



Research article

A two-stage deep learning-based hybrid model for daily energy consumption forecasting with variational mode decomposition and wavelet thresholding denoising

Bowen Hou^{1,2}, Yin Liu^{1,3}, Jiajun Chen⁴, Letian Qi⁵, Jinhao Zhang¹ and Honghao Zhao^{1,*}

¹ Department of Decision Sciences, School of Business, Macau University of Science and Technology, Macao 999078, China

² China Nanguang Group Co., Ltd.: Zhuhai, Guangdong-Macao In-Depth Cooperation Zone, 519000, China

³ School of Electronic Engineering, Huainan Normal University, Huainan 232038, China

⁴ Biomedical Engineering, The Hong Kong Polytechnic University, Hong Kong 999077, China

⁵ School of Computer Science and Engineering, Faculty of Innovation Engineering, Macau University of Science and Technology, Macao 999078, China

* **Correspondence:** Email: hhzhao@must.edu.mo; Tel: +85388972937.

Abstract: Global energy demand keeps rising amid economic expansion and population growth. Although renewables like wind and solar have grown fast, fossil fuels (especially petroleum) remain central to the global energy mix, underpinning industry, households, policy, and stability. This article presents an improved hybrid forecasting framework, denoted as the VMD-WTD-GRU model with mode-specific denoising, for daily average petroleum consumption, which adopts a two-stage decomposition-targeted denoising strategy. The framework combines the models of variational mode decomposition, wavelet threshold denoising, and the gated recurrent unit networks for petroleum energy forecasting, where different models are designed for time-series data decomposition, noise elimination, sub-series forecasting, and integration, respectively. In general, the proposed hybrid model has achieved excellent predictive performance on multiple evaluation indicators when compared with several candidate models, especially demonstrating that the mode-specific denoising strategy can effectively reduce noise interference while preserving useful trend information in low-frequency components. Such a framework can be easily implemented to forecast petroleum consumption in real application scenarios.

Keywords: daily energy consumption prediction; deep learning; two-stage hybrid model; variational mode decomposition; wavelet thresholding denoising

1. Introduction

At present, the global energy landscape is changing rapidly. As the problem of global climate change becomes increasingly severe, countries have been increasing their efforts to develop and use renewable energy, and the global energy consumption structure has been gradually optimized. Although renewable new energy such as solar power, wind power, and hydropower have developed quickly, petroleum, natural gas, and coal are still important supports of the global energy system. Their consumption accounts for more than 80% of total primary energy consumption [1]. At the same time, international energy cooperation and competition have become increasingly intense. The geopolitical pattern has also become more complex and changeable. In the international energy market, unstable factors such as petroleum price fluctuations and supply chain risks have increased, and the global distribution of energy remains uneven. In particular, crude petroleum demand in Asia has continued to rise, and the supply–demand gap has kept widening due to limits in production capacity and resource endowment. By contrast, the Atlantic Basin has faced the problem of excess production capacity. The global energy market's stability has been considerably impacted by these factors. Therefore, accurate forecasting of petroleum consumption has important practical significance. For government departments, energy forecasting can help ensure energy security and support the formulation of reasonable energy policies. For enterprises, it can provide a basis for production planning and procurement decisions. For people's livelihoods, it can help reduce the economic impact caused by fluctuations in energy prices.

In traditional energy consumption forecasting research, conventional time-series approaches including autoregressive integrated moving average (ARIMA), vector autoregression (VAR), and generalized autoregressive conditional heteroskedasticity (GARCH) are widely used, with notable advantages in handling stationary data. However, petroleum consumption time-series data usually contain problems such as nonlinearity and nonstationarity. Traditional time-series methods such as ARIMA generally require the mean and variance of the sequence to remain stable, and this requirement restricts their forecasting performance in practical applications. This limitation was especially evident during the COVID-19 pandemic, when global petroleum demand fell by about 9%, and the subsequent recovery speed differed greatly across regions. As a result, traditional models found it difficult to accurately analyze such structural changes. Although the GARCH model can describe volatility clustering to some extent, it still has limitations in analyzing market behavior when multiple factors, such as geopolitical conflicts and OPEC production adjustments, act at the same time [2].

Machine learning and deep learning have gained wide attention and have been applied in energy forecasting in the past few years. In particular, long short-term memory networks (LSTMs) can extract and learn changing trends and feature information from different periods of the data, thereby producing more accurate forecasting outputs. Plenty of existing studies have demonstrated the forecasting ability of LSTMs in different energy scenarios. For example, Zhang et al. [3] developed an LSTM network-based forecasting model, which was applied to predict the evolution of groundwater level in regions with intensive agricultural production. The model effectively captured hydrological cycle patterns over a multiyear scale, and its forecasting accuracy was 15% higher than that of ARIMA. Huang et al. [4] found that LSTM networks could control the output forecasting error of wind farms within 5% by modeling the interaction between wind turbine degradation processes and meteorological patterns.

Deng et al. [5] found that an LSTM-based forecasting framework could reduce the forecasting error of industrial electricity consumption by 33%.

However, the forecasting performance of directly applying deep learning models remains limited. This is mainly because actual energy consumption series are often affected by seasonal cycles, long-term trends, and random shocks, which lead to nonlinear, nonstationary, and highly fluctuating characteristics, such as high-frequency noise and irregular fluctuations [6]. These factors may weaken the ability of deep learning models to extract effective temporal features and generate accurate forecasts. To address this problem, hybrid forecasting frameworks that combine data preprocessing techniques with learning models have been widely investigated in energy forecasting. For example, Sun et al. [7] combined wavelet packet decomposition (WPD) with optimized artificial neural networks (ANNs) for short-term wind power forecasting and achieved lower forecasting errors than a single model. In addition, variational mode decomposition (VMD)-based hybrid models have been applied to energy-related forecasting tasks, such as electricity consumption forecasting and residential load forecasting [8, 9]. More recently, several studies have further combined wavelet-based denoising or decomposition with VMD to improve the quality of forecasting inputs in photovoltaic power forecasting and short-term power load forecasting [10, 11]. These studies indicate that decomposition-and denoising-based hybrid frameworks can improve the feature representation of complex energy time series and enhance forecasting performance. Nevertheless, existing decomposition denoising studies still have two key limitations. First, the residual component generated after VMD decomposition is often either ignored or directly fed into the forecasting model without specific treatment. Moreover, most studies do not adopt differentiated strategies for high-frequency, low-frequency, and residual components but instead apply a uniform processing pipeline to all modes. Such an undifferentiated processing strategy may overlook the heterogeneous characteristics of different components, thereby limiting the forecasting accuracy of the hybrid model.

To address the above problems, this paper proposes a variational modal decomposition-wavelet threshold denoising-gated recurrent unit (VMD-WTD-GRU) hybrid model with mode-specific denoising for regional petroleum consumption forecasting. First, VMD is used to decompose the original petroleum consumption series into multiple intrinsic mode function (IMF) components and one residual component, thereby extracting temporal patterns at different frequency scales. Then, WTD is selectively applied to high-frequency IMF components and the residual component, and low-frequency components are retained to preserve the main trend information. After that, each processed component is separately input into a GRU model for forecasting, and the forecasting outputs of all components are accumulated to reconstruct the final predicted value. The main contributions of this paper are summarized as follows. (1) A VMD-WTD-GRU hybrid forecasting model with mode-specific denoising is proposed for nonstationary and highly fluctuating daily petroleum consumption forecasting. (2) Different from conventional full-component denoising strategies, the proposed model selectively denoises high-frequency and residual modes, retaining low-frequency components to preserve the main trend of petroleum consumption. (3) The effectiveness of the proposed mode-specific denoising strategy is evaluated through comparative experiments with single models, variational mode decomposition (VMD)-based hybrid models, and denoising-enhanced hybrid models. Statistical testing is further conducted to verify the significance of forecasting improvements. (4) The proposed model is applied to regional daily petroleum consumption forecasting, demonstrating the applicability of decomposition–denoising-based hybrid

forecasting methods in petroleum consumption prediction and providing methodological support for petroleum demand forecasting and regional energy management.

The rest of this paper is organized as follows. In Section 2, the literature review summarizes the advantages and limitations of various time series prediction models and signal decomposition techniques. Section 3 introduces the overall structure and implementation process of the proposed VMD-WTD-GRU hybrid model with mode-specific denoising. Section 4 presents the experimental analysis, including data sources, statistical analysis, evaluation metrics, and comparative analysis of model forecasting results. Section 5 summarizes the whole paper, and discusses its application effects in energy management and future research directions.

2. Literature review

2.1. Traditional methods for energy consumption forecasting

Reliable consumption forecasts play an important role in supporting energy planning, infrastructure investment, and long-term urban development. At present, many simple and easy-to-interpret time-series and regression models are frequently used for forecasting. Among them, the ARIMA model has performed reasonably well in fitting the linear trend. For instance, Pappas et al. [12] found that the ARIMA model was better than other time-series models for forecasting electricity use in Greece. In practice, energy consumption data often contain nonlinear changes, seasonal effects, and random noise. These features are difficult for ARIMA to fully represent, especially when the dataset is incomplete or uneven in quality, as may occur in regions such as Iran [13].

At the same time, many neural network algorithms have been applied to the problem of energy consumption forecasting to improve accuracy. Azadeh et al. [14] developed an ANN-based method for renewable energy consumption estimation, and it achieved an accuracy of 99%. New forms of ANN have also shown better results in short-term forecasting. Bai et al. [15] improved the backpropagation neural network by adding a forecasting lead time adjustment, and both the accuracy and speed of forecasting were significantly increased. Ruiz et al. [16] built an optimized Elman neural network model for the area of public building energy forecasting; it can handle nonlinear changes in energy use better than the conventional methods.

Fuzzy logic is also widely used for energy consumption forecasting. Although ANN can model nonlinear relations, they are more prone to overfitting and local minima convergence with small datasets. The adaptive neuro-fuzzy inference system (ANFIS) offers a useful way to link data-driven learning with fuzzy rule-based reasoning, making it suitable for energy forecasting problems where both nonlinear behavior and uncertainty are present. Hybrid models that combine linear and nonlinear structures have also been explored to better represent complex consumption patterns. Zhang [17] first proposed a two-stage hybrid framework to separate the linear (ARIMA) and nonlinear (ANN) parts of time series. Wang et al. [18] developed an ARIMA-ANN model to better describe the mixed structure of time-series data. The model first used ARIMA to extract the linear structure of the series, and then applied an ANN to learn the nonlinear residual information. The model achieved better results on different datasets, thereby improving the representation of complex temporal dynamics. Barak and Sadegh [19] used ANFIS-based hybrids to reduce the forecasting error of ARIMA by fitting the residual nonlinearity.

2.2. Deep learning models for energy consumption forecasting

With the increasing complexity of energy consumption data, deep learning models have also gradually become widely used as an important research direction in the field of energy consumption prediction. In particular, recurrent neural networks (RNNs), LSTM, GRU, and their improved models have become important methodological approaches for energy consumption forecasting. Shabbir et al. [20] proposed an algorithm based on RNNs to predict energy consumption and production in Estonia, and its root mean square error (RMSE) was reduced by 12.3% compared with conventional forecasting methods. Ma et al. [21] designed a bidirectional GRU and Mogrifier LSTM that can dynamically adjust the intrinsic mode function, which improves the stability of port power load forecasting by 17.6%. Evolutionary optimization techniques have further improved the network performance. Sharma et al. [22] developed a deep learning model based on causality-based deep learning model and proved that the LSTM model is superior to other methods. Shin and Woo [23] studied and compared the performance of random forest, xgboost, and LSTM in the prediction of energy consumption in Korea and found that LSTM had better accuracy.

Apart from the previously stated RNN, and LSTM models, other deep learning models have also been explored for energy consumption forecasting. Dedinec et al. [24] developed a load forecasting model based on deep belief networks (DBN), which effectively improved the model's forecasting performance. Dong et al. [25] combined K-nearest neighbors (KNN) with DBN-based forecasting. In their method, KNN was first used to identify historical load patterns similar to the target forecast period, and then DBN was used for forecasting. Alipour et al. [26] proposed a deep learning model that combines an autoencoder and a cascaded deep neural network (DNN) to improve the accuracy of power load forecasting. Wang et al. [27] applied an improved quasi-recurrent neural network (QRNN) model to photovoltaic power generation forecasting. Similarly, Van der Meer et al. [28] used a probabilistic forecasting method to predict net load under the influence of photovoltaic power generation. Mei et al. [29] used a short-term net load forecasting model based on deep neural networks and achieved high forecasting accuracy under different climatic conditions and high photovoltaic penetration scenarios. Although the deep learning model has certain advantages, it still faces challenges in the aspects of model architecture optimization, hyperparameter tuning, and computational efficiency.

Generative learning is also a branch of deep learning. Generative adversarial networks (GANs) were first proposed by Goodfellow et al. [30] and are one of the most influential generative models. GANs can capture the latent distribution of data and generate synthetic samples similar to real observations. Wang and Hong [31] pointed out that GAN-based methods can be applied to data augmentation in building energy consumption forecasting. Fekri et al. [32] proposed a recurrent GAN (R-GAN), which can further improve forecasting accuracy by generating data with similar characteristics. In general, these methods are particularly suitable for energy consumption forecasting scenarios with limited historical data or high data acquisition costs.

2.3. Hybrid models for energy consumption forecasting

As single models often fail to capture the complex patterns in energy consumption data, recent studies have increasingly turned to hybrid models. Hybrid models combine decomposition techniques, feature engineering, ensemble learning, deep learning, and other intelligent methods,

applying the strengths of different approaches to improve forecasting accuracy, stability, and generalization ability. In recent years, hybrid forecasting models based on advanced deep learning architectures, such as Transformer, GRU, bidirectional LSTM (BiLSTM), convolutional neural network-LSTM (CNN-LSTM), and attention mechanisms, have become an important methodological stream in energy forecasting. For example, Ranjbar and Rahimzadeh [33] proposed a Transformer-LSTM-CNN ensemble model for gasoline consumption forecasting, where the self-attention mechanism was used to capture long-range dependence, and the CNN and LSTM were used to extract local and sequential patterns. He et al. [34] combined wavelet transform with the CNN-BiLSTM hybrid model to preprocess petroleum and coal consumption data and separate multiscale trend features from periodic change patterns. The model was trained using preprocessed hybrid features from multiple energy datasets and achieved an accuracy of 95%. Zhai et al. [35] developed a VMD-subquadratic sparse attention-Transformer (VMD-SSA-Transformer) model for photovoltaic power forecasting, and Venkateswaran et al. [36] proposed a VMD-Informer-LSTM architecture for solar radiation forecasting. Liu et al. [37] compared different neural network models based on VMD and an optimization algorithm for air-conditioning energy consumption prediction and showed that GRU outperforms LSTM, although both can benefit from VMD preprocessing. These studies show that deep learning models have been increasingly incorporated into hybrid forecasting models to enhance temporal feature extraction and improve forecasting performance.

Currently, a mainstream optimization approach for hybrid models in energy consumption forecasting is to combine signal decomposition methods with deep learning models to handle nonstationary, highly fluctuating, and noisy data. In particular, VMD has been widely applied in energy forecasting research. For example, Xu et al. [8] combined VMD and iterative Boruta selection with CNN-BiLSTM, and Ruan et al. [38] demonstrated that VMD preprocessing helps LSTM better extract trend and seasonal components from electricity consumption data. Jrhilifa et al. [9] further applied VMD-based two-stage decomposition with LSTM and an autoencoder for residential load forecasting. In addition to a single-decomposition hybrid model wavelet-based denoising has been increasingly combined with VMD to improve the quality of forecasting inputs. One route applies wavelet denoising before VMD decomposition, aiming to reduce noise in the original sequence and improve the stability of subsequent modal decomposition. For example, Liu et al. [10] proposed a hybrid prediction model based on WTD-VMD and applied it to photovoltaic power forecasting. Experiments verified that the framework of first reducing noise and then decomposing can significantly reduce various error indicators. Yang et al. [39] developed a DWT-VMD-GRU model for monthly runoff forecasting, where DWT was first used for denoising, and VMD was then applied to decompose the processed series. Guo et al. [11] developed a complete ensemble empirical mode decomposition with adaptive noise-wavelet transform-VMD (CEEMDAN-WT-VMD) combined denoising framework for short-term power load forecasting. It outperformed the benchmark model by a large margin, lowering the RMSE by 65.5%. These studies show that denoising before decomposition can reduce the interference of noise and improve the stability of subsequent component extraction. Another route first uses VMD to decompose the original series and then applies wavelet denoising to the decomposed components. Yu et al. [40] submitted a research work on a VMD-wavelet denoising-LSTM (VMD-WD-LSTM) model for short-term traffic flow forecasting. They first used VMD to decompose the original traffic flow time series and subsequently employed the LSTM model to predict traffic flow based on the decomposed subsequences. Multiple comparative

experiments demonstrated the effectiveness and reliability of this model. Wang et al. [41] further extended this VMD-wavelet denoising-LSTM (VMD-WD-LSTM) structure to an ensemble learning framework for short-term traffic flow forecasting. Some recent studies have also integrated decomposition and denoising into more complex multistage frameworks, such as the CEEMDAN-VMD-GRU wind power forecasting model proposed by Fang et al. [42] and the enhanced successive VMD denoising model proposed by Jiang et al. [43]. These studies indicate that hybrid frameworks combining mode decomposition, denoising, and deep learning have become an important research direction for time series forecasting.

Ensemble learning has also attracted attention, as it can combine multiple learners to improve prediction reliability. It integrates the forecasting results of multiple base models and makes full use of the strengths of different models in feature extraction, trend fitting, and nonlinear modeling. These features not only reduce the risk of overfitting of a single model, but also improve the overall stability and generalization ability of the forecasting performance. Xiao et al. [44] proposed an ensemble model for predicting China's energy consumption, which dynamically weights the outputs of ARIMA, SVM, and artificial neural network (ANN) models and uses a genetic algorithm to optimize the model weights. This method reduces the mean absolute error (MAE) by 10% compared to a single model. Some researchers have also begun to develop ensemble models based on deep learning. Shojaei and Mokhtar [45] developed a stacked ensemble model that combines CNN, LSTM, and a gradient boosting tree. The CNN is used to extract spatial features from historical load data, the LSTM is used to model temporal dependencies, XGBoost is used to improve residuals, and Bayesian optimization is added to adjust hyperparameters. Compared with a single model, the RMSE is reduced by 14%. Ranjbar and Rahimzadeh [33] proposed a Transformer-LSTM-CNN ensemble model for predicting gasoline consumption. A self-attention mechanism enables the transformer to capture long-range dependencies, and the CNN and LSTM layers extract local temporal patterns. Compared to a single LSTM, the ensemble model achieves a 22% improvement in R^2 , indicating that a multicomponent ensemble also has certain performance advantages.

3. Description of methods

To improve the forecasting performance of complex and nonstationary petroleum consumption series, this study constructs a VMD-WTD-GRU hybrid forecasting framework with mode-specific denoising.

3.1. Variational mode decomposition

Variational mode decomposition is an adaptive, nonrecursive signal processing technology, which is particularly suitable for processing complex nonstationary time series through its unique mathematical framework. It simultaneously extracts the pattern and its center frequency (IMF), reconstructs the input signal, and maintains smoothness after baseband demodulation [46], thereby generating a band limited IMF. Compared with empirical mode decomposition (EMD) and ensemble empirical mode decomposition (EEMD), VMD avoids frequency aliasing and end effects [47]. VMD defines the IMF under constraint conditions:

$$u_k(t) = A_k(t) \cos(\phi_k(t)), \quad (3.1)$$

where the $\phi_k(t) \geq 0$ phase function is nonmonotonically decreasing. The envelope is the non-negative $A_k(t) \geq 0$, and $u_k(t)$ is the k mode function.

Using the Hilbert transform to generate the analytic signal and its one-sided spectrum for each intrinsic mode function,

$$\left(\delta(t) + \frac{j}{\pi t}\right) * u_k(t), \quad (3.2)$$

where $\delta(t)$ is the Dirac function.

Add exponential terms to estimate the center frequency of each mode, and modulate the spectrum of each mode to the corresponding baseband:

$$\left[\left(\delta(t) + \frac{j}{\pi t}\right) * u_k(t)\right] e^{-j\omega_k t}. \quad (3.3)$$

Decompose the original signal into K finite-bandwidth modal components, with each component centered around a center frequency. During decomposition, we minimize the sum of the estimated bandwidths of all modalities, subject to the sum of all modal components being equal to the original signal f . The constrained variational model of VMD is thus expressed as:

$$\min_{\{u_k\}, \{\omega_k\}} \left\{ \sum_{k=1}^K \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t}\right) * u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \right\} \quad \text{s.t.} \quad \sum_{k=1}^K u_k = f, \quad (3.4)$$

where u_k is each modal function, ω_k is the central frequencies of each modal function, and “*” denotes convolution.

Transform the constrained optimization problem mentioned in Eq (3.4) into an unconstrained variational problem. Equation (3.5) employs quadratic penalty terms along with the Lagrange multiplier method to convert the problem into an unconstrained form, resulting in the augmented Lagrangian function [48]:

$$\mathcal{L}(\{u_k\}, \{\omega_k\}, \lambda) = \alpha \sum_{k=1}^K \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t}\right) * u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 + \left\| f(t) - \sum_{k=1}^K u_k(t) \right\|_2^2 + \left\langle \lambda(t), f(t) - \sum_{k=1}^K u_k(t) \right\rangle. \quad (3.5)$$

After the introduction of Lagrange operator λ and quadratic penalty factor α , each mode component and its central frequency are updated to $\hat{u}_k^{n+1}$, ω_k^{n+1} , and $\hat{\lambda}^{n+1}$ expressed as:

$$\hat{u}_k^{n+1}(\omega) = \frac{\hat{f}(\omega) - \sum_{i \neq k} \hat{u}_i(\omega) + \frac{\hat{\lambda}(\omega)}{2}}{1 + 2\alpha(\omega - \omega_k)^2}, \quad (3.6)$$

$$\omega_k^{n+1} = \frac{\int_0^\infty \omega |\hat{u}_k(\omega)|^2 d\omega}{\int_0^\infty |\hat{u}_k(\omega)|^2 d\omega}, \quad (3.7)$$

$$\hat{\lambda}^{n+1}(\omega) = \hat{\lambda}^n(\omega) + \tau \left(\hat{f}(\omega) - \sum_{k=1}^K \hat{u}_k^{n+1}(\omega) \right). \quad (3.8)$$

Repeat it n times until it converges when the condition $\sum_{k=1}^K \|u_k^{n+1} - u_k^n\|_2^2 / \|u_k^n\|_2^2 < \varepsilon$ is reached.

3.2. Wavelet thresholding denoising

The wavelet threshold denoising (WTD) method applies filtering to signals in the wavelet domain for the denoising task of time series data [49]. Different wavelet coefficients have significantly different effects on noise [50]. The signal containing noise is decomposed into multiple components after wavelet decomposition. Then, by setting an appropriate threshold, the method smooths the different characteristics of the nonlinear noise signal, retaining components that exceed the threshold and removing those below it, thereby achieving denoising. Two classical threshold functions exist, the hard threshold and the soft threshold [51]:

$$\delta_{\lambda}^{\text{hard}}(w) = \begin{cases} w, & |w| \geq \lambda, \\ 0, & |w| < \lambda, \end{cases} \quad (3.9)$$

$$\delta_{\lambda}^{\text{soft}}(w) = \begin{cases} \text{sgn}(w)(|w| - \lambda), & |w| \geq \lambda, \\ 0, & |w| < \lambda, \end{cases} \quad (3.10)$$

where w represents the wavelet coefficient obtained by wavelet decomposition, and λ denotes the optimal threshold determined by the threshold selection rule.

Finally, the inverse wavelet transform is performed on the threshold-processed components of each scale to complete signal reconstruction, and the final denoised pure signal without high-frequency noise interference is obtained.

3.3. Gated recurrent unit

As a simplified architecture of LSTM, GRU retains the approximate identity mapping for state updates and reduces the complexity of neural units. It not only achieves a comparable capacity to mitigate gradient problems as LSTM but also cuts down the number of parameters in computation, which shortens training time and reduces the risk of overfitting.

The GRU merges the cell state and hidden state into a single hidden state h_t adopting a two-gate architecture composed of an update gate and a reset gate. The specific mathematical formulations are presented in Eqs (3.11)–(3.14). The update gate specifically addresses the exponential decay of gradients in long sequences, and the reset gate controls the contribution of historical hidden information. This design is crucial for alleviating vanishing and exploding gradients, making GRU more applicable to sequence prediction tasks with temporal dependencies. The GRU unit is formulated as

$$z_t = \sigma(W_z[h_{t-1}, x_t] + b_z), \quad (3.11)$$

$$r_t = \sigma(W_r[h_{t-1}, x_t] + b_r), \quad (3.12)$$

$$\hat{h}_t = \tanh(W_h[r_t \odot h_{t-1}, x_t] + b_h), \quad (3.13)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \hat{h}_t, \quad (3.14)$$

where z_t is the update gate, r_t is the reset gate, $\hat{h}_t$ is the candidate hidden state, and h_t is the hidden state at the current time step.

3.4. The VMD-WTD-GRU hybrid model with mode-specific denoising

Hybrid forecasting models based on signal decomposition have been widely used in energy consumption forecasting. Among them, VMD is an effective preprocessing method for nonstationary time series because it can decompose a complex original sequence into several IMFs with different frequency characteristics. In a conventional VMD-GRU framework, the original sequence is first decomposed into multiple IMFs, each IMF is then modeled by a GRU network, and the final forecasting result is obtained by summing the forecasting outputs of all components.

However, this direct decomposition–forecasting strategy still has several limitations. VMD mainly separates the original sequence into different frequency bands, but it does not completely remove noise from the decomposed components. Some high-frequency IMFs may still contain dense oscillations, random disturbances, and local irregular fluctuations. If these components are directly used as GRU inputs, the model may learn noise-related patterns, which can reduce forecasting stability. In addition, many VMD-based forecasting models mainly focus on the decomposed IMFs, and the residual component generated during decomposition is often ignored or treated only as a minor reconstruction term. For petroleum energy consumption forecasting, the residual component may still contain local information or unexplained fluctuations. Ignoring it may lead to information loss in the final forecasting reconstruction.

To address these problems, this study proposes a variational modal decomposition-wavelet threshold denoising-gated recurrent unit (VMD-WTD-GRU) hybrid model with mode-specific denoising. Compared with conventional VMD-WTD-based forecasting models, the proposed framework is not a simple process of decomposing the original sequence, denoising all IMFs, and directly feeding them into GRU. Instead, it applies WTD mainly to the high-frequency IMFs and the residual mode, where noise-related disturbances and irregular fluctuations are more likely to occur. The low-frequency IMFs, which mainly contain long-term trend information and relatively stable variation patterns, are retained to avoid excessive smoothing and information loss. Therefore, the proposed framework jointly considers modal heterogeneity, residual information, and targeted denoising, thereby improving the stability and accuracy of petroleum energy consumption forecasting.

Let $x(t)$ denote the original petroleum energy consumption sequence. After VMD decomposition, $x(t)$ can be expressed as:

$$x(t) = \sum_{k=1}^K u_k(t) + r(t), \quad (3.15)$$

where $u_k(t)$ denotes the k th IMF component, K is the number of decomposed modes, and $r(t)$ denotes the residual component. In this study, the residual term is retained as an additional residual mode rather than being discarded.

For convenience, all decomposed components are denoted as:

$$z_j(t) = \begin{cases} u_j(t), & j = 1, 2, \dots, K, \\ r(t), & j = K + 1, \end{cases} \quad (3.16)$$

where $z_j(t)$ represents either an IMF component or the residual mode. For the decomposed components, this study does not apply WTD uniformly to all IMFs. Instead, WTD is mainly performed on the high-frequency IMF components and the residual mode. The denoising strategy can be expressed as:

$$\tilde{z}_j(t) = \begin{cases} \mathcal{WTD}(z_j(t)), & z_j(t) \in \Omega_H \cup r(t), \\ z_j(t), & z_j(t) \in \Omega_L, \end{cases} \quad (3.17)$$

where Ω_H denotes the set of high-frequency IMF components, Ω_L denotes the set of low-frequency IMF components, $r(t)$ denotes the residual mode, and $\mathcal{WTD}(\cdot)$ denotes the wavelet threshold denoising operation. This design enables the model to suppress noise-sensitive components while preserving the main trend information contained in low-frequency modes. After WTD processing, each processed component is modeled by a separate GRU network. For the j th component, the GRU model uses a sliding input window to generate the one-day-ahead forecasting result:

$$\hat{z}_j(t+1) = \text{GRU}_j(\tilde{z}_j(t-L+1), \tilde{z}_j(t-L+2), \dots, \tilde{z}_j(t)), \quad (3.18)$$

where L denotes the length of the sliding time window, and $\hat{z}_j(t+1)$ is the forecasting result of the j th component at time $t+1$. Finally, the forecasting outputs of all IMF components, and the residual mode are summed to obtain the final petroleum energy consumption forecasting result:

$$\hat{x}(t+1) = \sum_{j=1}^{K+1} \hat{z}_j(t+1), \quad (3.19)$$

where $\hat{x}(t+1)$ denotes the final one-day-ahead forecasting result. The proposed VMD-WTD-GRU model is implemented through four main steps, as shown in Figure 1.

Step 1: VMD decomposition. The original petroleum energy consumption sequence is decomposed into K IMF components and one residual component. The residual component is retained as an additional mode for subsequent processing.

Step 2: WTD denoising for high-frequency and residual modes. WTD is applied to the high-frequency IMF components and the residual mode, and the low-frequency components are retained to preserve the main trend information. This targeted denoising strategy avoids mechanically processing all decomposed components and helps reduce noise interference without weakening useful low-frequency temporal patterns.

Step 3: GRU forecasting. Each processed component is independently modeled by a GRU network. The GRU model learns the temporal dependence of each component and generates the one-day-ahead forecasting result.

Step 4: Forecasting reconstruction. The forecasting outputs of all IMF components, and the residual mode is aggregated to obtain the final forecasting value of petroleum energy consumption.

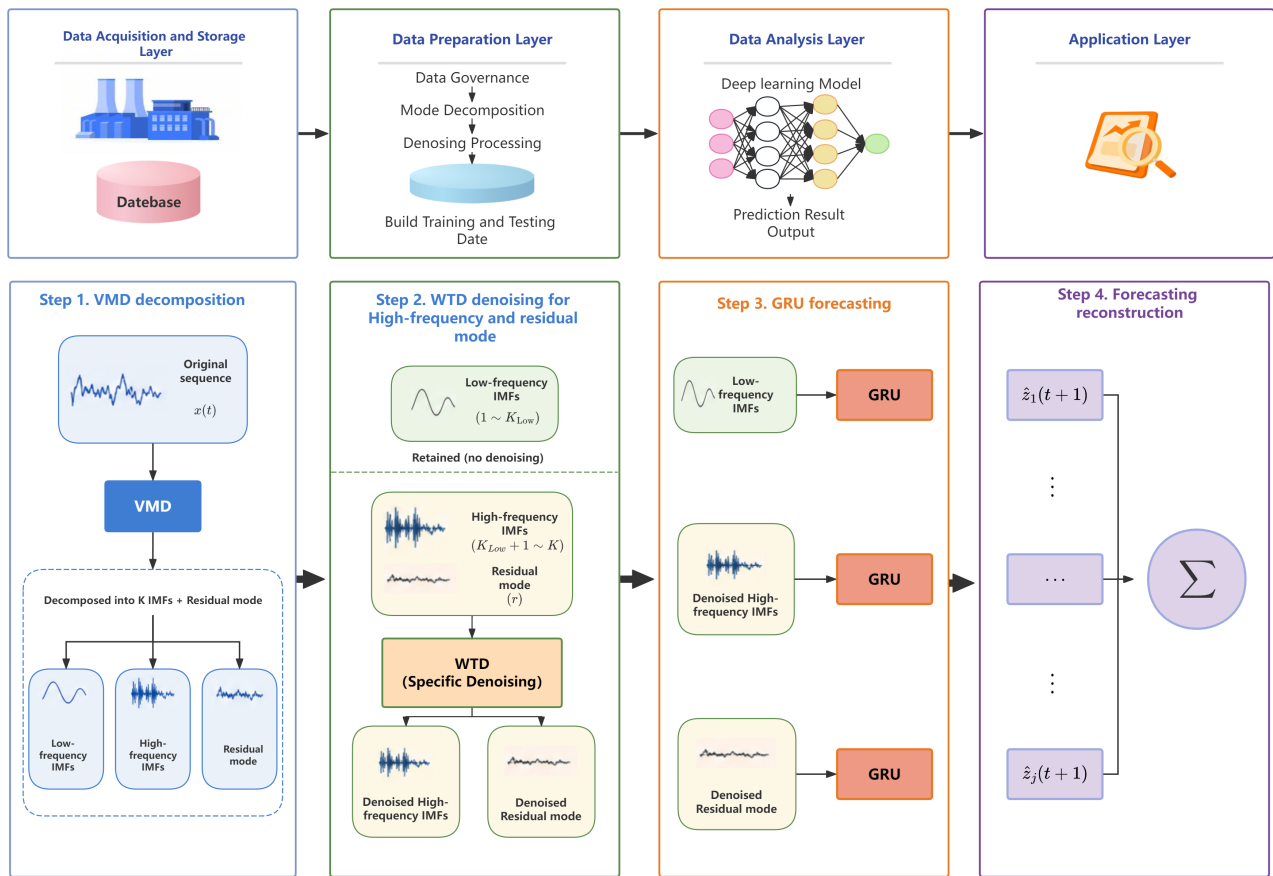


Figure 1. Framework of the proposed VMD-WTD-GRU model.

Compared with conventional VMD-WTD-based forecasting models, the proposed VMD-WTD-GRU framework is not a simple process of decomposing the original sequence, denoising all IMFs, and directly feeding them into GRU. Instead, it considers the different fluctuation characteristics of decomposed modes and applies WTD to noise-sensitive modal components in a more targeted way. In addition, the residual component generated by VMD is retained as an additional residual mode and included in the denoising, forecasting, and final reconstruction processes. This design helps reduce noise interference while avoiding information loss, thereby improving the stability and accuracy of petroleum energy consumption forecasting.

4. Experiments analysis

4.1. Description of data

The research is based on the daily outbound shipment records of three refined petroleum products (gasoline, kerosene, and diesel) sourced from every petroleum depot within a delimited region. The data span a 12-year period, extending from 1 January 2012 to 31 December 2023. This comprehensive temporal coverage is essential for capturing long-term trends and seasonal variations

in energy consumption patterns. Each record contains the following fields: petroleum type p : gasoline g , kerosene k , diesel d , outgoing volume (liters, L), and time t : year/month/day.

In order to unify the measurement units of different petroleum products, the volume data (liters) is converted into mass (tons) according to the density of each petroleum product. The conversion refers to the international industry standard ASTM D4052 (American Society for Testing and Materials), based on the standard densities of gasoline ASTM-D4814, kerosene ASTM-D975, and diesel ASTM-D975 at 15 °C. Gasoline $\rho_g = 0.74 \text{ g/cm}^3$, Kerosene $\rho_k = 0.81 \text{ g/cm}^3$, Diesel $\rho_d = 0.85 \text{ g/cm}^3$. 15 °C is the international standard temperature used in the petroleum industry for trade settlements and density corrections. The amount of each type of petroleum consumed per day (tons) is expressed as follows:

$$M_p(t) = V_p(t) \times \rho_p \times 10^{-3}. \quad (4.1)$$

The total regional daily petroleum energy consumption (tons) is:

$$C(t) = \sum_{p \in \{g,k,d\}} M_p(t), \quad (4.2)$$

where $M_p(t)$ is the mass of petroleum p at day t (tons), $V_p(t)$ is the outgoing volume of petroleum p at day t (liters), and $C(t)$ is the total regional petroleum energy consumption at day t (tons).

4.2. Preliminary analysis

Petroleum, as an important part of the regional energy mix, and its derivatives (including gasoline, kerosene and diesel) dominate the petroleum energy consumption structure. This study can reflect the consumption of petroleum energy in the region by constructing the regional petroleum energy daily consumption data. Table 1 details the descriptive statistics of the consumption of major petroleum products, with special emphasis on the quartiles (Q1, Q3) and interquartile range (IQR).

Table 1. Descriptive statistics.

Serial number	Petroleum	Maximum	Minimum	Average	Q1	Q3	IQR
1	Gasoline consumption	1862.30	14.20	306.79	280.50	335.80	55.30
2	Kerosene consumption	3855.10	0.00	438.11	2.90	1243.98	1241.07
3	Diesel consumption	5195.70	2.50	403.16	250.70	471.40	220.70
4	Total petroleum consumption	6336.60	21.10	1147.30	620.65	1921.65	1301.00

The daily total petroleum energy consumption in the region, as summarized in Figure 2, presents a pronounced asymmetric distribution (skewness = 1.19, kurtosis = 4.78), with a positive skewness coefficient of > 0 , indicating a right skewness in the distribution, and a spike characterizing a kurtosis of > 3 , implying that the data are highly concentrated at the center and that there are extremes in the tails. This distribution pattern implies that the distribution of energy consumption data is relatively concentrated around the mean but with occasional extreme deviations from the mean.

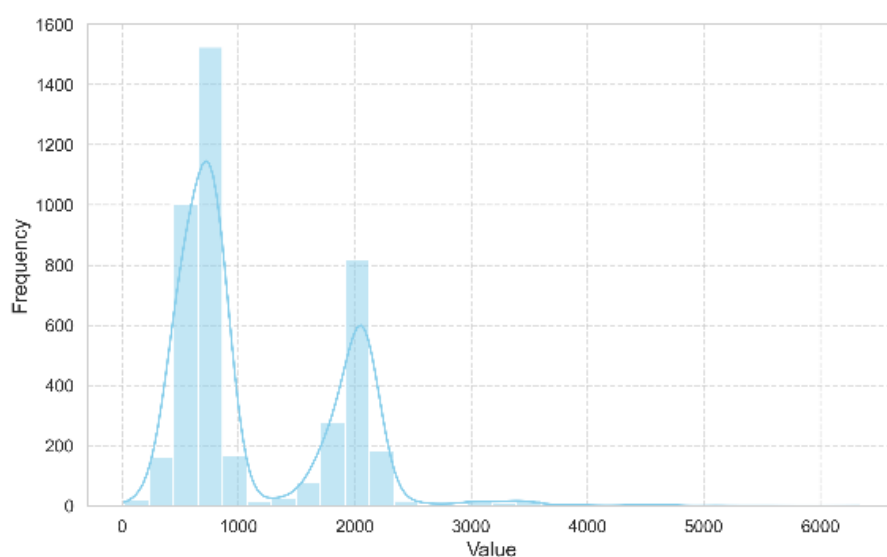


Figure 2. Time series of total petroleum consumption.

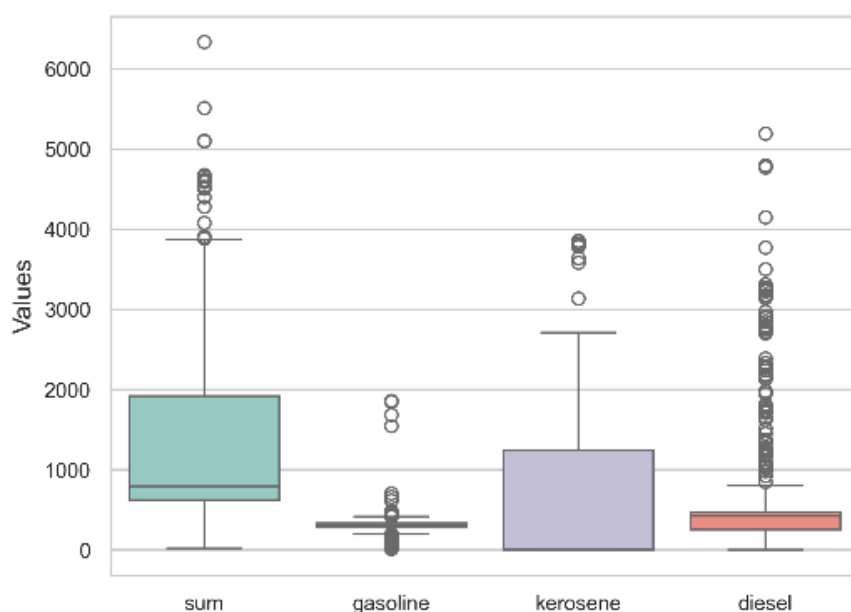


Figure 3. Box plot of petroleum consumption.

As can be seen through the box plot Figure 3, different petroleum products show differentiated distribution characteristics, with gasoline (IQR = 55.30) and diesel (IQR = 220.70) consumption showing low dispersion characteristics (coefficient of variation $CV < 0.75$) and kerosene consumption having a significantly high degree of dispersion (IQR = 1241.08, coefficient of variation $CV = 1.45$). This anomalous dispersion phenomenon mainly stems from the special demand structure of the air transportation industry, which is affected by factors such as business travel and peak travel seasons, and is inherently highly variable. It is worth noting that during the COVID-19 epidemic

(2020Q2-2022Q3), a large number of flights were canceled, resulting in a sharp drop in kerosene use due to strict regional epidemic prevention policies and travel restrictions.

4.3. Evaluation metrics

To evaluate the forecasting accuracy of the proposed model against alternative approaches, three commonly used evaluation metrics were selected for this study: root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE), as shown in Table 2.

Table 2. Three evaluation metrics.

Metric	Definition	Equation
RMSE	Root mean squared error	$RMSE = \sqrt{\frac{1}{N} \sum_{t=1}^N (X(t) - \widehat{X}(t))^2}$
MAE	Mean absolute error	$MAE = \frac{1}{N} \sum_{t=1}^N X(t) - \widehat{X}(t) $
MAPE	Mean absolute percentage error	$MAPE = \frac{1}{N} \sum_{t=1}^N \left \frac{X(t) - \widehat{X}(t)}{X(t)} \right \times 100\%$

To further examine the robustness of the out-of-sample forecasting results, the Diebold–Mariano (DM) test is used to compare the predictive ability of different models. The DM test was proposed by Diebold and Mariano [52] and is commonly used to test whether two forecasting models have significantly different forecasting accuracy. For model i and model j , the squared-error loss differential is defined as

$$d_t = (y_t - \hat{y}_{i,t})^2 - (y_t - \hat{y}_{j,t})^2, \quad (4.3)$$

where y_t is the actual value, and $\hat{y}_{i,t}$ and $\hat{y}_{j,t}$ are the forecasting values of model i and model j , respectively. A positive d_t indicates that model i has a smaller forecasting loss than model j . The null hypothesis is $E(d_t) = 0$, indicating equal forecasting accuracy between the two models. The DM statistic is calculated as

$$DM = \frac{\bar{d}}{\sqrt{\hat{\sigma}_d^2/T}}, \quad (4.4)$$

where $\bar{d}$ is the mean of the loss differential series d_t , $\hat{\sigma}_d^2$ is the estimated variance of d_t , and T is the number of test samples. A significant DM statistic indicates a statistically significant difference in forecasting accuracy between the two models.

4.4. Dataset division

The experimental dataset used in this study spans from January 1, 2012 to December 31, 2023, including 4383 daily observations of regional petroleum energy consumption. Because time-series data have clear chronological dependence, the dataset is divided in temporal order rather than by random sampling.

To convert the original sequence into supervised learning samples, a sliding time window is adopted. Let $C(t)$ denote the total apparent consumption of regional petroleum energy on day t , and let L be the window length. Considering the short-term fluctuation and periodic characteristics of petroleum energy consumption, L is set to 64. Therefore, the consumption values of the previous 64 consecutive days are used to forecast the next-day consumption, as shown in Eq (4.5):

$$X_t = [C(t - L), C(t - L + 1), \dots, C(t - 1)], \quad y_t = C(t), \quad (4.5)$$

where X_t denotes the input sequence consisting of 64 historical observations, and y_t denotes the forecasting target on day t . The sliding-window sample construction process is illustrated in Figure 4.

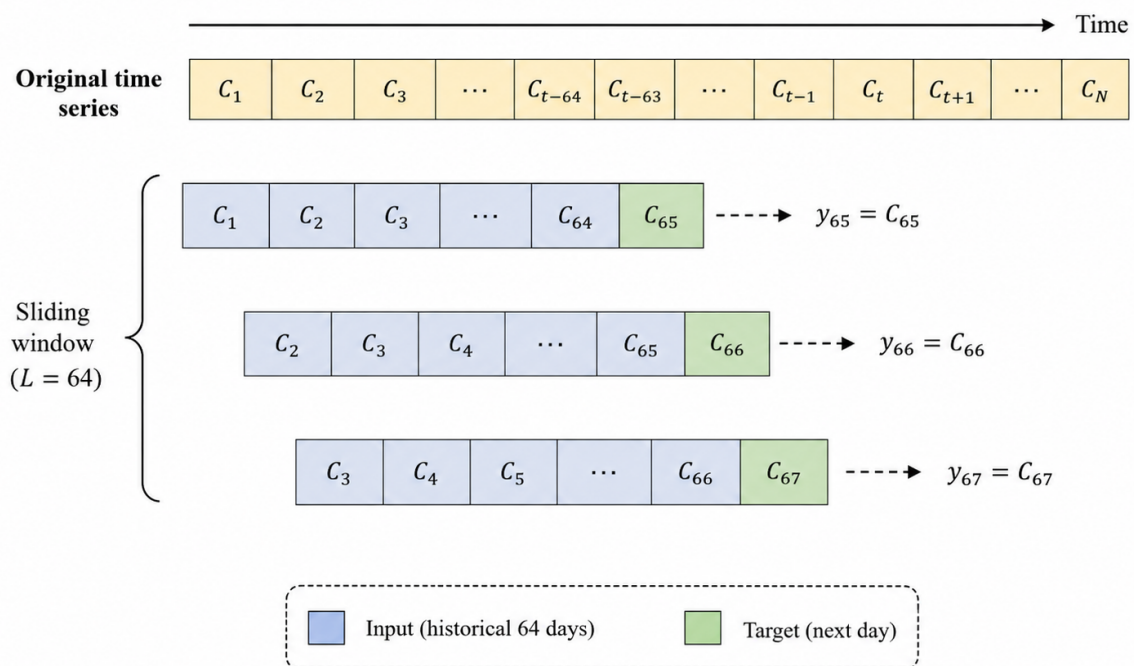


Figure 4. Sliding-window sample construction.

After sliding-window reconstruction, a total of 4319 supervised learning samples are obtained. Following the commonly used 8:2 division strategy for time-series forecasting, the first 80% of the reconstructed samples are used as the training set, and the remaining 20% are used as the test set. Specifically, 3455 samples are used for model training, parameter learning, and hyperparameter selection, and the remaining 864 samples are used only for final performance evaluation.

4.5. Experiments and results

This section evaluates the forecasting performance of different neural network models for regional petroleum energy consumption. The experiments are organized in a progressive manner. First, five baseline models, including BP, RNN, LSTM, GRU, and Transformer, are compared to examine their direct forecasting ability on the original petroleum energy consumption sequence. Second, VMD is

introduced as a modal decomposition method, and the corresponding VMD-based decomposition models, namely VMD-BP, VMD-RNN, VMD-LSTM, VMD-GRU, and VMD-Transformer, are constructed to evaluate the contribution of modal decomposition. Third, WTD is further applied after VMD decomposition to suppress high-frequency noise in the decomposed IMF components, forming the VMD-WTD-LSTM and VMD-WTD-GRU models.

This experimental design aims to examine the effects of model structure, modal decomposition, and denoising optimization on forecasting performance. The baseline models provide the initial reference for comparison. The VMD-enhanced models are used to test whether decomposing the original non-stationary sequence into several modal components can improve model learning ability. The VMD-WTD-based models further verify whether noise reduction can enhance the stability and accuracy of the decomposition-based forecasting framework.

To ensure fair comparison, all models are trained and tested under the same data division strategy described in Section 4.4. The main training settings are kept comparable as far as possible. The batch size is set to 64, the number of training epochs is set to 40, and the hidden-layer structure of the neural network models is set to (30, 30), where applicable. Each model is trained independently, and the test set is used only for final performance evaluation. RMSE, MAE, and MAPE are adopted as evaluation indicators, with MAPE used as the main criterion for relative forecasting accuracy. All experiments are implemented in Python 3.9.

4.5.1. Performance evaluation of baseline forecasting models

To establish a benchmark for subsequent hybrid model evaluation, five baseline forecasting models are first compared, including BP, RNN, LSTM, GRU, and Transformer. All models are trained and tested under the same data division strategy and comparable experimental settings. The forecasting performance is evaluated using RMSE, MAE, and MAPE, and the results are reported in Table 3.

Table 3. Forecasting performance of baseline models.

Model	RMSE	MAE	MAPE (%)
BP	615.36	471.37	74.87
RNN	575.76	431.91	64.98
LSTM	634.42	434.68	57.64
GRU	642.91	438.39	58.38
Transformer	765.29	533.68	70.59

As shown in Table 3, LSTM achieves the lowest MAPE of 57.64%, indicating the best relative forecasting accuracy among the baseline models, whereas GRU obtains a similar MAPE of 58.38%. RNN records the lowest RMSE of 575.76 and MAE of 431.91, demonstrating superior performance in reducing absolute forecasting errors. Overall, recurrent neural network models outperform BP, suggesting that petroleum energy consumption exhibits strong temporal dependence and is better modeled by sequence-learning architectures. Transformer achieves a lower MAPE than RNN but still performs worse than the recurrent models in terms of RMSE and MAE, indicating that the attention-based structure does not consistently improve forecasting stability under the current sample size and data.

To further compare the forecasting behavior of different baseline models, the forecasting curves and

actual values are shown in Figure 5.

As shown in Figure 5, the baseline models can follow the general movement of petroleum energy consumption to some extent, but obvious deviations remain in periods with sharp peaks, valleys, or short-term fluctuations. The recurrent models show stronger adaptability to local temporal changes than BP and Transformer, which is generally consistent with the quantitative results in Table 3. However, none of the baseline models can fully capture the irregular fluctuations in the original series.

Overall, the baseline comparison shows that single neural network models still have limitations in directly modeling petroleum energy consumption. Although LSTM and GRU achieve relatively better MAPE values, the remaining forecasting errors indicate that the original sequence contains complex nonstationary fluctuations and high-frequency variations. Therefore, VMD-based modal decomposition is introduced in the next subsection to reduce sequence complexity and improve forecasting performance.

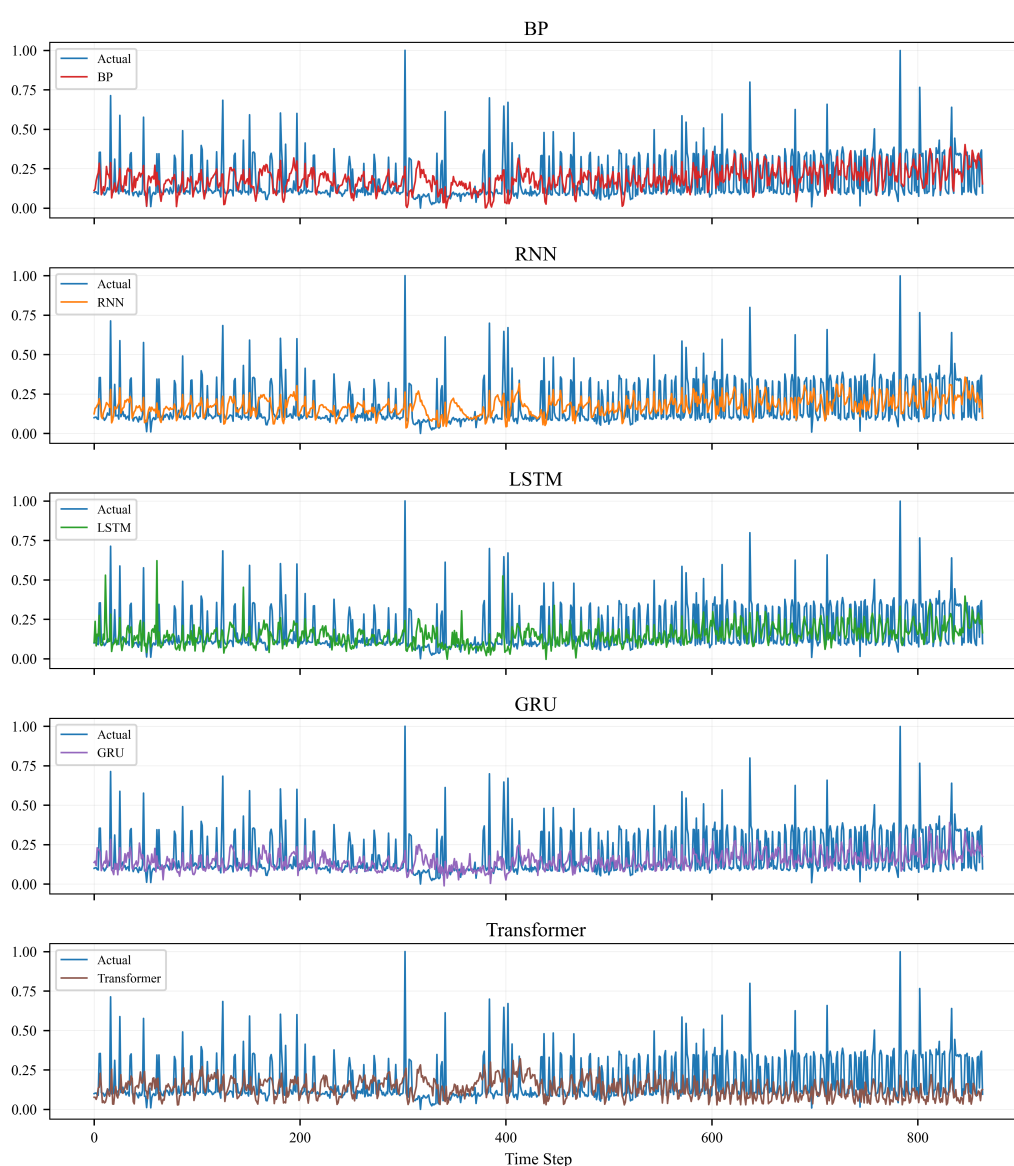


Figure 5. Forecasting curves of baseline models.

4.5.2. The role of VMD for baseline forecasting models

To investigate the influence of VMD on the network model, we compare different values of the decomposition layers K and penalty parameter α using a grid search approach. The decomposition performance is optimal when $K = 10$, and $\alpha = 2000$. We can clearly see that there are a total of 10 waveforms arranged from the top to the bottom, and these waveforms represent the 10 eigenmodal function components IMF1 to IMF10, respectively. From further observation of Figure 6, we can find that after the VMD processing, each component from IMF1 to IMF10 exhibits its own unique and easily distinguishable characteristics.

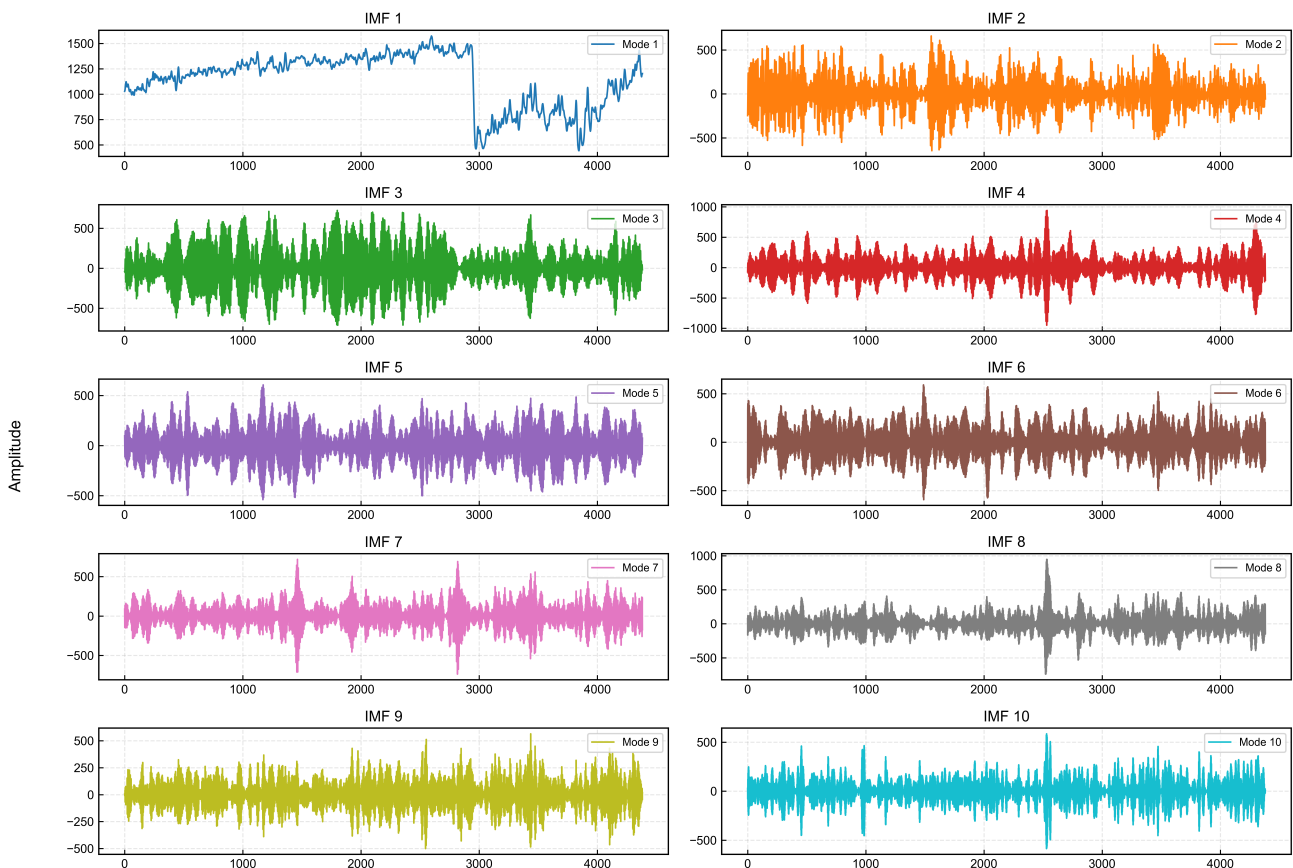


Figure 6. VMD decomposition results (IMF1–IMF10).

Furthermore, we employ VMD as the modal decomposition approach and combine it with five deep learning models, namely BP, RNN, LSTM, Transformer, and GRU, to establish hybrid forecasting models including VMD-BP, VMD-RNN, VMD-LSTM, VMD-Transformer, and VMD-GRU. The optimization effect of VMD on different deep learning architectures is analyzed, and the impacts of critical hyperparameters of VMD on the predictive performance of hybrid models are further explored. Baseline models are compared with their VMD-enhanced deep learning hybrid counterparts, and the prediction performance of various VMD-optimized hybrid models is listed in Table 4.

As shown in Table 4, VMD preprocessing can effectively enhance their prediction accuracy of

five models, namely BP, RNN, LSTM, Transformer, and GRU. In terms of the MAPE indicator, the MAPE of BP drops from 74.87% to 42.43%, RNN from 64.98% to 39.81%, LSTM from 57.64% to 36.33%, Transformer from 70.59% to 39.34%, and GRU from 58.38% to 33.13%. After VMD decomposition preprocessing, the RMSE, MAE, and MAPE evaluation indicators of each comparison model decrease synchronously, indicating that variational mode decomposition can achieve multiscale feature decomposition of nonstationary time series data. By stripping different frequency components of the sequence and mining hidden time series rules, the fitting effect of the time series prediction model can be effectively improved.

Table 4. Prediction performance comparison of deep learning hybrid models optimized by VMD.

Model	RMSE	MAE	MAPE (%)
BP	615.36	471.37	74.87
VMD-BP	368.74	274.31	42.43
RNN	575.76	431.91	64.98
VMD-RNN	367.93	267.08	39.81
LSTM	634.42	434.68	57.64
VMD-LSTM	284.97	214.16	36.33
Transformer	765.29	533.68	70.59
VMD-Transformer	328.77	248.96	39.34
GRU	642.91	438.39	58.38
VMD-GRU	273.97	201.86	33.13

In terms of optimization magnitude, BP obtains the largest MAPE reduction of 32.44%. Transformer comes next with a MAPE drop of 31.25%, followed by GRU, RNN, and LSTM with respective declines of 25.25%, 25.17%, and 21.31%. Despite not having the highest error reduction rate, VMD-GRU achieves the optimal MAPE alongside the lowest RMSE of 273.97 and MAE of 201.86, across all experimental groups, demonstrating superior overall prediction performance compared to other comparative models.

After confirming the favorable comprehensive performance of VMD-GRU, this study further explores the influence of critical VMD parameters on the predictive performance of hybrid models. The modal number K determines the total number of intrinsic mode functions (IMFs) decomposed via VMD. A too-small K results in mode aliasing and incomplete decomposition of the raw series: Trend terms, periodic fluctuations, and high-frequency noise with distinct frequencies are compressed into a single IMF and cannot be separated based on frequency-domain features. Conversely, an excessively large K causes over-decomposition of valid signals and generates redundant spurious modes that only contain random noise. Both improper parameter settings deteriorate VMD decomposition quality and degrade their prediction accuracy of hybrid models. To intuitively reflect how modal number K affects model performance, the MAPE values and variation tendency of VMD-GRU under different K are listed in Table 5 and illustrated in Figure 7.

As shown in Table 5, with the increase of modal number K , the MAPE of the model exhibits a trend of first decreasing and then increasing: When K rises from 5 to 10, the MAPE continuously drops from 46.70% to 33.13%; when K further increases to 11, the MAPE rebounds to 35.86%. For K

ranging from 5 to 9, the modal number is insufficient, leading to mode aliasing in decomposition—load fluctuations and noise of different frequencies cannot be separated, thus limiting the optimization effect of preprocessing. When K exceeds 10, over-decomposition occurs, where valid homologous signals are split into redundant spurious modes. Based on the comprehensive analysis of the above indicators, the optimal modal number is determined as $K = 10$ in this study.

Table 5. MAPE values of the VMD-GRU model at different K values.

K	5	6	7	8	9	10	11
MAPE (%)	46.70	41.11	41.17	38.14	35.23	33.13	35.86

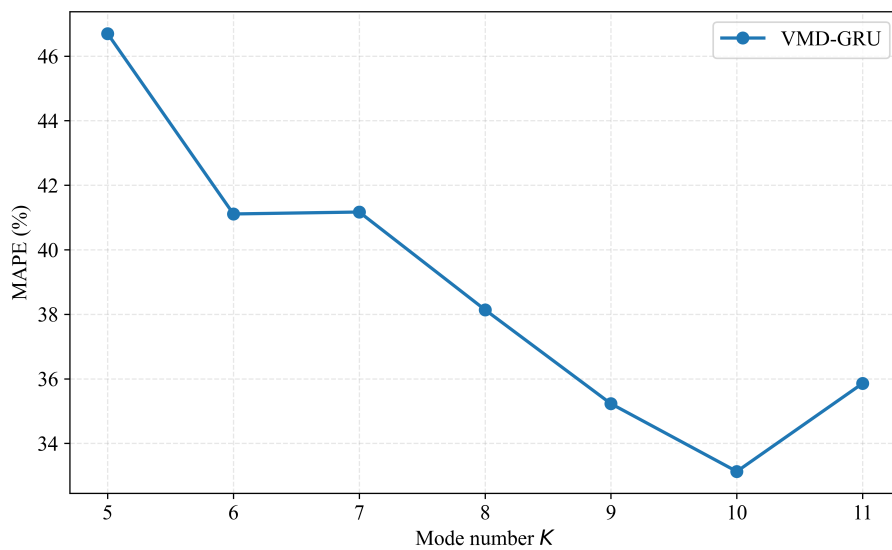


Figure 7. Variation trend of MAPE for the VMD-GRU model with various modal numbers K .

Using the optimal decomposition number $K = 10$ identified from the preceding screening test, we adopt grid search to optimize VMD's modal number K and penalty coefficient α . Restricted to the optimal range $K \in [5, 10]$ derived earlier, K is preliminarily searched within $[5, 10]$, whereas α takes values of 1500, 2000, and 2500 in comparative tests, with corresponding MAPE results of the VMD-GRU model detailed in Table 6.

Table 6. MAPE (%) of the VMD-GRU model under different combinations of K and α .

α	$K = 5$	$K = 6$	$K = 7$	$K = 8$	$K = 9$	$K = 10$
2500	50.26	43.37	47.30	37.34	36.20	36.32
2000	46.70	41.11	41.17	38.14	35.23	33.13
1500	45.87	43.57	38.08	36.61	30.61	33.88

To intuitively demonstrate the discrepancies in prediction performance of the VMD-GRU model under various combinations of K and α , a three-dimensional MAPE surface plot is plotted as shown in

Figure 8. The corresponding MAPE values of each parameter combination are marked on the figure to facilitate a clear comparison of prediction errors under diverse parameter configurations.

According to the results in Table 6 and Figure 8, the MAPE values of all parameter combinations range from 30.61% to 50.26%, with an overall span of nearly 20 percentage points. The prediction error of the VMD-GRU model presents a concave distribution with respect to VMD parameters K and α . This indicates that the selection of K and α in VMD has a significant impact on their prediction accuracy of the VMD-GRU model. Although the minimum MAPE within the grid search is obtained at $K = 9$ and $\alpha = 1500$, such a small value of α provides insufficient bandwidth constraint from the perspective of VMD decomposition principle and tends to induce implicit mode aliasing. Hence, this minimum is merely an accidental optimum at the search boundary and cannot be regarded as the optimal parameter combination. In contrast, the parameter set $K = 10$, $\alpha = 2000$ yields a competitively low MAPE. Comprehensively considering prediction accuracy, model robustness, and the VMD decomposition mechanism, the parameter combination of $K = 10$ and $\alpha = 2000$ is adopted for experiments in this research.

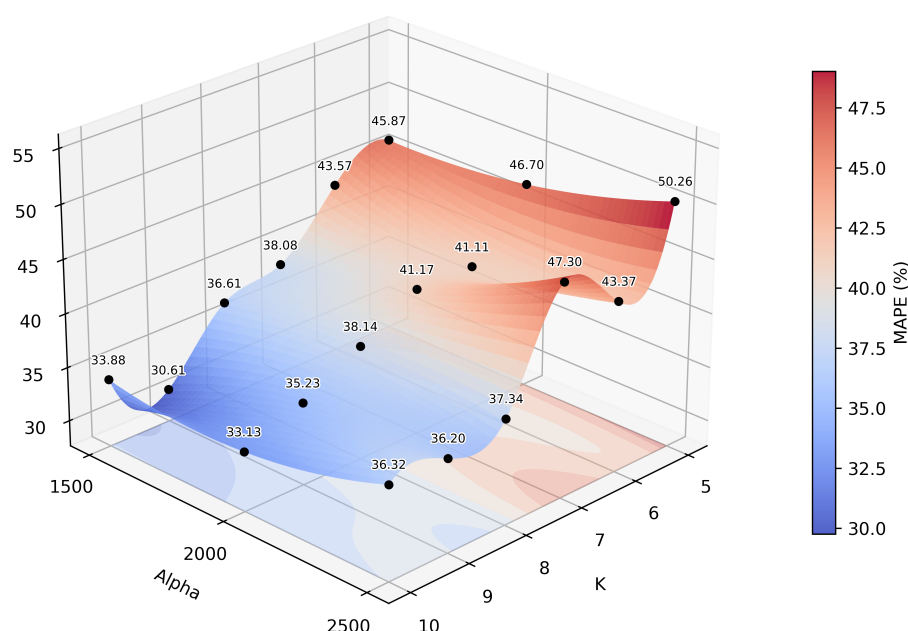


Figure 8. Three-dimensional surface plot of MAPE for the VMD-GRU model under different combinations of K and α .

4.5.3. Effectiveness analysis of the VMD-WTD hybrid forecasting framework

After evaluating the effect of VMD-based modal decomposition, WTD is further introduced to examine whether noise reduction can improve the forecasting performance of decomposition-based models. Because VMD-LSTM and VMD-GRU exhibit superior forecasting performance among the decomposition-based models, this subsection focuses on the comparison among GRU, LSTM, VMD-GRU, VMD-LSTM, VMD-WTD-GRU, and VMD-WTD-LSTM.

To visually examine the denoising effect of WTD, Figure 9 compares IMF 5 to IMF 10 before and after WTD processing. These components contain relatively stronger local fluctuations and are more

likely to carry noise information. The denoised curves generally maintain the main fluctuation pattern of the original IMF components, whereas part of the dense high-frequency oscillations is weakened.

The forecasting performance of the VMD-WTD-based models is then evaluated. To investigate the effects of modal decomposition and denoising on forecasting accuracy, the forecasting results of GRU and LSTM are compared with those of the corresponding VMD-based and VMD-WTD-based models. The results are reported in Table 7.

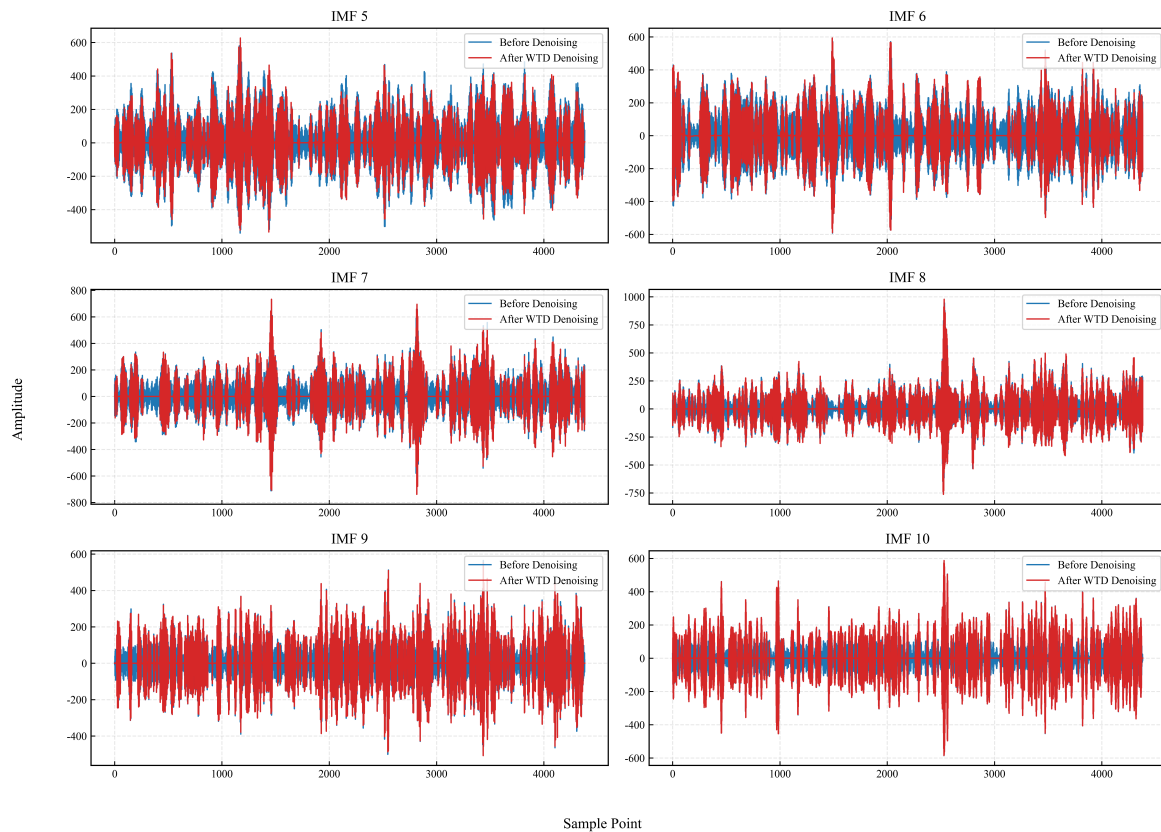


Figure 9. Comparison of IMF components before and after WTD denoising.

Table 7. Forecasting performance of VMD-WTD-based models.

Model	RMSE	MAE	MAPE (%)
GRU	642.91	438.39	58.38
VMD-GRU	273.97	201.86	33.13
VMD-WTD-GRU	258.14	193.97	31.45
LSTM	634.42	434.68	57.64
VMD-LSTM	284.97	214.16	36.33
VMD-WTD-LSTM	273.10	210.56	34.32

The introduction of WTD further improves the forecasting performance of the VMD-based models. VMD-WTD-GRU achieves the best overall results, with an RMSE, MAE, and MAPE of 258.14, 193.97, and 31.45%, respectively. Compared with VMD-GRU, the RMSE decreases

from 273.97 to 258.14, the MAE decreases from 201.86 to 193.97, and the MAPE decreases by 1.67 percentage points. Similarly, VMD-WTD-LSTM improves upon VMD-LSTM, reducing the RMSE from 284.97 to 273.10, the MAE from 214.16 to 210.56, and the MAPE from 36.33% to 34.32%. These results indicate that WTD can provide cleaner modal inputs and improve the forecasting stability of decomposition-based models.

Figure 10 shows the forecasting curves of LSTM, GRU, VMD-LSTM, VMD-GRU, VMD-WTD-LSTM, and VMD-WTD-GRU.

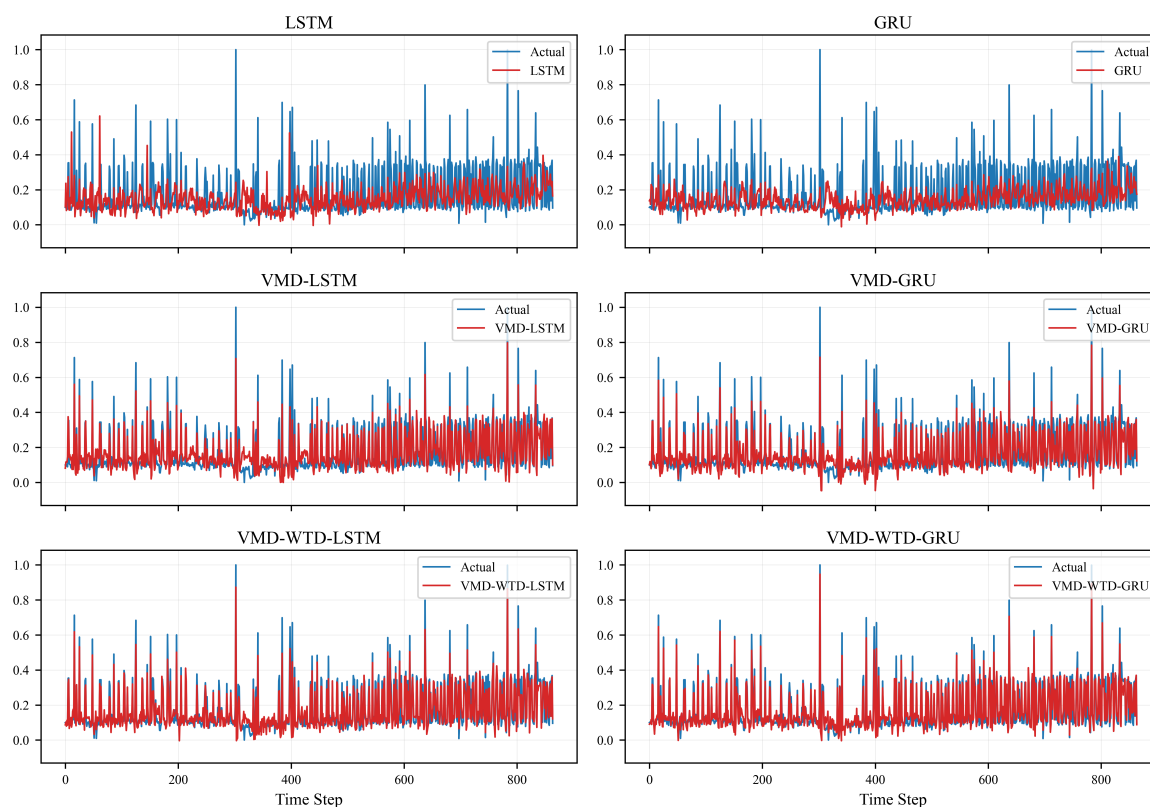


Figure 10. Forecasting performance of LSTM/GRU with VMD-WTD preprocessing.

As shown in Figure 10, the single GRU and LSTM models only capture the general fluctuation range of petroleum energy consumption, and their forecasting curves are relatively smooth when local peaks and short-term variations occur. After VMD decomposition is introduced, the forecasting results of VMD-GRU and VMD-LSTM follow the temporal variation trend of the actual series more closely, indicating that modal decomposition can reduce the complexity of the original nonstationary sequence. On this basis, the VMD-WTD-based models further improve the fitting performance. Among them, VMD-WTD-GRU presents the closest fit to the actual curve, with smaller deviations in most test intervals and a more stable response to local changes. These observations are consistent with the quantitative evaluation results reported in Table 7, where VMD-WTD-GRU achieves the lowest RMSE, MAE, and MAPE values of 258.14, 193.97, and 31.45%, respectively.

4.5.4. The effects of the VMD-WTD denoising scheme on different frequency components

In previous experiments, the IMF components obtained through VMD decomposition were used as inputs for WTD denoising and neural network prediction. However, the reconstructed signal obtained by summing all VMD components may still have residual differences from the original petroleum energy consumption sequence. These residual signals are usually considered as noise or unexplained disturbances that cannot be captured by the decomposed IMF. Therefore, this section further incorporates residual components into the VMD-WTD framework and compares different denoising combinations of low-frequency, high-frequency, and residual components through ablation experiments to evaluate the contribution of each component configuration to overall prediction performance.

According to the frequency characteristics of the decomposed components, IMF1–IMF5 are classified as low-frequency components, which mainly reflect relatively smooth trend information. IMF6–IMF10 are classified as high-frequency components, which mainly contain short-term fluctuations and local disturbances. The residual component IMF11 is used to supplement the remaining information between the original sequence and the reconstructed VMD signal.

To further examine the contribution of different components, five VMD-WTD processing strategies are designed: High+Low, High+Res, High+Low+Res, Low+Res, and Res. Components included in a given strategy are first processed by WTD before being input into the forecasting model, whereas components not selected for WTD are directly fed into the forecasting model in their original VMD-decomposed form. In all strategies, every IMF component and the residual mode are retained for forecasting and final reconstruction; the only difference is whether a component undergoes WTD processing prior to prediction.

Table 8. Forecasting performance of different VMD-WTD processing strategies.

Model	Strategies	RMSE	MAE	MAPE (%)
VMD-WTD-GRU	High+Low	258.14	193.97	31.45
	High+Res	184.56	141.80	23.15
	High+Low+Res	214.89	164.83	27.51
	Low+Res	219.53	166.05	25.82
	Res	195.15	148.81	24.34
VMD-WTD-LSTM	High+Low	273.10	210.56	34.32
	High+Res	213.16	160.83	26.56
	High+Low+Res	236.46	182.25	29.79
	Low+Res	219.15	164.66	26.83
	Res	240.85	183.94	29.06

Table 8 shows that the High+Res strategy achieves the best performance for both GRU and LSTM. For VMD-WTD-GRU, High+Res obtains the lowest RMSE, MAE, and MAPE values of 184.56, 141.80, and 23.15%, respectively. Compared with High+Low, the MAPE decreases by 8.30

The High+Low+Res strategy does not achieve the best accuracy, although it processes all component IMFs. This indicates that applying WTD to more IMFs does not necessarily improve forecasting performance. Because low-frequency components mainly contain trend information, excessive denoising may weaken useful smooth variations. Overall, High+Res is selected as the

preferred component-processing strategy. The corresponding forecasting curves are further compared in Figure 11.

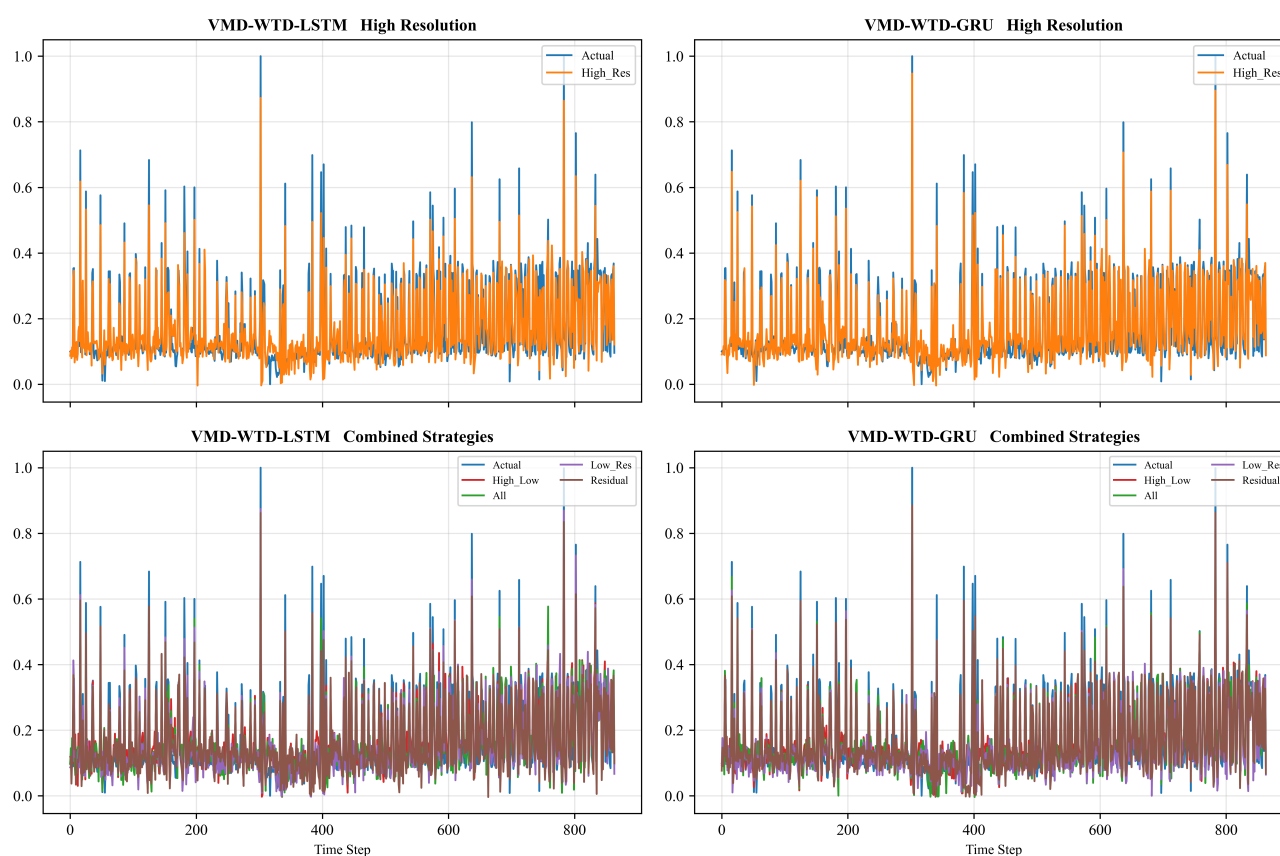


Figure 11. Forecasting performance of LSTM/GRU with different hybrid strategies.

4.5.5. Diebold–Mariano test

The prediction accuracy of different forecasting models has been compared in the previous sections. To further quantitatively distinguish whether the accuracy differences arise from inherent performance disparities among models or random errors, the DM test is employed to conduct significance tests on forecasting accuracy.

As shown in Table 9, most DM statistics are statistically significant, indicating that the forecasting differences among the compared models are robust. Among the baseline models, LSTM performs slightly better than GRU. The DM statistic for the comparison between these two models is -0.45, which is negative and statistically insignificant. This suggests that the two recurrent models have close forecasting ability on the original sequence. In contrast, Transformer performs worse than LSTM and GRU, with significant negative DM statistics of -3.51 and -4.71, respectively. This may be related to the limited sample size and strong local fluctuations of daily petroleum consumption data.

After VMD decomposition is introduced, all VMD-based models significantly outperform their corresponding baseline models. The DM statistics of VMD-BP, VMD-RNN, VMD-LSTM, VMD-GRU, and VMD-Transformer against their baseline models are 10.23, 7.66, 11.04, 8.76, and 5.42,

respectively, all significant at the 1% level. Among the VMD-based models, VMD-GRU performs best, significantly outperforming VMD-BP, VMD-RNN, VMD-LSTM, and VMD-Transformer, with DM statistics of 6.24, 4.79, 4.55, and 3.55, respectively. These results indicate that VMD improves the forecasting ability of neural networks, and GRU is more effective in modeling the decomposed IMF components.

The proposed VMD-WTD-GRU (High+Res) model achieves the best overall performance. Compared with VMD-GRU, it obtains a significant positive DM statistic of 4.10 at the 1% level, indicating that WTD-based denoising further improves forecasting accuracy after VMD decomposition. Moreover, VMD-WTD-GRU (High+Res) significantly outperforms VMD-WTD-LSTM (High+Res) at the 5% level, with a DM statistic of 1.96. These results confirm that applying WTD to high-frequency IMFs and the residual mode is an effective component-processing strategy, and they further verify the robustness, and superiority of the proposed VMD-WTD-GRU (High+Res) model.

Table 9. DM test results of different forecasting models.

Model	BP	RNN	LSTM	GRU	Transformer	VMD-BP	VMD-RNN	VMD-LSTM	VMD-GRU	VMD-Transformer	VMD-WTD-LSTM	VMD-WTD-GRU
BP	–											
RNN	7.04***	–										
LSTM	10.78***	3.90***	–									
GRU	11.15***	4.53***	-0.45	–								
Transformer	1.32	-1.90*	-3.51***	-4.71***	–							
VMD-BP	10.23***	7.00***	5.95***	4.64***	5.25***	–						
VMD-RNN	10.87***	7.66***	6.96***	5.36***	5.71***	7.89***	–					
VMD-LSTM	15.00***	10.26***	11.04***	7.97***	7.46***	3.56***	2.16**	–				
VMD-GRU	15.68***	11.21***	12.19***	8.76***	7.90***	6.24***	4.79***	4.55***	–			
VMD-Transformer	9.43***	6.82***	6.18***	4.83***	5.42***	1.76*	0.2887	-1.59	-3.55***	–		
VMD-WTD-LSTM	14.45***	11.15***	12.15***	8.96***	8.07***	10.60***	9.82***	7.63***	5.86***	9.67***	–	
VMD-WTD-GRU	11.05***	8.66***	8.84***	7.05***	6.83***	8.64***	7.98***	5.04***	4.10***	8.92***	1.96**	–

Note: The values in the table are DM statistics. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

4.5.6. Extended performance comparison with representative transformer-based forecasting models

To further assess the robustness of the proposed VMD-WTD-GRU model with mode-specific denoising, its performance is further compared with that of four recent Transformer-based forecasting models, which are Autoformer [53], PatchTST [54], FEDformer [55], and Informer [56], respectively. All models are evaluated under the same experimental conditions, and the results are presented in Table 10.

Table 10. Performance comparison of the proposed model with representative Transformer-based forecasting models.

Model	RMSE	MAE	MAPE (%)
Autoformer	675.82	449.66	57.73
PatchTST	701.93	453.38	65.38
FEDformer	673.53	475.43	66.47
Informer	668.65	413.26	51.02
Transformer	750.62	531.90	73.68
VMD-WTD-GRU (High+Res)	184.56	141.80	23.15

As shown in Table 10, Autoformer, PatchTST, FEDformer, and Informer are superior to vanilla Transformer in three evaluation indicators, which shows that these improvements can improve their prediction performance. Among them, Informer has the best performance, with an RMSE of 668.65,

MAE of 413.26, and MAPE of 51.02%. The proposed model achieves the optimal predictive performance. Compared with Informer, its RMSE, MAE, and MAPE are reduced by 72.40%, 65.69%, and 27.87%, respectively. Figure 12 shows their forecasting results. It can be observed that the Transformer-based models usually capture the overall time pattern but tend to smooth some local peaks and short-term changes, whereas the proposed model predicts these fluctuations more closely. These observations are consistent with the evaluation metrics in Table 10. The results indicate that the proposed VMD-WTD-GRU model with mode-specific denoising is well-suited to capturing the complex nonlinear and nonstationary fluctuations in petroleum consumption series. Moreover, recent Transformer variants specifically designed for time-series forecasting have shown clear improvements over the vanilla Transformer, and their potential for petroleum consumption forecasting warrants future research.

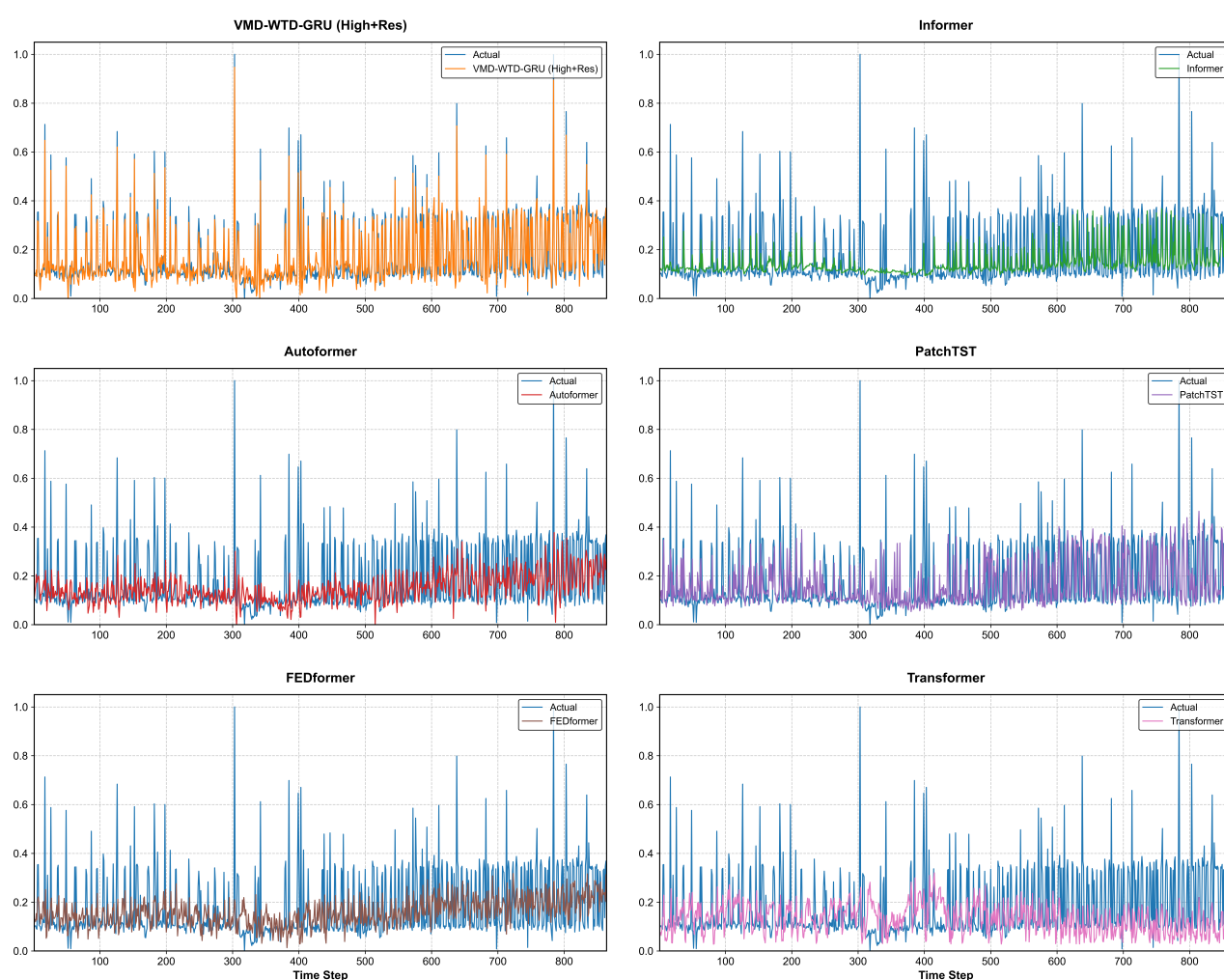


Figure 12. Forecasting result comparison between the proposed model and representative Transformer-based models.

5. Conclusions

This study investigates the effectiveness of decomposition- and denoising-based hybrid models for regional daily petroleum consumption forecasting. Several single deep learning models, VMD-based hybrid models, and denoising-enhanced hybrid models are constructed and compared to evaluate the role of modal decomposition, denoising optimization, and forecasting model selection. The experimental results show that directly applying deep learning models to the original petroleum consumption series still has limitations in capturing complex nonlinear and nonstationary fluctuations. After VMD decomposition, the forecasting performance of neural network models is generally improved, indicating that modal decomposition can effectively separate temporal patterns at different frequency scales and enhance the learnability of the input series.

Further analysis shows that the proposed VMD-WTD-GRU model with mode-specific denoising achieves the best overall forecasting performance. Compared with the conventional VMD-WTD-GRU model that applies wavelet threshold denoising to all decomposed components, the proposed mode-specific denoising strategy reduces the RMSE from 258.14 to 184.56, the MAE from 193.97 to 141.80, and the MAPE from 31.45% to 23.15%, corresponding to reductions of 28.50%, 26.90%, and 8.30%, respectively. These results indicate that denoising all decomposed modes may remove useful information together with noise, thereby limiting forecasting performance. In contrast, the proposed model selectively applies WTD to high-frequency IMF components and the residual component, and low-frequency components are retained to preserve the main trend information. This mode-specific denoising strategy can effectively suppress random noise and irregular disturbances without weakening useful low-frequency temporal patterns, leading to more accurate petroleum consumption forecasts.

In summary, the proposed VMD-WTD-GRU framework with mode-specific denoising achieves the best forecasting performance for regional petroleum energy consumption. The selective denoising of high-frequency IMF components and the residual component further improve forecasting accuracy while preserving important trend information. Comparative experiments and statistical tests confirm the effectiveness and reliability of the proposed approach. These findings provide useful methodological support for petroleum demand forecasting, regional energy planning, and intelligent energy management.

Besides the appealing performance of the proposed hybrid model, there still exist several potential drawbacks. In this research, the hyperparameter values selection is not systematically optimized according to the entire flow of the model. In this sense, metaheuristic optimization methods can be introduced to enable adaptive optimization of key hyperparameters, including the VMD decomposition parameters, WTD threshold parameters, and GRU network parameters. Furthermore, the deep learning forecasting module in this study mainly focuses on representative fundamental architectures, including LSTM, GRU, and the basic Transformer models. Other advanced deep learning models, particularly recent Transformer-based time-series forecasting methods, could also be incorporated into the proposed framework. Further investigation of how different forecasting architectures affect the overall performance of the hybrid framework is therefore a worthwhile direction for future research. In addition, the prediction model proposed by this research only relies on the energy consumption data. If more predictor values are introduced in the model, more sophisticated deep learning models can be adopted to further enhance the feature extraction capability

and forecasting precision of the proposed hybrid forecasting framework. Recent soft computing studies such as those by Wang et al. [57] and Zhao et al. [58] have also shown a broader tendency to combine data-driven modeling with domain knowledge, adaptive learning, and intelligent optimization mechanisms, such as knowledge-based population optimization, dual Q-learning, and heterogeneous graph reinforcement learning. All the above directions have huge potential to make meaningful research contributions and are worthy of further research.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Author contribution

Bowen Hou: Conceptualization, Methodology, Software, Investigation, Data curation, Formal analysis, Validation, Visualization, Writing – original draft, and Writing – review & editing; **Yin Liu:** Methodology, Investigation, and Writing – original draft; **Jiajun Chen:** Software, Validation, Visualization, and Writing – review & editing; **Letian Qi:** Software, Validation, Visualization, and Writing – review & editing; **Jinhao Zhang:** Investigation; **Honghao Zhao:** Conceptualization, Methodology, Supervision, and Writing – review & editing.

All authors have read and approved the final version of the manuscript.

Conflict of interest

The authors declare there are no conflicts of interest.

References

1. G. E. Halkos, E. C. Gkampoura, Reviewing usage, potentials, and limitations of renewable energy sources, *Energies*, **13** (2020), 2906. <https://doi.org/10.3390/en13112906>
2. A. Yi, M. Yang, Y. Li, Macroeconomic uncertainty and crude oil futures volatility—evidence from China crude oil futures market, *Front. Environ. Sci.*, **9** (2021), 636903. <https://doi.org/10.3389/fenvs.2021.636903>
3. J. Zhang, Y. Zhu, X. Zhang, M. Ye, J. Yang, Developing a Long Short-Term Memory (LSTM) based model for predicting water table depth in agricultural areas, *J. Hydrol.*, **561** (2018), 918–929. <https://doi.org/10.1016/j.jhydrol.2018.04.065>
4. Y. Huang, L. Yang, S. Liu, G. Wang, Multi-step wind speed forecasting based on ensemble empirical mode decomposition, long short term memory network and error correction strategy, *Energies*, **12** (2019), 1822. <https://doi.org/10.3390/en12101822>
5. Z. Deng, B. Wang, H. Guo, C. Chai, Y. Wang, Z. Zhu, Unified quantile regression deep neural network with time-cognition for probabilistic residential load forecasting, *Complexity*, **2020** (2020), 9147545. <https://doi.org/10.1155/2020/9147545>

6. J. Zhang, X. Xin, Y. Shang, Y. Wang, L. Zhang, Nonstationary significant wave height forecasting with a hybrid VMD-CNN model, *Ocean Eng.*, **285** (2023), 115338. <https://doi.org/10.1016/j.oceaneng.2023.115338>
7. S. Sun, Y. Wang, Y. Meng, C. Wang, X. Zhu, Multi-step wind speed forecasting model using a compound forecasting architecture and an improved QPSO-based synchronous optimization, *Energy Rep.*, **8** (2022), 9899–9918. <https://doi.org/10.1016/j.egyr.2022.07.164>
8. J. Xu, Y. Wei, P. Zeng, VMD-based iterative Boruta feature extraction and CNNA-BiLSTM for short-term load forecasting, *Electr. Power Syst. Res.*, **238** (2025), 111172. <https://doi.org/10.1016/j.epsr.2024.111172>
9. I. Jrhilifa, H. Ouadi, A. Jilbab, N. Mounir, A. Ouaguid, A VMD-deep learning approach for individual load monitoring and forecasting for residential buildings energy management, *e-Prime*, **8** (2024), 100624. <https://doi.org/10.1016/j.prime.2024.100624>
10. L. Liu, J. Zhang, S. Xue, Photovoltaic power forecasting: Using wavelet threshold denoising combined with VMD, *Renew. Energy*, **249** (2025), 123152. <https://doi.org/10.1016/j.renene.2025.123152>
11. X. Guo, Y. Gao, W. Song, Y. Zen, X. Shi, Short-term power load forecasting based on CEEMDAN-WT-VMD joint denoising and BiTCN-BiGRU-Attention, *Electronics*, **14** (2025), 1871. <https://doi.org/10.3390/electronics14091871>
12. S. S. Pappas, L. Ekonomou, D. C. Karamousantas, G. Chatzarakis, S. Katsikas, P. Liatsis, Electricity demand loads modeling using AutoRegressive Moving Average (ARMA) models, *Energy*, **33** (2008), 1353–1360. <https://doi.org/10.1016/j.energy.2008.05.008>
13. A. Azadeh, S. M. Asadzadeh, M. Saberi, A Neuro-fuzzy-stochastic frontier analysis approach for long-term natural gas consumption forecasting and behavior analysis: The cases of Bahrain, Saudi Arabia, Syria, and UAE, *Appl. Energy*, **88** (2011), 3850–3859. <https://doi.org/10.1016/j.apenergy.2011.04.027>
14. A. Azadeh, R. Babazadeh, S. Asadzadeh, Optimum estimation and forecasting of renewable energy consumption by artificial neural networks, *Renew. Sustain. Energy Rev.*, **27** (2013), 605–612. <https://doi.org/10.1016/j.rser.2013.07.007>
15. J. Bai, J. Wang, J. Ran, X. Li, C. Tu, An improved neural network algorithm for energy consumption forecasting, *Sustainability*, **16** (2024), 9332. <https://doi.org/10.3390/su16219332>
16. L. G. B. Ruiz, R. Rueda, M. P. Cuéllar, M. Pegalajar, Energy consumption forecasting based on Elman neural networks with evolutive optimization, *Expert Syst. Appl.*, **92** (2018), 380–389. <https://doi.org/10.1016/j.eswa.2017.09.059>
17. G. P. Zhang, Time series forecasting using a hybrid ARIMA and neural network model, *Neurocomputing*, **50** (2003), 159–175. [https://doi.org/10.1016/S0925-2312\(01\)00702-0](https://doi.org/10.1016/S0925-2312(01)00702-0)
18. L. Wang, H. Zou, J. Su, L. Li, S. Chaudhry, An ARIMA-ANN hybrid model for time series forecasting, *Syst. Res. Behav. Sci.*, **30** (2013), 244–259. <https://doi.org/10.1002/sres.2179>
19. S. Barak, S. S. Sadegh, Forecasting energy consumption using ensemble ARIMA–ANFIS hybrid algorithm, *Int. J. Electr. Power Energy Syst.*, **82** (2016), 92–104. <https://doi.org/10.1016/j.ijepes.2016.03.012>

20. N. Shabbir, L. Kutt, M. Jawad, M. N. Iqbal, P. S. Ghahfaroki, Forecasting of energy consumption and production using recurrent neural networks, *Adv. Electr. Electron. Eng.*, **18** (2020), 190–197.
21. K. Ma, X. Nie, J. Yang, L. Zha, G. Li, H. Li, A power load forecasting method in port based on VMD-ICSS-hybrid neural network, *Appl. Energy*, **377** (2025), 124246. <https://doi.org/10.1016/j.apenergy.2024.124246>
22. K. Sharma, Y. K. Dwivedi, B. Metri, Incorporating causality in energy consumption forecasting using deep neural networks, *Ann. Oper. Res.*, **339** (2024), 537–572. <https://doi.org/10.1007/s10479-022-04857-3>
23. S. Y. Shin, H. G. Woo, Energy consumption forecasting in Korea using machine learning algorithms, *Energies*, **15** (2022), 4880. <https://doi.org/10.3390/en15134880>
24. A. Dedinec, S. Filiposka, A. Dedinec, L. Kocarev, Deep belief network based electricity load forecasting: An analysis of Macedonian case, *Energy*, **115** (2016), 1688–1700. <https://doi.org/10.1016/j.energy.2016.07.090>
25. Y. Dong, X. Ma, T. Fu, Electrical load forecasting: A deep learning approach based on K-nearest neighbors, *Appl. Soft Comput.*, **99** (2021), 106900. <https://doi.org/10.1016/j.asoc.2020.106900>
26. M. Alipour, J. Aghaei, M. Norouzi, T. Niknam, S. Hashemi, M. Lehtonen, A novel electrical net-load forecasting model based on deep neural networks and wavelet transform integration, *Energy*, **205** (2020), 118106. <https://doi.org/10.1016/j.energy.2020.118106>
27. Y. Wang, N. Zhang, Q. Chen, D. S. Kirschen, P. Li, Q. Xia, Data-driven probabilistic net load forecasting with high penetration of behind-the-meter PV, *IEEE Trans. Power Syst.*, **33** (2018), 3255–3264. <https://doi.org/10.1109/TPWRS.2017.2762599>
28. D. W. Van der Meer, J. Munkhammar, J. Widén, Probabilistic forecasting of solar power, electricity consumption and net load: Investigating the effect of seasons, aggregation and penetration on prediction intervals, *Sol. Energy*, **171** (2018), 397–413. <https://doi.org/10.1016/j.solener.2018.06.103>
29. F. Mei, Q. Wu, T. Shi, J. Lu, Y. Pan, J. Zheng, An ultrashort-term net load forecasting model based on phase space reconstruction and deep neural network, *Appl. Sci.*, **9** (2019), 1487. <https://doi.org/10.3390/app9071487>
30. I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, et al., Generative adversarial nets, in *Advances in Neural Information Processing Systems 27*, Curran Associates, Inc., (2014), 2672–2680.
31. Z. Wang, T. Hong, Generating realistic building electrical load profiles through the Generative Adversarial Network (GAN), *Energy Build.*, **224** (2020), 110299. <https://doi.org/10.1016/j.enbuild.2020.110299>
32. M. N. Fekri, A. M. Ghosh, K. Grolinger, Generating energy data for machine learning with recurrent generative adversarial networks, *Energies*, **13** (2020), 130. <https://doi.org/10.3390/en13010130>
33. M. Ranjbar, M. Rahimzadeh, Advancing gasoline consumption forecasting: A novel hybrid model integrating transformers, lstm, and cnn, preprint, arXiv:2410.16336.

34. J. He, Y. Li, X. Xu, D. Wu, Energy consumption forecasting for oil and coal in China based on hybrid deep learning, *PLOS One*, **20** (2025), e0313856. <https://doi.org/10.1371/journal.pone.0313856>
35. C. Zhai, X. He, Z. Cao, M. Abdou-Tankari, Y. Wang, M. Zhang, Photovoltaic power forecasting based on VMD-SSA-Transformer: Multidimensional analysis of dataset length, weather mutation and forecast accuracy, *Energy*, **324** (2025), 135971. <https://doi.org/10.1016/j.energy.2025.135971>
36. D. Venkateswaran, R. Venkateswaran, C. Shin, W. Choi, Y. Cho, Novel VMD-Informer-LSTM architecture for solar radiation forecasting with signal decomposition and sparse attention for smart energy applications, *Expert Syst. Appl.*, **305** (2026), 130612. <https://doi.org/10.1016/j.eswa.2025.130612>
37. J. Liu, Y. Zhang, K. Wen, Y. Ding, Analysis of different neural network models based on variational mode decomposition and dung beetle optimizer algorithm for the prediction of air-conditioning energy consumption in multifunctional complex large public buildings, *Energy Build.*, **334** (2025), 115518. <https://doi.org/10.1016/j.enbuild.2025.115518>
38. Y. Ruan, G. Wang, H. Meng, F. Qian, A hybrid model for power consumption forecasting using VMD-based the long short-term memory neural network, *Front. Energy Res.*, **9** (2022), 772508. <https://doi.org/10.3389/fenrg.2021.772508>
39. Y. Yang, W. Li, D. Liu, Monthly runoff prediction for Xijiang River via gated recurrent unit, discrete wavelet transform, and variational modal decomposition, *Water*, **16** (2024), 1552. <https://doi.org/10.3390/w16111552>
40. Y. Yu, Q. Shang, T. Xie, A hybrid model for short-term traffic flow prediction based on variational mode decomposition, wavelet threshold denoising, and long short-term memory neural network, *Complexity*, **2021** (2021), 7756299. <https://doi.org/10.1155/2021/7756299>
41. L. Wang, M. An, W. Zhu, W. Xiang, R. He, Forecasting of short-term traffic flow using LSTM ensembled with variational mode decomposition and wavelet denoising, *Int. J. Intell. Syst. Technol. Appl.*, **24** (2026), 1–27. <https://doi.org/10.1504/IJISTA.2026.152033>
42. N. Fang, Z. Liu, S. Fan, Short-term wind power prediction method based on CEEMDAN-VMD-GRU hybrid model, *Energies*, **18** (2025), 1465. <https://doi.org/10.3390/en18061465>
43. W. Jiang, Q. Lu, Y. Xu, Z. Chen, T. Gong, Short-term wind speed prediction based on denoising algorithm of enhanced successive variational mode decomposition and integrated parallel prediction model, *Energy*, **338** (2025), 138857. <https://doi.org/10.1016/j.energy.2025.138857>
44. J. Xiao, Y. Li, L. Xie, D. Liu, J. Huang, A hybrid model based on selective ensemble for energy consumption forecasting in China, *Energy*, **159** (2018), 534–546. <https://doi.org/10.1016/j.energy.2018.06.161>
45. T. Shojaei, A. Mokhtar, Forecasting energy consumption with a novel ensemble deep learning framework, *J. Build. Eng.*, **96** (2024), 110452. <https://doi.org/10.1016/j.jobee.2024.110452>
46. K. Dragomiretskiy, D. Zosso, Variational Mode Decomposition, *IEEE Trans. Signal Process.*, **62** (2014), 531–544. <https://doi.org/10.1109/TSP.2013.2288675>

47. Y. Hou, X. Deng, Y. Xia, Precipitation prediction based on variational mode decomposition combined with the crested porcupine optimization algorithm for long short-term memory model, *AIP Adv.*, **14** (2024), 065315. <https://doi.org/10.1063/5.0204644>
48. D. P. Bertsekas, Multiplier methods: A survey, *Automatica*, **12** (1976), 133–145. [https://doi.org/10.1016/0005-1098\(76\)90077-7](https://doi.org/10.1016/0005-1098(76)90077-7)
49. M. Han, Y. Liu, J. Xi, W. Guo, Noise smoothing for nonlinear time series using wavelet soft threshold, *IEEE Signal Process. Lett.*, **14** (2007), 62–65. <https://doi.org/10.1109/LSP.2006.881518>
50. Y. M. Liu, Application of improved VMD and wavelet threshold denoising method in bolt detection, in *2020 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR)*, IEEE, (2020), 14–20. <https://doi.org/10.1109/ICWAPR51924.2020.9494619>
51. D. Valencia, D. Orejuela, J. Salazar, J. Valencia, Comparison analysis between rigrsure, sqtwolog, heursure and minimaxi techniques using hard and soft thresholding methods, in *2016 XXI Symposium on Signal Processing, Images and Artificial Vision (STSIVA)*, IEEE, (2016), 1–5. <https://doi.org/10.1109/STSIVA.2016.7743309>
52. F. X. Diebold, R. S. Mariano, Comparing predictive accuracy, *J. Bus. Econ. Stat.*, **13** (1995), 253–263. <https://doi.org/10.1080/07350015.1995.10524599>
53. Y. Yang, Y. Gao, H. Zhou, J. Wu, S. Gao, Y. G. Wang, Multi-Granularity Autoformer for long-term deterministic and probabilistic power load forecasting, *Neural Netw.*, **188** (2025), 107493. <https://doi.org/10.1016/j.neunet.2025.107493>
54. Z. Qu, Y. Meng, X. Hou, R. Chi, Y. Ai, Z. Wu, Integrated energy short-term multivariate load forecasting based on PatchTST secondary decoupling reconstruction for progressive layered extraction multi-task learning network, *Expert Syst. Appl.*, **269** (2025), 126446. <https://doi.org/10.1016/j.eswa.2025.126446>
55. W. Zhao, J. Zhao, X. Shen, J. Liu, S. Fan, A two-stage uncertainty forecasting method for offshore wind power based on FEDformer and SDCDM, *Energy*, **348** (2026), 140452. <https://doi.org/10.1016/j.energy.2026.140452>
56. Z. Yang, J. Li, H. Wang, C. Liu. An Informer model for very short-term power load forecasting, *Energies*, **18** (2025), 1150. <https://doi.org/10.3390/en18051150>
57. H. Wang, B. R. Sarker, J. Li, J. Li, Adaptive scheduling for assembly job shop with uncertain assembly times based on dual Q-learning, *Int. J. Prod. Res.*, **59** (2021), 5867–5883. <https://doi.org/10.1080/00207543.2020.1794075>
58. F. Zhao, Z. Fu, L. Wang, H. Sang, A heterogeneous graph reinforcement learning framework with question-aware neighborhood aggregation and interoption prompt attention for dynamic flexible job shop scheduling problem, *IEEE Trans. Ind. Inform.*, **22** (2026), 2863–2874. <https://doi.org/10.1109/TII.2025.3646962>

