution $u^i(t)$ is added to the original signal $s(t)$ in the first step. ε_0 denotes the noise coefficient, m refers to the times of white noise sequence, and the original sequence is expressed as follows:

$$s^i(t) = s(t) + \varepsilon_0 u^i(t), i = 1, 2, \cdots, m \tag{3}$$

The decomposing process is then performed to obtain the first $IMF_1(t) = \frac{1}{m}\sum_{i=1}^{m} IMF_1^i(t)$. After the first IMF component is acquired, we calculate the first residue signal $r_1(t) = s(t) - IMF_1(t)$.

After obtaining the residue, the second component is expressed as follows:

$$r_1(t) + \varepsilon_1 E_1(u^i(t)) = IMF_2^i(t) + r_2^i(t), i = 1, 2, \cdots, m \tag{4}$$

$$IMF_2(t) = \frac{1}{m}\sum_{i=1}^{m} IMF_2^i(t) \tag{5}$$

where $E_1(u^i(t))$ denotes the first IMF component obtained by EMD, and sequence $r_1(t) + \varepsilon_1 E_1(u^i(t))$ is decomposed by EMD to get $IMF_2^i(t)$.

And for the remaining process ($j = 2, \ldots, n$), $(j-1)_{th}$ residue $r_{j-1}(t) = r_{j-2}(t) - IMF_{j-1}(t)$ and we can get j_{th} IMF component $IMF_j(t) = \frac{1}{m}\sum_{i=1}^{m}(E_{j-1}(r_{j-1}(t)) + \varepsilon_{j-1}E_{j-1}(u^i(t)))$. Repeat the steps above until the last IMF of CEEMDAN cannot be decomposed. Finally, the original sequence is decomposed as

$$s(t) = \sum_{j=1}^{n} IMF_j + r_n(t) \tag{6}$$

LSTM

As is mentioned above, LSTM is a developed type of RNNs. Compared with other neutral networks, RNN allows information to persist for a long time, making it possible to use historical information. But RNN has the problem of long-term dependencies, which means the gradient vanishing would occur when the time sequence is long (Bengio et al., 1994). LSTM, which is capable of learning long-term series, was proposed to settle down the problem, (Hochreiter and Schmidhuber, 1997). It adds memory units to each neural unit of hidden layer, so that the memory information of time series can be controlled. Because of its unique structure, it is more suitable for processing and predicting time series problems. Thus, it is widely applied in analyzing financial markets (Kim and Won, 2018; Livieris et al., 2020; Vidal and Kristjanpoller, 2020). The calculation process can be separated into the following steps.

First, put data into forget gate and determine what should be discarded. At time t, the forget gate will get the input x_t and the previous output h_{t-1}. The inputs are processed by the corresponding weight matrix W plus the corresponding bias vector b, then we use a sigmoid layer to get rid of some information.

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{7}$$

At the same time, calculate the value of the input gate to determine the new inputs that need to be retained. We use the input gate to update the value and a tanh layer to create a vector of new candidate values.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{8}$$

$$\tilde{C}_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{9}$$

Then the memory cell C_{t-1} will get a new state value after forgetting the previous memory and absorbing in new inputs.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{10}$$

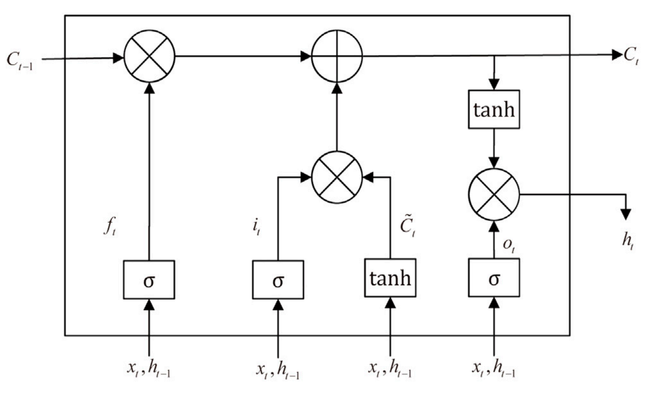


FIGURE 2 | The structure of the LSTM cell.

Finally, get the final outputs through the forget gate. We use the output gate to determine what kind of state values will be output. Then we use a tanh activation function to calculate the candidate value of current state value C_t and it is multiplied by the output of the sigmoid gate.

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

$$h_t = o_t * tanh(C_t) \tag{12}$$

The cell structure of LSTM is shown in **Figure 2**.

Grey relational analysis

The Grey relational analysis method is used to measure the correlations among factors according to the degree of similarity. It has no requirements on sample size and statistical rules, so it is a common method when studying the relationship between variables (Feng et al., 2009). For example, Malinda and Chen (2021) use the Grey relational analysis method in predicting consumer exchange-traded funds (ETFs) and find four main factors influencing consumer ETFs from eight different variables, which are EUR/USD exchange rate, Commodity Research Bureau Index, New York Stock Exchange Composite Index, and put/call ratio. In addition, Chen et al. (2014) utilize Grey relational analysis and artificial neural network to predict the return of real estate investment trust and find that Grey relational analysis is of great significance in correcting predication errors.

The calculation process of this method is as follows. First, determine the reference sequence that reflects the characteristics of the system behavior and the comparison sequences that affect the system behavior. Then, make the reference sequence and comparison sequence dimensionless and calculate the Grey relational coefficient between reference sequence $X_0 = \{x_0(1), x_0(2), \cdots, x_0(n)\}$ and comparison sequences $X_i = \{x_i(1), x_i(2), \cdots, x_i(n)\}, i = 1, 2, \cdots, m.$

The Grey relational coefficient $\xi(x_i)$ is calculated as follows:

$$\xi_{0i} = \frac{\underset{i}{min}\ \underset{t}{min}|x_0(t) - x_i(t)| + \rho\, \underset{i}{max}\ \underset{t}{max}|x_0(t) - x_i(t)|}{|x_0(t) - x_i(t)| + \rho\, \underset{i}{max}\ \underset{t}{max}|x_0(t) - x_i(t)|} \tag{13}$$

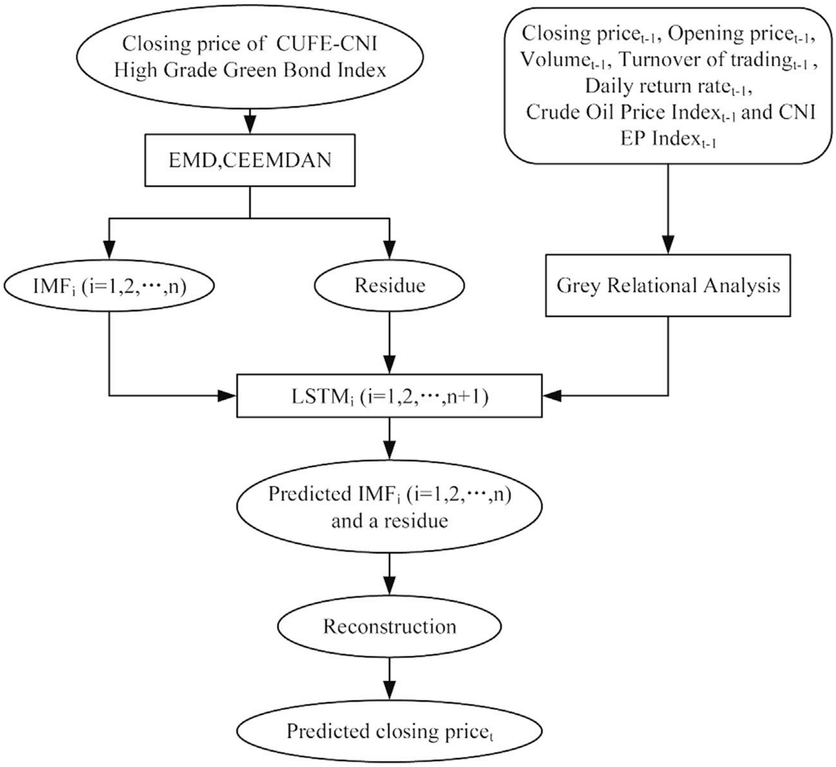

Closing price of CUFE-CNI High Grade Green Bond Index
Closing price$_{t-1}$, Opening price$_{t-1}$, Volume$_{t-1}$, Turnover of trading$_{t-1}$, Daily return rate$_{t-1}$, Crude Oil Price Index$_{t-1}$ and CNI EP Index$_{t-1}$
EMD,CEEMDAN
IMF$_i$ (i=1,2,⋯,n)
Residue
Grey Relational Analysis
LSTM$_i$ (i=1,2,⋯,n+1)
Predicted IMF$_i$ (i=1,2,⋯,n) and a residue
Reconstruction
Predicted closing price$_t$

FIGURE 3 | The flowchart of our integrated forecasting model.

where ρ denotes the identification coefficient (in the range of 0–1, usually $\rho = 0.5$), $\underset{i}{min}\,\underset{t}{min}|x_0(t) - x_i(t)|$ denotes the two-stage minimum difference, and $\underset{i}{max}\,\underset{t}{max}|x_0(t) - x_i(t)|$ denotes the two-stage maximum difference.

After that, we calculate the correlation level $r_i = \frac{1}{n}\sum_{t=1}^{n}\xi_i(t)$ to make comparisons. The correlation level is higher when the value of r_i is closed to 1. Finally, we rank these correlation levels by finding out the maximum r_i.

Integrated forecasting model

Most previous literature have confirmed that the combined CEEMDAN-LSTM model have excellent performance in financial series prediction (Hu, 2021; Wang et al., 2021; Weng et al., 2021). In this paper, we also apply Grey relational analysis, CEEMDAN, and LSTM into the prediction of the green bond Index. The flow chart of our integrated forecasting model is presented in **Figure 3**. First, based on the previous literature, we have chosen five trading indicators of the green bond index as the predictors (closing price, opening price, volume, turnover of trading volumes, and daily return rate). Considering the dynamic correlations between the green bond market and other financial markets, we innovatively introduce the crude oil index and green stock index as predictors as well. After that, the Grey relational analysis is used to testify whether the predictors we choose are capable of predicting the green bond index. In this part, we use the closing price as the reference sequence and the lagged predictors as the comparison sequences to examine the correlations between them. If the correlations are significant, it means that the closing price is closely related to the first lag of predictors, suggesting our predictors can be used to predict the bond index ahead of time.

At the same time, in order to stabilize the green bond index, we use CEEMDAN and EMD to decompose the index, respectively. This index is decomposed into several IMFs with different signal frequency and a residue that stands for the trend. These sequences are then used as the inputs of LSTM model. After the predicted sequences are output from LSTM, the forecasted green bond index can be obtained by summing up the predicted IMF sequences and the predicted residue sequence, as follows:

$$prediction(t) = \sum_{i=1}^{n} IMF_i(t) + r_n(t), i = 1, 2, 3, \cdots, n \tag{14}$$

where $prediction(t)$ is the result of forecasting model, and n stands for the number of IMFs.

Before testing the prediction ability of our LSTM model, we have to train the model first. In this study, we divide our dataset into two parts: 80% of the data serve as the training data and the remaining 20% are used for out-of-sample prediction. In order to select the optimal model, we have built up three models in this paper: CEEMDAN-LSTM, EMD-LSTM, and LSTM model. These three models differ in the data processing of the green bond index. In CEEMDAN-LSTM and EMD-LSTM models, the predicted sequences are the IMFs and residue decomposed by CEEMDAN or EMD, respectively, while in the LSTM model, the original green bond index is used as the predicted sequence directly.

Evaluation criteria

Prediction accuracy means the similarity between the predicted value and the actual value. The closer the predicted value is to the actual value, the higher the prediction accuracy is. Following (Huang et al., 2005; Cao et al., 2019), we adopt four loss functions (MSE, RMSE, MAE, and MAPE) to evaluate the accuracy of different prediction models.

We use $\hat{y}$ to represent the predicted value and y to represent the real value. $h + 1$ is the start date of prediction, and $h + n$ is the end date of prediction. n refers to the total number of days.

1) Mean square error (MSE) represents the mean of squares of the distances between each predicted value and the actual value. The greater the MSE is, the greater the errors are.

$$MSE = \frac{1}{n}\sum_{i=h+1}^{h+n}(y_i - \hat{y}_i)^2 \tag{15}$$

2) RMSE stands for root mean square error. The relationship between MSE and RMSE is similar to the difference between variance and standard deviation.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=h+1}^{h+n}(y_i - \hat{y}_i)^2} \tag{16}$$

3) Mean absolute error (MAE) is similar to RMSE and represents the mean of absolute value of the distances between each predicted value and the actual value.

$$MAE = \frac{1}{n}\sum_{i=h+1}^{h+n}|y_i - \hat{y}_i| \tag{17}$$

4) Mean absolute percentage error (MAPE) compares the difference between the predicted value and the actual value to the actual value to see how much it is accounted for.

$$MAPE = \frac{1}{n}\sum_{i=h+1}^{h+n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{18}$$

EMPIRICAL RESULTS

The result of Grey relational analysis

In order to evaluate whether the predictors we utilize are reasonable, we apply the method of Grey relational analysis to testify the relationships between one-day-lagged predictors and the closing price of the green bond index. **Table 2** shows the result of Grey relational analysis, presenting that the Grey relational grades of these seven indicators are all above 0.75. It proves that the lagged variables are closely related to the closing price. Among the predictors, the lagged closing price and the lagged opening price are highly correlated with the closing price. It means that there exists first-order auto correlation in the closing price series. Therefore, we are able to predict the future trends of the green bond index based on historical data. Besides, the relationships between the closing price and other trading indicators are also demonstrated. It is worth noticing that the lagged crude oil index and lagged stock index are both closely related to the green bond index, indicating that the price changes in the crude oil market and green stock market will exert influence on the next day's green bond market. Therefore, it is reasonable to dig out relationships between green bond markets and other

TABLE 2 | The results of Grey relational analysis on predictors

Predictors	Lagged closing price	Opening price	Volume	Turnover	DRR	Crude Oil Price Index	CNI EP Index
Grey relational grade	0.9986	0.9976	0.8424	0.8424	0.7994	0.7544	0.8425

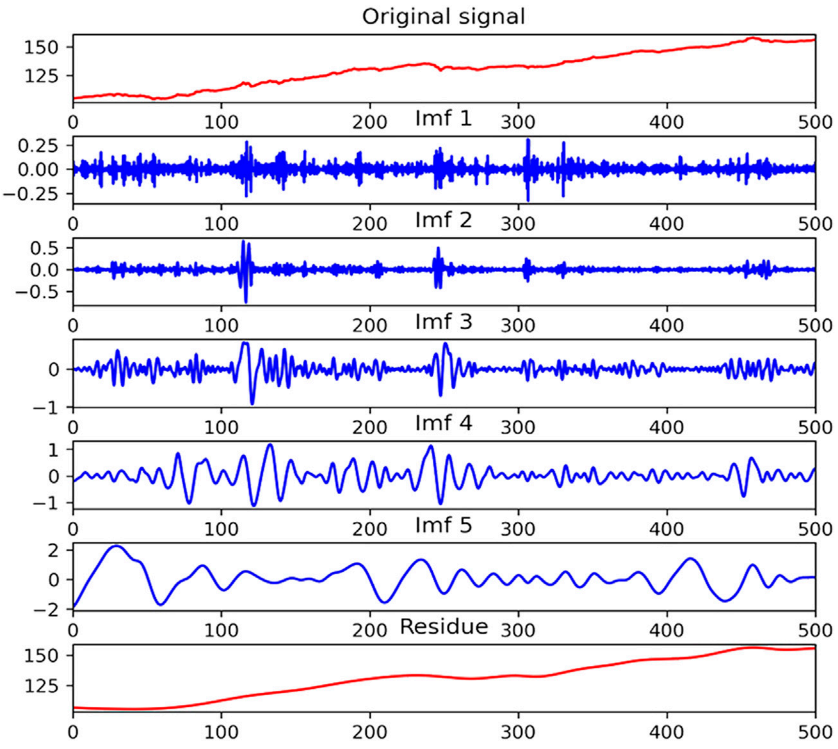


FIGURE 4 | The IMFs decomposed by CEEMDAN.

financial markets. In a word, the predictors we select are highly related to the next day's closing price of green bond index, which means they would lead the green bond market to a certain extent and can be used in the prediction of CUFE-CNI High Grade Green Bond Index.

EMD and CEEMDAN methods

The EMD and CEEMDAN methods are widely applied in decomposing time series. Since (Huang et al., 2005) first introduced EMD to predict the stock price, many studies have utilized it to decompose the financial time series into several sequences for predicting various financial markets. CEEMDAN is later raised based on EMD to mitigate mode mixing and reduce noise. Compared with the original sequence, the decomposed sequences are more stable and smoother, which can be predicted more accurately. In our study, we use CEEMDAN and EMD, respectively, to decompose the CUFE-CNI High Grade Green Bond Index into five IMFs and a residue. These components are later used to predict the index through LSTM.

As is shown in **Figures 4**, **5**, IMFs with higher frequency are placed at the higher places. In fact, the high-frequency sequences are often viewed as the noise in the green bond market, while the low-frequency sequences represent the fluctuations. Also, the residue representing the basic trend of the green bond index is arranged at the bottom of two figures. **Table 3** gives a brief statistical description of the IMFs decomposed by CEEMDAN. It can be seen that all IMFs pass the Augmented Dickey-Fuller test under the statistical significance of 1%, suggesting they are all stationary series.

Training process and the results

After decomposing, the IMFs are supposed to be normalized before training. In this study, we use the following normalization equation:

$$X_{Norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{19}$$

where X_{Norm} refers to the data after normalization, X represents the original data, X_{max} is the maximum of the series, and X_{min} is the minimum.

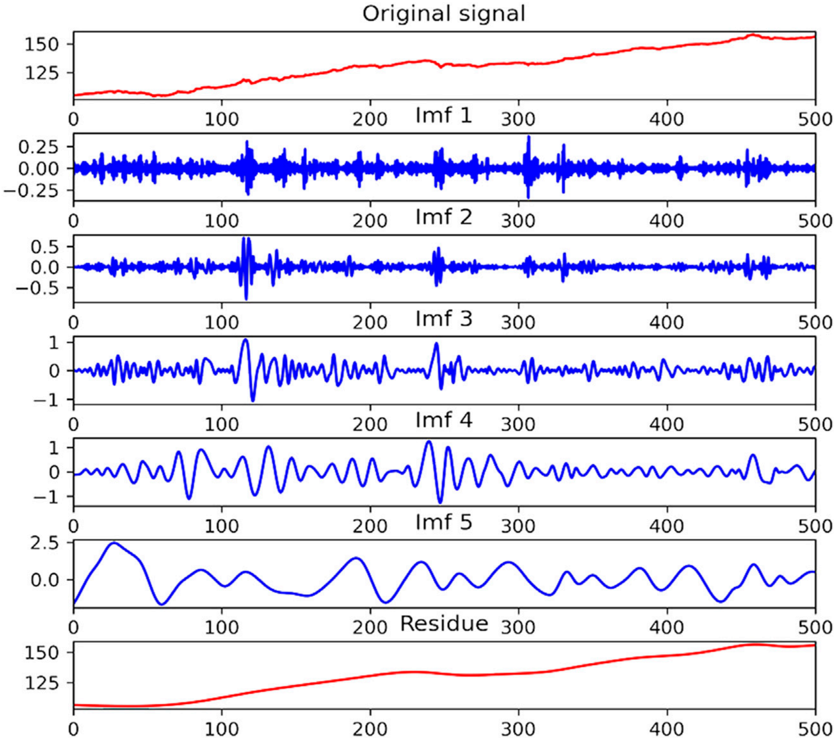


FIGURE 5 | The IMFs decomposed by EMD.

TABLE 3 | The statistical description of IMFs

Components	Count	Min	Max	Mean	Standard deviation	Skewness	Kurtosis	ADF test
IMF1	1944	−0.3195	0.2988	−0.0001	0.0519	0.1104	5.7492	−14.2975***
IMF2	1944	−0.5221	0.5248	−0.0003	0.0663	−0.0955	15.1253	−18.7529***
IMF3	1944	−0.5385	0.5044	0.0026	0.1359	0.0371	1.6438	−16.8431***
IMF4	1944	−1.6634	1.4910	0.0196	0.3877	0.2593	4.1055	−10.1823***
IMF5	1944	−2.3431	2.1464	0.0378	0.7417	0.1651	0.8597	−6.3438***

*Notes: *** represents the statistical significance of 1%.*

To get the optimal predicted green bond index, we have trained the model many times, and the parameters are set as follows. In the process of CEEMDAN decomposing, the number of white noise trials is 50. As for the hyper-parameters in LSTM, after many experiments, we finally choose the number of epochs, which is the total times of training, as 175; and the figure of batch size, which refers to the samples captured in one training session is 64. In the first of hidden layer, the number of neurons is 50 and there is 1 neuron in the output layer to predict the closing price.

Figures 6–8 present the predicted curves of the CUFE-CNI High Grade Green Bond Index, through CEEMDAN-LSTM, EMD-LSTM, and LSTM models, respectively. It is transparent that, compared with original series, IMFs obtained through the CEEMDAN and EMD methods could be used to predict the future trend better. From **Figure 8**, we can see that the predicted curve of LSTM is of high volatility, suggesting there are more noise signals that weaken the accuracy of predicted value. Meanwhile, CEEMDAN-LSTM and EMD-LSTM all perform relatively well in forecasting, which means the prediction accuracy can be significantly enhanced when the original series is decomposed according to the signal frequency.

However, since CEEMDAN-LSTM and EMD-LSTM all have good performance in predicting, we apply four loss functions to compare these two models' predicting abilities, and the result is suggested in **Table 4**. The four forecasting criteria we have chosen are MSE, RMSE, MAE, and MAPE. By contrasting forecasting errors of the CEEMDAN-LSTM and EMD-LSTM models, we can easily draw the conclusion that CEEMDAN-LSTM is the optimal method. No matter what kind of loss functions are used to evaluate the prediction performance, the CEEMDAN-LSTM model has the highest forecasting accuracy rate. In addition, the RMSE of the

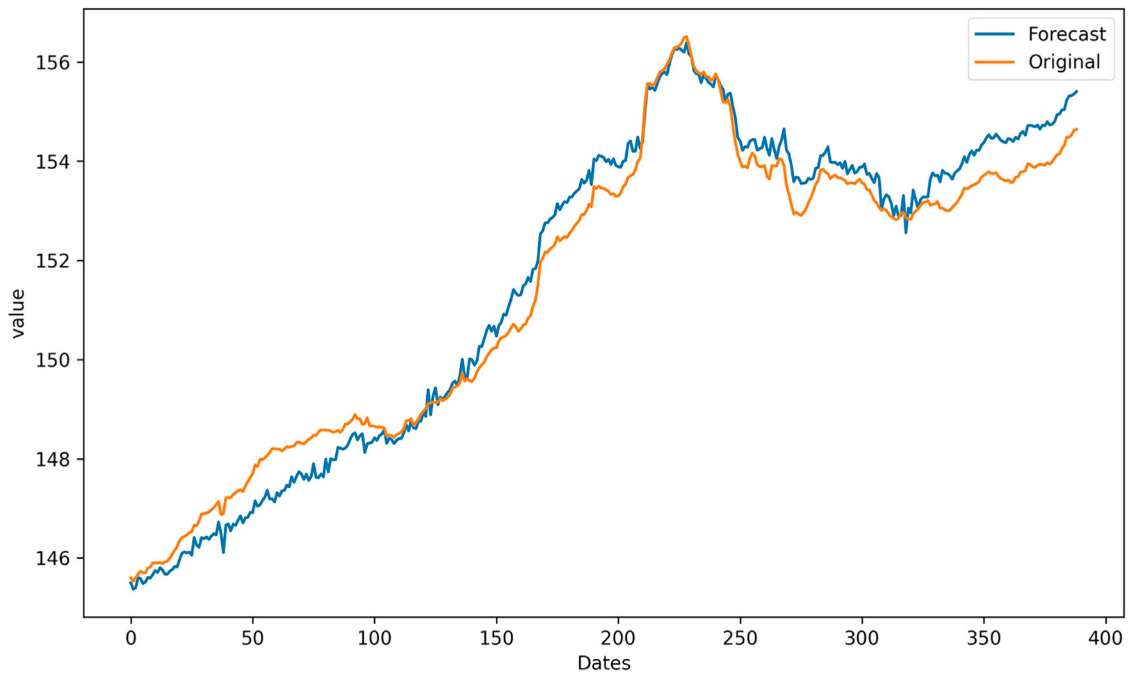


FIGURE 6 | The forecasting results based on CEEMDAN-LSTM.

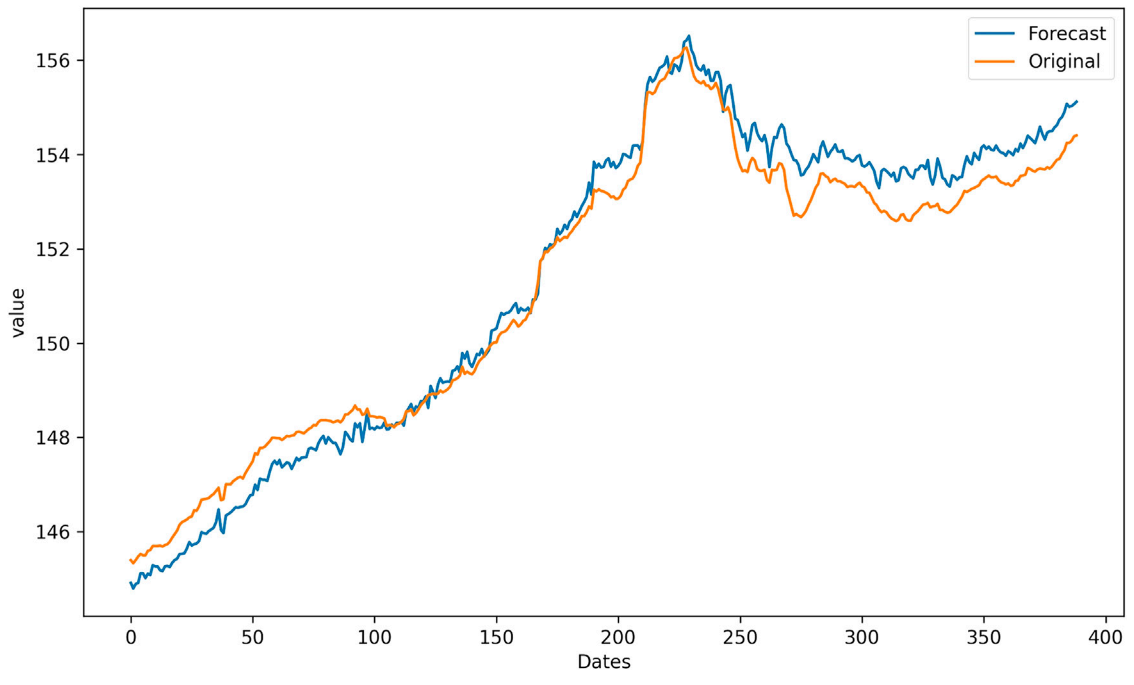


FIGURE 7 | The forecasting results based on EMD-LSTM.

CEEMDAN-LSTM model suggests that on average, the difference between the predicted value and the actual value of the CUFE-CNI High Grade Green Bond Index is 0.267635. Therefore, our CEEMDAN-LSTM model can predict the closing price of the CUFE-CNI High Grade Green Bond Index to a large extent. **Figure 9** shows the predicted closing price of the CUFE-CNI High Grade Green Bond Index based on the CEEMDAN-LSTM model. By contrasting the predicted

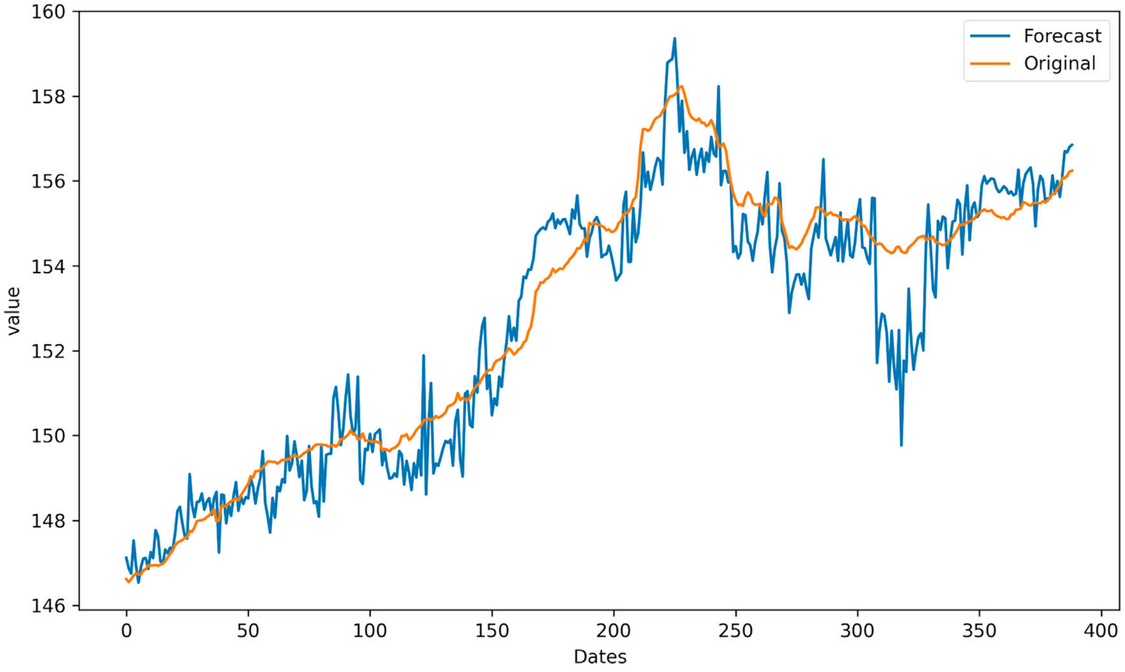


FIGURE 8 | The forecasting results based on LSTM.

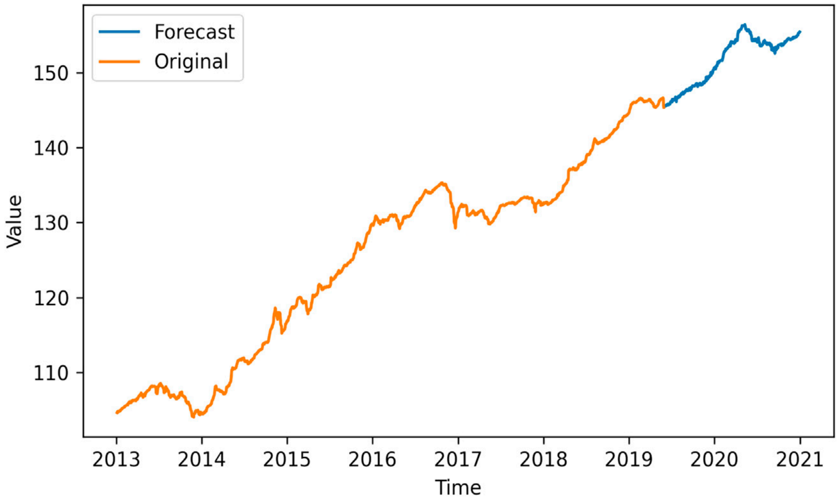


FIGURE 9 | The predicted CUFE-CNI High Grade Green Bond Index based on CEEMDAN-LSTM.

TABLE 4 | The results of prediction using LSTM, EMD-LSTM, and CEEMDAN-LSTM

Model	MSE	RMSE	MAE	MAPE
CEEMDAN-LSTM	0.267635	0.517335	0.446490	0.294982
EMD-LSTM	0.315517	0.561709	0.499066	0.329675
LSTM	0.830592	0.911368	0.745647	0.486262

value and the actual closing price presented in **Figure 1**, we find our model fit the CUFE-CNI High Grade Green Bond Index well. Therefore, the CEEMDAN-LSTM model performs well in forecasting the green bond index.

Further discussion

The relationships between the green bond market and other financial markets have been heatedly discussed by lots of studies. Some papers suggest the correlation is weak (Reboredo and Ugolini, 2020), while others hold the opposite opinion (Dutta and Noor, 2021). Our work also partly answers the question by testing the predicting ability of other markets'

TABLE 5 | The results of prediction when the Crude oil Price Index and CNI EP Index are used or not

Model	Oil	Stock	MSE	RMSE	MAE	MAPE
CEEMDAN-LSTM	Yes	Yes	0.267635	0.517335	0.446490	0.294982
CEEMDAN-LSTM	No	Yes	0.494996	0.703560	0.643508	0.423646
CEEMDAN-LSTM	Yes	No	0.798809	0.893761	0.807019	0.532829
CEEMDAN-LSTM	No	No	1.211728	1.100785	1.023617	0.673265

TABLE 6 | The improvements when different indexes are added in the model

Improvement of loss functions	MSE	RMSE	MAE	MAPE
Improvement with oil Δ_{oil}	0.412919	0.207024	0.216598	0.140436
Improvement with stock Δ_{stock}	0.716732	0.397225	0.380109	0.249619
Improvement with oil and stock $\Delta_{oil+ stock}$	0.944093	0.58345	0.577127	0.378283

indices. When choosing predictors, we also take the price indices of the crude oil market and green stock market into account. **Table 5** compares the forecasting results when the Crude Oil Price Index and CNI EP index are employed or not. It is shown that when Crude Oil Price Index and CNI EP Index are used as predictors, the forecasting performance of CEEMDAN-LSTM can be greatly improved.

In **Table 6**, we give a further analysis on the prediction performance that improved when the Crude Oil Price Index and the CNI EP Index served as the indicators. We can find that if the two indices are not included, the values of the loss functions MSE, RMSE, MAE, and MAPE of the CEEMDAN-LSTM model are 1.211728, 1.100785, 1.023617, and 0.673265 correspondingly. A new indicator Δ is defined as the improvement of the loss function after adding a new predictor. It is shown that when the Crude Oil Price Index or the CNI EP Index is used separately, the forecasting performance of CEEMDAN-LSTM can both be improved. Moreover, there exists $\Delta_{stock} > \Delta_{oil}$ for all loss functions, which illustrates that compared to crude oil index, environment-friendly stocks index is a more important factor to predict the green bond market. Besides, after the two indices are introduced into the model together, we can see that $\Delta_{oil+stock} > \Delta_{oil}$ and $\Delta_{oil+stock} > \Delta_{stock}$, showing that the accuracy of the CEEMDAN-LSTM model, would increase greatly after adding two predictors. This is in coincidence with the previous study, which implies the green bond markets could receive sizable influence from other financial markets.

CONCLUSION

As ESG is attached with greater importance, many financial products designed to help environment-friendly projects are emerging. Green bond, serving as an innovative fixed-income asset, has been appealing to a majority of investors. However, this kind of financial instrument does not exist until 2007, which leads to the limited studies about it. Up to now, the prediction of the green bond market is still a research gap.

Previous literature has illustrated that machine methods (especially LSTM) perform better in terms of financial market prediction, and the prediction accuracy would be enhanced when the data are decomposed by CEEMDAN. Motivated by these studies, our paper proposes a hybrid CEEMDAN-LSTM model to forecast the green bond index (represented by the CUFE-CNI High Grade Green Bond Index). As for the predictor variables, we mainly utilize several technical predictors, including the closing price, the opening price, the trading volume, the turnover of trading volumes, and daily return rate. Considering the potential correlations between green bond market and other financial markets, we also try to use indices from the crude oil market and environmental stock market to forecast the green bond index. To examine the performance of our mixed CEEMDAN-LSTM model, we also apply the EMD-LSTM model and the LSTM model to predict the index.

Our empirical results suggest that compared with the other two models, our CEEMDAN-LSTM model is optimal with considerably high prediction accuracy, which demonstrates the powerful prediction ability of machine learning methods. Besides, our study also shows that the crude oil market and the environmental stock market could exert influence on the green bond market, implying the correlations among these three markets. The indices of these two markets can be used in green bond market forecasting.

Given our findings, several policy implications are put forward as follows:

(a) The rapid development of green bonds will definitely fuel the sustainable economy, which is especially significant for a developing country like China to seize the opportunity and promote the green investment.

(b) Our study has shown that the green bond index could be predicted considerably accurately through historical information. However, the trading information of the bond market is relatively inadequate compared with the stock market, resulting in the difficulty to forecast the future trend. Therefore, it is of great importance to set up the comprehensive information disclosure mechanism in the green bond market, which would also enable investors to green their portfolios effectively.

(c) As demonstrated in our work and previous literature, the green bond market could receive sizeable influence from other markets (e.g., crude oil market, stock market), suggesting the potential risk contagion among financial markets. Thus, more suitable government regulations are supposed to be implemented in order to monitor the financial contagion.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**; further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

JW: Conceptualization, Data Curation, Methodology, Software, Formal analysis, Investigation, Writing-original draft, Visualization. JT: Conceptualization, Investigation, Writing-original draft, Visualization. KG: Conceptualization, Methodology, Investigation, Writing-review and editing, Supervision.

FUNDING

This research was funded by the University of Chinese Academy of Sciences and the Fundamental Research Funds for the Central Universities.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fenrg.2021.793413/full#supplementary-material

REFERENCES

Akita, R., Yoshihara, A., Matsubara, T., and Uehara, K. (2016). "Deep Learning for Stock Prediction Using Numerical and Textual Information," in International Conference on Computer and Information Science (ICIS) (Okayama, Japan: IEEE), 1–6. doi:10.1109/ICIS.2016.7550882

Al-Thelaya, K. A., El-Alfy, E.-S. M., and Mohammed, S. (2019). "Forecasting of bahrain Stock Market with Deep Learning: Methodology and Case Study," in 2019 8th International Conference on Modeling Simulation and Applied Optimization (ICMSAO), 1–5. doi:10.1109/ICMSAO.2019.8880382

Assis, C. A. S., Pereira, A. C. M., Carrano, E. G., Ramos, R., and Dias, W. (2018). "Restricted Boltzmann Machines for the Prediction of Trends in Financial Time Series," in 2018 International Joint Conference on Neural Networks (IJCNN), 1–8. doi:10.1109/IJCNN.2018.8489163

Bauer, M. D., and Rudebusch, G. D. (2017). Resolving the Spanning Puzzle in Macro-Finance Term Structure Models*. *Rev. Finance* 21 (2), 511–553. doi:10.1093/rof/rfw044

Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning Long-Term Dependencies with Gradient Descent Is Difficult. *IEEE Trans. Neural Netw.* 5 (2), 157–166. doi:10.1109/72.279181

Besley, T., and Ghatak, M. (2007). Retailing Public Goods: The Economics of Corporate Social Responsibility. *J. Public Econ.* 91 (9), 1645–1663. doi:10.1016/j.jpubeco.2007.07.006

Bezerra, P. C. S., and Albuquerque, P. H. M. (2017). Volatility forecasting via SVR–GARCH with mixture of Gaussian kernels. *Comput. Manag. Sci.* 14 (2), 179–196. doi:10.1007/s10287-016-0267-0

Bianchi, D., Büchner, M., and Tamoni, A. (2021). Bond Risk Premiums with Machine Learning. *Rev. Financial Stud.* 34 (2), 1046–1089. doi:10.1093/rfs/hhaa062

Bracking, S. (2015). Performativity in the Green Economy: How Far Does Climate Finance Create a Fictive Economy? *Third World Q.* 36 (12), 2337–2357. doi:10.1080/01436597.2015.1086263

Cao, J., Li, Z., and Li, J. (2019). Financial Time Series Forecasting Model Based on CEEMDAN and LSTM. *Physica A: Stat. Mech. its Appl.* 519, 127–139. doi:10.1016/j.physa.2018.11.061

Chen, J.-H., Chang, T.-T., Ho, C.-R., and Diaz, J. F. (2014). Grey Relational Analysis and Neural Network Forecasting of REIT Returns. *Quantitative Finance* 14 (11), 2033–2044. doi:10.1080/14697688.2013.816765

Chen, L., Chi, Y., Guan, Y., and Fan, J. (2019). "A Hybrid Attention-Based EMD-LSTM Model for Financial Time Series Prediction," in 2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD), 113–118. doi:10.1109/ICAIBD.2019.8837038

Chen, Y., and Zhao, Z. J. (2021). The Rise of green Bonds for Sustainable Finance: Global Standards and Issues with the Expanding Chinese Market. *Curr. Opin. Environ. Sustainability* 52, 54–57. doi:10.1016/j.cosust.2021.06.013

Choi, J., and Kim, Y. (2018). Anomalies and Market (Dis)integration. *J. Monetary Econ.* 100, 16–34. doi:10.1016/j.jmoneco.2018.06.003

Chordia, T., Goyal, A., Nozowa, Y., Subrahmanyam, A., and Tong, Q. (2014). Is the Cross-Section of Expected Bond Returns Influenced by Equity Return Predictors? *Res. Collection Lee Kong Chian Sch. Business*. Available at: https://ink.library.smu.edu.sg/lkcsb_research/4521.

Christophers, B. (2019). Environmental Beta or How Institutional Investors Think about Climate Change and Fossil Fuel Risk. *Ann. Am. Assoc. Geogr.* 109 (3), 754–774. doi:10.1080/24694452.2018.1489213

Climate Bonds Initiative (2018). Bonds and Climate Change: State of the Market. Available at: http://www.climatebonds.net/resources/reports/green-bonds-state-market-2018 (Accessed October 10, 2021).

Climate Bonds Initiative (2020). China State of the Market 2020 Report. Available at: http://www.climatebonds.net/resources/reports/china-state-market-2020-report (Accessed November 25, 2021).

Climate Bonds Initiative (2021). Sustainable Debt Highlights H1 2021. Available at: http://www.climatebonds.net/resources/reports/sustainable-debt-highlights-h1-2021 (Accessed November 25, 2021).

Colominas, M. A., Schlotthauer, G., and Torres, M. E. (2014). Improved Complete Ensemble EMD: A Suitable Tool for Biomedical Signal Processing. *Biomed. Signal Process. Control.* 14, 19–29. doi:10.1016/j.bspc.2014.06.009

Connolly, R., Stivers, C., and Sun, L. (2005). Stock Market Uncertainty and the Stock-Bond Return Relation. *J. Financ. Quant. Anal.* 40 (1), 161–194. doi:10.1017/S0022109000001782

Devpura, N., Narayan, P. K., and Sharma, S. S. (2021). Bond Return Predictability: Evidence from 25 OECD Countries. *J. Int. Financial Markets, Institutions Money* 75, 101301. doi:10.1016/j.intfin.2021.101301

Dingli, A., Fournier, K. S., and Fournier, K. S. (2017). Financial Time Series Forecasting - A Deep Learning Approach. *Int. J. Machine Learn. Comput.* 7 (5), 118–122. doi:10.18178/ijmlc.2017.7.5.632

Dutta, A., Bouri, E., and Noor, M. H. (2021). Climate Bond, Stock, Gold, and Oil Markets: Dynamic Correlations and Hedging Analyses during the COVID-19 Outbreak. *Resour. Pol.* 74, 102265. doi:10.1016/j.resourpol.2021.102265

Fama, E. F. (1965). The Behavior of Stock-Market Prices. *J. Bus* 38 (1), 34–105. doi:10.1086/294743

Febi, W., Schäfer, D., Stephan, A., and Sun, C. (2018). The Impact of Liquidity Risk on the Yield Spread of green Bonds. *Finance Res. Lett.* 27, 53–59. doi:10.1016/j.frl.2018.02.025

Feng, D., Qingmei, T., and Xiaohui, L. (2009). "The Relationship between Chinese Energy Consumption and GDP: An Econometric Analysis Based on the Grey Relational Analysis(GRA)," in 2009 IEEE International Conference on Grey

Systems and Intelligent Services (GSIS 2009) (Nanjing, China: IEEE), 153–157. doi:10.1109/GSIS.2009.5408333

Flammer, C. (2021). Corporate green Bonds. *J. Financial Econ.* 142, 499–516. doi:10.1016/j.jfineco.2021.01.010

Fong, T. P. W., and Wu, S. T. (2020). Predictability in Sovereign Bond Returns Using Technical Trading Rules: Do Developed and Emerging Markets Differ? *North Am. J. Econ. Finance* 51, 101105. doi:10.1016/j.najef.2019.101105

Friede, G., Busch, T., and Bassen, A. (2015). ESG and Financial Performance: Aggregated Evidence from More Than 2000 Empirical Studies. *J. Sustain. Finance Investment* 5 (4), 210–233. doi:10.1080/20430795.2015.1118917

Ganguli, S., and Dunnmon, J. (2017). Machine Learning for Better Models for Predicting Bond Prices. *arXiv preprint arXiv:1705.01142*. Available at: https://arxiv.abs/1705.01142.

Gao, T., and Chai, Y. (2018). Improving Stock Closing price Prediction Using Recurrent Neural Network and Technical Indicators. *Neural Comput.* 30 (10), 2833–2854. doi:10.1162/neco_a_01124

Ghoddusi, H., Creamer, G. G., and Rafizadeh, N. (2019). Machine Learning in Energy Economics and Finance: A Review. *Energ. Econ.* 81, 709–727. doi:10.1016/j.eneco.2019.05.006

Giacoletti, M., Laursen, K. T., and Singleton, K. J. (2021). Learning from Disagreement in the U.S. Treasury Bond Market. *J. Finance* 76 (1), 395–441. doi:10.1111/jofi.12971

Gite, S., Khatavkar, H., Kotecha, K., Srivastava, S., Maheshwari, P., and Pandey, N. (2021). Explainable Stock Prices Prediction from Financial News Articles Using Sentiment Analysis. *PeerJ Comput. Sci.* 7, e340. doi:10.7717/peerj-cs.340

Gormus, A., Nazlioglu, S., and Soytas, U. (2018). High-yield Bond and Energy Markets. *Energ. Econ.* 69, 101–110. doi:10.1016/j.eneco.2017.10.037

Gu, S., Kelly, B., and Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. *Rev. Financial Stud.* 33 (5), 2223–2273. doi:10.1093/rfs/hhaa009

Hachenberg, B., and Schiereck, D. (2018). Are green Bonds Priced Differently from Conventional Bonds? *J. Asset Manag.* 19 (6), 371–383. doi:10.1057/s41260-018-0088-5

Hammoudeh, S., Ajmi, A. N., and Mokni, K. (2020). Relationship between green Bonds and Financial and Environmental Variables: A Novel Time-Varying Causality. *Energ. Econ.* 92, 104941. doi:10.1016/j.eneco.2020.104941

Henrique, B. M., Sobreiro, V. A., and Kimura, H. (2019). Literature Review: Machine Learning Techniques Applied to Financial Market Prediction. *Expert Syst. Appl.* 124, 226–251. doi:10.1016/j.eswa.2019.01.012

Hochreiter, S., and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Comput.* 9 (8), 1735–1780. doi:10.1162/neco.1997.9.8.1735

Hou, X., Zhu, S., Xia, L., and Wu, G. (2018). "Stock price Prediction Based on Grey Relational Analysis and Support Vector Regression," in 2018 Chinese Control and Decision Conference (CCDC) (Shenyang, China: IEEE), 2509–2513. doi:10.1109/CCDC.2018.8407547

Hsu, M.-W., Lessmann, S., Sung, M.-C., Ma, T., and Johnson, J. E. V. (2016). Bridging the divide in Financial Market Forecasting: Machine Learners vs. Financial Economists. *Expert Syst. Appl.* 61, 215–234. doi:10.1016/j.eswa.2016.05.033

Hu, Z. (2021). Crude Oil price Prediction Using CEEMDAN and LSTM-Attention with News Sentiment index. *Oil Gas Sci. Technol. - Rev. IFP Energies Nouvelles* 76, 28. doi:10.2516/ogst/2021010

Huang, D., Jiang, F., Tong, G., Tong, G., and Zhou, G. (2020). "Real Time Macro Factors in Bond Risk Premium, SSRN Journal," in Asian Finance Association (AsianFA) 2018 Conference. doi:10.2139/ssrn.3107612

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., et al. (19981971). The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Non-stationary Time Series Analysis. *Proc. R. Soc. Lond. A.* 454, 903–995. doi:10.1098/rspa.1998.0193

Huang, W., Nakamori, Y., and Wang, S.-Y. (2005). Forecasting Stock Market Movement Direction with Support Vector Machine. *Comput. Operations Res.* 32 (10), 2513–2522. doi:10.1016/j.cor.2004.03.016

Huynh, H. D., Dang, L. M., and Duong, D. (2017). "A New Model for Stock Price Movements Prediction Using Deep Neural Network," in Proceedings of the Eighth International Symposium on Information and Communication Technology, 57–62. doi:10.1145/3155133.3155202

Jiang, W. (2021). Applications of Deep Learning in Stock Market Prediction: Recent Progress. *Expert Syst. Appl.* 184, 115537. doi:10.1016/j.eswa.2021.115537

Jothimani, D., and Yadav, S. S. (2019). Stock Trading Decisions Using Ensemble-Based Forecasting Models: a Study of the Indian Stock Market. *J. Bank Financ. Technol.* 3, 113–129. doi:10.1007/s42786-019-00009-7

Kamal, I. M., Bae, H., Sunghyun, S., and Yun, H. (2020). DERN: Deep Ensemble Learning Model for Short- and Long-Term Prediction of Baltic Dry Index. *Appl. Sci.* 10 (4), 1504. doi:10.3390/app10041504

Khan, M., Serafeim, G., and Yoon, A. (2016). Corporate Sustainability: First Evidence on Materiality. *Account. Rev.* 91 (6), 1697–1724. doi:10.2308/accr-51383

Kim, H. Y., and Won, C. H. (2018). Forecasting the Volatility of Stock price index: A Hybrid Model Integrating LSTM with Multiple GARCH-type Models. *Expert Syst. Appl.* 103, 25–37. doi:10.1016/j.eswa.2018.03.002

Kim, J.-M., Kim, D. H., and Jung, H. (2021). Applications of Machine Learning for Corporate Bond Yield Spread Forecasting. *North Am. J. Econ. Finance* 58, 101540. doi:10.1016/j.najef.2021.101540

Kochetygova, J., and Jauhari, A. (2014). Climate Change, green Bonds and index Investing: the New Frontier. Retrieved, 20, 2017. Available at: https://www.spglobal.com/spdji/en/documents/research/research-climate-change-green-bonds-and-index-investing-the-new-frontier.pdf.

Kumar, M., and Thenmozhi, M. (2014). Forecasting Stock index Returns Using ARIMA-SVM, ARIMA-ANN, and ARIMA-Random forest Hybrid Models. *Int. J. Banking Account. Finance* 5 (3), 284–308. doi:10.1504/IJBAAF.2014.064307

Lesmond, D. A., Ogden, J. P., and Trzcinka, C. A. (1999). A New Estimate of Transaction Costs. *Rev. Financ. Stud.* 12 (5), 1113–1141. doi:10.1093/rfs/12.5.1113

Li, J., Hao, J., Sun, X., and Feng, Q. (2021). Forecasting China's sovereign CDS with a decomposition reconstruction strategy. *Appl. Soft. Comput.* 105, 107291. doi:10.1016/j.asoc.2021.107291

Li, T., Zhou, Y., Li, X., Wu, J., and He, T. (2019). Forecasting Daily Crude Oil Prices Using Improved CEEMDAN and ridge Regression-Based Predictors. *Energies* 12 (19), 3603. doi:10.3390/en12193603

Li, Y., and Ma, W. (2010). "Applications of Artificial Neural Networks in Financial Economics: a Survey," in 2010 International symposium on computational intelligence and design (Hangzhou, China: IEEE), 211–214. doi:10.1109/ISCID.2010.70

Liaw, K. T. (2020). Survey of Green Bond Pricing and Investment Performance. *J. Risk Financial Manage.* 13 (9), 193. doi:10.3390/jrfm13090193

Lin, H., Wu, C., and Zhou, G. (2018). Forecasting Corporate Bond Returns with a Large Set of Predictors: An Iterated Combination Approach. *Manage. Sci.* 64 (9), 4218–4238. doi:10.1287/mnsc.2017.2734

Lin, Y., Yan, Y., Xu, J., Liao, Y., and Ma, F. (2021). Forecasting Stock index price Using the CEEMDAN-LSTM Model. *North Am. J. Econ. Finance* 57, 101421. doi:10.1016/j.najef.2021.101421

Livieris, I. E., Pintelas, E., and Pintelas, P. (2020). A CNN-LSTM Model for Gold price Time-Series Forecasting. *Neural Comput. Applic* 32 (23), 17351–17360. doi:10.1007/s00521-020-04867-x

Malinda, M., and Chen, J.-H. (2021). The Forecasting of Consumer Exchange-Traded Funds (ETFs) via Grey Relational Analysis (GRA) and Artificial Neural Network (ANN). *Empir Econ.* 2021, 1–45. doi:10.1007/s00181-021-02039-x

Nazlioglu, S., Gupta, R., and Bouri, E. (2020). Movements in International Bond Markets: The Role of Oil Prices. *Int. Rev. Econ. Finance* 68, 47–58. doi:10.1016/j.iref.2020.03.004

Orlitzky, M., Siegel, D. S., and Waldman, D. A. (2011). Strategic Corporate Social Responsibility and Environmental Sustainability. *Business Soc.* 50 (1), 6–27. doi:10.1177/0007650310394323

Partridge, C., and Medda, F. (2018). The Creation and Benchmarking of a green Municipal Bond index. *SSRN J.* Available at SSRN 3248423. doi:10.2139/ssrn.3248423

Pham, L., and Luu Duc Huynh, T. (2020). How Does Investor Attention Influence the green Bond Market? *Finance Res. Lett.* 35, 101533. doi:10.1016/j.frl.2020.101533

Piñeiro-Chousa, J., López-Cabarcos, M. Á., Caby, J., and Šević, A. (2021). The Influence of Investor Sentiment on the green Bond Market. *Technol. Forecast. Soc. Change* 162, 120351. doi:10.1016/j.techfore.2020.120351

Reboredo, J. C. (2018). Green Bond and Financial Markets: Co-movement, Diversification and price Spillover Effects. *Energ. Econ.* 74, 38–50. doi:10.1016/j.eneco.2018.05.030

Reboredo, J. C., and Ugolini, A. (2020). Price Connectedness between green Bond and Financial Markets. *Econ. Model.* 88, 25–38. doi:10.1016/j.econmod.2019.09.004

Rezaei, H., Faaljou, H., and Mansourfar, G. (2021). Stock price Prediction Using Deep Learning and Frequency Decomposition. *Expert Syst. Appl.* 169, 114332. doi:10.1016/j.eswa.2020.114332

Sadorsky, P. (2021). A Random Forests Approach to Predicting Clean Energy Stock Prices. *J. Risk Financial Manag.* 14 (2), 48. doi:10.3390/jrfm14020048

Sanboon, T., Keatruangkamala, K., and Jaiyen, S. (2019). Singapore: IEEE, 757–760. doi:10.1109/CCOMS.2019.8821776 A Deep Learning Model for Predicting Buy and Sell Recommendations in Stock Exchange of Thailand Using Long Short-Term Memory2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)

Sethia, A., and Raut, P. (2019). "Application of LSTM, GRU and ICA for Stock price Prediction," in *Information and Communication Technology for Intelligent Systems* (Singapore: Springer), 479–487. doi:10.1007/978-981-13-1747-7_46

Shah, D., Isah, H., and Zulkernine, F. (2019). Stock Market Analysis: A Review and Taxonomy of Prediction Techniques. *Int. J. Financial Stud.* 7 (2), 26. doi:10.3390/ijfs7020026

Sheng, Q., Zheng, X., and Zhong, N. (2021). Financing for Sustainability: Empirical Analysis of green Bond Premium and Issuer Heterogeneity. *Nat. Hazards* 107, 2641–2651. doi:10.1007/s11069-021-04540-z

Sun, J., Xiao, K., Liu, C., Zhou, W., and Xiong, H. (2019). Exploiting Intra-day Patterns for Market Shock Prediction: A Machine Learning Approach. *Expert Syst. Appl.* 127, 272–281. doi:10.1016/j.eswa.2019.03.006

Tang, D. Y., and Zhang, Y. (2020). Do shareholders Benefit from green Bonds? *J. Corporate Finance* 61, 101427. doi:10.1016/j.jcorpfin.2018.12.001

Torres, M. E., Colominas, M. A., Schlotthauer, G., and Flandrin, P. (2011). "A Complete Ensemble Empirical Mode Decomposition with Adaptive Noise," in 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). doi:10.1109/ICASSP.2011.5947265

Trinks, A., Scholtens, B., Mulder, M., and Dam, L. (2018). Fossil Fuel Divestment and Portfolio Performance. *Ecol. Econ.* 146, 740–748. doi:10.1016/j.ecolecon.2017.11.036

Vidal, A., and Kristjanpoller, W. (2020). Gold Volatility Prediction Using a CNN-LSTM Approach. *Expert Syst. Appl.* 157, 113481. doi:10.1016/j.eswa.2020.113481

Vlasenko, A., Rashkevych, Y., Vlasenko, N., Peleshko, D., and Vynokurova, O. (2020). "A Hybrid EMD - Neuro-Fuzzy Model for Financial Time Series Analysis," in 2020 IEEE Third International Conference on Data Stream Mining & Processing (DSMP) (Lviv, Ukraine: IEEE), 112–115. doi:10.1109/DSMP47368.2020.9204179

Wang, J., Sun, X., Cheng, Q., and Cui, Q. (2021). An Innovative Random forest-based Nonlinear Ensemble Paradigm of Improved Feature Extraction and Deep Learning for Carbon price Forecasting. *Sci. Total Environ.* 762, 143099. doi:10.1016/j.scitotenv.2020.143099

Wang, J., Sun, X., and Li, J. (2017). How Does Economic Policy Uncertainty Interact with Sovereign Bond Yield? Evidence from the US. *Proced. Comput. Sci.* 122, 154–158. doi:10.1016/j.procs.2017.11.354

Weng, Y., Wang, Z., and Zhou, L. (2021). LSTM Framework Design and Volatility Research on Intelligent Forecasting Model for Solving the Parallel Dislocation Problem. *J. Phys. Conf. Ser.* 1982 (1), 012028. doi:10.1088/1742-6596/1982/1/012028

World Bank Group (2015). What Are green Bonds? (English). Available at: https://documents.worldbank.org/en/publication/documents-reports/documentdetail/400251468187810398/what-are-green-bonds (Accessed November 24, 2021).

Wu, Z., and Huang, N. E. (2009). Ensemble Empirical Mode Decomposition: a Noise-Assisted Data Analysis Method. *Adv. Adapt. Data Anal.* 01 (01), 1–41. doi:10.1142/s1793536909000047

Xian, L., He, K., Wang, C., and Lai, K. K. (2020). Factor Analysis of Financial Time Series Using EEMD-ICA Based Approach. *Sustainable Futures* 2, 100003. doi:10.1016/j.sftr.2019.100003

Yu, L., Wang, S., and Lai, K. K. (2008). Forecasting Crude Oil price with an EMD-Based Neural Network Ensemble Learning Paradigm. *Energ. Econ.* 30 (5), 2623–2635. doi:10.1016/j.eneco.2008.05.003

Zerbid, O. D. (2019). The Effect of Pro-environmental Preferences on Bond Prices: Evidence from green Bonds. *J. Bank. Financ.* 98, 39–60. doi:10.1057/s41260-018-0088-5

Zhang, H. (2020). Regulating green Bond in China: Definition Divergence and Implications for Policy Making. *J. Sustain. Finance Investment* 10 (2), 141–156. doi:10.1080/20430795.2019.1706310

Zhang, N., Lin, A., and Shang, P. (2017). Multidimensionalk-nearest Neighbor Model Based on EEMD for Financial Time Series Forecasting. *Physica A: Stat. Mech. its Appl.* 477 (1), 161–173. doi:10.1016/j.physa.2017.02.072

Zhang, X., Yu, L., Wang, S., and Lai, K. K. (2009). Estimating the Impact of Extreme Events on Crude Oil price: An EMD-Based Event Analysis Method. *Energ. Econ.* 31 (5), 768–778. doi:10.1016/j.eneco.2009.04.003

Zhang, Y., Xiong, R., He, H., and Pecht, M. G. (2018). Long Short-Term Memory Recurrent Neural Network for Remaining Useful Life Prediction of Lithium-Ion Batteries. *IEEE Trans. Veh. Technol.* 67 (7), 5695–5705. doi:10.1109/TVT.2018.2805189

Zhong, X., and Enke, D. (2017). Forecasting Daily Stock Market Return Using Dimensionality Reduction. *Expert Syst. Appl.* 67, 126–139. doi:10.1016/j.eswa.2016.09.027

Zhou, Y., Li, T., Shi, J., and Qian, Z. (2019). A CEEMDAN and XGBOOST-Based Approach to Forecast Crude Oil Prices. *Complexity* 2019, 1–15. doi:10.1155/2019/4392785

Zhou, Z., Lin, L., and Li, S. (2018). International Stock Market Contagion: A CEEMDAN Wavelet Analysis. *Econ. Model.* 72, 333–352. doi:10.1016/j.econmod.2018.02.010

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 25 February 2022
doi: 10.3389/fenrg.2021.807881

Overnight-Intraday Mispricing of Chinese Energy Stocks: A View from Financial Anomalies

Min Zhou[1] *and Xiaoqun Liu*[2]*

[1] *School of Design and Art, Hunan Institute of Technology, Hengyang, China,* [2] *School of Economics, Hainan University, Haikou, China*

We verify the existence of firm-level "intraday return vs. overnight return" pattern and overnight-intraday effect of nine financial anomalies of Chinese energy industry stocks of the Chinese stock market. Though energy finance has been an independent research area, we also take Chinese A-shares stocks as samples for empirical analysis to avoid the so-called sample selection bias. Specifically, it verifies that the overnight returns are strongly negative and intraday returns are positive for energy industry stocks, which is totally contrary to the American stock markets. In addition, alphas of the zero-cost strategies based on nine classic financial anomalies are almost earned at night for energy industry stocks. Finally, it is risk-related anomalies that occur overnight for energy industry stocks, while both four risk-related anomalies and two firm characteristics related anomalies occur at night for all A-shares stocks. Our empirical findings based on Chinese financial markets enrich the existing research on the mispricing of financial anomaly and shed a new sight on the asset pricing in energy finance.

Keywords: trading strategies, overnight-intraday effect, energy industry, financial anomaly, fama-macbeth cross-sectional regression

OPEN ACCESS

Edited by:
Jian Chai,
Xidian University, China

Reviewed by:
Fenghua Wen,
Central South University, China
Xu Gong,
Xiamen University, China

****Correspondence:***
Xiaoqun Liu
liuxiaoqun22@163.com

Specialty section:
This article was submitted to
Smart Grids,
a section of the journal
Frontiers in Energy Research

Received: *02 November 2021*
Accepted: *07 December 2021*
Published: *25 February 2022*

Citation:
Zhou M and Liu X (2022) Overnight-Intraday Mispricing of Chinese Energy Stocks: A View from Financial Anomalies.
Front. Energy Res. 9:807881.
doi: 10.3389/fenrg.2021.807881

1 INTRODUCTION

The empirical analysis of the existence and mechanism of financial anomalies have always attracted much attention from scholars. In recent years, extant papers mainly focus on pinning down which model is optimal to simultaneously depict financial anomalies, and how to make profits through those financial anomalies. Admittedly, financial anomaly represents an access to arbitrage, and investors can fully exploit this mispricing opportunity by constructing zero-costing trading strategies and exactly quantify the risk exposure of the anomaly by Fama and MacBeth (1973) regression.

Numerous financial anomalies are usually documented by various factor models, including the three-factor model of Fama and French (1993), the five-factor model of Fama and French (2015), the four-factor model of Fama and French (1993) and Carhart (1997), the four-factor "*q*-factor" model of Hou et al. (2015a), and the four-factor mispricing-factor model of Stambaugh et al. (2015). For instance, Stambaugh et al. (2015) construct two mispricing factors from the set of 11 prominent anomalies examined by Stambaugh et al. (2012), Stambaugh et al. (2014), Stambaugh et al. (2015). In addition, Hou et al. (2015a), Hou et al. (2015b) examined 73 anomalies, such as the total volatility, idiosyncratic volatility, and systematic volatility in Ang et al. (2006), the failure probability in Campbell et al. (2008), the dispersion of analysts' earnings forecasts in Diether et al. (2002), the total accrual in Richardson et al. (2005), and the illiquidity in Amihud (2002). They concluded that their *q*-factor model consisting of the market factor, a size factor, an investment factor, and a profitability factor largely summarizes the cross-section of average stock returns, while one-half of the anomalies are insignificant in the broad cross-section.

Anomalies mean that we could earn profits by a long-short strategy based on the specific anomaly, that is, a chance of arbitrage exists. Taking the Ang et al. (2006) idiosyncratic volatility and Campbell et al. (2008) distress risk, for example, both high idiosyncratic stocks and financially distressed stocks have delivered anomalously low return, named idiosyncratic volatility puzzle and financial distress puzzle by researchers, respectively. These two puzzles represent that the investors could earn profits by trading strategies of long (short) high (low) idiosyncratic stocks and long (short) high (low) financially distressed stocks.

In recent years, the literature documents unique characteristics of the components of close-to-close return among different financial markets (Cliff et al., 2008; Cai and Qiu 2009; Kelly and Clark 2011; Aboody et al., 2018; Lou et al., 2019; Muravyev and Ni 2020; Hendershott et al., 2020; Qiao and Dam 2020). Specifically, Cliff et al. (2008) document that strongly positive return at night and negative return during the day holds for individual stocks, equity indexes, and future contracts on equity indexes Cai and Qiu (2009) find that overnight non-trading period returns are significantly higher than both trading period returns and close-to-close daily returns in 23 countries at the stock index level, and they asserted that short selling contributed to this phenomenon. Kelly and Clark (2011) also report the positive overnight and negative intraday risk premium at index exchange-traded funds level. Aboody et al. (2018) examine the suitability of using overnight returns to measure firm-specific investor sentiment by analyzing whether they possess characteristics expected of a sentiment measure. Lou et al. (2019) verify the overnight-intraday effect with 14 anomalies, demonstrating that risk-adjusted alphas are either totally overnight effect or totally intraday effect. Muravyev and Ni (2020) decompose option returns into intraday and overnight components, finding a pattern of positive intraday returns and negative overnight returns. Hendershott et al. (2020) find that stock returns are positively related to beta overnight, whereas returns are negatively related to beta during the trading day. Qiao and Dam (2020) document the average overnight return in the Chinese stock market is negative and argue that the "T+1" trading rule contributes significantly to this overnight return puzzle.

Energy finance is interdisciplinary, setting up a bridge on two most important industries, that is, finance and energy, in real life. In recent years, topics on asset pricing, financial risk management, investment, and so on have been widely applied in the energy industry area (Wen et al., 2021a; Wen et al., 2021b; Cao et al., 2022; Farouq et al., 2021; Liu et al., 2021; Peng et al., 2021; Tian et al., 2021; Zheng et al., 2021). In this paper, we verify the existence of firm-level "intraday return vs. overnight return" pattern and overnight-intraday effect of nine financial anomalies in the energy industry market. Though energy finance has been an independent research area, we also take A-shares stocks as samples for empirical analysis to avoid the so-called sample selection bias. The empirical results show that strong persistence of overnight and intraday firm-level return emerge both in energy industry stocks and A-shares stocks. In addition, the overnight returns are strongly negative and intraday returns are positive, which is totally contrary to the American stock markets. However, Qiao and Dam (2020) also provide evidence of negative overnight returns in the Chinese stock market, and they argue that the "T+1" trading rule contributes significantly to this overnight return puzzle. Finally, profits are almost earned entirely overnight among nine trading strategies in energy industry stocks and A-shares stocks, which is sharply in contrast to the results of Lou et al. (2019).

The organization of this paper is as follows. **Section 2** puts forward the motivation, a brief summary of energy finance, and potential contributions. **Section 3** describes the data and methodology. **Section 4** presents the empirical analysis of the anomaly strategies at levels of portfolio analysis and Fama-MacBeth regression. **Section 5** concludes.

2 MOTIVATION AND CONTRIBUTION

Motivated by the literature on non-trading hour vs. trading hour return patterns at individual stock, index stock, and index fund levels, the overnight-intraday effect from a portfolio strategies view of Lou et al. (2019), and the special overnight effect in the Chinese stock market of Qiao and Dam (2020), we attempt to examine overnight-intraday trading strategies in the cross-section of energy industry stock returns in China. As for the construction of trading strategies, we choose the basic five financial anomalies according to the five-factor model of Fama and French (2015) and the "q-factor" model of Hou et al. (2015a), that is, size, value, momentum, reversal, and profitability. In addition, we choose the following most influential anomalies, named idiosyncratic volatility puzzle, beta, and turnover.

Energy is the foundation and driving force for the progress of human civilization. Energy and resource constraints, together with climate change, environmental risks, and challenges have become severe global problems in the modern world. How to develop a clean and low-carbon energy strategy matters to world energy security, to addressing global climate change, and to boosting global economic growth. Energy finance mainly focuses on the significant connection between energy and finance. As the asset pricing and financial risk management are the frontiers in finance (Güngör and Tastan 2021; Huong et al., 2021; Liow et al., 2021; Mao and Zhang 2021; Umutlu and Bengitöz 2021), in recent years, the literatures on energy-based asset pricing, and energy financial risk management have received considerable attention (Gong and Lin 2017, Gong and Lin 2018, Gong and Lin 2021; Gong et al., 2021).

Energy finance is an interdisciplinary, setting up a bridge between two most important industries in real life. In recent years, topics on asset pricing, financial risk management, investment, and so on have been widely applied in the energy industry area (Lian et al., 2020; Ye et al., 2020; Zolfaghari et al., 2020; Dai et al., 2021; Ghoddusi and Wirl, 2021; Si et al., 2021; Wang et al., 2021). Specifically, Lian et al. (2020) examine how the tail behavior of various risk factors affects the tail behavior of individual oil stock returns; Ye et al. (2020) investigate the interaction between crude oil prices and investor sentiment from the time and the frequency domains; Zolfaghari et al. (2020) verify that the energy market and the stock market have stronger co-volatility spillover than foreign currency market; Ghoddusi and Wirl (2021) discuss the risk-hedging feature of the refinery industry when the crude oil market faces supply vs. demand shocks; Dai et al. (2021) demonstrate that the skewness of oil price return can predict the aggregate stock market returns; Si et al. (2021) investigate the effects of financial deregulation on the energy enterprises' operational risks in China; Wang et al. (2021) examine the impact of equity concentration on the investment efficiency of Chinese energy companies based on the shock that the shareholding ratio restriction of qualified foreign institutional investors (QFIIs) is relaxed.

The potential contributions of this paper are as follows: at first, it is a nature point that we investigate the overnight-intraday effect of trading strategies in the energy industry, and as far as we know, we are the first to empirically test which financial anomaly belongs to overnight effect or intraday effect in the energy industry stocks as well as in A-shares stocks. Second, we demonstrate the pattern of overnight anomaly vs. intraday anomaly both in the Chinese energy industry market and Chinese A-shares markets, thus, it could not only avoid the sample bias, but also contribute to the empirical asset pricing in the area of energy finance. Last, it is beneficial to well understand the financial anomalies in a particular industry and in all Chinese A-shares stock markets from the perspective of components of return, as we do find differences between these two samples. That is, in cases of trading strategies based on market risk and liquidity risk, profits are earned entirely overnight for energy industry stocks, while there are much more financial anomalies belonging to overnight effect for A-shares stocks beside the risk-related anomalies.

3 DATA AND METHODOLOGY

The data is collected from China Stock Market & Accounting Research (CSMAR) database and Wind database from January 2001 to December 2019 for all energy stocks and A-shares stocks traded in Chinese stock markets. Stocks with prices below ¥1 a share are excluded from the sample. For each firm i, at one day d, we decompose daily close-to-close return ($r^{i}_{\text{close-to-close},d}$) into close-to-open return ($r^{\text{overnight}}_{\text{close-to-open},i,d}$) and open-to-close return ($r^{\text{intraday}}_{\text{open-to-close},i,d}$) as Lou et al. (2019). The specific formula is as follows:

$$r^{i}_{\text{close-to-close},d} = \frac{P^{i}_{\text{close},d}}{P^{i}_{\text{close},d-1}} - 1$$

$$r^{\text{intraday}}_{\text{open-to-close},i,d} = \frac{P^{i}_{\text{close},d}}{P^{i}_{\text{open},d-1}} - 1$$

$$r^{\textit{overnight}}_{\text{close-to-open},i,d} = \frac{P^{i}_{\text{open},d}}{P^{i}_{\text{close},d-1}} - 1$$

Then, we calculate the monthly overnight return and intraday return by accumulating the above daily return components across days in each month d, namely $r^{\text{intraday}}_{\text{open-to-close},i,m}$, $r^{\text{overnight}}_{\text{close-to-open},i,m}$.

Because we conduct our trading strategies at the portfolio level, so we also calculate the following three components of the portfolio, p,

$$r^{\text{intraday}}_{\text{open-to-close},p,m} = \sum\nolimits_{i} w^{i}_{t-1} r^{\text{intraday}}_{\text{open-to-close},i,m}.$$
$$r^{\text{overnight}}_{\text{close-to-open},p,m} = \sum\nolimits_{i} w^{i}_{t-1} r^{\text{overnight}}_{\text{close-to-open},i,m}.$$

where w stands for weights of constructing the portfolio, and in this paper, we use market capitalization value-weight portfolios.

The main object of this paper is to analyze which financial anomalies belong to overnight effect or intraday effect, so it is of interest to explain how to measure these anomalies as well. Based on the influential five-factor model of Fama and French (2015) and the "q-factor" model of Hou et al. (2015a), in this paper, we construct nine trading strategies according to the size, value, momentum, reversal, profitability, idiosyncratic volatility, beta, turnover, and the corresponding definition or measurement, which are introduced in Section 4.

4 EMPIRICAL RESULTS

In this section, we first examine the persistence of overnight/intraday return for the anomaly alphas pattern of the anomalies by portfolio analysis and Fama-MacBeth regression. More specifically, we decompose the abnormal returns earned by a range of trading strategies based on the following anomalies, including size, value, momentum, reversal, profitability, idiosyncratic volatility, beta, and turnover into their overnight and intraday components.

4.1 Persistence of Overnight/Intraday Abnormal Returns for Trading Anomaly Strategy

Trading strategies are constructed to capture the alpha associated with trading at night or during the day, and thereby we should first test the persistence in the components of close-to-close return. We conduct the test by calculating the raw excess decomposed returns and the risk-adjusted excess decomposed returns based on overnight return-sorted and intraday return-sorted portfolios, respectively. Specifically, given that the existence of the persistence of overnight returns or intraday returns, we could see a positive (negative) long-short overnight (intraday) alpha of the overnight return-based portfolio, and a negative (positive) long-short overnight (intraday) alpha of the intraday return-based portfolio, respectively.

We verify the overnight-intraday continuation and reversal patterns by rebalancing the portfolios in the current month and calculating the components of the close-to-close return in the next month both in the energy industry stocks and China A-shares stocks.

Table 1 shows the basic descriptive statistics of the main variables for energy industry stocks. First, there are 74 energy industry stocks over the period from 2002 to 2019. The mean (media) monthly overnight return of the Chinese energy stock market is −3.71% (−3.67%), while the mean monthly intraday

TABLE 1 | Descriptive statistics

	Mean	std	5%	25%	Median	75%	95%	Count
Overnight return	–0.0371	0.0143	–0.0656	–0.0448	–0.0367	–0.0276	–0.0157	74
Intraday return	0.0283	0.0114	0.0108	0.0203	0.0293	0.0349	0.0465	74
Close to close return	–0.0088	0.0103	–0.0284	–0.0116	–0.0055	–0.0025	0.0001	74
Turnover	56.6461	33.4854	18.3256	34.5915	46.3667	70.9966	122.3308	74
PE	62.1596	109.5379	–37.1063	24.0591	44.2928	80.6917	192.0541	74
ROE	0.9706	16.5457	–11.8013	1.6297	3.4302	6.4195	9.8928	74
Idiosyncratic volatility	0.0241	0.0052	0.0155	0.0203	0.0239	0.0267	0.0334	74
beta	0.0675	0.1425	–0.1186	–0.0267	0.0832	0.1392	0.2859	74

This table provides a brief description of the main variables that are used in this study. The variables are monthly overnight return, monthly intraday return, monthly close-to-close return, the main firm characteristics (i.e., turnover, PE, ROE), the main market-risk related characteristics (i.e., idiosyncratic volatility, market beta). The summary statistics includes the number of observations, mean, median, standard deviation (STD), the percentiles (5 and 95%), and quartiles (25 and 75%) distribution of the variables. The definition of daily intraday return is the price appreciation between market open and close of the same day, while the daily overnight return is the price appreciation between market open price of the current day and close of the past day. Daily close-to-close return is the price appreciation between market close of the current day and close of the past day. We calculate the monthly components of returns by accumulating corresponding daily intraday return and overnight return. Monthly turnover is the number of shares traded in the current month scaled by the number of shares outstanding. Price earnings ratio (PE) is the stock price divided by the earnings per share (EPS). ROE is the return on equity. Monthly betas of the stock with respect to the Shanghai Composite Index estimated following CAPM, that is, we estimate time-varying monthly betas using daily returns over rolling 12-months windows. Idiosyncratic volatility is the standard deviation of daily residuals based on the Fama-French-Carhart four-factor model over the preceding 1 year. The sample period is 2002–2019.

TABLE 2 | Persistence of overnight-intraday return for energy industry stocks

Panel A: Portfolios sorted by lagged 1-month overnight returns

Quintile	Overnight return			Intraday return		
	Excess	CAPM	Three-Factor	Excess	CAPM	Three-Factor
1	–4.17%***	–4.40%***	–4.41%***	3.11%***	2.69%***	2.67%***
	(–14.62)	(–10.87)	(–10.58)	(8.68)	(5.71)	(5.92)
5	–1.93%***	–2.09%***	–2.11%***	0.77%***	0.35%	0.45%
	(–4.99)	(–6.61)	(–6.66)	(2.96)	(0.86)	(1.14)
5–1	2.24%***	2.31%***	2.30%***	–2.34%***	–2.34%***	–2.22%***
	(6.23)	(6.40)	(5.83)	(–4.39)	(–4.27)	(–4.24)

Panel B: Portfolios sorted by lagged 1-month intraday returns

Quintile	Overnight return			Intraday return		
	Excess	CAPM	Three-Factor	Excess	CAPM	Three-Factor
1	–1.94%***	–2.11%***	–2.05%***	1.05%***	0.60%	0.63%*
	(–4.14)	(–7.08)	(–6.97)	(3.86)	(1.63)	(1.76)
5	–4.72%***	–4.90%***	–4.86%***	2.78%***	2.33%***	2.41%***
	(–17.70)	(–8.64)	(–8.36)	(7.81)	(4.75)	(4.71)
5–1	–2.77%***	–2.79%***	–2.81%***	1.73%***	1.73%***	1.77%***
	(–5.62)	(–5.53)	(–5.46)	(3.20)	(3.16)	(3.10)

This table reports overnight-intraday return persistence and reversal patterns for Energy industry stocks. In Panel A, all stocks are sorted into quantile based on their lagged 1-month overnight returns. In Panel B, stocks are sorted based on their lagged 1-month intraday returns. We then go long the value-weight winner quantile and short the value-weight loser quantile. The first three columns show the overnight return in the subsequent month of the two short-term reversal strategies, and the next three columns show the intraday returns in the subsequent month. We report monthly raw excess portfolio returns, alphas adjusted by the CAPM and by the Fama-French three-factor model. t*-statistics are calculated by correcting standard errors for serial-dependence with 12 lags. *,**,*** represent that the results are 10, 5, 1% statistically significant, respectively. Sample period is 2001–2019.*

return is 2.83% (2.93), further confirming the pattern of overall negative overnight return and positive intraday return in the Chinese stock market, as documented by Qiao and Dam (2020). Meanwhile, the corresponding monthly standard deviations of the monthly overnight return and intraday return are 0.0143 and 0.0114, consistent with the classic hypothesis of return-risk tradeoff. The average monthly turnover, price earnings ratio (PE), and return on equity (ROE) are 56.65, 62.16, and 0.97, respectively. In addition, the mean monthly idiosyncratic volatility and market beta are 0.024 and 0.068, respectively.

For energy industry stocks, due to the limited sample data, we construct zero-cost trading strategies of longing (shorting) the current month value-weight top (bottom) quintile for the past 1-month overnight return-sorted portfolios (Panel A) and the intraday return-sorted portfolio (Panel B), respectively. **Table 2** reports monthly portfolio raw excess return, and two risk-adjusted abnormal overnight return and intraday return.

Correspondingly, as for all A-shares stocks, we construct zero-cost trading strategies of longing (shorting) the current month value-weight top (bottom) decile for the past 1-month overnight return-sorted portfolios (Panel A) and the intraday return-sorted

TABLE 3 | Persistence of overnight/intraday return for A-shares stocks

Panel A: Portfolios sorted by past 1-month overnight returns

Decile	Overnight return			Intraday return		
	Excess	CAPM	Three-Factor	Excess	CAPM	Three-Factor
1	−4.19%***	−4.37%***	−4.38%***	4.56%***	4.08%***	4.06%***
	(−14.62)	(−17.70)	(−17.54)	(8.68)	(12.18)	(12.91)
10	−1.36%***	−1.52%***	−1.50%***	1.49%***	1.01%***	1.11%***
	(−4.99)	(−5.97)	(−5.95)	(2.96)	(3.44)	(3.67)
10–1	2.83%***	2.85%***	2.88%***	−3.07%***	−3.06%***	−2.95%***
	(22.32)	(20.89)	(21.30)	(2.96)	(−9.91)	(−9.71)

Panel B: Portfolios sorted by past 1-month intraday returns

Decile	Overnight			Intraday		
	Excess	CAPM	Three-Factor	Excess	CAPM	Three-Factor
1	−1.09%***	−1.26%***	−1.22%***	1.92%***	1.43%***	1.46%***
	(−4.14)	(−5.55)	(−5.37)	(3.86)	(4.49)	(4.75)
10	−4.37%***	−4.55%***	−4.53%***	4.07%***	3.62%***	3.79%***
	(−17.70)	(−15.36)	(−15.08)	(7.81)	(10.30)	(10.15)
10–1	−3.28%***	−3.29%***	−3.30%***	2.15%***	2.18%***	3.79%***
	(−18.42)	(−16.44)	(−15.65)	(5.67)	(5.93)	(6.19)

This table reports overnight-intraday return persistence and reversal patterns for A-shares stocks. In Panel A, all stocks are sorted into deciles based on their lagged 1-month overnight returns. In Panel B, stocks are sorted based on their lagged 1-month intraday returns. We then go long the value-weight winner decile and short the value-weight loser decile. The first three columns show the overnight return in the subsequent month of the two short-term reversal strategies, and the next three columns show the intraday returns in the subsequent month. We report monthly raw excess portfolio returns, alpha adjusted by the CAPM and by the Fama-French three-factor model. t *-statistics are calculated by correcting standard errors for serial-dependence with 12 lags. *,**,*** represent that the results are 10, 5, 1% statistically significant, respectively. Sample period is 2001–2019.*

portfolio (Panel B), respectively. **Table 3** reports monthly portfolio raw excess return, and two risk-adjusted abnormal overnight and intraday returns.

The empirical results in both the energy stocks and China A-shares stocks totally contrast to Lou et al. (2019). When using the energy stocks as the sample, a long-short strategy based on the past 1-month overnight returns earns a strongly significant raw excess overnight return of 2.24% as well as significantly CAPM or the three-factor adjusted alphas, reaching 2.31% or 2.30% in the current month, respectively. Meanwhile, a significantly average CAPM (three-factor)-adjusted intraday alpha of −2.34% (−2.22%) per month is earned based on the overnight-return hedge portfolio, verifying a reversal in the intraday period as well.

We almost get the same results when utilizing the Chinese A-shares stocks as a sample, a long-short strategy based on the past 1-month overnight returns yields a strongly significant average raw excess overnight return of 2.83% in the current month. When adjusted by CAPM and Fama-French 3 factor models, the overnight alphas are still significantly positive. Meanwhile, a significantly average CAPM (three-factor)-adjusted intraday alpha of −3.06% (−2.95%) per month is earned based on the overnight-return hedge portfolio, that is, an intraday reversal effect exists.

On the other hand, when we go long the top quantile of the energy stocks and go short the bottom ones based on the past 1-month intraday return, we could get significant positive (negative) raw excess intraday (overnight) returns and two-risk adjusted intraday (overnight) alphas. Again, the above finding continues to hold when sorting stocks using the China A-shares stocks as our sample.

In all, **Table 2** and **Table 3** confirm that there are striking overnight/intraday momentum and reversal patterns both in China A-shares stocks and the energy stocks. In addition, we depict these momentum and reversal patterns in overnight return and intraday return in **Figure 1** and **Figure 2**.

4.2 Cross-Sectional Overnight-Intraday Alphas of the Anomalies

In this part, we use a list of popular financial anomies to understand the overnight-intraday pattern. We decompose the abnormal returns earned by a range of trading strategies based on the following anomalies, including size, value, momentum, reversal, profitability, idiosyncratic volatility, beta, and turnover, into their overnight and intraday components.

We empirically test the cross-sectional overnight/intraday alphas of the anomalies by portfolio analysis and Fama-MacBeth regressions. In the portfolio analysis, for each financial anomaly, we calculate the CAPM-adjusted overnight/intraday alphas of zero-cost trading strategies, that is, we get the risk-adjusted alpha by long (or short) the top portfolio and short (or long) the bottom portfolio according to the characteristics of the anomies. For instance, as for idiosyncratic volatility anomaly, we go long low idiosyncratic volatility quantile and short high idiosyncratic volatility quantile. However, for the well-known momentum anomaly, we will go long the top winner cumulative returns of the portfolio and go short the bottom loser cumulative ones. **Section 4.2.1** and **Section 4.2.2** present the overnight anomalies and intraday anomalies with the portfolio analysis, respectively.

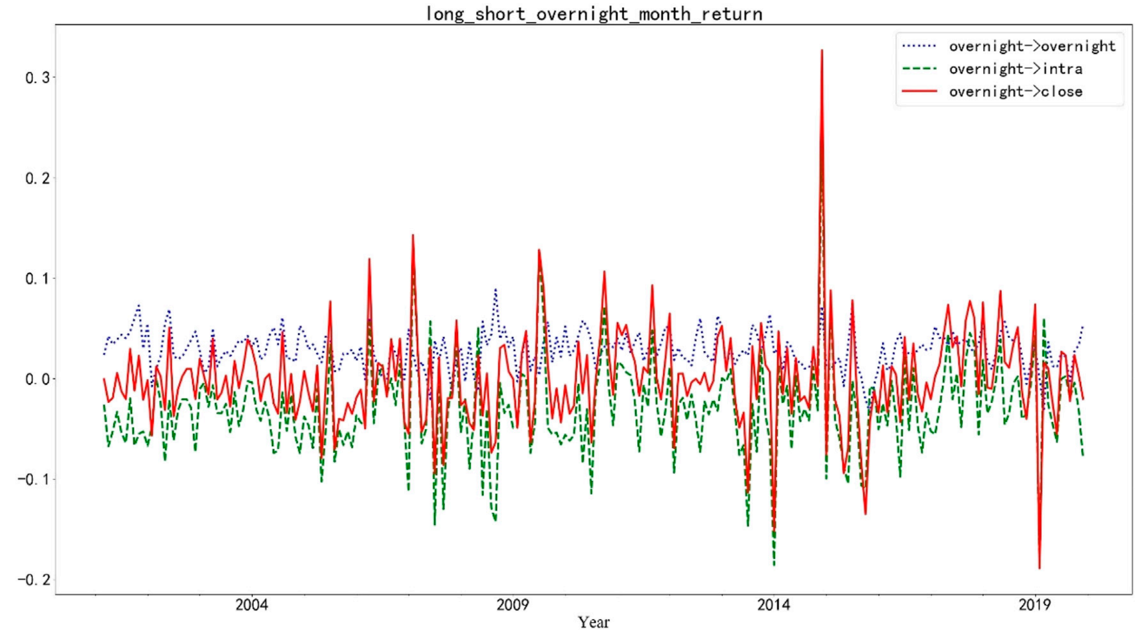


FIGURE 1 | This figure depicts the overnight-intraday persistence and reversal pattern return for energy industry stocks, as reported in Panel A of **Table 2**. The blue dotted curve corresponds to using 1-month lagged overnight returns to forecast the current overnight return. The green dashed curve corresponds to using lagged overnight returns to forecast the current intraday return. The red solid curve corresponds to using lagged overnight returns to forecast the current close-to-close return.

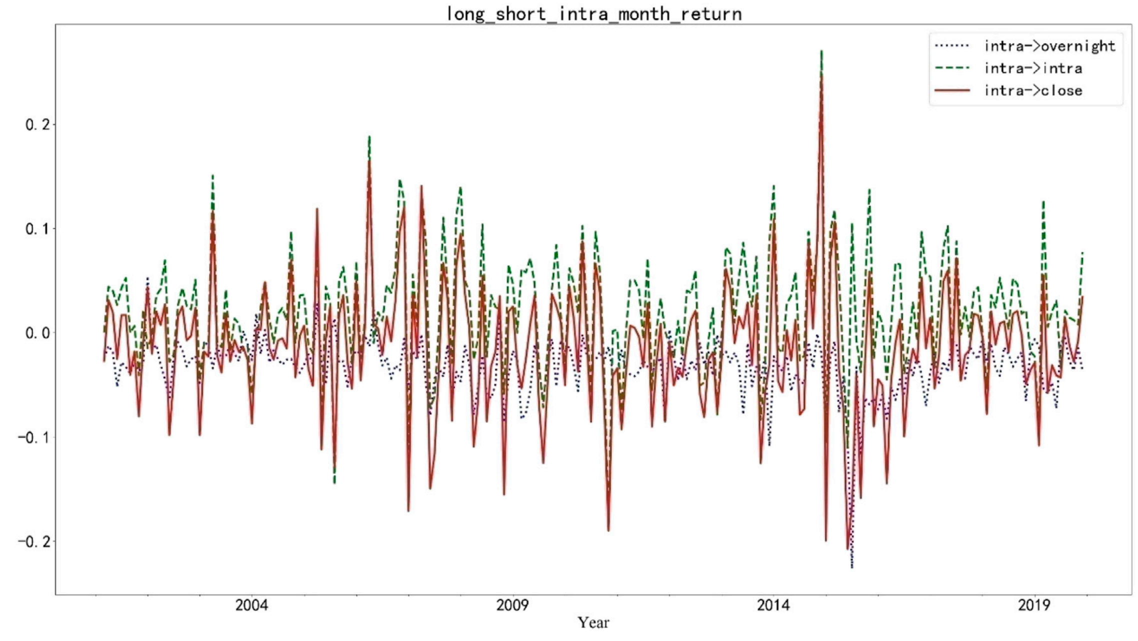


FIGURE 2 | This figure depicts the intraday-overnight persistence and reversal pattern return for energy industry stocks, as reported in Panel B of **Table 2**. The blue dotted curve corresponds to using 1-month lagged intraday returns to forecast the current overnight return. The green dashed curve corresponds to using lagged intraday returns to forecast the current intraday return. The red solid curve corresponds to using lagged intraday returns to forecast the close-to-close return.

4.2.1 Overnight Anomalies

We verify that four out of nine financial anomalies in energy industry stocks, including idiosyncratic volatility, beta, turnover, and reversal, belong to the overnight effect. **Table 4** and **Table 5** report overnight alphas and intraday alphas based on equity premium and eight cross-sectional anomaly-based strategies for

TABLE 4 | Overnight-intraday abnormal return of anomalies for energy industry stocks

	Overnight alpha	Intraday alpha		Overnight alpha	Intraday alpha
Panel A overnight anomalies					
Beta	0.98%*** (2.65)	−0.50% (−0.93)	Ivol	2.23%*** (5.55)	−0.79% (−1.29)
Turnover	2.54%*** (5.08)	−1.29%** (−2.18)	Reversal	1.45%*** (2.99)	−0.51% (−0.99)
Panel B intraday anomalies					
	Overnight alpha	**Intraday alpha**		**Overnight alpha**	**Intraday alpha**
Index	−2.28%*** (−6.77)	1.23%*** (2.56)	Size	−1.57%*** (−5.85)	1.90%*** (3.47)
Panel C others					
	Overnight alpha	**Intraday alpha**		**Overnight alpha**	**Intraday alpha**
BM	−2.63%*** (−5.49)	−0.51% (−0.94)	Mom	−1.30%*** (−2.82)	0.52% (0.83)
ROE	0.56% (1.28)	−0.65% (−1.28)			

This table reports abnormal return to of various cross-sectional strategies during the day vs. at night for energy industry stocks. In Panel A, we examine the overnight/intraday abnormal return of four risk-related financial anomalies, including beta, idiosyncratic volatility, turnover, and short-term reversal. In Panel B, we examine the overnight/intraday abnormal return of equity premium and size anomaly. In Panel C, we examine the overnight/intraday abnormal return of two firm characteristics related anomalies (value and probability) and momentum anomaly. The definition of the financial anomalies and the detailed zero-strategies based on these anomalies are explained in ***Section 4.2.1*** *and* ***Section 4.2.2****.* t *-statistics are calculated by correcting standard errors for serial-dependence with 12 lags. *,**,*** represent t.*

TABLE 5 | Overnight/intraday abnormal return of anomalies for A-shares stocks. This table reports abnormal return to of various cross-sectional strategies during the day vs. at night for A-shares stocks. In Panel A, we examine the overnight/intraday abnormal return of four risk-related financial anomalies (beta, idiosyncratic volatility, turnover, short-term reversal) and two firm characteristics related financial anomalies (value and probability). In Panel B, we examine the overnight/intraday abnormal return of the momentum anomaly. The definition of the financial anomalies and the detailed zero-strategies based on these anomalies are explained in **Section 4.2.1** and **Section 4.2.2**. *t* -statistics are calculated by correcting standard errors for serial-dependence with 12 lags. *,**,*** represent that the results are 10, 5,1% statistically significant, respectively. The sample period is 2001–2019.

	Overnight alpha	Intraday alpha		Overnight alpha	Intraday alpha
Panel A overnight alpha adjusted by CAPM					
BM	1.61%*** (9.33)	−0.80%* (−1.85)	ROE	1.36%*** (8.20)	−0.91%** (−2.27)
Beta	1.10%*** (6.31)	−0.38% (−0.86)	Ivol	2.80%*** (14.33)	−1.84%*** (−3.75)
Turnover	1.97%*** (9.89)	−1.53%*** (−3.19)	Reversal	1.72%*** (8.05)	−0.64% (−1.61)
Panel B intraday alpha adjusted by CAPM					
	Overnight alpha	**Intraday alpha**		**Overnight alpha**	**Intraday alpha**
Index	−1.68%*** (−7.10)	2.03%*** (4.72)	Size	−0.88%*** (−5.24)	3.17%*** (5.70)
Panel C others					
	Overnight alpha	**Intraday alpha**			
Mom	−1.15%*** (−5.46)	0.29% (0.62)			

the energy industry stocks and A-shares stocks, respectively. The details are shown in the following part. Energy industry stocks and A-shares stocks are sorted into quantile and decile, respectively.

Specifically, the first discussed two zero-cost trading strategies are risk-related financial anomies. The most traditional one is that high market beta stocks should have high return original from the CAPM of Sharpe (1964), Lintner (1965), and Black (1972), while be challenged by empirical evidence which shows that the security market line is not volatile (Frazzini and Pedersen 2014); the other is the well-known "idiosyncratic volatility puzzle" found by Ang et al. (2006), which argued that high idiosyncratic volatility stocks had abnormally low return.

We analyze the beta strategy that goes long the low-beta quintile (decile) and short the high-beta quintile (decile) for energy stocks (A-shares stocks). According to Lou et al. (2019) and Dimson (1979), we measure beta using daily returns over the last 12 months with three lags in the market model regression for each stock, the beta is the sum of the four coefficients, which is beneficial to taking non-synchronous trading issues into account. Panel A of **Table 4** and **Table 5** demonstrate that the beta alphas are totally overnight phenomenon in energy industry and all A-shares stocks, which is in sharp contrast to the Lou et al. (2019). Specifically, the overnight CAPM alpha is 0.98% (1.10%) with an associated *t*-statistics of 2.65 (6.31) for energy stocks (all A-shares stocks), and the corresponding intraday CAPM alpha is −0.50% (−0.38%) with an associated *t*-statistics of −0.93 (−0.86) for energy stocks (all A-shares stocks).

As for the idiosyncratic strategy (Ivol), we go long the low idiosyncratic volatility quintile (decile) and short the high idiosyncratic volatility quintile (decile) for the energy stocks (all A-shares stocks). We measure the idiosyncratic volatility as the volatility of the residual from a daily Fama-French-Carhart four-factor regression estimated over the last year. Panel A of **Table 4** and **Table 5** report that the Ivol alphas are a totally overnight phenomenon in energy stocks (all A-shares stocks) inconsistent with the result of Lou et al. (2019). Specifically speaking, the overnight CAPM alpha is 2.23% (2.80%) with an associated *t*-statistics of 5.55 (14.33) for all energy stocks (all A-shares stocks), and the corresponding intraday CAPM alpha is −0.27% (−1.84%) with an associated *t*-statistics of −1.29 (−3.75) for energy stocks (all A-shares stocks).

Then, we examine the turnover and short-term reversal anomalies, which is related to the liquidity risk. We first analyze the turnover strategy that goes long the lowest turnover quintile (decile) and short the highest turnover quintile (decile) for energy stocks (all A-shares stocks) based on the previous findings of Datar et al. (1998) and Lee and Swaminathan (2000), who show that turnover could negatively explain the cross-section average returns. We measure the turnover as the average daily volume over the last 12 months following Lee and Swaminathan (2000). Again, Panel A of **Table 4** and **Table 5** report that the turnover premiums are a totally overnight phenomenon in the energy industry and all A-shares stocks as there is significant negative expected intraday return, inconsistent with the result of Lou et al. (2019). In particular, the strongly statistically significant overnight CAPM alpha is 2.54% (1.97%) for energy stocks (all A-shares stock), and the corresponding insignificant intraday CAPM alpha is −1.29% (−1.53%) for energy stocks (all A-shares stocks).

At last, we measure short-term reversal as 1-month return and analyze this strategy (reversal) that goes long the low past 1-month return quintile (decile) and short the high turnover quintile (decile) for the energy stocks (A-shares stocks). Panel A of **Table 4** and **Table 5** report that the STR premiums are totally overnight phenomenon in energy industry and all A-shares stocks, inconsistent with the result of Lou et al. (2019). Specifically, the highly significant overnight CAPM alpha is 1.45% (1.72%) for energy stocks (all A-shares stocks), and the corresponding insignificant intraday CAPM alpha is −1.29% (−0.64%) for energy stocks (A-shares stocks).

Except the above four risk-related financial anomalies, value and probability anomalies documented in Fama-French five-factor model belong to overnight effect for all A-shares stocks. We investigate the value strategy that goes long the highest book-to-market quintile (decile) and short the lowest book-to-market quintile (decile) for energy stocks (A-shares stocks). It is found that for A-shares stocks, essentially, the value alpha occurs overnight, which is totally inconsistent with Lou et al. (2019), while the value alpha does not exist for energy stocks. Specifically, as for energy stocks, both the overnight and intraday CAPM alphas are negative, −2.63% with a *t*-statistics of −5.49 and −0.51% with a *t*-statistics of −0.94, respectively. However, as for all A-shares stocks, the overnight CAPM alphas are 1.61% with a *t*-statistics of 9.33, while the intraday CAPM alpha is slightly negative, −0.80%, and statistically significant with an associate *t*-statistic of −1.85.

Besides the classic firm characteristics, the literature documented that profitability could be another anomaly in cross-sectional stock markets (Haugen and Baker 1996; Vuolteenaho 2002; Novy-Marx 2013), and the latest and famous one is Fama and French (2015) who proposed and tested that the profitability could help capture the cross-section of average returns based on the Fama-French 3-factor. In this paper, we measure the profitability by the return on equity (ROE), then conduct profitability strategy that goes long the highest profitability quintile (decile) and short the lowest profitability quintile (decile) for energy stocks (all A-shares stocks). Panel C of **Table 4** and Panel A of **Table 5** report that the profitability alphas are overnight phenomenon for all A-shares stocks. Specifically, both the overnight and intraday CAPM alpha are not statistically significant in energy stocks, while the overnight and intraday CAPM alpha are 1.36% with an associated *t*-statistics of 8.20 and −0.91% with an associated *t*-statistics of 8.20, respectively, for all A-shares stocks. In all, essentially, the profitability alpha in China belongs to overnight, which is contrary to the result of Lou et al. (2019), implying a huge difference in these two markets exists.

4.2.2 Intraday Anomalies

We verify that two out of nine financial anomalies both in energy industry stocks and all A-shares stocks, including equity premium and size belong to the intraday effect.

First, we analyze the basic equity overnight premium and intraday premium at an index level, we could interpret it as the anomaly related to CAPM. Panel B of **Table 4** shows that the market portfolio as measured by the value-weight energy stocks has an average monthly overnight raw excess return of −2.28% and an average intraday raw excess return of 1.23%. Panel B of **Table 5** shows that the market portfolio as measured by the value-weight all A-shares stocks has an average monthly overnight alpha of 2.03% and an average intraday excess return of −1.68%. These findings mean that the equity premium is an overnight phenomena both in energy industry stocks and A-shares stocks, which is consistent with previous work done by Qiao and Dam (2020), who document the average overnight return in the Chinese stock market is negative and argue that the “T+1” trading rule contributes significantly to this overnight return puzzle.

It is well acknowledged that size effect and value effect, proposed by Fama and French (1992), are the first two anomalies in empirical asset pricing. Then, momentum effect, first proposed by Jegadeesh and Titman (1993), and formally put forward as an asset pricing factor by Carhart (1997) along with the size and value effect were widely used as a risk-adjustment benchmark in empirical studies. Specifically, as for the size anomaly, we go buying the smallest quintile (decile) and selling the largest quintile (decile) for energy stocks (A-shares stocks). Panel B of both **Table 4** and **Table 5** report the Size’s CAPM-adjusted overnight and intraday abnormal returns for energy stocks and all A-shares stocks, respectively. We find the size alpha occurs intraday for both samples, which is quite consistent with Lou et al. (2019). As we can see, the CAPM-adjusted intraday alphas are 1.90% with an associated *t*-statistics of 3.47 and 3.17% with *t*-statistics of 5.70, while the CAPM-adjusted overnight alphas are −1.57% with an associated *t*-statistics of 3.47 and −0.88% with an associated *t*-statistics of −5.24 for energy stocks and all A-shares stocks, respectively.

We then pin down the abnormal overnight returns and intraday returns of the momentum strategy founded by Jegadeesh and Titman (1993) and developed by Carhart (1997). We find the momentum premium almost does not occur in overnight and intraday for the two samples, specifically, the intraday CAPM alpha is statistically insignificant, 0.52% with an associated *t*-statistics of 0.62 (0.29% with *t*-statistics of 0.83) for the winner quintile (decile) minus the loser quintile (decile) for energy stocks (all A-shares stocks), and the corresponding overnight CAPM alpha is significantly negative, −1.15% with an associated *t*-statistics of 0.62 (with *t*-statistics of −1.30%). These results are quite different from Lou et al. (2019), implying there are obvious discrepancies of the investors’ behavior between the United States and China’s stock market.

In all, the results of Panel A of **Table 4** and **Table 5** show unique characteristics of the overnight and intraday abnormal profits of a series of trading strategies in China energy industry stocks and A-shares stocks, which is quite different from the America stock market. On the one hand, only the alpha of the Size strategy occurs within the day, and the premia of the other seven strategies all occur overnight. However, in Lou et al. (2019), the alphas of three momentum (momentum, price momentum, industry momentum) and reversal strategy mainly occur overnight, while others all occur within the day. Why do such significant differences exist between these two countries? We think that the potential reason lies in the “T+1” trading mechanism in China. “T+1” trading rule prohibits traders to sell shares they bought on the same day, leading to asymmetric effects for buyers and sellers and making most of the investors prefer to trade at the close rather than at the open. Qiao and Dam (2020) verify that the “T+1” trading rule produces a discount on opening prices due to this asymmetric effect, in essence a liquidity discount, and it could explain the negative overnight return, named overnight return puzzle. On the other hand, it is risk-related anomalies that occur during overnight for energy industry stocks, while both four risk-related anomalies and two firm characteristics related anomalies occur within the day for all A-shares stocks.

4.2.3 Fama-MacBeth Regressions

We further conduct the Fama-MacBeth regressions to test the cross-section of intraday and overnight expected abnormal returns. The advantage of this econometric method is that we could control for a list of characteristics and thus make the result more precise, while we could only do one-dimension, double-dimension, or at most three-dimension portfolio sorts. We carry out five cross-sectional regressions as does Lou et al. (2019).

The five cross-sectional regressions conclude: the dependent variables are close-to-close return (regression 1), overnight return (regression 2), intraday return (regression 3), overnight return minus intraday return (regression 4), scaled overnight return minus intraday return (the coefficient in this regression is the difference between the overnight coefficient ∗24/18.5 and intraday coefficient ∗24/5.5 (regression 5), respectively. In each regression, we include the above anomalies except for short-term reversal (Str), as we also control the most recent 1-month intraday/overnight return (ret_intraday/ret_overnight), the exponentially weighted moving averageovernight/intraday return (ewma_overnight/ewma_intraday).

As for energy stocks, (Regression 1) in **Table 6** shows that ewma_overnight, Mom, and Turnover are statistically significant. (Regression 2) shows that ret_overnight, Size, Turnover, ROE, and Ivol are statistically significant, while (Regression 3) reveals that all the ret_overnight, ewma_overnight, Size, Turnover, ROE are significant. It is worth noting that the sign of the coefficient of the same significant independent variables are totally different for (Regression 2) and (Regression 3). (Regression 4) and (Regression 5) indicate that the overnight and intraday partial alpha for each anomaly is not equal, and the scaled difference is obvious.

As for A-shares stocks, the results are essentially same with Energy stocks. Regression (1) in **Table 7** shows that ret_intraday, Size, BM, and Turnover are statistically significant. Regression (2) shows that ewma_overnight, ewma_intraday, Size, Turnover, and ROE are statistically significant, while Regression (3) reveals no independent variables are significant. Regression (4) and Regression (5) indicate that the overnight and intraday partial alpha for each anomaly is not equal, and the scaled difference is obvious.

In summary, the predictive power of these characteristics to the overnight return is better than that of the intraday component, which is consistent with the fact that most characteristic strategies of **Table 4** occur in the overnight to a certain extent.

5 CONCLUSION

Financial anomaly is one of the most important topics in empirical asset pricing and financial risk management. The extant papers show that there is a sharp distinction in the financial anomalies between the Chinese stock market and the Western stock market due to their market structure, market institution, investor’s structure, and so on. For instance, Hou et al. (2020) confirm that 65% of the 452 anomalies fail to hold up at the *t*-value of 1.96, while Qiao (2019) constructs 231 anomalies in Chinese A-shares stock market and find only 41 anomalies are significant at the 5% significance level. Enlightened by Lou et al. (2019) who link investor heterogeneity to the persistence of the overnight return and intraday returns, this paper attempts to

TABLE 6 | Fama-MacBeth return regression for Energy industry stocks. This table reports Fama-MacBeth regressions of monthly excess components of stock returns on lagged firm characteristics. The dependent variable in the first column is the close-to-close return in the following month; the dependent variable in the second column is the overnight return in the following month, and the dependent variable in the third column is the intraday return in the following month. In Column 4, we report the difference between the coefficients in Columns 2 and 3 (i.e., overnight-intraday). In Column 5, we report the difference between the overnight coefficient ∗24/18.5 and intraday coefficient ∗24/5.5. The independent variables are the same with **Table 5** *t* -statistics are calculated by correcting standard errors for serial-dependence with 12 lags. *,**,*** represent that the results are 10, 5,1% statistically significant, respectively. Sample period is 2001–2019.

	Close-to-close	Overnight	Intraday	Overnight-intraday	Scaled difference
	[1]	[2]	[3]	[4]	[5]
Intercept	−0.0096	0.0186***	−0.0264***	−0.0449***	0.1392***
	(−1.22)	(5.08)	(−3.80)	(5.71)	(4.53)
ret_overnight	−0.0028	0.0022**	−0.0049***	0.0070***	0.0241***
	(−1.50)	(2.15)	(−3.06)	(3.68)	(3.39)
ret_intraday	−0.0009	−0.0006	−0.0003	−0.0003	0.0004
	(−0.98)	(−1.01)	(−0.27)	(−0.23)	(0.09)
ewma_overnight	−0.0069*	0.0025	−0.0089***	0.0114***	0.0421***
	(−1.93)	(1.42)	(−2.70)	(2.92)	(2.84)
ewma_intraday	−0.0008	−0.0010	0.0004	−0.0014	−0.0028
	(−0.53)	(−0.97)	(0.24)	(−0.68)	(−0.42)
mom	0.0007*	0.0001	0.0006	−0.0005	−0.0023
	(1.72)	(0.57)	(1.42)	(−0.90)	(−1.26)
Size	0.0004	−0.0008***	0.0012***	−0.0020***	−0.0062***
	(1.17)	(−5.13)	(3.82)	(−5.77)	(−4.57)
Bm	−7.74e-05	−8.06e-05	6.07e-06	−8.67e-05	−0.0001
	(−1.31)	(−1.29)	(0.14)	(−0.96)	(−0.56)
Ivol	0.0067	0.0112*	−0.0038	0.0150	0.0310
	(0.57)	(1.75)	(−0.35)	(1.15)	(0.65)
beta	0.0002	−0.0003	0.0006	−0.0009*	−0.0029*
	(0.61)	(−1.10)	(1.57)	(−1.69)	(−1.69)
turnover	5.23e-06*	−7.44e-08	5.17e-06*	−5.24e-06	−2.27e-05*
	(1.75)	(−0.05)	(1.75)	(−1.40)	(−1.66)
Roe	−4.857e-06	2.7e-05***	−2.9e-05***	−5.649e-05***	0.0002***
	(−0.38)	(2.99)	(−2.77)	(3.72)	(3.28)
No.obs	10135	10135	10135	10135	10135

*As for A-shares stocks, the results are essentially same with Energy stock. (Regression 1) in **Table 6** shows that ret_intraday, Size, BM, and Turnover are statistically significant. (Regression 2) shows that ewma_overnight, ewma_intraday, Size, Turnover, and ROE are statistically significant, while (Regression 3) reveals no independent variables are significant. (Regression 4) and (Regression 5) indicate that the overnight and intraday partial alpha for each anomaly is not equal, and the scaled difference is obvious.*

TABLE 7 | Fama-MacBeth return regression for A-shares stocks.

	Close-to-close	Overnight	Intraday	Overnight-intraday	Scaled difference
	[1]	[2]	[3]	[4]	[5]
Intercept	0.0071***	0.0252***	0.0338***	−0.0608***	−0.1876***
ret_overnight	0.0004	0.0015	−0.0002	0.0004	−0.0001
ret_intraday	−0.0024**	−0.0014	0.0004	0.0000	0.0006
ewma_overnight	0.0000	0.0000***	0.0000	0.0***	0.0000
ewma_intraday	0.0000	−0.000***	0.0000	0.0000	0.0000
mom	0.0009	0.0000	0.0009	−0.0004	0.0005
Size	0.0131**	0.0231**	0.0084	0.0056	0.0097
Bm	−0.0002***	−0.0002	0.0000	0.0001	0.0000
Ivol	0.0000	0.0000	0.0000	0.0000	0.0000*
beta	0.0002	−0.0002	−0.0003	−0.0004	−0.0004
turnover	0.1249***	0.1585***	0.0782	0.1072*	0.0952*
Roe	0.0019	−0.0026**	−0.0011	−0.0009	−0.0009

*This table reports Fama-MacBeth regressions of monthly excess components of stock returns on lagged firm characteristics. The dependent variable in the first column is the close-to-close return in the following month; the dependent variable in the second column is the overnight return in the following month, and the dependent variable in the third column is the intraday return in the following month. In Column 4, we report the difference between the coefficients in Columns 2 and 3 (i.e., overnight-intraday). In Column 5, we report the difference between the overnight coefficient ∗24/18.5 and intraday coefficient ∗24/5.5. The independent variables are the same with **Table 5**.* t *-statistics are calculated by correcting standard errors for serial-dependence with 12 lags. *,**,*** represent that the results are 10, 5,1% statistically significant, respectively. Sample period is 2001–2019.*

explore the financial anomalies in the energy industry from the perspective of the component of close-to-close return. The paper demonstrates a unique characteristic of the overnight and intraday abnormal profits of a series of trading strategies in Chinese energy industry stocks and A-shares stocks, which is quite contrary to Lou et al. (2019).

More specifically, we first verify that there are overnight/intraday return persistence and reversal patterns. Contrary to the developed countries, the overnight premium is negative for energy industry stocks and A-shares stocks. In addition, by using portfolio analysis and Fama-MacBeth regression, it is found that risk-adjusted alphas earned by seven trading strategies based on the value, profitability, beta, turnover, idiosyncratic volatility, and reversal are actual overnight effects, while only size trading strategy is an intraday effect. However, in Lou et al. (2019), the premia of three types of momentum and reversal strategy mainly occur overnight, while others all occur within the day. We think it possible that overnight return puzzle caused by "T+1" trading rule in Qiao and Dam (2020) might contribute to the above results. Finally, the energy industry has its own uniqueness, that is, it is risk-related anomalies that occur during overnight for energy industry stocks, while both four risk-related anomalies and two firm characteristics related anomalies occur within the day for all A-shares stocks.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**. Further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

Conceptualization: MZ and XL Methodology: MZ and XL Software: MZ and XL Validation: MZ Formal analysis: XL Investigation: MZ and XL Resources: XL Data curation: MZ Writing—Original Draft: XL Writing—Review; Editing: MZ and XL Visualization: MZ Supervision: XL Project administration: MZ and XL Funding acquisition: MZ and XL.

FUNDING

MZ acknowledges financial support from the National Nature Science Foundation of China (grant number #72073058). XL acknowledges financial support from the National Nature Science Foundation of China (grant numbers #71761010, #72101067).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fenrg.2021.807881/full#supplementary-material

REFERENCES

Aboody, D., Even-Tov, O., Lehavy, R., and Trueman, B. (2018). Overnight Returns and Firm-specific Investor Sentiment. *J. Financ. Quant. Anal.* 53 (2), 485–505. doi:10.1017/s0022109017000989

Amihud, Y. (2002). Illiquidity and Stock Returns: Cross-Section and Time-Series Effects. *J. Financial Markets* 5, 31–56. doi:10.1016/s1386-4181(01)00024-6

Ang, A., Hodrick, R. J., Xing, Y., and Zhang, X. (2006). The Cross-Section of Volatility and Expected Returns. *J. Financ.* 61, 259–299. doi:10.1111/j.1540-6261.2006.00836.x

Black, F. (1972). Capital Market Equilibrium with Restricted Borrowing. *J. Busin.* 45, 444–454. doi:10.1086/295472

Branch, B., and Ma, A. (2012). Overnight Return, the Invisible Hand behind Intraday Returns. *J. Financ. Mark.* 2, 90–100.

Branch, B., and Ma, A. (2008). *The Overnight Return, One More Anomaly*. Unpublished working paper. Amherst: University of Massachusetts.

Cai, T. T., and Qiu, M. (2008). *International Evidence on Overnight Return Anomaly*. Unpublished working paper. Wellington: Massey University.

Campbell, J. Y., Hilscher, J., and Szilagyi, J. (2008). In Search of Distress Risk. *J. Financ.* 63, 2899–2939. doi:10.1111/j.1540-6261.2008.01416.x

Cao, J., Wen, F., Zhang, Y., Yin, Z., and Zhang, Y. (2022). Idiosyncratic Volatility and Stock price Crash Risk: Evidence from china. *Financ. Res. Lett.* 44, 102095. doi:10.1016/j.frl.2021.102095

Carhart, M. M. (1997). On Persistence in Mutual Fund Performance. *J. Finance* 52, 57–82. doi:10.1111/j.1540-6261.1997.tb03808.x

Cliff, M. T., Cooper, M. J., and Gulen, H. (2008). *Return Differences between Trading and Non-trading Hours: Like Night and Day*. Unpublished working paper. Blacksburg: Virginia Tech.

Dai, Z. F., Zhou, H. T., Kang, J., and Wen, F. H. (2021). The Skewness of Oil price Returns and Equity Premium Predictability. *Energy Econ* 94, 1–11. doi:10.1016/j.eneco.2020.105069

Datar, V. T., Y. Naik, N., and Radcliffe, R. (1998). Liquidity and Stock Returns: An Alternative Test. *J. Financial Markets* 1, 203–219. doi:10.1016/s1386-4181(97)00004-9

Diether, K. B., Malloy, C. J., and Scherbina, A. (2002). Differences of Opinion and the Cross Section of Stock Returns. *J. Finance* 57, 2113–2141. doi:10.1111/0022-1082.00490

Dimson, E. (1979). Risk Measurement when Shares Are Subject to Infrequent Trading. *J. Financial Econ.* 7, 197–226. doi:10.1016/0304-405x(79)90013-8

Fama, E. F., and French, K. R. (2015). A Five-Factor Asset Pricing Model. *J. Financial Econ.* 116, 1–22. doi:10.1016/j.jfineco.2014.10.010

Fama, E. F., and French, K. R. (1993). Common Risk Factors in the Returns on Stocks and Bonds. *J. Financial Econ.* 33, 3–56. doi:10.1016/0304-405x(93)90023-5

Fama, E. F., and French, K. R. (1992). The Cross-Section of Expected Stock Returns. *J. Finance* 47, 427–465. doi:10.1111/j.1540-6261.1992.tb04398.x

Fama, E. F., and MacBeth, J. D. (1973). Risk, Return, and Equilibrium: Empirical Tests. *J. Polit. Economy* 81, 607–636. doi:10.1086/260061

Farouq, I. S., Sambo, N., Umar Sambo, N., Ahmad, A. U., Jakada, A. H., and Danmaraya, I. i. A. (2021). Does Financial Globalization Uncertainty Affect CO2 Emissions? Empirical Evidence from Some Selected SSA Countries. *Quant. Financ. Econ.* 5 (2), 247–263. doi:10.3934/qfe.2021011

Frazzini, A., and Pedersen, L. H. (2014). Betting against Beta. *J. Financial Econ.* 111, 1–25. doi:10.1016/j.jfineco.2013.10.005

Ghoddusi, H., and Wirl, F. (2021). A Risk-Hedging View to Refinery Capacity Investment in OPEC Countries. *Energy J* 42 (1), 67–92. doi:10.5547/01956574.42.1.hgho

Gong, X., and Lin, B. (2021). Effects of Structural Changes on the Prediction of Downside Volatility in Futures Markets. *J. Futures Markets* 41, 1124–1153. doi:10.1002/fut.22207

Gong, X., and Lin, B. (2017). Forecasting the Good and Bad Uncertainties of Crude Oil Prices Using a HAR Framework. *Energ. Econ.* 67, 315–327. doi:10.1016/j.eneco.2017.08.035

Gong, X., and Lin, B. (2018). The Incremental Information Content of Investor Fear Gauge for Volatility Forecasting in the Crude Oil Futures Market. *Energ. Econ.* 74, 370–386. doi:10.1016/j.eneco.2018.06.005

Gong, X., Liu, Y., and Wang, X. (2021). Dynamic Volatility Spillovers across Oil and Natural Gas Futures Markets Based on a Time-Varying Spillover Method. *Int. Rev. Financial Anal.* 76, 101790. doi:10.1016/j.irfa.2021.101790

Güngör, A., and Taştan, H. (2021). On Macroeconomic Determinants of Co-movements Among International Stock Markets: Evidence from DCC-MIDAS Approach. *Quant. Financ. Econ.* 5 (1), 19–39. doi:10.3934/qfe.2021002

Haugen, R. A., and Baker, N. L. (1996). Commonality in the Determinants of Expected Stock Returns. *J. Financial Econ.* 41, 401–439. doi:10.1016/0304-405x(95)00868-f

Hendershott, T., Livdan, D., and Rösch, D. (2020). Asset Pricing: A Tale of Night and Day. *J. Financial Econ.* 138 (3), 635–662. doi:10.1016/j.jfineco.2020.06.006

Hou, K., Xue, C., and Zhang, L. (2015b). *A Comparison of New Factor Models*. Working paper. Columbus: Ohio State University, University of Cincinnati, and Ohio State University.

Hou, K., Xue, C., and Zhang, L. (2015a). Digesting Anomalies: An Investment Approach. *Rev. Financ. Stud.* 28, 650–705. doi:10.1093/rfs/hhu068

Hou, K., Xue, C., and Zhang, L. (2020). Replicating Anomalies. *Rev. Financ. Stud.* 33 (5), 2019–2133. doi:10.1093/rfs/hhy131

Huong, T. T. X., Nga, T., Nga, T. T. T., and Oanh, T. T. K. (2021). Liquidity Risk and Bank Performance in Southeast Asian Countries: a Dynamic Panel Approach. *Quant. Financ. Econ.* 5 (1), 111–133. doi:10.3934/qfe.2021006

Jegadeesh, N., and Titman, S. (1993). Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency. *J. Finance* 48, 65–91. doi:10.1111/j.1540-6261.1993.tb04702.x

Kelly, M. A., and Clark, S. P. (2011). Returns in Trading versus Non-trading Hours: the Difference Is Day and Night. *J. Asset Manag.* 12, 132–145. doi:10.1057/jam.2011.2

Lee, C. M. C., and Swaminathan, B. (2000). Price Momentum and Trading Volume. *J. Finance* 55, 2017–2069. doi:10.1111/0022-1082.00280

Lian, Z. Y., Cai, J., and Webb, R. I. (2020). Oil Stocks, Risk Factors, and Tail Behavior. *Energ. Econ* 91, 1–19. doi:10.1016/j.eneco.2020.104932

Lintner, J. (1965). The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets. *Rev. Econ. Stat.* 47, 13–37. doi:10.2307/1924119

Liow, K. H., Song, J., Song, J., and Zhou, X. (2021). Volatility Connectedness and Market Dependence across Major Financial Markets in China Economy. *Quant. Financ. Econ.* 5 (3), 397–420. doi:10.3934/qfe.2021018

Liu, R., Chen, J., and Wen, F. (2021). The Nonlinear Effect of Oil price Shocks on Financial Stress: Evidence from China. *North Am. J. Econ. Finance* 55, 101317. doi:10.1016/j.najef.2020.101317

Lou, D., Polk, C., and Skouras, S. (2019). A Tug of War: Overnight versus Intraday Expected Returns. *J. Financial Econ.* 134 (1), 192–213. doi:10.1016/j.jfineco.2019.03.011

Mao, L., Zhang, Y., and Zhang, Y. (2021). Robust Optimal Excess-Of-Loss Reinsurance and Investment Problem with P-Thinning Dependent Risks under CEV Model. *Quant. Financ. Econ.* 5 (1), 134–162. doi:10.3934/qfe.2021007

Muravyev, D., and Ni, X. (2020). Why Do Option Returns Change Sign from Day to Night? *J. Financial Econ.* 136, 219–238. doi:10.1016/j.jfineco.2018.12.006

Novy-Marx, R. (2013). The Other Side of Value: the Gross Profitability Premium. *J. Financial Econ.* 108, 1–28. doi:10.1016/j.jfineco.2013.01.003

Peng, Q., Wen, F., and Gong, X. (2021). Time-dependent Intrinsic Correlation Analysis of Crude Oil and the US Dollar Based on CEEMDAN. *Int. J. Fin Econ.* 26 (1), 834–848. doi:10.1002/ijfe.1823

Qiao, F. (2019). *Replicating Anomalies in China*. Working paper. Elsevier.

Qiao, K., and Dam, L. (2020). The Overnight Return Puzzle and the "T+1" Trading Rule in Chinese Stock Markets. *J. Financ. Markets* 50, 1–13. doi:10.1016/j.finmar.2020.100534

Robert, F., Stambaugh, R. F., and Y Yuan, Y. (2017). Mispricing Factors. *Rev. Financ. Stud.* 30 (4), 1270–1315.

Sharpe, W. F. (1964). Capital Asset Prices: a Theory of Market Equilibrium under Conditions of Risk*. *J. Finance* 19, 425–442. doi:10.1111/j.1540-6261.1964.tb02865.x

Si, D. K., Li, X. L., and Huang, S. J. (2021). Financial Deregulation and Operational Risks of Energy enterprise: The Shock of Liberalization of Bank Lending Rate in China. *Energ. Econ* 93, 1–10. doi:10.1016/j.eneco.2020.105047

Stambaugh, R. F., Yu, J., and Yuan, Y. (2015). Arbitrage Asymmetry and the Idiosyncratic Volatility Puzzle. *J. Finance* 70, 1903–1948. doi:10.1111/jofi.12286

Stambaugh, R. F., Yu, J., and Yuan, Y. (2014). The Long of it: Odds that Investor Sentiment Spuriously Predicts Anomaly Returns. *J. Financial Econ.* 114, 613–619. doi:10.1016/j.jfineco.2014.07.008

Stambaugh, R. F., Yu, J., and Yuan, Y. (2012). The Short of it: Investor Sentiment and Anomalies. *J. Financial Econ.* 104, 288–302. doi:10.1016/j.jfineco.2011.12.001

Tian, M., Li, W., and Wen, F. (2021). The Dynamic Impact of Oil price Shocks on the Stock Market and the USD/RMB Exchange Rate: Evidence from Implied Volatility Indices. *North Am. J. Econ. Finance* 55, 101310. doi:10.1016/j.najef.2020.101310

Umutlu, M., Bengitöz, P., and Bengitöz, P. (2021). Return Range and the Cross-Section of Expected index Returns in International Stock Markets. *Quant. Financ. Econ.* 5 (3), 421–451. doi:10.3934/qfe.2021019

Vuolteenaho, T. (2002). What Drives Firm-Level Stock Returns. *J. Finance* 57, 233–264. doi:10.1111/1540-6261.00421

Wang, J. Y., Wang, H., and Wang, D. (2021). Equity Concentration and Investment Efficiency of Energy Companies in China: Evidence Based on the Shock of Deregulation of QFIIs. *Energ. Econ* 93, 1–8. doi:10.1016/j.eneco.2020.105032

Wen, F., Cao, J., Liu, Z., and Wang, X. (2021a). Dynamic Volatility Spillovers and Investment Strategies between the Chinese Stock Market and Commodity Markets. *Int. Rev. Financial Anal.* 76, 101772. doi:10.1016/j.irfa.2021.101772

Wen, F., Zhang, K., and Gong, X. (2021b). The Effects of Oil price Shocks on Inflation in the G7 Countries. *North Am. J. Econ. Finance* 57, 101391. doi:10.1016/j.najef.2021.101391

Ye, Z. K., Hu, C. Y., Oy, G. D., and W, F. H. (2020). The Dynamic Time-Frequency Relationship between International Oil Prices and Investor Sentiment in China: A Wavelet Coherence Analysis. *Energ. J* 41 (5), 251–270. doi:10.5547/01956574.41.5.fwen

Zheng, Y., Zhou, M., and Wen, F. (2021). Asymmetric Effects of Oil Shocks on Carbon Allowance price: Evidence from China. *Energ. Econ.* 97, 105183. doi:10.1016/j.eneco.2021.105183

Zolfaghari, M., Ghoddusi, H., and Faghihian, F. (2020). Volatility Spillovers for Energy Prices: A diagonal BEKK Approach. *Energ. Econ* 92, 1–17. doi:10.1016/j.eneco.2020.104965

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Energy Research

METHODS
published: 07 July 2022
doi: 10.3389/fenrg.2022.741018

Research on Risk Features and Prediction of China's Crude Oil Futures Market Based on Machine Learning

Yaoqi Guo[1], *Shuchang Zhang*[1] *and Yanqiong Liu*[2]*

[1]School of Mathematics and Statistics, Central South University, Changsha, China, [2]School of Mathematics and Statistics, Hunan First Normal University, Changsha, china

Facing the rapidly changing domestic and foreign futures markets, how to accurately and immediately predict the price trend of crude oil futures in order to avoid the risks caused by price fluctuations is very important for all participants in the crude oil futures market. Based on the 5-min high-frequency trading data of China's crude oil futures market in recent 3 years, this paper uses the EMD-MFDFA model combined with multifractal detrended fluctuation analysis (MF-DFA) and empirical mode decomposition unsupervised K-means clustering and Gaussian mixture model (GMM) to identify the risk status of each trading day. Further, Support vector machine (SVM), extreme gradient lifting (XGBoost) and their improved algorithms are used to predict the risk state of China's crude oil futures market. The empirical results are as follows: first, There are obvious multifractal features in the return rate series of China's crude oil futures market and its single trading day; Second, compared with the traditional SVM model, the improved Twin Support Vector Machine (TWSVM) based on solving the sample imbalance issue has better prediction ability for China's crude oil futures risk.; Third, The XGBoost has a great impact on the prediction of China's crude oil risk, and the Focal-XGBoost with focal loss function performs the best in predicting the risk of China's crude oil futures market.

Keywords: China's crude oil futures, multifractal, clustering, sample imbalance, risk prediction

OPEN ACCESS

Edited by:
Xun Zhang,
Academy of Mathematics and Systems Science (CAS), China

Reviewed by:
Vidya C. T.,
Centre for Economic and Social Studies (CESS), India
Narottam Das,
Central Queensland University, Australia
Wendong Yang,
Shandong University of Finance and Economics, China

****Correspondence:***
Yanqiong Liu
sxyliuyq@hnfnu.edu.cn

Specialty section:
This article was submitted to Sustainable Energy Systems and Policies,
a section of the journal Frontiers in Energy Research

Received: *14 July 2021*
Accepted: *08 June 2022*
Published: *07 July 2022*

Citation:
Guo Y, Zhang S and Liu Y (2022) Research on Risk Features and Prediction of China's Crude Oil Futures Market Based on Machine Learning. Front. Energy Res. 10:741018. doi: 10.3389/fenrg.2022.741018

INTRODUCTION

With the rapid development of economy, energy issues have become the focus of the world. Energy is indispensable to the world economic development, and crude oil plays an important role in the energy market. According to the 2019–2020 Blue Book of China's Oil and Gas Industry Development Analysis and Prospect Report, China ranks among the top in both crude oil imports and consumption. Specifically, China's crude oil imports reached 506 million tons in 2019, with a year-on-year growth of 9.5%, and its external dependence reached 70.8%. In terms of crude oil consumption, China consumed 696 million tons in 2019, with a year-on-year growth of 6.8%. The data indicate that the crude oil market has a huge impact on China's energy economic market, and its price fluctuation often brings huge consequences.

With the rapid development of economy, energy issues have become the focus of the world. Energy is indispensable to the world economic development, and crude oil plays an important role in the energy market. According to the 2019–2020 Blue Book of China's Oil and Gas Industry Development Analysis and Prospect Report, China ranks among the top in both crude oil imports and consumption. Specifically, China's crude oil imports reached 506 million tons in 2019, with a year-on-year growth of 9.5%, and its external dependence reached 70.8%. In terms of crude oil consumption, China consumed 696 million tons in 2019, with a year-on-year growth of 6.8%. With the sustained and rapid growth of China's economy, the demand for crude oil import and consumption is increasing, the fluctuation of crude oil price has an increasing impact on China.

After years of development, the crude oil market, which is closely related to the economic development of each country, has formed a relatively authoritative price system, and its supply and demand as well as trade are carried out in the global scope. Before the official launch of Chinese crude futures, West Texas Intermediate (WTI) of the United States and Brent of the United Kingdom dominated the pricing system for global oil prices. After 17 years of careful planning, China's crude oil futures market was officially listed on the Shanghai International Energy Exchange on 26 March 2018, denominated in RMB, and filled the gap of domestic crude oil futures market. Less than half a year after listing, China's crude oil futures trading volume has reached 17 million contracts, accounting for 12% of the global crude oil futures market volume, and its accumulated trading volume reached 8.57 trillion yuan, ranking among the top three in the world. As can be seen from the data, China's crude oil futures market is developing rapidly. Up to now, China's crude oil futures market has exceeded 6% of the international market share, and the market activity has been continuously improved, becoming the third largest crude oil futures variety in the world after WTI and Brent.

China's oil futures is of great significance to the global oil futures market. It sets a benchmark for Asian oil futures markets and provides a channel for Chinese companies to hedge their oil consumption and avoid risks. At the same time, the establishment of a crude oil price benchmark level that reflects the relationship between demand and supply in China and the Asia-Pacific market has filled the gap in the existing international crude oil pricing system and increased China's participation in the international market. However, compared with the mature crude oil futures market, China's crude oil futures market, which has been established for a short time, has many aspects to be improved and the demand for risk aversion has become increasingly urgent. Therefore, it is necessary to study the risk status of China's crude oil futures market from the perspective of market price fluctuation.

However, traditional risk research models, such as VaR (Value at Risk), are mainly based on the efficient market hypothesis (EMH) proposed by Fama. EMH believes that investors can respond to information rationally and linearly, so market prices can timely and fully reflect information changes in the system, that is, prices in the financial market have no long-term memory, and price fluctuations are unpredictable. However, a lot of research found that financial market usually shows nonlinear structural characteristics, and its complex operation mechanism, which cannot reflect the actual situation of the market, is contrary to the efficient market hypothesis. Therefore (Altman, 1967), proposed the nonlinear fractal theory for measuring financial investment risk. Further (Peters, 1994a), proposed the Fractal Market Hypothesis (FMH) on the basis of Mandelbrot's theory. From the practical point of view, he regarded the capital market as a complex nonlinear dynamic system with the characteristics of interaction and self-adaptability. Therefore, FMH, with the characteristics of interaction and self-adaptability, can better describe the complexity of the market, analyze the nonlinear dynamic characteristics of market price fluctuations, measure the impact of information on prices, and explore the predictability of the market. A large number of studies also show that fractal features are indeed universal in financial markets.

Furthermore, with the development of computer technology, machine learning algorithms, such as Decision Tree, Support Vector Machine (SVM) and Artificial Neural Network (ANN) came into being. With the further development of technology, the integration algorithm, which combines several weak learners into strong learners, has received more and more attention. The main ways to synthesize weak learners are bagging, boosting and stacking. For example, Random Forest is the representative of bagging algorithm, and Extreme Gradient Boosting (XGBoost) is a boosting algorithm. Machine learning models have been widely used in the research of risk prediction due to their outstanding advantages in dealing with nonlinear complex systems.

Taking China's crude oil futures market as the research object, this paper introduces the multifractal feature parameters into the machine learning model, and carries out risk status recognition and prediction of China's crude oil futures market. In the turbulent economic situation, futures with its unique hedging function is favored by more and more investors, and has become a crisis management means to deal with the economic recession. By predicting the risk of China's crude oil futures market, relevant investors can find the potential risk in advance and formulate preventive and control measures in time, so as to avoid the risk reasonably and reduce the loss to a large extent.

LITERATURE REVIEW

Existing relevant literature mainly focuses on four aspects, namely, the characteristics of crude oil futures, multifractal method, multifractal spectrum parameters and financial market risk prediction.

The first is to study the risk features of crude oil futures. At present, more and more scholars study China's crude oil futures, China's crude oil market environment and oil policy. Sun et al. (2018) used GARCH and TARCH models to study the fluctuation characteristics of China's crude oil futures returns rate based on high-frequency data, and they found that the changes of China's crude oil futures returns rate in the current as well as the lag period were mainly influenced by itself, and the influence coefficient of one period lag was larger and the influence time was longer. Ji and Zhang (2019) analyzed the initial characteristics of China's crude oil futures market, laying a good foundation for subsequent studies. Li et al. (2019) proposes a new, novel crude oil price forecasting method based on online media text mining, with the aim of capturing the more immediate market antecedents of price fluctuations, the empirical results suggest that the proposed topic-sentiment synthesis forecasting models perform better than the older benchmark models. Liu et al. (2019a) constructed Copula-POT-CoVaR model to study the Risk Spillover Effect of crude oil market on BRIC stock markets, and found that there was significant risk spillover. Özdurak (2021) constructed DCC-GARCH model to study the spillover effect of crude oil price on clean energy investment, and found that with the rise of oil price, renewable energy investment will also tend to decrease. Weng et al. (2021) proposed a modeling framework, genetic algorithm regularization online extreme learning machine with forgetting factor (GA-RFOS-ELM), to estimate the effects of news during the COVID-19 pandemic on the volatility of crude oil futures which could be effective and efficient in volatility forecasting of crude oil futures.

The second is to study the multifractal method. Since the traditional efficient market theory does not conform to the objective facts, Mandelbrot (Altman, 1967) first proposed the concept of fractal in the 1970s. On this basis, Peters (1994a) proposed the fractal market hypothesis (FMH). R/S method was first proposed by Hurst in hydrological analysis in 1951, and was first used in the analysis of financial time series by Mandelbrot (Mandelbrot and Wheeler, 1983) in 1983. However, the research of Lo (1989) and Peters (1994b), Peters (1996) found that the length of sample interval and the short-term correlation of samples will affect the analysis results of R/s method. In order to solve this defect, Peng et al. (1994) proposed detrended fluctuation analysis (DFA) when studying the chimeric tissue of DNA, which distinguishes local correlation from long-term correlation, so as to remove the pseudo correlation phenomenon, and can effectively analyze the long-term power-law correlation of unstable time series, which is widely used in financial time series analysis. On this basis (Kantelhardt et al., 2002), generalized the DFA method and obtained the multifractal detrended fluctuation analysis (MF-DFA) method. In 2008, podobnik and Stanley (Podobnik and Stanley, 2008; Podobnik et al., 2009) formed detrended cross correlation analysis (DCCA) on DFA method, which expanded it into a method that can measure the long-term correlation of two non-stationary time series. Jiang and Zhou (2011) and others further improved the MF-DCCA method and proposed multifractal detrended moving average correlation analysis (MF-X-DMA) (Wang et al., 2012). Combined statistical moment with multifractal cross-correlation analysis to test the cross multifractality between the two sequences. Ruan et al. (2016) used the price and trading volume data of gold spot and futures to study the cross-correlation and time-varying characteristics of price and trading volume. Zhang et al. (2019) and others studied the multifractal characteristics of bitcoin market with MF-DCCA, and further analyzed the multifractal correlation between bitcoin price and other financial market prices. Feng and Cao (2022) used multifractal detrended cross-correlation analysis (MF-X-DFA) and multifractal detrended partial cross-correlation analysis (MF-DPXA) to explore the fluctuation characteristics of cross-correlation between China and the United States agricultural futures market before and after canceling the price of West Texas medium crude oil futures, as well as the impact and cross-correlation on the market.

The third is to study the multifractal spectrum parameters. In the field of engineering, multifractals are mostly used to extract the characteristics of signals, and then the extracted parameters are used in the research of signal recognition and classification. Li and Xie (2013) identified the multifractal spectrum characteristics of radar signals and discussed the identification mechanism of multifractal spectrum parameters. The empirical study shows that the feature parameters are effective to recognize signals. Li et al. (2020a) verified the validity of multifractal spectral parameters by analyzing the multifractal features of friction signals and quantitatively describing the friction vibration characteristics under different friction states through the calculated spectral parameters. In the field of finance, multifractal parameters have also been widely used. Sun et al. (2001) found that the main parameter $\Delta f(\alpha)$ of the multifractal spectrum was directly related to the daily return rate of Hang Seng Index. In order to make better use of the statistical information in the multifractal spectrum (Wei and Huang, 2005), constructed a new market risk measurement method, which contains the comprehensive information of the multifractal spectrum parameters $\Delta\alpha$ and $\Delta f(\alpha)$. After theoretical and empirical research, they believe that the multifractal parameter method is a powerful tool for studying price fluctuations in financial markets, from which a large amount of statistical information can be obtained, which is helpful for us to understand the complexity of financial markets. Yuan et al. (2009) used the MF-DFA to study the multifractal features of daily returns of Shanghai Composite Index, and they also used the range (Δh) and standard deviation (σ_h) of the generalized Hurst index to measure the risk of the securities market. They believe that the greater Δh and σ_h are, the greater the multifractal intensity is, and the greater the market risk is. The empirical results show that this risk measurement index is reasonable to the Chinese stock market risk measurement. Zhu and Zhang (2018) analyzed the multifractal structure of China's stock market by using the MF-DFA, and they found that the shape and width of the multifractal spectrum were related to the order. Through further study, they found that the multifractal parameters played an important role in risk prediction.

The fourth aspect is the financial market risk prediction research. At present, the risk prediction models of financial market can be divided into two categories: one is the statistical approach, which mainly includes linear models such as univariate, multivariate and logistic regression. The idea of multivariate linear early warning model was first proposed by (Altman, 1967), whose Z-score model is the most classic and representative linear risk prediction model at present. Dong et al. (2019) use the CAViaR method to forecast the oil return risks, and further depict the dynamic and heterogeneous features during the crisis (or non-crisis) period, as well as in different markets via DCC-GARCH models. Latunde et al. (2020) uses the CAPM and some statistical tools (variance, covariance and mean) to study risks on the expected return of investing in four common Deutsche Bank (DB) crude oil assets, the result reveals that DTO-DB Crude oil Double Short has the highest beta risk and highest expected return. And the higher the risk, the higher the expected return, and vice versa, that is, the risk is directly proportional to the expected return. Liu et al. (2019b) extend the Copula-CoVaR models by introducing the Peak-over-Threshold and construct the Copula-POT-CoVaR model to investigate the risk spillover effect from crude oil market to BRICS stock markets. By using the crude oil market and BRICS stock market data from 2006 to 2016 as the sample, the empirical study results show that: there is a significant risk spillover from crude oil market to BRICS stock markets, and the risk of crude oil market explains more than 50 percent of BRICS stock markets' risk. Li et al. (2021) use the Conditional Autoregressive Value at Risk models (CAViaR) approach to forecast the risk of Bitcoin's returns, the results show that Bitcoin's volatility is significantly related to the volatility of the crypto-asset's return and the main determinants of volatility are speculation, investor attention, market interoperability and the interaction between speculation and market interoperability. Li et al. (2020b) measure the return risks of the cryptocurrencies by using the CAViaR model, the results show that they have similar risk tendencies, the risk spillover directions are highly correlative with the market capitalizations of the cryptocurrencies. However, the statistical approach, which mainly includes linear models, is difficult to describe the nonlinear relationship in the financial market. The second category is the machine learning approach. With the rapid development of machine learning algorithms, many scholars begin to combine computer technology with relevant knowledge of financial markets to do interdisciplinary research on the risk prediction of financial markets, with algorithms such as Support Vector Machine (SVM) and Extreme Gradient Boost (XGBoost). Tam and Kiang (1990) compared the neural network model and the traditional statistical model in predicting the risks of banks, and he found that the prediction accuracy of the BP (back propagation) neural network was higher. Later, in order to make the results more reliable (Tam, 1991), compared the prediction results of the BP neural network with those of other algorithms (such as logistic regression, decision tree and feed-forward artificial neural network), and he found that the prediction effect of the BP neural network was the best. Uthayakumar et al. (2020) proposed a cluster-based classification model, including improved K-means clustering and fitness-scaling chaotic genetic ant colony algorithm (FSCGACA) classification model to predict financial crises. Zhao et al. (2018) used least squares support vector machine (LSSVM) to predict systemic financial risks, and Particle Swarm Optimization (PSO) was used to optimize the parameters of the model, and the results show that LSSVM is better at accurate prediction and generalization. Ma and Lv (2019) took the objective function of machine learning algorithms such as support vector machine and neural network as the basis function to carry out the weighted average, and used the constructed Multi-Lingual Information Access (MLIA) model to predict the credit risk of Internet finance. The empirical results show that this model has a higher prediction accuracy compared with logistic regression. Li and Quan (2019) used BP neural network to predict the financial risks of manufacturing enterprises, optimized the model parameters by using improved particle swarm optimization (IPSO), and established a financial risk prediction model based on the IPSOBP model.

Throughout the above literature, although the existing literature has carried out a large number of studies on the multifractal theory and analysis methods, multifractal spectrum parameters and risk prediction models, there is still room for further research. ① Since China's crude oil futures market is an emerging market, there are few studies on it at present. Most of the existing research focus on price fluctuations of China's crude oil futures, or comparison with other markets through econometric models by studying the co-integration relationship, Granger causality relationship or linkage effect between markets. Although (Wang et al., 2011) introduced the multifractal method into the research of China's crude oil futures market, they did not study the risk of this market from the perspective of multifractal spectrum parameters. ② A large number of existing studies focus on the confirmation and generation mechanism of multifractal features of financial markets, but the achievements of fractal theories applied to financial markets are relatively scattered. Although some scholars have substituted the fractal indirect index (fractal spectral parameter) for variance to measure the financial market risk, there are few studies that combine the multi-fractal parameters with clustering algorithm to carry out pattern recognition of market risk. ③ Although the machine learning method has been introduced into the research of financial market risk prediction, it mainly focuses on the analysis and measurement of the overall risk of the market, instead of using the multi-fractal parameters to predict the risk status of the financial market from the perspective of the multifractal features. In this paper, therefore, with China's crude oil futures as the research object, we employ the multifractal theory framework and introduce multifractal feature parameters into the machine learning model to identify and predict China's oil futures market risk, so as to provide relevant investors a more effective reference for risk management by helping them identify potential risks in advance and promptly formulate prevention and control measures.

The marginal contribution of this paper is mainly reflected in the following two aspects. First, this paper studies the multifractal features of China's crude oil futures market from the perspective of high frequency. This paper calculates the intra-day multifractal spectrum parameters through the improved EMD-MFDFA method, and combines it with the unsupervised clustering algorithm to identify as well as define the risk status of the market in each trading day. Second, this paper adopts SVM and XGBoost as well as their improved algorithms based on sample imbalance issue to predict the risk status of China's crude oil futures market, so that relevant investors can identify potential risks in advance and formulate prevention and control measures in time.

The overall framework of this paper is as follows: **Section 3** analyzes the risk characteristics of China's crude oil futures market, providing sample data for the risk prediction of energy futures market; **Section 4** identifies and measures the risk of China's crude oil futures market; **Section 5** is about the risk prediction of China's crude oil futures market. The main conclusions of this paper are in **Section 6**.

RISK FEATURES OF CHINA'S CRUDE OIL FUTURES MARKET

Data Sources and Basic Analysis of China's Crude Oil Futures Market

This paper selects China's crude oil futures issued in March 2018 as the research object, and the sample data time span is from 26 March 2018 to 1 March 2021, with a total of 73,575 5-min high-frequency trading records of 712 trading days (excluding weekends and holidays; data are from Shanghai Futures Exchange). Data collection starts at 21:00 p.m. on the day before trading and ends at 15:00 p.m. on the day of trading, recording once every 5 min, then 111 pieces of data can be

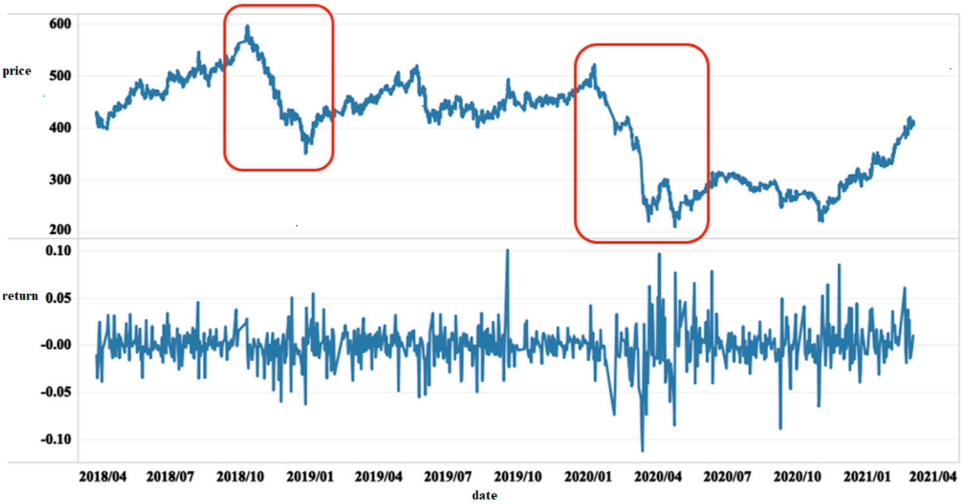


FIGURE 1 | Time series of China's crude oil futures market price and return rate.

collected on each trading day (Note: the trading time of each trading day is 21:00-02:30, 09:00-11:30, 13:30-15:00).

This paper defines the logarithmic return rate as: $Return = lnP(t+1) - lnP(t)$, where $P(t)$ represents the closing price of China's crude oil futures market at time t. **Figure 1** shows the fluctuation situation of the closing price of China's crude oil futures and the corresponding return rate. As can be seen from the figure, the price of China's crude oil futures dropped significantly at the end of 2018, even erasing all the gains since the beginning of the year. The possible reason for this situation is that the growth of international crude oil demand is weak, but the supply is greatly increased, leading to the imbalance between supply and demand. Secondly, the rapid rise of oil price at the early stage has a negative impact on economy and society (high oil price leads to economic recession, which in turn leads to a series of social unrest), thus leading to the continuous decline of oil price. Similarly, from the end of 2019 to the beginning of 2020, affected by the global COVID-19 epidemic, the export and storage of crude oil were blocked, leading to a continuous and significant decline in the oil price, and the corresponding returns fluctuation increased significantly compared with other periods, and there was an obvious fluctuation aggregation phenomenon.

In addition, **Table 1** displays the descriptive statistics of sample data. The series skewness and kurtosis shown in the table are obviously not zero, indicating obvious non-normality of both the price series and the return rate series. Specifically, the skewness values of price and returns are both less than 0, and the kurtosis values are greater than 0. According to the skewness value, the distribution of the return rate series is slightly to the left. The kurtosis value indicates that the return rate series presents the characteristic of sharp peak and thick tail. What's more, the Jarque-Bena (JB) statistic is used to test the normality of the sequence, and it is found that the JB statistic is relatively large, which indicates that the hypothesis of the sequence obeying normal distribution is rejected at the 1% confidence level.

Multifractal Features of China's Crude Oil Futures Market

Although the MF-DFA method can effectively analyze the multifractal features of non-stationary time series, there are still some shortcomings in this method. Firstly, the MF-DFA method requires the time series to be detrended. Specifically, it is found that when the MF-DFA method is used to segment the whole sequence, the segmented interval length is not always an integral multiple of the original sequence length, so the segmented interval is not always continuous. This uncertainty will lead to the discontinuity of the fitting polynomials of adjacent segmented intervals, which may produce new pseudo-random fluctuation error, and then make the fluctuation function produce a certain deviation, resulting in the distortion of the scale index. Therefore, in this paper, the sliding-window method is adopted to improve the discontinuity problem of the segmented interval, so that the segmentation of the non-overlapping interval is optimized into continuous overlapping interval, and the error caused by the discontinuity of the segmented interval is avoided. Secondly, in the MF-DFA method, the polynomial fitting method is used to estimate the local trend of the sequence, and each interval should be de-trended. But the polynomial fitting needs to determine the order of polynomial artificially in advance, and there is no certain standard for the choice of order, so it is subject to great random interference. Therefore, this paper combines the empirical mode decomposition (EMD) with multifractal detrended fluctuation analysis to improve the shortcomings of MF-DFA. The improved EMD-MFDFA method eliminates the trend term extracted by empirical mode decomposition from the original series, so as to eliminate the trend in the time series and avoid the error caused by the unfixed order of polynomial fitting.

To sum up, this paper combines the advantages of the sliding-window technology and the EMD method to improve the original

TABLE 1 | Descriptive statistics of data.

	Mean	Maximum	Minimum	Standard Deviation	Skewness	Kurtosis	JB statistics
Price	408.02	597.60	209.70	87.88	–0.46	1.99	5785.73
Returns rate	–6.99E-07	0.09	–0.09	0.00	–0.11	302.11	2.74E+08

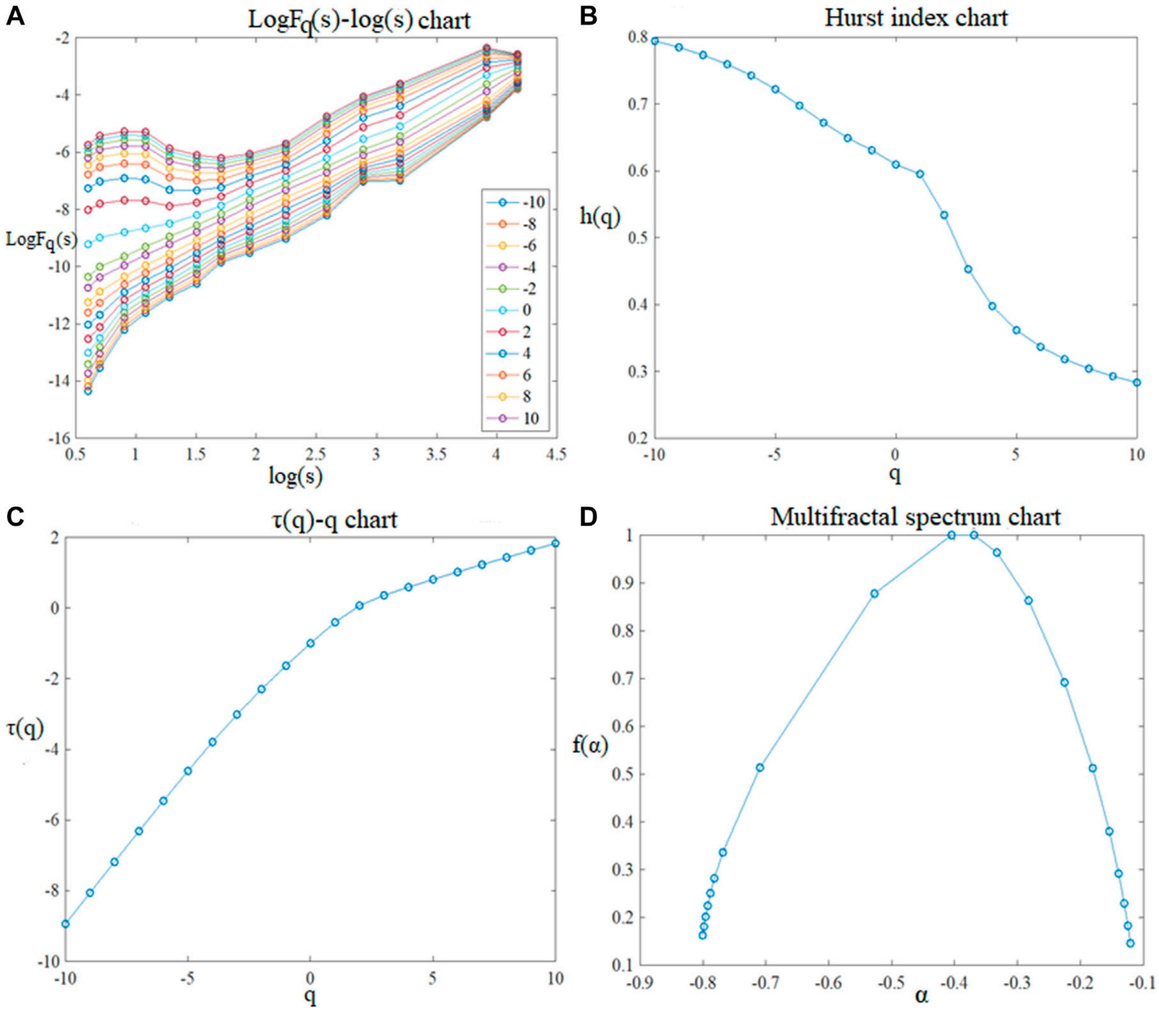


FIGURE 2 | Multifractal features analysis of China's crude oil futures market with the EMD-MFDFA.

MF-DFA method, and uses the improved EMD-MFDFA method to analyze the multifractal features of the return rate series of China's crude oil futures market, which are shown in **Figure 2**.

The following conclusions can be drawn from **Figure 2**:

① Figure**2A** shows the double logarithm relationship between the scale s and the fluctuation function $F_q(s)$ (q-order wave function) at different values of q. It is obvious that when s increases to a certain extent, the fluctuation function $F_q(s)$ increases roughly linearly, which indicates that the return rate series of China's crude oil futures market has obvious power-law correlation and long-term correlation. It should be noted that the above linear relationship changes when $s = 23$, and that 23 corresponds to about 1 month, which is consistent with the results of most financial markets.

② As is known to all, when the value of $h(q)$ (Generalized Hurst index) changes with the value of q, the sequence will show a multifractal feature, otherwise, it will show a single fractal feature. As can be seen from **Figure 2B**, when the value of q changes from -10 to 10, the return rate series $h(q)$ decreases from 0.7932 to 0.2748, indicating that the return rate series of China's crude oil futures market has obvious multifractal features. Specifically, when the order q is a large positive number, it reflects the behavioral information of large

TABLE 2 | Multifractal parameters Δh and $\Delta\alpha$.

	Max	min	Δ
h	0.7932	0.2748	0.5184
α	–0.1142	–0.8134	0.6993

fluctuation components of the price series. In this case, $h(q) < 0.5$, which indicates that the large fluctuation presents anti-persistence characteristics and is more prone to trend changes. However, when the order number q is small or negative, the small fluctuation component of the price series is amplified, and at this point, $h(q) > 0.5$, indicating that the small fluctuation shows a certain degree of persistence. In addition, when $q = 2$, $h(q)$ at this time is the traditional Hurst index. According to the experimental results, $h(2) = 0.5320$, which is greater than 0.5, indicating that the market has long-term memory characteristics. Therefore, China's crude oil futures market has relatively obvious long-term memory characteristics.

③ It can also be seen from **Figure 2C** that there is an obvious nonlinear relationship between the Renyi index $\tau(q)$ (Multifractal scaling index) and q of the return rate series of China's crude oil futures market; the image is presented as an increasing convex function, which further verifies the multifractal features of the series.

④ **Figure 2D** shows the multifractal spectrum of the sequence. It can be seen from the figure that the multifractal spectrum changes with α, showing an obvious arch shape, and the values of α are between –0.8134 and –0.1142, indicating the existence of multifractal features in this sequence.

The above analysis of the generalized Hurst index and multifractal spectrum is only a direct and qualitative analysis on multifractal features. On this basis, we also need to carry out quantitative analysis to accurately describe the multifractal degree. Because the multifractal parameters Δh and $\Delta\alpha$ can reveal the fluctuating state of the series, and measure the intensity of the multifractal features of the series, we use these two indicators to quantify the multifractal degree of the trading rate series of China's crude oil futures market. The calculation formulas of Δh and $\Delta\alpha$ are as follows:

$$\boldsymbol{\Delta h = \max[h(q)] - min[h(q)]} \tag{1}$$

$$\boldsymbol{\Delta\alpha = \max[\alpha] - min[\alpha]} \tag{2}$$

Since Δh can be used to reflect the fluctuation mode and relative amplitude of the series, and $\Delta\alpha$ represents the dispersion degree of the trend distribution of the financial time series, they can be used to measure the absolute range of the series fluctuation. As can be seen from **Table 2**, the values of the multifractal parameters Δh and $\Delta\alpha$ are 0.5184 and 0.6993, respectively, indicating that the relative as well as the absolute amplitude of the fluctuation change of this series is large, that is, the multifractal degree of the series is large.

RISK IDENTIFICATION AND MEASUREMENT OF CHINA'S CRUDE OIL FUTURES MARKET

Risk Identification of China's Crude Oil Futures Market Based on Fractal Characteristics

The price fluctuation of China's crude oil futures market has obvious multifractal characteristics. On this basis, this paper divides the whole sample data into daily trading data and calculates the daily multifractal spectrum parameters, so as to effectively identify the daily risk pattern of the market. In order to make the research more rigorous, this paper first analyzes the multifractal features of each trading-day series with the EMD-MFDFA by selecting a trading day at random, and the results are shown in **Figure 3**.

Taking the 5-min high-frequency trading data on 24 May 2018 as an example, the double logarithm graph and multifractal spectrum are drawn. It can be seen from **Figure 3A** that the daily return rate series of China's crude oil futures market has obvious power-law relationship under different q values, that is, it has multifractal features. In addition, the multifractal spectrum, **Figure 3B**, also shows an obvious arch shape, which is consistent with the overall multifractal results. It should be noted that other trading days have similar performance. Therefore, we find that the daily price fluctuation of China's crude oil futures also has multifractal features. It is worth mentioning that the multifractal parameters are calculated based on the 5-min high-frequency data of the day's trading, so they can cover most of the trading information of the day. Compared with the return rate corresponding to the daily closing price, the risk state defined by the multifractal parameters is more real and reliable. Therefore, this paper further analyzes the daily multifractal spectrum parameters.

The definition of $\Delta\alpha$, the width of the fractal spectrum, has been given above, and the corresponding parameter Δf is also defined. According to the partition function method, α_{min} and α_{max} represent the minimum probability measure and the maximum probability measure respectively. The larger $\Delta\alpha$ is, the wider the multifractal spectrum is, indicating that the price distribution of the day is more uneven and the absolute range of fluctuation is greater. Due to the same probability measure of α_{min} and α_{max}, there exist corresponding parameters $f(\alpha_{min})$ and $f(\alpha_{\max})$. $f(\alpha_{min})$ represents the possibility that the sequence trend is above the average, and $f(\alpha_{\max})$ represents the possibility that the sequence trend is below the average, so $\Delta f = f(\alpha_{min}) - f(\alpha_{\max})$ can be used to measure the uniformity and complexity of the sequence in a certain period of time. Since Δf has its own sign, when $\Delta f > 0$, it indicates that the price stays above the average for a long time, and investors believe that the price trend is good; otherwise, when $\Delta f < 0$, prices are below the average, investors perceive the market as weak. Generally speaking, the larger the absolute value of Δf is, the more uneven the time series distribution is and the more complex the fluctuation state is.

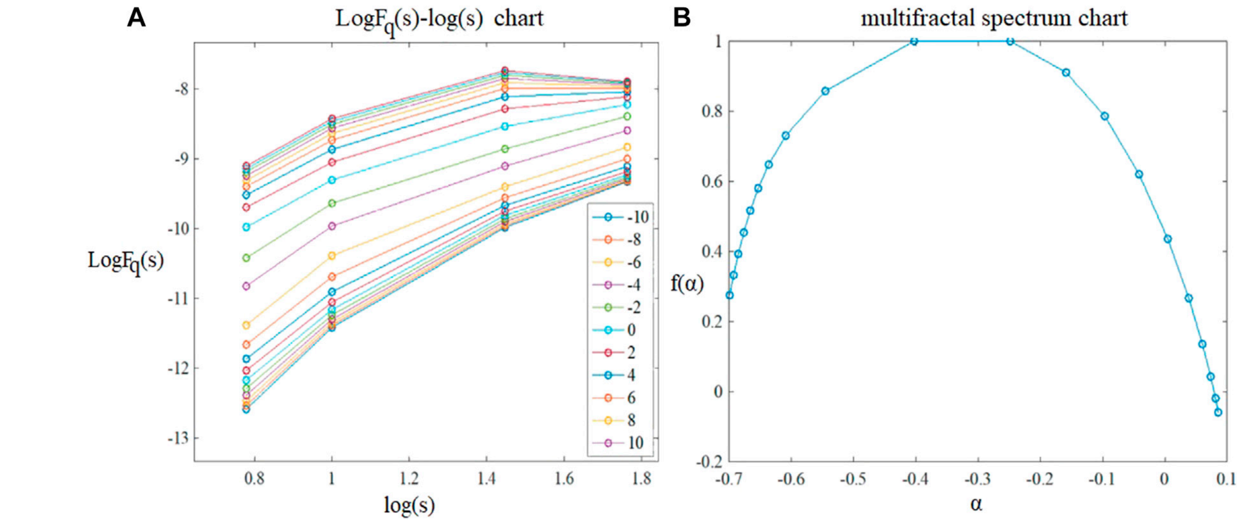


FIGURE 3 | Multifractal analysis of the return rate of China's crude oil futures market under day granularity (taking 2018/05/24 as an example).

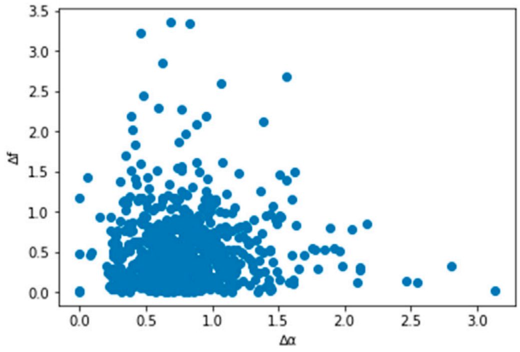


FIGURE 4 | Scatter diagram of $\Delta\alpha$ ~ Δf.

To sum up, $\Delta\alpha$ can be used to measure the absolute amplitude of price fluctuation in a day, and Δf can be used to measure the relative trend height and complexity of price fluctuation. Therefore, this paper will further analyze the daily multifractal characteristics of the return rate series of China's crude oil futures, so as to provide data support for accurately defining the normal state and risk state of the market. After calculating the multifractal spectrum parameters of each trading day, the scatter diagram is drawn, as shown in **Figure 4**.

Obviously, the data distribution in the lower left corner of the figure is relatively concentrated. In combination with the above theoretical analysis, it can be seen that the larger the values of $\Delta\alpha$ and Δf are, the greater the fluctuation of the sequence is and the higher the complexity of the fluctuation is, and vice versa. Therefore, the sample points in the lower left corner of **Figure 4** indicate that the market is in a normal state on the trading day. In order to make the identification of market daily risk status more accurate, this paper introduces the unsupervised clustering algorithm, without setting the threshold value for $\Delta\alpha$ and Δf, the impact of artificial random interference on risk identification is avoided.

In this paper, the K-means clustering and the Gaussian Mixture Model (GMM) are used to cluster the parameters $\Delta\alpha$ and Δf calculated above.

In short, the Gaussian Mixture Model (GMM) can be regarded as an optimization of the K-means algorithm. It is not only a kind of technical means commonly used in industry, but also belongs to a generation model. The GMM is to mix the probability distribution of multi-dimensional Gaussian model, so as to fit different sample data sets, so it has strong generalization ability and good fitting effect. In the K-means algorithm, the probability that the sample belongs to each cluster is qualitative, only "yes" or "no," and the corresponding probability value cannot be output. The GMM method, on the other hand, gives the probability of these sample data points being assigned to each cluster, and it can assign samples to different clusters according to artificial threshold values. Therefore, the information obtained by the GMM method is more. **Figure 5** shows the risk pattern recognition results with K-means clustering and GMM clustering algorithms for China's crude oil futures market. It is also obvious from the clustering results in the figure that the results gathered by the GMM are more accurate and more in line with the actual situation of the market. Therefore, this paper uses the GMM algorithm to identify the risks of China's crude oil futures market and defines the market risk status into two categories: the normal status and the risk status, providing a label basis for subsequent risk prediction model.

Selection of Risk Feature Indicators

After obtaining the risk status indicator variables of China's crude oil futures market, it is also necessary to select appropriate feature indicator variables for the market risk prediction model. Since there are many factors that affect market volatility, in order to get as much information as possible, this paper selects the risk feature

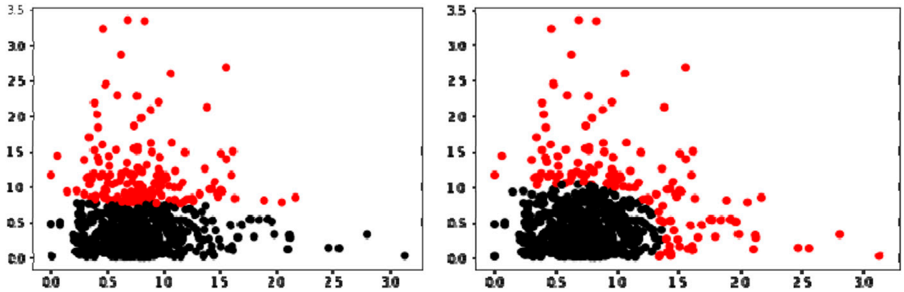


FIGURE 5 | Risk identification of China's crude oil futures market by K-means clustering (left) and GMM clustering (right).

TABLE 3 | Risk feature indicators.

Basic indicators	**Indicators explanation**
open	Opening price
high	Highest price
low	Lowest price
close	Closing price
volume	Trading volume
settle	Settlement price
pre_settle	Pre settlement price
return	Logarithmic rate of return
Technical indicators	Indicators explanation
MA	Moving average
MACD	Moving average convergence divergence
SAR	Stop and reverse indicator
BOP	Balance of power indicator
ATR	Average true range indicator
MFI	Money flow index
MOM	Momentum index
KDJ	Stochastic oscillator indicators
ROCP	Return of capital indicator
CCI	Commodity Channel Index
RSI	Relative strength index
OBV	On balance volume
WILLR	Williams %R

indicators from two aspects: basic indicators and technical indicators.

To be specific, This paper selects eight basic indicators (open, high, low, close, volume, settle, pre_settle, return) and 16 technical indicators (MA5, MA10, MACD, SAR, BOP, ATR, MFI, MOM, K, D, J, ROCP, CCI, RSI, OBV, WILLR) as the eigenvectors of the prediction model. Among them, most of the technical indicators in this paper are calculated from the quantified transaction package Ta-Lib in Python. The basic meanings of indicators are shown in **Table 3**.

Data Processing

Through the above analysis, this paper transforms and processes eight basic indicators to calculate 17 technical indicators, obtaining the feature indicator variables of China's crude oil futures market in each trading day from 26 March 2018 to 1 March 2021; then, this paper combines the variables with the risk pattern recognition results (label index) in **Section 4.2** to form a sample data set of risk prediction model. The feature indicators and the label indicators can be expressed as $x_t^{(i)}$ and y_t respectively. Specifically, $x_t^{(i)}$ is the i-th feature indicator corresponding to trading day t; y_t indicates the risk status indicator corresponding to the t-th trading day, and its value is 0 or 1 (where 0 indicates that the market is in a normal state and 1 indicates that the market is in a risk state). Therefore, the feature indicator variables and the status indicator variables constitute the sample point $(x_t^{(i)}, y_t)$ of this paper. And because this paper is to predict the risk of China's crude oil futures market, that is, to predict the status indicator variables of the next moment through the feature indicator variables of the current moment, then the sample data set used in the prediction model is $(x_t^{(i)}, y_{t+1})$.

Because the selected feature indicators have different orders of magnitude, if they are not processed, the information extraction of the data will be incomplete, and the effect of the model will also be greatly affected. In order to narrow the magnitude gap among feature data and improve the accuracy of model prediction, this paper adopts the Min-Max method to normalize the sample feature data, that is, to make linear changes to the original feature data so that the processed data results can be mapped to a unified interval. The specific formula is as follows:

$$\boldsymbol{x}^{(i)'} = \frac{\boldsymbol{x}^{(i)} - \mathbf{min}\,(\boldsymbol{x}^{(i)})}{\mathbf{max}\,(\boldsymbol{x}^{(i)}) - \mathbf{min}\,(\boldsymbol{x}^{(i)})} \tag{3}$$

After data normalization, the prediction accuracy and convergence speed of the model can be improved.

Screening of Risk Feature Indicators

The prediction model is a complex model with multiple indicators. Only by accurately extracting the feature vectors that affect market risks can we make the risk prediction more accurate. It is worth noting that many technical indicators are calculated based on the basic indicators, so the feature indicators we select may have obvious correlation between each other, and the information contained in one indicator may be relatively similar with another. Therefore, choosing more indicators will not make the model better, but will reduce the learning efficiency and increase the time cost of the model. At the same time, there may be some unclassifiable feature indicators in the initial ones. Thus, in order to simplify the complexity of the model and

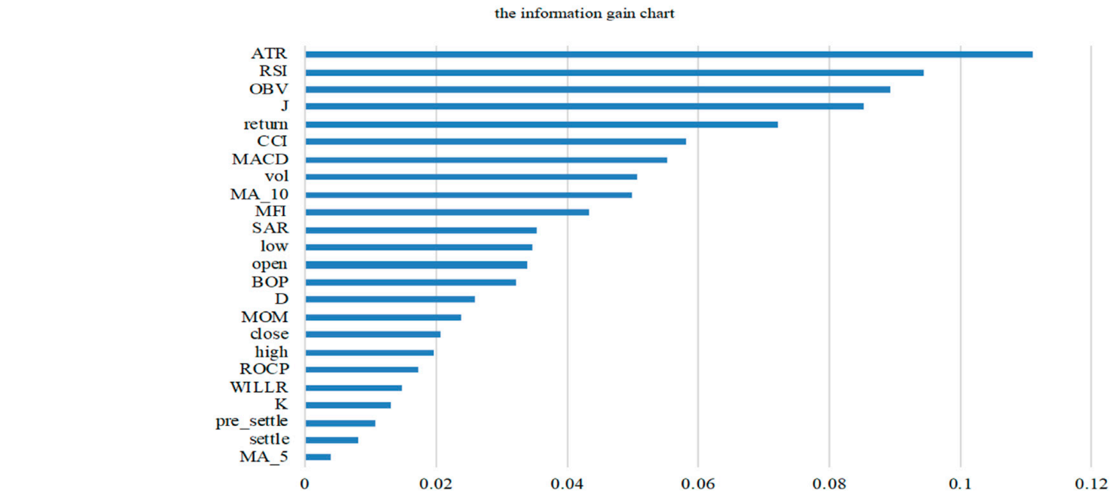


FIGURE 6 | The importance ranking of feature indicator variables based on information gain.

improve its prediction efficiency and accuracy, we need to further screen the selected initial feature indicators, so that the selected ones can contain the information of the majority of features, and achieve the effect of dimensionality reduction and denoising. Therefore, in order to extract the representative risk measurement indicators of China's crude oil futures market from the initial indicators, this paper adopts the decision tree algorithm and calls the feature_importances interface in the decision tree model to obtain the importance of the features. This method mainly measures whether a feature is important or not from two aspects: first, the total number of features split; Second, the total (average) information gain from features. The more the total number of feature splits or the greater the total (average) information gain, the higher the importance of the feature is, and vice versa. In this paper, information gain will be used to calculate the importance of each feature, and the results are shown as follows:

In the decision tree construction, the larger the information gain of a feature, the stronger the ability of classification, that is, the higher the importance of the feature. Therefore, we need to select features with large information gain from the original features as the feature indicator variables. As can be seen from **Figure 6**, the top 10 variables of information gain are ATR, RSI, OBV, J, return, CCI, MACD, vol, MA_10 and MFI, so this paper takes them as the feature indicators in the risk prediction model of China's crude oil futures market.

RISK PREDICTION OF CHINA'S CRUDE OIL FUTURES MARKET

Risk Prediction Evaluation Criteria

After the above data processing, this paper obtained a complete data set for risk prediction, including 10 characteristic indicators and label indicators obtained by multi-fractal spectral parameter clustering. According to the statistics, among the 691 trading days included in the sample, 550 trading days are in the normal state and 141 trading days are in the risk status. The proportion of risk samples and normal samples is close to 1:4, so the samples are unbalanced. Therefore, the accuracy of classification can not be used as an evaluation criterion of the quality of the model, and some other evaluation criteria are necessary to measure the training ability and generalization ability of the classification model. Based on the confusion matrix, this paper calculates two comprehensive evaluation indexes as model evaluation criteria to solve the problem of sample imbalance in this paper. The specific meaning of confusion matrix is shown in **Table 4**:

TABLE 4 | The confusion matrix table.

	Actual minority class	Actual majority class
Actual minority class	True Positive (TP)	False Negative (FN)
Actual majority class	False Positive (FP)	True Negative (TN)

According to the results of the confusion matrix, the accuracy, precision, recall rate and specificity of the risk prediction model can be calculated. The specific meanings and formulas are as follows:

Accuracy: The proportion of all correctly predicted samples to the total number of samples.

$$Accuracy = \frac{\boldsymbol{TP + TN}}{\boldsymbol{TP + FP + TN + FN}} \tag{4}$$

Precision: the proportion of true minorities in all samples predicted to be minorities.

$$precision = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FP}} \tag{5}$$

Recall rate: The percentage of a sample that is actually a minority category that is predicted to be a minority category.

$$Recall = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \tag{6}$$

Specificity: a measure of how many samples that are actually in the majority class are correctly predicted to be majority.

$$Specificity = \frac{\boldsymbol{TN}}{\boldsymbol{TN+FP}} \tag{7}$$

To sum up, there is a trade-off between accuracy and recall rate, and the balance between the two means that we should try not to miss the majority class while capturing the minority. Therefore, in order to meet the above requirements, the harmonic mean of the two is calculated as a comprehensive index and expressed by F1. According to the characteristics of the harmonic mean which tends to favor the index with a smaller value, when the accuracy and recall rate are both large, the closer the value of F1 is to 1, the better the classification effect of the model. The specific formula of F1 is as follows:

$$F1 = \frac{2}{\frac{1}{\boldsymbol{Precision}} + \frac{1}{\boldsymbol{Recall}}} = \frac{2*\boldsymbol{Precision}*\boldsymbol{Recall}}{\boldsymbol{Precision} + \boldsymbol{Recall}} \tag{8}$$

In addition, according to the calculation formulas of recall rate and specificity, recall rate can be used to measure the classification accuracy of the minority class, while specificity can represent the classification accuracy of most classes. Similarly, in order to take both recall rate and specificity into account, the geometric mean of both are constructed as a comprehensive evaluation index G, that is, only when both recall rate and specificity are high, the corresponding G value will be relatively ideal.

$$G = \sqrt{\boldsymbol{Recall}*\boldsymbol{Specificity}} \tag{9}$$

To sum up, F1 and G, the two comprehensive evaluation indexes, can be used to measure the prediction ability of the model for samples of the minority class and the comprehensive prediction ability for two classes of samples, respectively. The larger the F1 is, the better the prediction ability of the model is in predicting the minority class samples, and vice versa. If G is large, it indicates that the model has high accuracy in predicting both classes of samples. Therefore, this paper measures the effect of the classification model of unbalanced samples by using two comprehensive evaluation indexes, F1 and G, which are calculated from the confusion matrix.

Selection of Risk Prediction Methods and Comparison of Prediction Results

Based on the sample data set constructed by the feature indicator variables and the label indicator variables constructed above, and considering the advantages of the support vector machine (SVM) model in dealing with such problems, this paper firstly uses the SVM model to forecast the risks of China's crude oil futures market. The empirical process is completed in Python, mainly using Numpy, Pandas, Sklearn and other libraries. At the same time, in order to make the experimental prediction results more accurate, this paper also uses the five-fold cross validation method, and adopts the StratifiedKFold sampling method when dividing the training set and the test set to ensure that the proportion of normal samples and risk samples in the training set and the test set is consistent with the original data set. In the empirical study, the function SVC in Sklearn library, which is used to classify support vectors, is used to process the sample data in this paper. Considering the imbalance of samples in this paper, the class_weight parameter in the SVC function is set to balanced to make the results of the model more accurate.

After empirical adjustment, the values of F1 and G are 0.1356 and 0.1387, respectively, both of which are relatively small, indicating that the prediction ability of the model is poor. Although the class_weight parameter has been processed, the decision hyperplane of SVM will still automatically bias to the minority class when processing asymmetric data sets, which will result in weak prediction ability of the model and failure to accurately identify the risk samples in this paper. Therefore, twin support vector machine (TWSVM) is introduced in this paper on the basis of SVM. One decision hyperplane in SVM is extended into two decision hyperplanes, making each hyperplane close to the sample points of this class and far away from the sample points of the other class, so as to overcome the defect of SVM when dealing with the problem of sample imbalance.

The Twin Support Vector Machine (TWSVM) method was first proposed by Khemchandani and Chandra (2007) Its basic idea is similar to the traditional SVM algorithm. It transforms a large classification problem into two small classification problems, so that the constraints of each quadratic programming problem become half of the original. Specifically, two non-parallel decision hyperplanes are determined by solving two related SVM classification problems, and samples are classified according to the closest decision hyperplane of a given sample point. This improvement not only solves the error caused by the sample imbalance to some extent, but also improves the generalization ability and iteration speed of the model.

In order to make the prediction results of the two models comparable, the same training set and test set are also adopted in TWSVM, and the prediction results of SVM and TWSVM are compared, as shown in **Table 5**:

TABLE 5 | Comparison of prediction results of SVM and TWSVM.

Model	F1	G
SVM (RBF)	0.1356	0.1387
TWSVM (RBF)	0.3860	0.6425

It can be found from the results in **Table 5** that the F1 and G values of the TWSVM in the test set are significantly higher than those of traditional SVM model, that is, the prediction ability of the TWSVM model for samples of the minority class as well as the comprehensive class are better than that of SVM, indicating that TWSVM can effectively solve the problem of sample imbalance to some extent, and has high prediction accuracy.

TABLE 6 | Comparison of prediction results of XGBoost series models.

Model	F1	G
XGBoost	0.2752	0.2766
Weighted-XGBoost	0.4173	0.5935
Focal-XGBoost	0.4362	0.6473

Risk Prediction Algorithm Selection and Prediction Results Comparison

For the problem of sample imbalance, most existing studies start from the data set level and solve the sample imbalance by over-sampling and under-sampling. However, over-sampling will lead to the problem of over-fitting, and under-sampling will lose important information in the data, so they are not advisable. At the algorithm level, in addition to changing the decision-making ideas of the algorithm (such as the TWSVM method mentioned above), we can start from the loss function of the algorithm. Lin et al. (2017) introduced the Focal loss function and weighted loss functions on the basis of XGBoost, an extreme gradient lifting algorithm proposed by Chen Tianqi, and proposed an algorithm of Imbalance- XGBOOST for unbalanced samples. On this basis, Wang et al. (2020) derived the theory in detail, and verified that the method could effectively solve the problem of sample imbalance through practical application, and expanded the use scenarios of XGBoost. For the convenience of understanding, two loss functions used in the improved algorithm are listed. It should be noted that since this paper is aimed at classification problems, the activation functions are all sigmoid functions.

For the focal loss function:

$$\boldsymbol{L_{focal}} = -\sum_{i=1}^{m}\left[\boldsymbol{y_i}(1-\hat{\boldsymbol{y}}_i)^{\gamma}\log(\hat{\boldsymbol{y}}_i) + \hat{\boldsymbol{y}}_i^{\gamma}(1-\boldsymbol{y}_i)log(1-\hat{\boldsymbol{y}}_i)\right] \quad (10)$$

For the weighted loss function:

$$\boldsymbol{L_{weighted}} = -\sum_{i=1}^{m}\left[\boldsymbol{\alpha y_i}\log(\hat{\boldsymbol{y}}_i) + (1-\boldsymbol{y}_i)log(1-\hat{\boldsymbol{y}}_i)\right] \quad (11)$$

Where y_i is the actual label; $\hat{y}_i = \frac{1}{1+\exp(-z_i)}$, α, γ are parameters.

In the empirical study of this section, this paper mainly calls the integrated libraries such as Sklearn and Imbalance-XGboost in Python to predict the risks of China's crude oil futures market. Similarly, the samples used in this section are the same as those in the previous section, and the training set and the test set are also the same. When adjusting the parameters of the model, GridSearch is used to optimize the parameters of the above loss function within the range of (Altman, 1967; Li et al., 2019), and the optimal parameters ($\alpha = 3$, $\gamma = 1.5$) are returned through the best_estimator_ interface. Further, we compared the values of F1 and G of XGBoost and its improved models under the optimal parameters, and the results are shown in **Table 6**.

According to the empirical results, both F1 and G values of the original XGBoost are low, indicating that the non-equilibrium samples have a great impact on the prediction effect of the XGBoost algorithm. After the improvement of its loss function, the values of F1 and G are significantly improved, and when the focal loss function is used, the F1 and G of Focal-XGBoost are the best, indicating that Focal-XGBoost could effectively solve the problem of sample imbalance existing in this paper and improve the prediction accuracy of the model.

CONCLUSION

This paper takes the return rate series of China's crude oil futures market as the research object, and uses the EMD-MFDFA method to study the multifractal characteristics based on 5-min high-frequency trading data. At the same time, the multifractal analysis is carried out on 111 trading data generated in each trading day, and the calculated daily multifractal spectral parameters are used to analyze the risk status of each trading day. The unsupervised clustering algorithms K-means and Gaussian Mixture Model (GMM) are further used to cluster the obtained spectral parameters. Each trading day is identified as the risk status or the normal status, and the identified risk status is used as the label data and combined with the corresponding technical indicators. SVM, XGBoost and their improved algorithms are used to predict the risks of China's crude oil futures market, Based on the calculation results of confusion matrix, the prediction effects of each model are compared, and the optimal model is selected to predict the risks of China's crude oil futures market, so that relevant investors can identify potential risks in advance and formulate prevention and control measures in time. The following conclusions are drawn:

① There are obvious multifractal characteristics in the return rate series of both China's crude oil futures market and its single trading day, and the calculated daily multifractal parameters can effectively show the fluctuation of the series.
② Due to the imbalance of sample data, twin support vector machine (TWSVM) model has better prediction ability than the traditional support vector machine (SVM) model for the risk prediction of China's crude oil futures market.
③ The XGBoost algorithm has a great impact on the risk prediction, and the Focal-XGBoost is better for China's crude oil market risk prediction.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

Conceptualization, YG, SZ, and YL; Data curation, SZ and YL; Formal analysis, YG, SZ, and YL; Methodology, YG, SZ, and YL; Software, SZ; Validation, YG, SZ, and YL; Writing–original draft, YG, SZ, and YL; Writing–review and editing, YG, SZ, and YL.

FUNDING

This research was funded by Natural Science Foundation of Hunan Province of China (2021JJ30175).

REFERENCES

Altman, E. I. (1967). *The Prediction of Corporate Bankruptcy: A Discriminant Analysis*. Los Angeles: University of California.

Dong, H., Liu, Y., Liu, Y., and Chang, J. (2019). The Heterogeneous Linkage of Economic Policy Uncertainty and Oil Return Risks. *Green Financ.* 1, 46–66. doi:10.3934/gf.2019.1.46

Feng, Y. S., and Cao, B. M. (2022). Multifractal Fluctuation Analysis of Correlations between Agricultural Futures Markets in China and the US Based on MF-X-DFA and MF-DPXA Methods[J]. *Fluctuation Noise Lett.* 21 (01). doi:10.1142/s0219477522500067

Ji, Q., and Zhang, D. (2019). China's Crude Oil Futures: Introduction and Some Stylized Facts. *Finance Res. Lett.* 28, 376–380. doi:10.1016/j.frl.2018.06.005

Jiang, Z. Q., and Zhou, W. X. (2011). Multifractal Detrending Moving-Average Cross-Correlation Analysis. *Phys. Rev. E Stat. Nonlin Soft Matter Phys.* 84, 016106. doi:10.1103/PhysRevE.84.016106

Kantelhardt, Jan W., Zschiegner, Stephan A., Koscielny-Bunde, Eva, Havlin, Shlomo, Bunde, Armin, and Eugene Stanley, H. (2002). Multifractal Detrended Fluctuation Analysis of Nonstationary Time Series. *Phys. A Stat. Mech. its Appl.* 316 (1). doi:10.1016/s0378-4371(02)01383-3

Khemchandani, R., and Chandra, S. (2007). Twin Support Vector Machines for Pattern Classification. *IEEE Trans. Pattern Anal. Mach. Intell.* 29 (5), 905–910.

Latunde, T., Akinola, L. S., Shina Akinola, L., and Deborah Dare, D. (2020). Analysis of Capital Asset Pricing Model on Deutsche Bank Energy Commodity. *Green Financ.* 2 (1), 20–34. doi:10.3934/gf.2020002

Li, J. M., Wei, H. J., Wei, L. D., Zhou, D. P., and Qiu, Y. (2020). Extraction of Frictional Vibration Features with Multifractal Detrended Fluctuation Analysis and Friction State Recognition. *Symmetry-Basel* 12 (2), 22. doi:10.3390/sym12020272

Li, Q., and Xie, W. (2013). Classification of Aircraft Targets with Low-Resolution Radars Based on Multifractal Spectrum Features. *J. Electromagn. Waves Appl.* 27 (16), 2090–2100. doi:10.1080/09205071.2013.832394

Li, S., and Quan, Y. (2019). Financial Risk Prediction for Listed Companies Using IPSO-BP Neural Network. *Int. J. Perform. Eng.* 15 (4), 1209. doi:10.23940/ijpe.19.04.p16.12091219

Li, X., Shang, W., and Wang, S. (2019). Text-based Crude Oil Price Forecasting: A Deep Learning Approach. *Int. J. Forecast.* 35 (4), 1548–1560. doi:10.1016/j.ijforecast.2018.07.006

Li, Z. H., Dong, H., Floros, C., Charemis, A., and Failler, P. (2021). Re-examining Bitcoin Volatility: A CAViaR-Based Approach. *Emerg. Mark. Financ. Trade*, 1–19. doi:10.1080/1540496x.2021.1873127

Li, Z., Wang, Y., and Huang, Z. (2020). Risk Connectedness Heterogeneity in the Cryptocurrency Markets. *Front. Phys.* 8, 243. doi:10.3389/fphy.2020.00243

Lin, T. Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). "Focal Loss for Dense Object Detection," in Proceedings of the IEEE International Conference on Computer Vision, 2980–2988.

Liu, K., Luo, C., Luo, C., and Li, Z. (2019). Investigating the Risk Spillover from Crude Oil Market to BRICS Stock Markets Based on Copula-POT-CoVaR Models. *Quantitative Finance Econ.* 3 (4), 754–771. doi:10.3934/qfe.2019.4.754

Liu, K., Luo, C. Q., Luo, C., and Li, Z. (2019). Investigating the Risk Spillover from Crude Oil Market to BRICS Stock Markets Based on Copula-POT-CoVaR Models. *Quant. Financ. Econ.* 3, 754–771. doi:10.3934/qfe.2019.4.754

Lo, W. C . (1989). Long-term Memory in Stock Market Prices. *Work. Pap.* doi:10.3386/w2984

Ma, X., and Lv, S. (2019). Financial Credit Risk Prediction in Internet Finance Driven by Machine Learning. *Neural Comput. Applic* 31 (12), 8359–8367. doi:10.1007/s00521-018-3963-6

Mandelbrot, B. B., and Wheeler, J. A. (1983). The Fractal Geometry of Nature. *Am. J. Phys.* 51 (3). doi:10.1119/1.13295

Özdurak, C. (2021). Nexus between Crude Oil Prices, Clean Energy Investments, Technology Companies and Energy Democracy. *Gf* 3 (3), 337–350. doi:10.3934/gf.2021017

Peng, C. K., Buldyrev, S. V., Havlin, S., Simons, M., Stanley, H. E., and Goldberger, A. L. (1994). Mosaic Organization of DNA Nucleotides. *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.* 49 (2), 1685–1689. doi:10.1103/physreve.49.1685

Peters, E. E . (1994). *Fractal Market Analysis : Applying Chaos Theory to Investment and Economics*. John Wiley & Sons. Vol. 24.

Peters, E. E. (1996). *Chaos and Order in the Capital Markets: A New View of Cycles, Prices,and Market Volatility*. John Wiley & Sons.

Peters, E. E. (1994). *Fractal Market Analysis: Applying Chaos Theory to Investment and Economics*. John Wiley & Sons.

Podobnik, B., Horvatic, D., Petersen, A. M., and Stanley, H. E. (2009). Cross-correlations between Volume Change and Price Change. *Proc. Natl. Acad. Sci. U. S. A.* 106 (52), 22079–22084. doi:10.1073/pnas.0911983106

Podobnik, B., and Stanley, H. E. (2008). Detrended Cross-Correlation Analysis: a New Method for Analyzing Two Nonstationary Time Series. *Phys. Rev. Lett.* 100 (8), 084102. doi:10.1103/PhysRevLett.100.084102

Ruan, Q., Jiang, W., and Ma, G. (2016). Cross-correlations between Price and Volume in Chinese Gold Markets. *Phys. A Stat. Mech. its Appl.*, 451. doi:10.1016/j.physa.2015.12.164

Sun, H. G., and Li, W. H. (2018). "Analysis of the Fluctuation of Chinese Crude Oil Futures- Based on GARCH-type Model," in *Proceedings of the 2018 3rd International Conference on Modelling, Simulation and Applied Mathematics*. Editors A. LuevanosRojas, G. Ilewicz, D. J. Jakobczak, and K. Weller (Paris: Atlantis Press), 160, 110–112. doi:10.2991/msam-18.2018.25

Sun, X., Chen, H. P., Wu, Z. Q., and Yuan, Y. Z. (2001). Multifractal Analysis of Hang Seng Index in Hong Kong Stock Market. *Phys. A* 291 (1-4), 553–562. doi:10.1016/s0378-4371(00)00606-3

Tam, K. (1991). Neural Network Models and the Prediction of Bank Bankruptcy. *Omega* 19 (5), 429–445. doi:10.1016/0305-0483(91)90060-7

Tam, K. Y., and Kiang, M. (1990). Predicting Bank Failures: A Neural Network Approach. *Appl. Artif. Intell.* 4 (4), 265–282. doi:10.1080/08839519008927951

Uthayakumar, J., Metawa, N., Shankar, K., and Lakshmanaprabu, S. K. (2020). Intelligent Hybrid Model for Financial Crisis Prediction Using Machine Learning Techniques. *Inf. Syst. E-Bus Manage* 18 (4), 617–645. doi:10.1007/s10257-018-0388-9

Wang, C., Deng, C., and Wang, S. (2020). Imbalance-XGBoost: Leveraging Weighted and Focal Losses for Binary Label-Imbalanced Classification with XGBoost. *Pattern Recognit. Lett.* 136, 190–197.

Wang, J., Shang, P., and Weijie, G. E. (2012). Multifractal Cross-Correlation Analysis Based on Statistical Moments. *Fractals* 20. doi:10.1142/s0218348x12500259

Wang, Y., Wei, Y., and Wu, C. (2011). Detrended Fluctuation Analysis on Spot and Futures Markets of West Texas Intermediate Crude Oil. *Phys. A Stat. Mech. its Appl.* 390 (5), 864–875. doi:10.1016/j.physa.2010.11.017

Wei, Y., and Huang, D. S. (2005). Multifractal Analysis of SSEC in Chinese Stock Market: A Different Empirical Result from Heng Seng Index. *Phys. A* 355 (2-4), 497–508. doi:10.1016/j.physa.2005.03.027

Weng, F., Zhang, H., and Yang, C. (2021). Volatility Forecasting of Crude Oil Futures Based on a Genetic Algorithm Regularization Online Extreme Learning Machine with a Forgetting Factor: The Role of News during the COVID-19 Pandemic. *Resour. Policy* 73, 102148. doi:10.1016/j.resourpol.2021.102148

Yuan, Y., Zhuang, X.-t., and Jin, X. (2009). Measuring Multifractality of Stock Price Fluctuation Using Multifractal Detrended Fluctuation Analysis. *Phys. A Stat. Mech. its Appl.* 388 (11), 2189–2197. doi:10.1016/j.physa.2009.02.026

Zhang, X., Yang, L., and Zhu, Y. (2019). Analysis of Multifractal Characterization of Bitcoin Market Based on Multifractal Detrended Fluctuation Analysis. *Phys. A Stat. Mech. its Appl.*, 523. doi:10.1016/j.physa.2019.04.149

Zhao, D. D., Ding, J. C., and Chai, S. C. (2018). Systemic Financial Risk Prediction Using Least Squares Support Vector Machines. *Mod. Phys. Lett. B* 32 (17), 15. doi:10.1142/s021798491850183x

Zhu, H., and Zhang, W. (2018). Multifractal Property of Chinese Stock Market in the CSI 800 Index Based on MF-DFA Approach. *Phys. A Stat. Mech. its Appl.* 490, 497–503. doi:10.1016/j.physa.2017.08.060

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership