



Multi-step-ahead and interval carbon price forecasting using transformer-based hybrid model

Wang Yue¹ · Wang Zhong¹ · Wang Xiaoyi¹ · Kang Xinyu¹

Received: 10 November 2022 / Accepted: 2 August 2023 / Published online: 9 August 2023
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2023

Abstract

Accurate and stable carbon price forecasts serve as a reference for assessing the stability of the carbon market and play a vital role in enhancing investment and operational decisions. However, realizing this goal is still a significant challenge, and researchers usually ignore multi-step-ahead and interval forecasting due to the non-linear and non-stationary characteristics of carbon price series and its complex fluctuation features. In this study, a novel hybrid model for accurately predicting carbon prices is proposed. The proposed model combines multi-step-ahead and interval carbon price forecasting based on the Hampel identifier (HI), time-varying filtering-based empirical mode decomposition (TVFEMD), and transformer model. First, HI identifies and corrects outliers in carbon price. Second, TVFEMD decomposes carbon price into several intrinsic mode functions (imfs) to reduce the non-linear and non-stationarity of carbon price to obtain more regular features in series. Next, these imfs are reconstructed by sample entropy (SE). Subsequently, the orthogonal array tuning method is used to optimize the transformer model's hyperparameters to obtain the optimal model structure. Finally, after hyperparameter optimization and quantile loss function, the transformer is used to perform multi-step-ahead and interval forecasting on each part of the reconstruction, and the final prediction result is obtained by summing them up. Five pilot carbon trading markets in China were selected as experimental objects to verify the proposed model's prediction performance. Various benchmark models and evaluation indicators were selected for comparison and analysis. Experimental results show that the proposed HI-TVFEMD-transformer hybrid model achieves an average MAE of 0.6546, 1.3992, 1.6287, and 2.2601 for one-step, three-step, five-step, and ten-step-ahead forecasting, respectively, which significantly outperforms other models. Furthermore, interval forecasts almost always have a PICI above 0.95 at a confidence interval of 0.1, thereby indicating the effectiveness of the hybrid model in describing the uncertainty in the forecasts. Therefore, the proposed hybrid model is a reliable carbon price forecasting tool that can provide a dependable reference for policymakers and investors.

Keywords Multi-step-ahead forecasting · Interval forecasting · TVFEMD · OATM · Transformer · Carbon price

Introduction

Global warming stands as the greatest challenge confronting humanity in the twenty-first century, presenting a great threat to human survival and development. Greenhouse gas emissions (mainly carbon dioxide) from human industrialization activities are the direct cause of global warming. Many measures have been proposed to reduce emissions

and control the trend of global warming. The emission trading system (ETS) was first proposed in the Kyoto Protocol, which assigns a price to carbon and is recognized as an effective policy tool to control emissions. In the 26th UN Climate Change Conference of the Parties (COP26), the importance of ETS for climate change was emphasized, and the global carbon trading system was initially finalized. Different from other financial markets, the carbon market emerged late and is characterized by an immature market system that is highly susceptible to other external factors, such as market regulation, energy, and environmental policies (Sun and Zhang 2018). These factors lead to dramatic fluctuations in carbon price. Therefore, building an accurate carbon price forecasting model that can help policymakers and enterprises understand the fluctuation pattern of carbon

Responsible Editor: Ilhan Ozturk

✉ Wang Zhong
wangzhong2012@cdut.edu.cn

¹ College of Management Science, Chengdu University of Technology, Chengdu 610059, China

price and thus develop relevant policies and investment strategies is necessary (Hao et al. 2020). However, the non-linear, uncertain, and complex nature of carbon price fluctuations (Lutz et al. 2013) makes the accurate forecasting of carbon price series a significant challenge.

The importance of carbon prices has garnered substantial attention among researchers in recent years, thereby leading to a growing interest in the field of carbon price forecasting. Many scholars have proposed advanced carbon price forecasting models, and various models have been applied to carbon price forecasting. In general, these models can be divided into three main categories (Zhu et al. 2017), namely, (1) traditional time series models, (2) artificial intelligence models, and (3) hybrid models. Traditional time series models are primarily based on statistical methods. For example, autoregressive integrated moving average (ARIMA) (Zhu and Wei 2013), autoregressive conditional heteroscedasticity (GARCH) (Byun and Cho 2013); (Benz and Trück 2009), and other models. These statistical methods are based on the significance of constructing models that can capture certain statistical features of carbon price fluctuations, such as heteroscedasticity, fat-tailedness, and leverage effects. Although the construction of statistical models is simple, easy to implement, widely applied, and has achieved specific results, statistical models based on the assumption of linearity cannot effectively address these characteristics (Sun and Zhang 2018) because of the non-linear and non-stationarity characteristics of carbon prices (Tian and Hao 2020). Therefore, achieving satisfactory accuracy in carbon price forecasting using these traditional time series models is difficult.

In recent years, the rise of artificial intelligence (AI) has prompted the utilization of many AI models for carbon price forecasting. Compared with traditional time series models, these AI models exhibit better robustness and generalization ability and can effectively deal with non-linear and non-stationary time series (Han et al. 2019). The most commonly used models include support vector machines (SVMs), extreme learning machines (ELMs), and several types of artificial neural networks (ANNs). Yi et al. (2017) predicted carbon prices through back propagation neural network (BPNN). The experimental results proved that this model exhibited higher prediction accuracy than the statistical model. Du et al. (2022) employed a BPNN to predict prices in Fujian's carbon market. They verified that BPNN could make effective carbon price predictions. These studies illustrate that AI models significantly improved prediction accuracy and could be adapted to various situations. Among these AI models, ANN-based models exhibit exceptional performance. Zhang and Wen (2022) predicted carbon price by using temporal convolutional neural network (TCN), and their experimental findings demonstrated the superior performance of TCN compared with traditional statistical models and some machine learning models, such as random

forest (RF), XGBoost, and SVM. Therefore, ANN-based carbon price prediction models, such as LSTM, GRU, and TCN, have emerged as prominent contenders in the field. These sequential ANN models possess excellent ability for time series modeling. However, the failure to effectively handle long-term dependencies and interrelatedness in time series has left room for further improvement in carbon price forecasting accuracy. Transformer model based on attention mechanisms is more effective in the long-term dependence and interactions of time series data than other ANN structures (Wu et al. 2020), thereby appearing in time series modeling. Bommidi et al. (2023) utilized the transformer model for short-term wind speed prediction, and the experimental results proved that the prediction ability of the transformer model is higher than the commonly used sequential ANN structure. Wang et al. (2022) constructed a temporal fusion transform (TFT) to predict carbon prices in the Chinese pilot market, and experiments showed that TFT has superior prediction effects compared with LSTM and GRU. Therefore, the transformer model holds promising application prospects in carbon price forecasting.

Nevertheless, each single model has corresponding defects in the face of the current highly volatile and chaotic carbon price (Huang et al. 2021). The prediction accuracy of using single models is not the best and most ideal (Niu and Wang 2019). Many advanced hybrid models have been proposed and used. Hybrid models can be mainly classified into two categories, namely, by using decomposition techniques and integrated optimization algorithms. Many researchers employ signal decomposition techniques to decompose carbon price series, which reduces non-stationarity and extracts the different scale features of carbon prices (Zhu et al. 2018), thereby improving the predictive accuracy. Many decomposition methods that are used, commonly include empirical mode decomposition (EMD-Type), variational mode decomposition (VMD), and wavelet transform. Yun et al. (2022) constructed a hybrid carbon price prediction model that included complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) and long short-term memory (LSTM). The experimental results show that the hybrid model exhibited better prediction performance than the single model. Sun et al. (2016) predicted the carbon price based on VMD and spiking neural networks (SNNs) and achieved good results. In experiments, VMD has demonstrated better feature extraction performance than EMD, which resulted in higher prediction accuracy. Liu and Shen (2019) combined the empirical wavelet transform (EWT) and gated recurrent unit (GRU) neural network to establish a hybrid model for carbon price prediction and verified through experiments that the model was superior to a single ARIMA, BPNN, and GRU. The decomposition algorithms, including EMD-type, VMD, and WT, offer effective ways to decompose the carbon price series. However, they possess

certain drawbacks. VMD and WT require manual selection of decomposition level and wavelet basis function, respectively (Yang et al. 2021). EMD-type is an adaptive decomposition algorithm but is susceptible to mode aliasing phenomenon and be affected by end effects (Wu and Huang 2004). These issues can affect the quality of decomposition and thus lead to unstable forecasting outcomes. Time-varying filtering-based empirical mode decomposition (TVFEMD) was proposed by Li et al. 2017. TVFEMD is an adaptive method that can effectively overcome mode mixing aliasing and has high computational efficiency. Due to its superior performance, it has been applied in some time series forecasting, such as non-ferrous metal price forecast (Wang et al. 2021) and wind speed prediction (Xiong et al. 2021). Therefore, this study applies TVFEMD to decompose carbon price. Many researchers use optimization algorithms to optimize the model's hyperparameters and thus improve its processing power. Considering that the performance of some AI models relies on the configuration of hyperparameters, especially ANN models (Zhang et al. 2017) and the accuracy of the model may fluctuate greatly with different hyperparameter settings, many scholars have employed intelligent optimization algorithms to optimize the configuration of model hyperparameters. Sun and Xu (2021) utilized linearly decreasing weight particle swarm optimization (LDWPSO) to optimize the hyperparameters of the wavelet least squares SVM (wLSSVM) and applied it to carbon price prediction. The optimized wLSSVM model demonstrated higher prediction accuracy than the non-optimized model. Sun and Zhang (2020) utilized an improved bat algorithm (IBA) to search for the optimal hyperparameters of extreme learning machine (ELM), which further enhanced ELM's predictive performance. The selection of a proper optimization algorithm is crucial for model configuration. However, heuristic optimization algorithms often struggle to ensure the stability of their optimization results and can easily fall into local optima. In this study, an orthogonal array-based hyperparameter optimization technique was used to identify the optimal hyperparameter configuration of the model.

In previous discussions, the majority of forecasting methods primarily emphasize point forecast, thereby ignoring the importance of interval forecasting, which cannot be disregarded in forecasting. Interval forecasts quantify the uncertainty in carbon price forecasts and therefore contain more information than point forecasts (Zhang et al. 2016). Interval forecast is based on a certain level of significance of a set of upper and lower bounds. Compared with point forecast, it can reflect the possibility of result variation caused by uncertainty in the prediction. When time series exhibit high instability and non-linear trends, interval forecast is a powerful tool to help decision makers. Quantile regression (QR) is a regression analysis model in statistics that estimates the conditional quantile relationship between variables without

knowing the type of variable distribution (He and Li 2018). QR can be well combined with ANN models to produce interval prediction results. Wang et al. (2020a, 2020b) combined LSTM with QR for short-term wind speed prediction, and the interval prediction results constructed by QR could cover the uncertainty of wind speed prediction well. Lim et al. (2021) integrated QR and transformer-based model and achieved remarkable interval forecasts. Therefore, this study combines the transformer model with QR for the interval forecasting of carbon prices. Furthermore, most models tend to overlook the presence of outliers in carbon price. In reality, carbon prices are commonly disturbed by non-controllable and unexpected non-repetitive information, such as the enactment of new market regulation policies, the upcoming date of carbon quota submission, or extreme weather. These factors can result in outliers within the carbon price series, potentially leading the model to learn erroneous information and leads to overfitting. Consequently, the generalization ability of the model is compromised. Sun et al. (2021) demonstrated that outliers in the carbon price series could negatively affect forecasts by using the box plot method to remove outliers from the carbon price. Hence, the identification and correction of outliers play a significant role in carbon price forecasting.

In summary, to fill the gap of the current research, this study constructs a hybrid carbon price forecasting method that combines Hampel identifier (HI), time-varying filtering for empirical mode decomposition (TVFEMD), transformer, and optimization with orthogonal array tuning (OATM) for multi-step-ahead and interval forecasting of carbon prices. First, HI is used to identify and correct outliers present in the original carbon price series to eliminate their negative impact on carbon price data. Subsequently, TVFEMD decomposes the processed carbon price series into multiple intrinsic mode functions (imfs), which are reconstructed through sample entropy (SE). Finally, based on the transformer model after the hyperparameters are optimized by OATM, each imfs is predicted, and the results are summed to obtain the deterministic prediction results. The quantile loss function is used to obtain the prediction intervals at various confidence levels. The main innovations and contributions of this study are outlined as follows:

- (1) Identification and correction of outliers. Carbon prices are irregular and fluctuate dramatically. By identifying and correcting the outliers in the original carbon price through HI, the basic information in the original data can be retained better, and the training of the model is not affected by the outliers, thus improving the prediction performance of the hybrid model.
- (2) An advanced data decomposition strategy is used. TVFEMD improves EMD. TVFEMD can eliminate the mode aliasing phenomenon in EMD and can adaptively decompose the original data into clearer and more

detailed sub-sequences, thereby effectively reducing the non-linear and non-stationarity characteristics in carbon price series.

- (3) Advanced deep learning architecture is used. The deep learning (DL) model is widely used in carbon price forecasting. This study is the first to attempt to use the transformer model in carbon price forecasting, explore the potential of the transformer’s application in carbon pricing forecasting, and expand on related theoretical methods because this model has better modeling ability in terms of long-term dependence and interaction of series. Subsequently, because the performance of deep learning models is very sensitive to the configuration of hyperparameters, OATM is used in this study to optimize the hyperparameters in the transformer model. OATM makes a trade-off between time consumption and model accuracy to select the optimal hyperparameter configuration.
- (4) Multi-step-ahead and interval forecasts of carbon prices are constructed. Considering that the results of one-step-ahead forecasts are not sufficient to provide reliable information for investors and policymakers, the multi-step-ahead and interval of carbon price are predicted, which can better reflect the changing trend and fluctuation pattern of carbon price, based on the transformer’s excellent modeling ability of long-term dependence of series and quantile loss function. Multi-step-ahead prediction can represent the long-term trend of carbon price, whereas interval prediction can comprehensively reflect the fluctuation information of carbon price.

The remainder of this paper is presented as follows: “Methodology” presents the methodology used in this study. “Framework of the proposed forecasting system” describes the framework flow of the proposed hybrid model. “Data collection and preprocessing” contains the description of the data and the evaluation metrics. “Experimental analysis” presents the analysis and discussion of the experiments. Finally, “Conclusion” presents the conclusions.

Methodology

In this section, the main methods used in the proposed model of this study, including HI, TVFEMD, transformer, OATM, and quantile loss, are introduced.

Hampel identifier

HI is an effective outlier identification and correction technique that utilizes median deviation and absolute median deviation as the criteria for determining outliers with good robustness (Yao et al. 2019). For the data series: $X = \{x_1, x_2, x_3, \dots, x_n\}$, k is the number of adjacent points on each side of x_i in a given

window, the size of the moving window is $2k+1$, the local median m_i , and the scale estimate of the median estimated deviation σ_i are calculated as follows:

$$m_i = \text{median}(x_{i-k}, x_{i-k+1}, \dots, x_i, \dots, x_{i+k-1}, x_{i+k}), \tag{1}$$

$$\sigma_i = k \cdot \text{median}(|x_{i-k} - m_i|, \dots, |x_{i+k} - m_i|), \tag{2}$$

where $k = \left(1 / \left(\sqrt{2} \text{erfc}^{-1}(1/2)\right)\right) \approx 1.4826$ is the unbiased estimate of the Gaussian distribution.

$$y_i = \begin{cases} x_i, & |x_i - m_i| \leq n_\sigma \sigma_i \\ \text{outlier}, & |x_i - m_i| > n_\sigma \sigma_i \end{cases} \tag{3}$$

According to Eq. 3, if the absolute value of the difference between the evaluated data x_i and the local median m_i is greater than n_σ times σ_i , then the evaluated data x_i are considered an outlier, and the outlier is replaced by the local median m_i , and n_σ is usually set to 3 (Wang et al. 2020a, 2020b).

Time-varying filter empirical mode decomposition

TVFEMD, which is an improvement of the empirical modal decomposition (EMD), is proposed by Li et al. (2017). The decomposition of the imfs in EMD and EEMD cannot guarantee the existence of only one amplitude mode, and the local mean values determined by EMD through the third spline interpolation of the upper and lower envelopes are difficult to represent by a strict analytical expression, which may lead to meaningless imfs obtained by EMD and EEMD. TVFEMD improves the performance of frequency separation and stability at low sampling rates, solves the problem of modal aliasing, and preserves the time-varying characteristics of the time series. The meaning of the TVFEMD parameter is clear and simple to choose (Ma and Zhang 2020), which allows the imfs obtained by TVFEMD to represent the pattern features in the time series better.

The specific calculation steps of TVFEMD are presented as follows:

The instantaneous amplitude $A(t)$ and frequency $\varphi'(t)$ of the original sequence $x(t)$ are obtained by Hilbert transform; then, the local maximum value $A(t_{\max})$ and the local minimum value $A(t_{\min})$ of $A(t)$ are solved. The interpolation of $A(t_{\max})$ and $A(t_{\min})$ yields $\beta_1(t)$ and $\beta_2(t)$; the instantaneous mean $\alpha_1(t) = (\beta_1(t) + \beta_2(t)) / 2$ and the instantaneous envelope $\alpha_2(t) = (\beta_1(t) - \beta_2(t)) / 2$ are calculated. $\eta_1(t)$ and $\eta_2(t)$ are obtained by the interpolation of $\varphi'(t_{\max})A^2(t_{\max})$ and $\varphi'(t_{\min})A^2(t_{\min})$.

$$\eta_1(t) = \varphi'_1(t)[\alpha_1^2(t) - \alpha_1(t)\alpha_2(t)] + \varphi'_2(t)[\alpha_2^2(t) - \alpha_1(t)\alpha_2(t)], \tag{4}$$

$$\eta_2(t) = \varphi'_1(t)[\alpha_1^2(t) + \alpha_1(t)\alpha_2(t)] + \varphi'_2(t)[\alpha_2^2(t) + \alpha_1(t)\alpha_2(t)]. \tag{5}$$

Subsequently, $\varphi'_1(t)$ and $\varphi'_2(t)$ are calculated by Eqs. 6 and 7, respectively.

$$\varphi'_1(t) = \frac{\eta_1(t)}{2\alpha_1^2(t) - 2\alpha_1(t)\alpha_2(t)} + \frac{\eta_2(t)}{2\alpha_1^2(t) + 2\alpha_1(t)\alpha_2(t)}, \quad (6)$$

$$\varphi'_2(t) = \frac{\eta_1(t)}{2\alpha_2^2(t) - 2\alpha_1(t)\alpha_2(t)} + \frac{\eta_2(t)}{2\alpha_2^2(t) + 2\alpha_1(t)\alpha_2(t)}. \quad (7)$$

Calculate the local cutoff probability $\varphi'_{\text{bis}}(t)$.

$$\varphi'_{\text{bis}}(t) = \frac{\varphi'_1(t) + \varphi'_2(t)}{2} = \frac{\eta_2(t) - \eta_1(t)}{4\alpha_1(t)\alpha_2(t)}. \quad (8)$$

The signal $h(t)$ can be extracted by $\varphi'_{\text{bis}}(t)$, and $h(t) = \cos \left[\int \varphi'_{\text{bis}}(t) dt \right]$. The signal $x(t)$ is approximated by B-sample interpolation of $h(t)$ extreme points $h(t_{\text{max}})$ and $h(t_{\text{min}})$, and the result of the approximation is $m(t)$.

Calculate the stopping criterion $\theta(t)$. When $\theta(t)$ is less than the threshold ξ , $x(t)$ is treated as an imf, otherwise, set $x(t) = x(t) - m(t)$, repeat the steps above until the stopping criterion is satisfied.

$$\theta(t) = \frac{B_{\text{Loughlim}}(t)}{\varphi_{\text{avg}}(t)}, \quad (9)$$

$$B_{\text{Loughlim}}(t) = \sqrt{\frac{\alpha_1^2(t) + \alpha_2^2(t)}{\alpha_1^2(t) + \alpha_2^2(t)} + \frac{\alpha_1^2(t)\alpha_2^2(t)(\varphi'_1(t) - \varphi'_2(t))}{(\alpha_1^2(t) + \alpha_2^2(t))^2}}, \quad (10)$$

$$\varphi_{\text{avg}}(t) = \frac{\alpha_1^2(t)\varphi'_1(t) + \alpha_2^2(t)\varphi'_2(t)}{\alpha_1^2(t) + \alpha_2^2(t)}. \quad (11)$$

The original sequence $x_{\text{raw}}(t)$ is finally decomposed into several imfs and a residual term, $x_{\text{raw}}(t) = \sum \text{IMF}_k + \text{Resid}$.

Transformer

The transformer model has gained significant attention in time series modeling because of its ability to efficiently capture long-term dependencies in sequences and interactions between data. Unlike TCN or LSTM that rely on recursive and convolutional layers, transformer models the information in time series through the encoder–decoder structure (Wen et al. 2022). The encoder maps the input sequence $X=[x_1, x_2, \dots, x_n]$ into a continuous representation $Z=[z_1, z_2, \dots, z_n]$, and Z inputs the decoder to generate the output sequence $Y=[y_1, y_2, \dots, y_m]$. The encoder and decoder are stacked with N identical modules, and each module mainly contains two parts: the multi-head self-attention layer and the feed-forward network (FFN). The structure of the Transformer model is shown in Fig. 1.

The attention mechanism uses the query-key-value (QKV) ($Q = hW_m^Q$, $K = hW_m^K$, $V = hW_m^V$, $h \in \mathbb{R}^{N \times d}$ is the input, $W_m^Q, W_m^K \in \mathbb{R}^{d \times d_k}$, and $W_m^V \in \mathbb{R}^{d \times d_v}$ is the trainable weight matrix model to generate the attention of the scaled dot product.

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V. \quad (12)$$

To improve the learning ability of single attention, the transformer model applies multi-head attention (Vaswani et al. 2017). The QKV is linearly projected for m times. In each projection, the attention function is executed in parallel to generate the output results of d_v dimension. The projected output results are spliced and projected again to obtain the final output results. The QKV is linearly projected for m times. In each projection, the attention function Eq. 13 is executed in parallel to generate the output results of d_v dimension. The projected output results are stacked and projected again to obtain the final output results.

$$\text{MultiHeadAtt}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_m) W^O, \quad (13)$$

$$\text{head}_i = \text{Attention} \left(QW_i^Q, KW_i^K, VW_i^V \right), \quad (14)$$

where $W_i^Q, W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$, in multi-head attention $d_k = d_v = d_{\text{model}}/m$.

Carbon price series have long-term memory (Fan et al. 2019), and for better extraction of historical information, long time steps must be covered. However, not all time steps are relevant. In multi-head attention, the Softmax activation function is used to calculate the attention score equation (12), and Softmax cannot assign exactly zero scores because all scores sum to one, which reduces the attention allocation in the relevant time steps (Wiegrefe and Pinter 2019) and extracts information from the irrelevant time steps to add noise to the model, thereby affecting the performance of the transformer model. Therefore, this study uses the sparse mapping α -entmax (Peters et al. 2019) instead of Softmax, which is defined as follows:

$$\alpha - \text{entmax}(h) = [(\alpha - 1)h - \tau \mathbf{1}]_+^{1/\alpha - 1}, \quad (15)$$

where $[\]_+$ is the Relu activation function, $\mathbf{1}$ is a vector of all ones, and τ is the Lagrange multiplier. When $\alpha = 1$ is equal to the Softmax function, when $\alpha > 1$ can be sparse mapping, and $\alpha = 1.5$ is a reasonable point (Martins and Fernandez Astudillo 2016). Therefore, this study set $\alpha = 1.5$.

Orthogonal array tuning method

Zhang et al. (2019a) proposed OATM. Compared with the highly time-consuming and over-dependent configurations

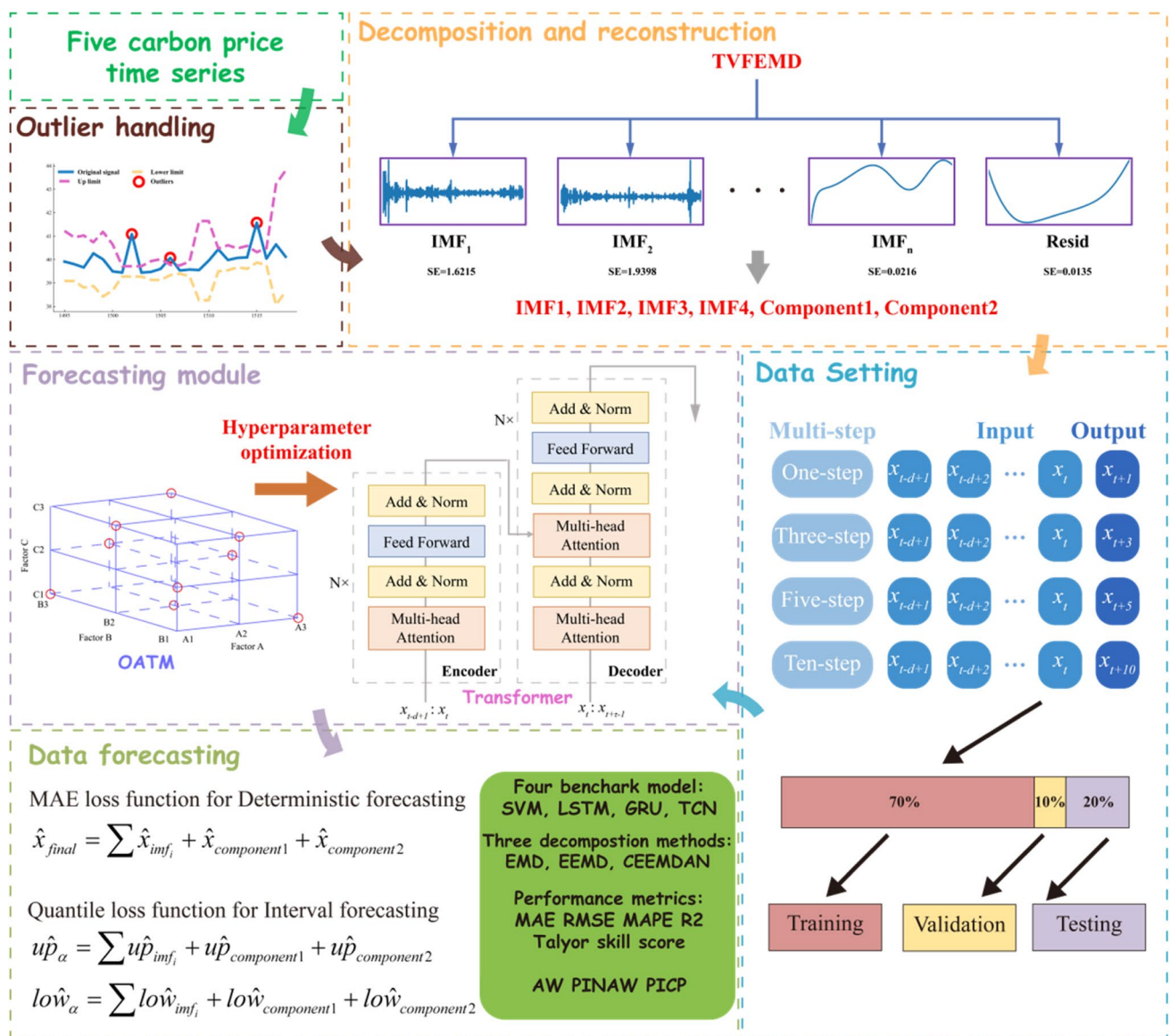


Fig. 1 Flow chart of the proposed hybrid forecasting system

of grid search, random search, and Bayesian optimization, OATM uses an orthogonal list to extract the most representative and balanced experimental combinations from all possible combinations. Furthermore, the use of these experimental combinations to optimize hyperparameters can achieve a trade-off between optimization time and model performance. The specific steps of OATM are presented as follows:

- Step 1. An orthogonal array is constructed to optimize the hyperparameters. The orthogonal array consists of a finite number of factors, with each factor containing the same finite number of levels. The arrangement of factors in the array ensures that each pair of different factors appears together in the same number of ordered combinations. For example,

Table 1 shows an orthogonal array with three factors and three levels each. The total number of combinations in the set of all factors and levels \mathcal{S} , $\text{Card}(\mathcal{S}) = 3 \times 3 \times 3 = 27$, and the set of orthogonal array is \mathcal{O} , $\text{Card}(\mathcal{O}) = 9$. The two sets are displayed in a cube, as shown in Fig. 1, where A1, A2, and A3 are the three levels of factor A, and B and C are also a set of factors. The 27 nodes on the surface of the cube are set \mathcal{S} , and the 9 red circles are set \mathcal{O} . In Fig. 1, each red circle is uniformly distributed, one exists on each edge, and three exist on each face. Thus, \mathcal{O} can be used as a representative subset of the set \mathcal{S} , $\mathcal{O} \subseteq \mathcal{S}$. The different hyperparameters in optimization are the factors in the orthogonal array, and the levels correspond to the values that can be chosen in the hyperparameters.

Table 1 Orthogonal array with three factors and three levels

Row no.	Factor 1	Factor 2	Factor3
1	1	1	1
2	1	2	2
3	1	3	3
4	2	1	2
5	2	2	3
6	2	3	1
7	3	1	3
8	3	2	1
9	3	3	2

- Step 2. Experiments are performed for all combinations of hyperparameters in the orthogonal array.
- Step 3. Range analysis. Range analysis plays a pivotal role in OATM. The experimental results obtained from Step 2 are analyzed using arrange analysis to determine the optimal level and importance of each hyperparameter. The importance of a hyperparameter is calculated by its influence on the experimental results. Through range analysis, each hyperparameter is optimized, and the optimal levels are combined to create an optimized combination of hyperparameters, implying that the optimized hyperparameter combination may not be found among the existing orthogonal array.
- Step 4. Run the experiments on models with optimized hyperparameter combinations.

OATM utilizes a representative smaller subset of the hyperparameters for optimization, leading to higher efficiency, as depicted in Fig. 1. The grid search method requires 27 experiments, whereas OATM only requires 9 experiments, saving approximately 67% of time.

Quantile loss

In this study, quantile regression (QR) is used for forecasting the carbon price interval. QR uses the data to conduct regression analysis at different quantiles, which can show the relationship between variables more comprehensively (Zhang et al. 2019a, 2019b). By improving the loss function of deterministic forecasting to a quantile loss function (Wen et al. 2017), the model can be used for quantile regression, and its loss function is expressed as follows:

$$\mathcal{L}(\Omega, W) = \sum_{y_t \in \Omega} \sum_{q \in Q} \sum_{\tau=1}^{\tau_{\max}} \frac{QL(y_t, \hat{y}(q, t - \tau, t), q)}{M \tau_{\max}}, \tag{16}$$

$$QL(y, \hat{y}, q) = q(y - \hat{y})_+ + (1 - q)(\hat{y} - y)_+, \tag{17}$$

where Ω is the training domain containing M samples; W is the weight of the model; Q is the set of quantile outputs $Q = \{\alpha/2, 1 - \alpha/2\}$; $1 - \alpha$ is the confidence interval if $\alpha=1$, that is, the same loss as the point forecast is the MAE loss function; and $(\cdot)_+ = \max(0, \cdot)$.

Evaluation indicators

This study selects several evaluation indicators to reflect the model’s performance in terms of precision and accuracy from various aspects. For deterministic forecasts, the evaluation indicators include mean absolute value error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE), goodness of fit (R^2), and Taylor skill score (TSS). MAE represents the actual deviation; RMSE can represent the discrete degree of the sample; MAPE represents the overall level of error; R^2 represents the overall fit of the forecast; and TSS is a comprehensive index that reflects the correlation coefficient, standard deviation, and centered root-mean-square (CRMS) of the prediction results (Taylor 2001). The smaller MAE, RMSE, and, MAPE are, the smaller the prediction error is, and the closer R^2 and TSS are to 1, which indicates an accuracy of the model. For interval forecast, three indicators are selected: average width (AW), prediction interval coverage probability (PICP), and prediction interval normalized average width (PINAW), where AW reflects the interval width of the interval forecast, PICP represents the coverage of interval forecast, and PINAW represents the percentage of average width in the data range. Generally, the prediction interval constructed is expected to have a small interval width but a high coverage rate. Thus, AW, PICP, and PINAW must be integrated to measure the effects of interval prediction. The evaluation index is expressed as follows:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|, \tag{18}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}, \tag{19}$$

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|, \tag{20}$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2}, \tag{21}$$

$$TSS = \frac{4(1+r)^4}{\left(SDR + \frac{1}{SDR} \right)^2 (1+r_0)^4}, \tag{22}$$

$$AW = \frac{1}{n} \sum_{i=1}^n (Ud_i - Ld_i), \quad (23)$$

$$PICP = \frac{1}{n} \sum_{i=1}^n C_i, C_i = \begin{cases} 1, C_i \in [Ld_i, Ud_i] \\ 0, C_i \notin [Ld_i, Ud_i] \end{cases}, \quad (24)$$

$$PINAW = \frac{1}{n} \sum_{i=1}^n [(Ud_i - Ld_i)/(y_{\max} - y_{\min})], \quad (25)$$

where y_i and \hat{y}_i are the actual and predicted values, respectively; \hat{y}_i represents the mean of the predicted values; n is the length of the sequence; r and SDR are the correlation coefficient and variance ratio between the predicted sequence and the actual sequence, respectively; and r_0 is set to one. Ud_i and Ld_i are the upper and lower boundaries of the interval prediction, respectively; y_{\max} is the max value in the actual sequence; and y_{\min} is the min value.

Framework of the proposed forecasting system

Figure 1 presents the framework flow of the designed hybrid carbon price forecasting system based on the methodology in the previous section. As seen in Fig. 1, the forecasting steps can be divided into the following parts:

- (1) Identification and correction of outliers. Owing to the complexity of carbon price, its fluctuation and fluctuation pattern become very drastic and irregular, respectively. Therefore, HI is used to process the original carbon price series to remove and correct outliers, thereby weakening the effects of outliers on model training.
- (2) Decomposition and reconstruction of carbon price. TVFEMD is applied to decompose the complex carbon price series to obtain several imfs, which contain the features with varying frequencies in series, thereby simplifying model learning. TVFEMD guarantees the existence of only one amplitude mode in the imfs. Thus, the imfs by TVFEMD are more detailed and greater in terms of quantity than those decomposed by EMD; however, more imfs extend computation time and increase accumulation of errors (Zhao et al. 2022). Hence, imfs need to be reconstructed, and the generally used method is to calculate the entropy value among different imfs (Sun et al. 2021); the complexity of imfs is classified according to the entropy value, and imfs that are classified into the same class are summed and merged (Wang and Qiu 2021). However, some studies (Sun and Huang 2020a, 2020b) suggest that high-frequency imfs contain complex fluctuations, and a secondary decomposition is implemented to further

improve the prediction accuracy, thereby generating more imfs, which contradict the concept of reconstruction. TVFEMD can effectively solve this conflict because the imfs obtained by its decomposition are more detailed, especially in the high-frequency part. Hence, the use of TVFEMD for the secondary decomposition of the imfs in the high-frequency part is unknott necessary. Therefore, this study uses SE (Richman and Moorman 2000) to estimate the complexity of each imfs. On the one hand, SE is retained for the imfs with high complexity. On the other hand, imfs with low complexity and the imfs with similar sample SE values are summed and combined to reduce computational time consumption and error accumulation and to retain complex imf features. Normalization is performed to map the data to [0,1] before each component is inputted into the forecasting model.

- (3) Optimization of hyperparameters. The transformer model includes a large number of hyperparameters, and different hyperparameter settings can have a huge impact on the prediction performance. In this study, OATM is used to tune the hyperparameters of the model and determine the appropriate hyperparameter configurations to improve the performance of the model.
- (4) Forecasting of carbon price. First, the Transformer model is used to perform one-step-ahead and multi-step-ahead forecasting on the decomposed and reconstructed subsequents. In multi-step-ahead forecasting, the transformer model applies the autoregressive method to generate multi-step-ahead results (Graves 2013). Next, the model takes the previously generated prediction results as the input of the next prediction and then sums the prediction results of each subsequent. Finally, the final prediction results are obtained. Based on the deterministic prediction, the interval prediction of carbon price is made by using the quantile loss function to measure the uncertainty of the deterministic prediction.
- (5) Comparison and analysis. In this study, different single models, including machine learning models (SVM) and commonly used deep learning models (LSTM, GRU, and TCN) and decomposition methods (EMD, EEMD, and CEEM-DAN) are selected for comprehensive comparison with the proposed hybrid model under a variety of evaluation indicators to prove the effectiveness of the proposed model.

Data collection and preprocessing

In this section, the original dataset of the carbon prices and its statistical features are described briefly. In addition, the results of the carbon price decomposition and reconstruction and hyperparameter optimization are presented.

Table 2 Statistical description of the experimental data

Markets	Number	Max	Min	Mean	Median	Std	Kurt	Skew	ADF(<i>p</i>)
Guangdong	1819	95.26	8.1	29.3530	23.06	19.5789	4.0223	1.4236	0.6248
Hubei	1921	61.48	10.07	26.5407	25.50	8.5651	3.1186	0.7448	0.6446
Beijing	1281	107.26	24	61.0160	53.50	16.8594	2.2420	0.5695	0.2142
Shanghai	1234	63	4.2	35.5784	38	12.4960	3.8618	-0.3972	0.4977
Shenzhen	1909	130.90	3.03	33.7840	31.83	19.9392	3.8329	0.8746	0.0010

Data collection

The closing price of the five pilot carbon trading markets in China, namely, Guangdong, Hubei, Beijing, Shanghai, and

Shenzhen, were selected as experimental subjects to verify the effectiveness of the proposed model in this study. The annual carbon quota in the Shenzhen carbon market is separated from the previous carbon quota, such as SZ2013 and SZ2014.

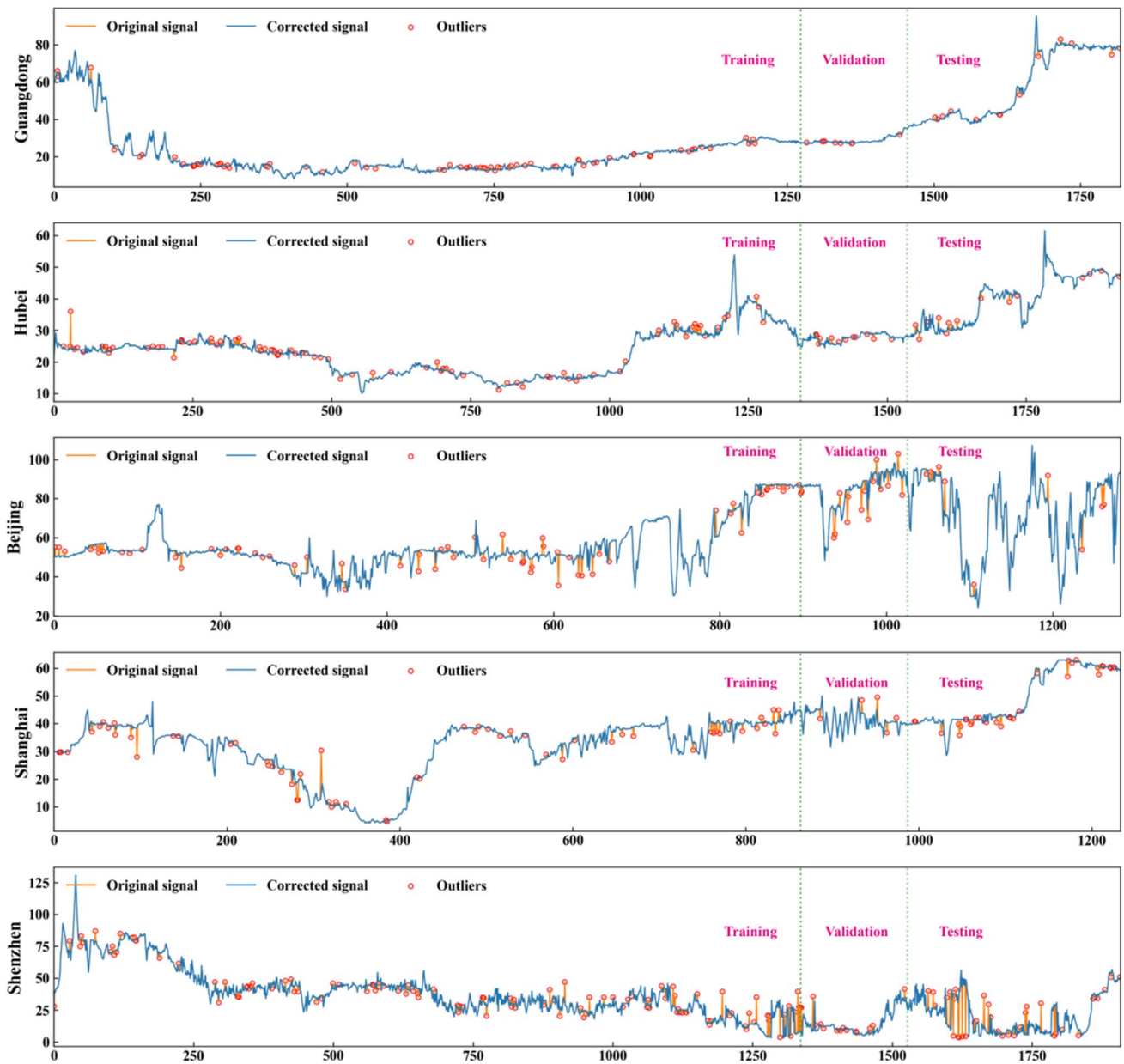


Fig. 2 Original and corrected series of five carbon markets

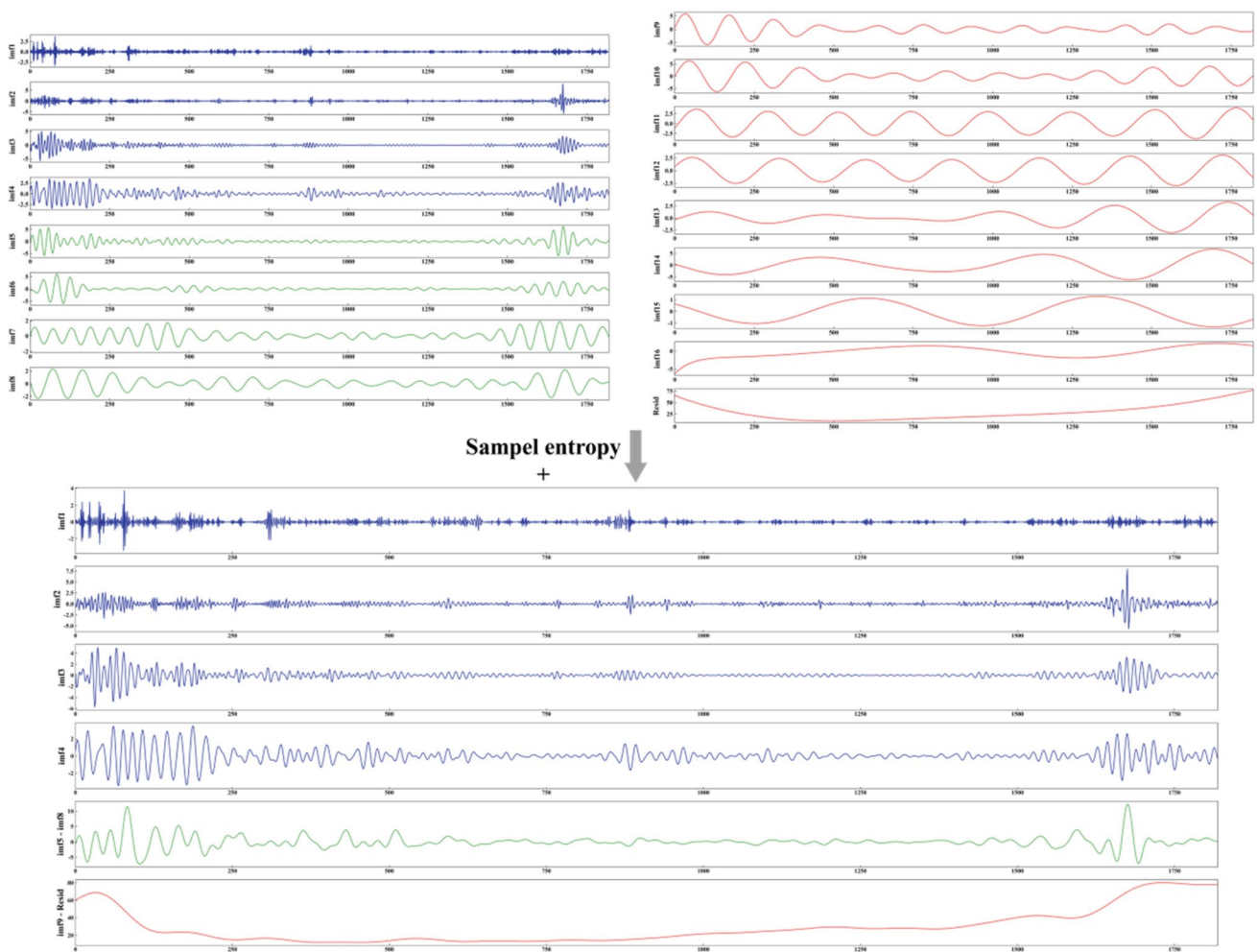


Fig. 3 Results of carbon price decomposition and reconstruction in Guangdong

Hence, the Shenzhen carbon price is represented by weighting and averaging the closing prices of different carbon quotas. The date when the trading volume of the carbon market is 0 is excluded, and the selected period is from the establishment time of each market to September 2, 2022. The data are collected from <http://www.tanpaifang.com/>. The dataset of relevant statistical information (Table 2) selects several typical statistical indexes, including the max, min, mean, median, standard deviation, kurtosis and skewness, and ADFtest, to determine

whether the Hubei carbon market is the most stable one among the aforementioned carbon markets. The standard deviation and the gap between the maximum and minimum values of all other carbon markets are relatively large, indicating that the carbon price series fluctuates dramatically. ADFtest shows that the *P* values of nearly all carbon markets, except Shenzhen carbon price, are much more significant than 0.05, indicating that the price data of these markets are non-stationary. The price series of these several carbon markets are shown in Fig. 2, where

Table 3 Sample entropy of each imf

	Sample entropy								
Retaining	imf1	imf2	imf3	imf4					
	1.6215	1.9398	1.1880	0.9827					
Component1	imf5	imf6	imf7	imf8					
	0.4954	0.3311	0.5786	0.3869					
Component2	imf9	imf10	imf11	imf12	imf13	imf14	imf15	imf16	Resid
	0.2007	0.2087	0.1291	0.1166	0.0823	0.0583	0.0406	0.0216	0.0135

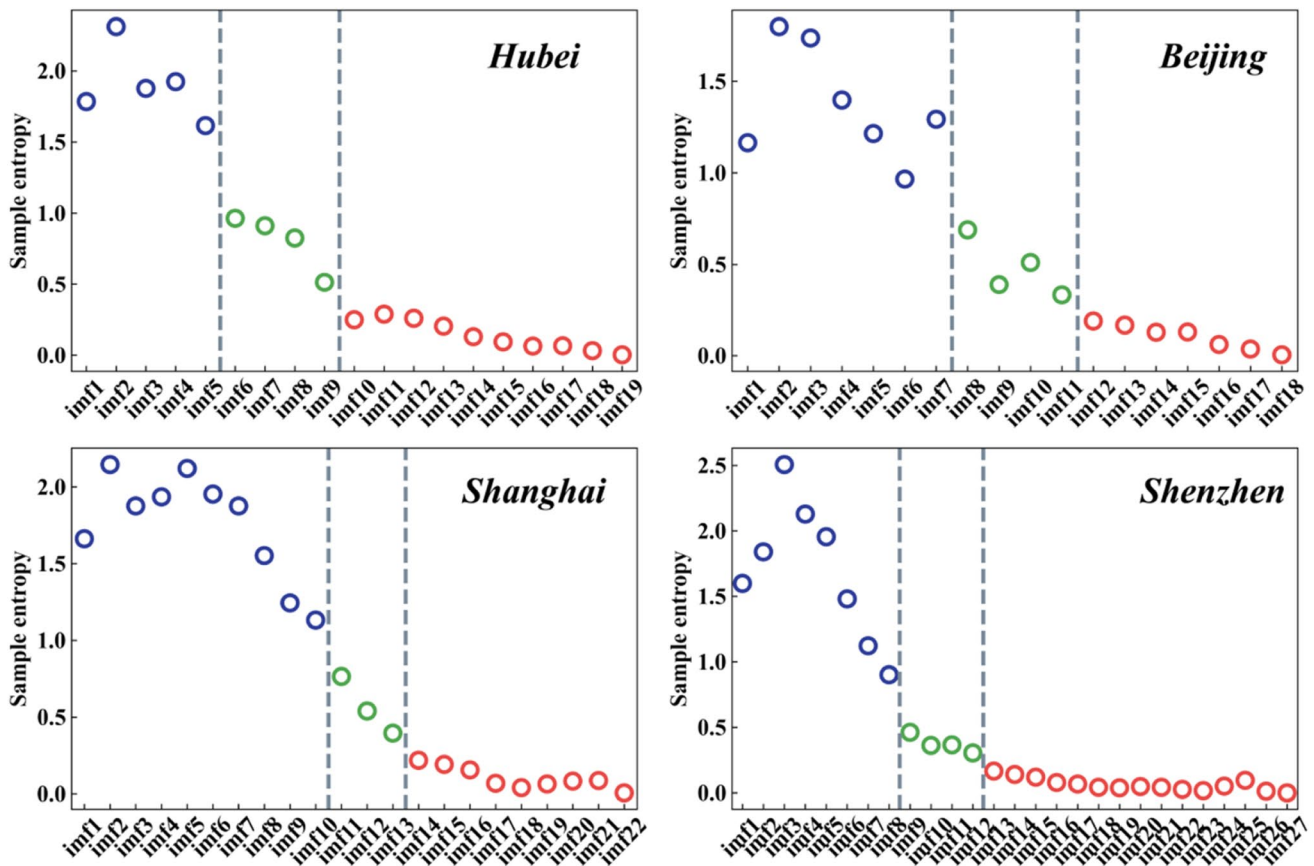


Fig. 4 Sample entropy and reconstruction results for the rest of the carbon market

many sudden upward and downward fluctuations in price series are observed. For the convenience of comparison, the carbon price series after outlier identification and correction are shown in Fig. 2, and the comparison shows that the processed series eliminates some points of sharp fluctuations.

In the experiment, the first 70%, the middle 10%, and the last 20% of the data are used as the training set, the validation set, and the test set, respectively. The training set is used to train the model, the validation set is used to optimize the model’s hyperparameters, and the test set is used to verify the model’s effectiveness.

Data decomposition and reconstruction

Guangdong carbon price is selected as an example to illustrate decomposition and reconstruction. According to the

designed hybrid model, the carbon price series after removing outliers is decomposed into several imfs and a residual term through TVFEMD. Each imf represents different amplitude patterns in the original signal, as shown in Fig. 3. According to Fig. 3, 17 imfs significantly increase the calculation amount of the hybrid model. Therefore, SE is used to identify the complexity of different imfs. In other words, the higher the SE, the higher the complexity, and imfs with high complexity are retained one by one. This phenomenon retains the complex features in time series to avoid repetitive secondary decompositions while other imfs are reconstructed according to their SE values. Other imfs are also reconstructed according to their SE values, which not only reduces the calculation of the model to avoid overfitting but also retains the fluctuating features with higher complexity.

Table 4 Comparison results before and after reconstruction

		Guangdong	Hubei	Beijing	Shanghai	Shenzhen
Running time(s)	Unreconstructed	2150.0437	2580.9971	1729.7639	2082.0854	3996.0947
	Reconstructed	798.3508	982.7669	885.4591	1168.1885	1569.1607
MAE	Unreconstructed	0.2534	0.3126	1.6902	0.4770	2.9084
	Reconstruction	0.2363	0.2947	1.5034	0.466	2.5681

Table 5 Factor-level table for the Transformer model

Factors	Levels			
	1	2	3	4
b_n	1	2	3	4
n_n	16	32	64	128
t_s	15	20	25	30
b_s	16	32	64	128
h_s	1	2	3	4

SE division does not follow stringent rules and regulations, and it is divided by the overall distribution of SE (Zhou et al. 2022; Wang et al. 2022; Sun et al. 2021). According to Table 3 and Fig. 3, the complexity of imfs has a decreasing distribution, and imfs can be roughly divided into three parts. The SE values of imf1–imf4 are almost all over 1 and fluctuate sharply. Hence, these imfs are retained separately. In contrast, the SE values of imf5–imf8 are relatively small and concentrated between 0.3 and 0.5. Thus, these imfs are

combined into a component, while the SE values of the remaining imfs are almost all below 0.2, combining these remaining imfs into another component. The reconstructed results are shown in Fig. 3. The rest of the carbon markets are also reconstructed based on similar division criteria for imfs, as shown in Fig. 4.

In this subsection, simple experiments are conducted to illustrate the effectiveness of the adopted decomposition–reconstruction method in terms of running time and forecasting accuracy. The optimized transformer model from “Interval forecasting” is used as the benchmark model. Running time and MAE are chosen as the criterion, and the results of the experiments are shown in Table 4. Table 4 demonstrates that a significant amount of running time is saved after reconstruction, without compromising its prediction accuracy. Thus, this decomposition–reconstruction method is used for all decomposition techniques in the following multi-step-ahead and interval prediction experiments.

Table 6 Range analysis of the transformer model

Row no.	b_n	n_n	t_s	b_s	h_n	RMSE
1	2	32	20	32	1	1.2753
2	2	128	25	16	3	1.1889
3	1	32	25	128	2	1.2017
4	3	64	25	64	1	1.1428
5	1	16	15	16	1	1.2047
6	1	128	20	64	4	1.1176
7	3	16	20	128	3	1.2121
8	1	64	30	32	3	1.2494
9	4	64	20	16	2	1.2499
10	2	64	15	128	4	1.1580
11	3	128	15	32	2	1.2272
12	2	16	30	64	2	1.2479
13	4	16	25	32	4	1.2447
14	3	32	30	16	4	1.1743
15	4	128	30	128	1	1.1160
16	4	32	15	64	3	1.3356
R_{level1}	4.7734	4.9094	4.9255	4.8178	4.7388	
R_{level2}	4.8702	4.9868	4.8550	4.9966	4.9267	
R_{level3}	4.7563	4.8002	4.7781	4.8440	4.9859	
R_{level4}	4.9463	4.6498	4.7875	4.6878	4.6947	
M_{level1}	1.1933	1.2273	1.2314	1.2045	1.1847	
M_{level2}	1.2175	1.2467	1.2137	1.2491	1.2317	
M_{level3}	1.1891	1.2000	1.1945	1.2110	1.2465	
M_{level4}	1.2366	1.1624	1.1969	1.1719	1.1737	
Lowest RMSE	1.1891	1.1624	1.1945	1.1719	1.1737	
Height RMSE	1.2366	1.2467	1.2314	1.2491	1.2465	
Range	0.0475	0.0843	0.0369	0.0772	0.0728	
Importance	$n_n > b_s > h_n > b_n > t_s$					
Best level	Level3	Level4	Level3	Level4	Level4	
Suggested optimal values	3	128	0.3	128	4	1.1075

Table 7 Evaluation indicators of comparison models and the proposed model in one-step-ahead forecasting

		HI-SVM	HI-LSTM	HI-GRU	HI-TCN	HI-transformer	HI-EMD-transformer	HI-EEMD-transformer	HI-CEEM-DAN-transformer	HI-TVFEMD-transformer
GD	MAE	2.5738	1.1227	1.5120	2.0968	0.9525	0.8356	0.6629	0.4869	0.2151
	RMSE	3.5037	1.6094	2.0520	2.8663	1.5204	1.3852	1.2485	1.0263	0.4596
	MAPE	0.0455	0.0185	0.0240	0.0305	0.0161	0.0140	0.0103	0.0077	0.0036
	R^2	0.9617	0.9919	0.9868	0.9743	0.9928	0.9940	0.9951	0.9967	0.9993
	TSS	0.9801	0.9939	0.9917	0.9846	0.9937	0.9940	0.9957	0.9968	0.9993
HB	MAE	1.5238	0.9158	0.9177	0.9590	0.7627	0.4979	0.4877	0.4364	0.1542
	RMSE	1.9624	1.4286	1.4097	1.4306	1.2484	0.7435	0.7507	0.7767	0.4306
	MAPE	0.0411	0.0237	0.0238	0.0244	0.0200	0.0131	0.0127	0.0116	0.0041
	R^2	0.9320	0.9640	0.9649	0.9639	0.9725	0.9902	0.9901	0.9894	0.9967
	TSS	0.9535	0.9656	0.9665	0.9653	0.9731	0.9904	0.9907	0.9894	0.9968
BJ	MAE	7.4643	6.0155	6.0476	5.9804	5.9106	3.5075	3.3378	3.4090	0.7582
	RMSE	9.1927	8.1469	8.1943	8.0369	8.0701	4.9323	4.7006	4.7668	1.9730
	MAPE	0.1182	0.0968	0.0971	0.0974	0.0950	0.0540	0.0517	0.0531	0.0110
	R^2	0.7820	0.8288	0.8268	0.8334	0.8320	0.9373	0.9430	0.9414	0.9900
	TSS	0.7734	0.8364	0.8367	0.8372	0.8400	0.9348	0.9407	0.9391	0.9872
SH	MAE	2.4002	1.0634	0.9509	0.8855	0.7993	0.7306	0.5076	0.4743	0.3814
	RMSE	2.8363	1.4692	1.4554	1.4103	1.3777	1.1675	0.8774	0.8533	0.4317
	MAPE	0.0473	0.0220	0.0203	0.0194	0.0175	0.0157	0.0108	0.0103	0.0077
	R^2	0.9157	0.9774	0.9778	0.9792	0.9801	0.9857	0.9919	0.9924	0.9980
	TSS	0.9505	0.9797	0.9784	0.9791	0.9803	0.9857	0.9920	0.9924	0.9994
SZ	MAE	4.7057	4.5306	4.3546	4.2990	4.2138	3.4482	3.1897	3.2412	1.7639
	RMSE	7.5539	7.4815	7.4793	7.2251	7.4225	6.7375	6.3657	6.5823	2.9226
	MAPE	0.4171	0.4494	0.3664	0.4077	0.3658	0.3124	0.3021	0.3184	0.1439
	R^2	0.7313	0.7365	0.7366	0.7542	0.7406	0.7863	0.8092	0.7960	0.9598
	TSS	0.7354	0.7619	0.7671	0.7592	0.7530	0.7989	0.8174	0.8117	0.9592

Hyperparameter optimization

In deep learning, the setting of hyperparameters plays a crucial role in determining the experimental results. In this study, OATM is used to optimize the model's hyperparameters. OATM can effectively balance the optimization time and accuracy through a representative subset. In consideration of relevant studies (Meka et al. 2021; Vaswani et al. 2017; Zhang et al. 2019a), five hyperparameters in the transformer model, including the number of blocks (b_n), the number of neurons (n_n), the time step (t_s), the batch size (b_s), and the number of heads (h_n), are selected as factors for the OATM. Each factor contains four levels, as shown in Table 5. The carbon price in Guangdong is selected as the experimental data. All experiments are conducted using TensorFlow 2.7.0 under Python 3.8.8, and the number of cycles is set to 1000. To prevent overfitting, an adaptive reducing learning rate Adam optimizer is used, along with the early stop mechanism. The dropout rate is set to 0.2, and the activation function is Relu.

The hyperparameter orthogonal array is constructed, and the experimental results are shown in Table 6. This table includes the validation set RMSE for 16 experiments. In comparison with the grid search method, which requires 1024 (4^5) experiments, OATM can save 98.43% of the computational time.

The range analysis enables us to determine the optimal level and the most critical factor according to Table 6. $R_{\text{level}i}$ and $M_{\text{level}i}$ are the sum and mean of RMSE at different levels, respectively. The difference between the max $M_{\text{level}i}$ and the min $M_{\text{level}i}$ is used to calculate range, which is used to represent the importance of different factors. The greater the range is, the greater the importance of the factor will be. That is, the more sensitive the Transformer model is to this factor. The number of neurons is the most sensitive factor expressed in terms of RMSE, followed by the batch size, number of heads, number of blocks, and the time step. The optimal level under each factor is determined by the lowest $M_{\text{level}i}$, from which the optimal hyperparameter combination of the Transformer model can be obtained ($b_n = 3$,

Table 8 Evaluation indicators of comparison models and the proposed model in three-step-ahead forecasting

		HI-SVM	HI-LSTM	HI-GRU	HI-TCN	HI-transformer	HI-EMD-transformer	HI-EEMD-transformer	HI-CEEM-DAN-transformer	HI-TVFEMD-transformer
GD	MAE	3.2670	3.0199	2.9434	2.1075	1.5736	0.9683	1.1403	0.7613	0.4776
	RMSE	4.4541	4.2378	4.1527	2.8744	2.9148	1.3650	2.0079	1.4223	0.7587
	MAPE	0.0493	0.0453	0.0464	0.0306	0.0253	0.0157	0.0174	0.0123	0.0083
	R^2	0.9379	0.9438	0.9460	0.9741	0.9734	0.9942	0.9874	0.9937	0.9982
	TSS	0.9630	0.9635	0.9653	0.9846	0.9739	0.9959	0.9878	0.9938	0.9982
HB	MAE	1.8135	1.4200	1.6819	1.6414	1.2937	0.7447	0.5680	0.8719	0.5088
	RMSE	2.5900	2.3406	2.5843	2.6687	2.2439	1.0230	0.9190	1.3580	0.6848
	MAPE	0.0441	0.0358	0.0413	0.0415	0.0334	0.0190	0.0148	0.0222	0.0129
	R^2	0.8811	0.9029	0.8816	0.8738	0.9108	0.9815	0.9850	0.9673	0.9917
	TSS	0.9045	0.9091	0.8974	0.8759	0.9146	0.9850	0.9856	0.9684	0.9936
BJ	MAE	10.1693	9.5036	9.7706	9.5026	9.4384	4.8003	4.3384	4.8267	2.4346
	RMSE	13.3139	12.9573	13.0996	12.6158	12.8502	6.4583	5.8488	6.4674	3.4405
	MAPE	0.1723	0.1598	0.1683	0.1592	0.1582	0.0757	0.0701	0.0798	0.0372
	R^2	0.5418	0.5660	0.5564	0.5886	0.5732	0.8922	0.9116	0.8919	0.9694
	TSS	0.5622	0.6035	0.5763	0.6075	0.6171	0.8906	0.9071	0.8887	0.9697
SH	MAE	2.7538	1.5926	1.4752	1.4763	1.2254	0.7862	0.6504	0.6242	0.4114
	RMSE	3.4919	2.4283	2.3833	2.3307	2.1254	1.1009	1.0756	1.0548	0.5819
	MAPE	0.0558	0.0333	0.0315	0.0315	0.0272	0.0167	0.0140	0.0136	0.0086
	R^2	0.8723	0.9382	0.9405	0.9431	0.9527	0.9873	0.9879	0.9883	0.9965
	TSS	0.9176	0.9408	0.9429	0.9451	0.9521	0.9871	0.9878	0.9883	0.9965
SZ	MAE	6.3385	5.5242	5.3552	5.3089	5.3335	4.2079	4.1057	3.7910	3.1638
	RMSE	8.8349	8.5532	8.4282	8.2661	8.3336	7.1768	6.8049	6.0630	5.2237
	MAPE	0.5012	0.4384	0.4320	0.4620	0.4543	0.3621	0.3786	0.3157	0.2757
	R^2	0.6333	0.6563	0.6663	0.6790	0.6737	0.7580	0.7825	0.8273	0.8718
	TSS	0.6230	0.6972	0.6976	0.6930	0.6855	0.7704	0.7872	0.8277	0.8679

$n_n = 128, t_s = 25, b_s = 128, b_n = 4$). Finally, the obtained optimal hyperparameter combinations are experimented, and the RMSE is obtained as 1.1075, which is better than the optimal result of 1.1160 in the orthogonal array, indicating that the OATM method can obtain the approximate global optimal solution of the hyperparameter to ensure the robustness of the model structure. All other models included in the comparison use OATM to determine their optimal structure. In other words, all models in the experiment are optimized by OATM.

Experimental analysis

Deterministic forecasting

The deterministic forecasts of carbon prices in the five carbon markets, including one-step, three-step, five-step, and ten-step-ahead forecasting, are analyzed. In addition, different graphs are used to compare the forecasting results

more visually. For example, histograms are used to represent MAPE and R^2 , radar chart is used to represent MAE and RMSE, and Taylor diagram is used to show the prediction results of these models. The forecasting results of each model are shown in Tables 7, 8, 9, and 10 (the optimal model is marked in bold, for presentation purposes, these tables only contain the results after using HI) and Figs. 5, 6, 7, 8, and 9 (containing the results of the model with and without HI). According to the results in Tables 7, 8, 9, and 10 and Figs. 5, 6, 7, 8, and 9, the fitted curves of the proposed HI-TVFEMD-transformer model are the closest to the actual values. More importantly, the RMSE, MAE, and MAPE of this model are significantly lower than those of other models, whereas their R^2 and TSS are almost always the highest, indicating that the forecasting performance of the proposed model is better than those of other benchmark models.

The following comparison was made according to Tables 7, 8, 9, and 10 and Figs. 5, 6, 7, 8, and 9.

Comparison I: the comparison of single model. Single model is the basis for hybrid models to achieve high

Table 9 Evaluation indicators of comparison models and the proposed model in five-step-ahead forecasting

		HI-SVM	HI-LSTM	HI-GRU	HI-TCN	HI-transformer	HI-EMD-transformer	HI-EEMD-transformer	HI-CEEM-DAN-transformer	HI-TVFEMD-transformer
GD	MAE	6.4257	5.1009	4.6058	4.7375	4.3889	2.2112	1.1854	1.0837	0.5364
	RMSE	7.7316	6.6878	6.1519	6.5092	6.5946	2.7013	2.2334	1.6703	0.8988
	MAPE	0.1044	0.0776	0.0701	0.0695	0.0649	0.0376	0.0182	0.0174	0.0090
	R^2	0.8130	0.8601	0.8816	0.8675	0.864	0.9772	0.9844	0.9913	0.9975
	TSS	0.9333	0.9277	0.9357	0.9201	0.9071	0.9896	0.9844	0.9918	0.9975
HB	MAE	2.2369	2.0287	2.0383	2.0080	1.8263	0.9898	0.7591	1.1168	0.5483
	RMSE	3.1514	3.2540	3.0454	2.9649	2.8399	1.3735	1.1502	1.8460	0.7210
	MAPE	0.0542	0.0506	0.0500	0.0491	0.0459	0.0253	0.0197	0.0285	0.0141
	R^2	0.8240	0.8123	0.8356	0.8442	0.8571	0.9666	0.9766	0.9396	0.9908
	TSS	0.8643	0.8238	0.8550	0.8653	0.8764	0.9672	0.9763	0.9460	0.9914
BJ	MAE	12.7391	12.0009	12.3936	12.1031	11.0786	6.1981	5.7118	6.0340	3.2400
	RMSE	16.3706	15.5445	16.3819	16.0117	13.7315	8.0682	7.6197	7.8158	4.4684
	MAPE	0.2235	0.2055	0.2118	0.2088	0.1900	0.0979	0.0948	0.0998	0.0496
	R^2	0.3073	0.3754	0.3063	0.3373	0.5126	0.8317	0.8499	0.8421	0.9484
	TSS	0.4161	0.4437	0.4424	0.4675	0.4912	0.8345	0.8394	0.8327	0.9489
SH	MAE	3.2780	2.0868	1.7643	1.7643	1.6039	0.9198	0.9124	0.8510	0.5057
	RMSE	4.1335	3.1460	2.9624	2.9499	2.7886	1.3253	1.3322	1.2185	0.6873
	MAPE	0.0666	0.0433	0.0383	0.0382	0.0357	0.0195	0.0199	0.0178	0.0108
	R^2	0.8210	0.8963	0.9081	0.9089	0.9186	0.9816	0.9814	0.9844	0.9951
	TSS	0.8923	0.9007	0.9084	0.9123	0.9193	0.9818	0.9828	0.9850	0.9953
SZ	MAE	7.2551	6.4302	6.3367	6.3720	6.3434	4.4167	4.5961	4.0707	3.3132
	RMSE	9.5872	9.2676	9.1529	9.1218	9.0928	6.6028	6.6636	6.2980	5.3436
	MAPE	0.5583	0.5654	0.5246	0.5331	0.4971	0.3590	0.3757	0.3309	0.2773
	R^2	0.5682	0.5965	0.6064	0.6091	0.6116	0.7952	0.7914	0.8137	0.8659
	TSS	0.5448	0.6225	0.6167	0.6164	0.6188	0.7915	0.7806	0.8133	0.8619

accuracy. Therefore, this study selects several single models that are commonly used in the mainstream of carbon price forecasting. These models include a machine learning model (SVM) and three deep learning models (e.g., LSTM, GRU, and TCN) that can address the time sequence problem. A comprehensive comparison of Tables 6, 7, 8, and 9 and Figs. 5, 6, 7, 8, and 9 shows that the deep learning model outperforms the machine learning model in all five metrics (e.g., MAE, RMSE, MAPE, R^2 , and TSS), suggesting that the deep learning model is more suitable for carbon price forecasting than the machine learning model. Among the deep learning models, the performance of LSTM, GRU, and TCN varies in different situations. Their respective strengths and weaknesses are shown in the histogram, radar chart, and Taylor diagram of each carbon market. For example, the MAPE of one-step-ahead prediction LSTM in Guangdong and Hubei are 0.0185 and 0.0237, respectively, which are better than GRU's 0.0240 and 0.0238. Meanwhile, in Guangdong and Hubei, the MAPE of LSTM in

the five-step prediction is 0.0776 and 0.0506, respectively, which is inferior to 0.0701 and 0.0500 of GRU. In the ten-step-ahead prediction in Shanghai, TCN is better than LSTM and GRU, and its MAPE is 0.0480, which is better than GRU's 0.0584 and LSTM's 0.0662. However, in most cases, the metrics for the transformer are the best among the single models, indicating that in the experiments, transformer's learning ability is better than those of SVM, LSTM, GRU, and TCN for capturing the complexity and nonlinearity in the carbon price.

Comparison II: the comparison of the efficacy of HI. The validity of HI is evaluated by comparing the model's predictive performance with and without the use of HI. According to Figs. 5, 6, 7, 8, and 9, the radar chart is in the inner circle without HI after using HI. Among the 180 ($5 \times 9 \times 4$) experiments, the number of experiments in which HI has improved all evaluation indicators is 113, and approximately 62.78% of the experiments have been improved while the remaining experiments have improved a few indicators. We calculated the average improvement rate of HI on MAE,

Table 10 Evaluation indicators of comparison models and the proposed model in ten-step-ahead forecasting

		HI-SVM	HI-LSTM	HI-GRU	HI-TCN	HI-transformer	HI-EMD-transformer	HI-EEMD-transformer	HI-CEEM-DAN-trans-former	HI-TVFEMD-transformer
GD	MAE	9.0425	6.1958	5.781	6.9974	3.4316	1.6663	1.5623	1.6110	0.7812
	RMSE	10.8861	8.0314	8.1871	9.1274	5.9156	2.5773	2.4168	2.3849	1.2325
	MAPE	0.1432	0.0972	0.0997	0.1045	0.0547	0.0269	0.0252	0.0259	0.0133
	R^2	0.6288	0.7980	0.7901	0.7391	0.8904	0.9792	0.9817	0.9822	0.9952
	TSS	0.8578	0.8936	0.8936	0.8545	0.9056	0.9801	0.9825	0.9850	0.9954
HB	MAE	4.2611	2.5055	3.5641	2.8745	2.4839	2.2862	1.5603	2.2082	0.8479
	RMSE	5.5580	3.9389	4.8511	4.1105	3.7714	3.0184	2.1589	2.9660	1.1986
	MAPE	0.0982	0.0609	0.0831	0.0688	0.0614	0.0547	0.0391	0.0553	0.0215
	R^2	0.4504	0.7240	0.5813	0.6994	0.7470	0.8379	0.9171	0.8435	0.9744
	TSS	0.6450	0.7518	0.6973	0.7603	0.7869	0.8528	0.9254	0.8946	0.9756
BJ	MAE	14.6715	15.3538	15.3522	15.1229	14.7375	9.1432	8.9363	13.3105	5.0567
	RMSE	18.8293	20.9114	21.4275	19.3549	18.7053	11.6185	11.6938	17.6211	6.5018
	MAPE	0.2698	0.2650	0.2830	0.2687	0.2672	0.1559	0.1383	0.2177	0.0784
	R^2	0.0817	-0.1326	-0.1892	0.0297	0.0937	0.6504	0.6458	0.1958	0.8905
	TSS	0.2301	0.2651	0.222	0.2651	0.2496	0.6313	0.6956	0.5036	0.8912
SH	MAE	4.2691	3.4307	2.9327	2.3130	2.250	1.1532	1.1111	0.9699	0.8969
	RMSE	5.2073	4.6453	4.2042	3.7082	3.6801	1.7475	1.6695	1.4465	1.2072
	MAPE	0.0852	0.0662	0.0584	0.0480	0.0475	0.0252	0.0246	0.0214	0.0194
	R^2	0.7160	0.7740	0.8149	0.8560	0.8582	0.968	0.9708	0.9781	0.9847
	TSS	0.8570	0.8227	0.8566	0.8757	0.8769	0.9694	0.9714	0.9782	0.9868
SZ	MAE	8.9157	8.3254	7.5796	7.8806	7.8557	4.888	5.5744	4.5682	3.7176
	RMSE	11.321	11.2627	10.5551	10.4368	11.0522	7.0452	7.6386	6.7684	5.7451
	MAPE	0.6491	0.6105	0.546	0.6052	0.5681	0.3754	0.3956	0.3672	0.3231
	R^2	0.3994	0.4056	0.478	0.4896	0.4276	0.7674	0.7266	0.7853	0.8453
	TSS	0.3900	0.4539	0.5118	0.4852	0.4948	0.7620	0.7124	0.7751	0.8426

RMSE, MAPE, R^2 , and TSS in the Guangdong experiment (the value was obtained by averaging the improvement rates in each experiment), and HI improved by 20.56%, 11.95%, 19.77%, 1.54%, and 0.44%, respectively. The average improvement rate of HI at different steps in the Guangdong experiment was also calculated. HI improved the one-step-ahead forecast by 19.2959%, 1.2028%, 19.2068%, 0.1511%, and - 0.0558%, respectively; three-step-ahead forecast by 28.65%, 21.55%, 28.23%, 1.93%, and 0.90%, respectively; for the five-step ahead forecast by 12.40%, 9.52%, 11.53%, 1.01%, and 0.29%, respectively; and for the ten-step-ahead forecast by 21.89%, 15.55%, 20.13%, 3.06%, and 0.62%, respectively. HI improves MAE, RMSE, and MAPE more significantly. However, it does not improve R^2 and TSS significantly or even negatively. The reason is that the outliers in the series are identified and corrected by HI to eliminate the abnormal fluctuations, which make the fluctuation characteristics more obvious and reduce the model error. Meanwhile, MAE, RMSE, and MAPE are the metrics used to measure the error between the prediction and the actual models. R^2 and TSS are metrics that are based on

the proportion of variance in the true values that can be explained by the model, the presence of outliers leads to increased variance in the series, and the model cannot learn this information after outliers are identified and corrected. Yet, by retaining these outliers, the model does not learn more useful information either. This instance prevents the model from fitting this part of the variance, which leads to no significant difference in R^2 and TSS before and after outlier processing. The above analysis demonstrates that outliers in carbon prices can negatively affect the training of the model, and, therefore, HI can improve the prediction ability of the model to some extent.

Comparison III: the comparison of different decomposition methods. According to Figs. 5, 6, 7, 8, and 9, the fitted curves for each carbon market demonstrate a closer fit to the actual values after incorporating the decomposition algorithm. Notably, EMD, EEMD, CEEMDAN, and TVFEMD all improve the prediction accuracy of the transformer model. However, the comparison of TVFEMD suggests that it has a more significant improvement than EMD, EEMD, and CEEMDAN. According to Table 7,

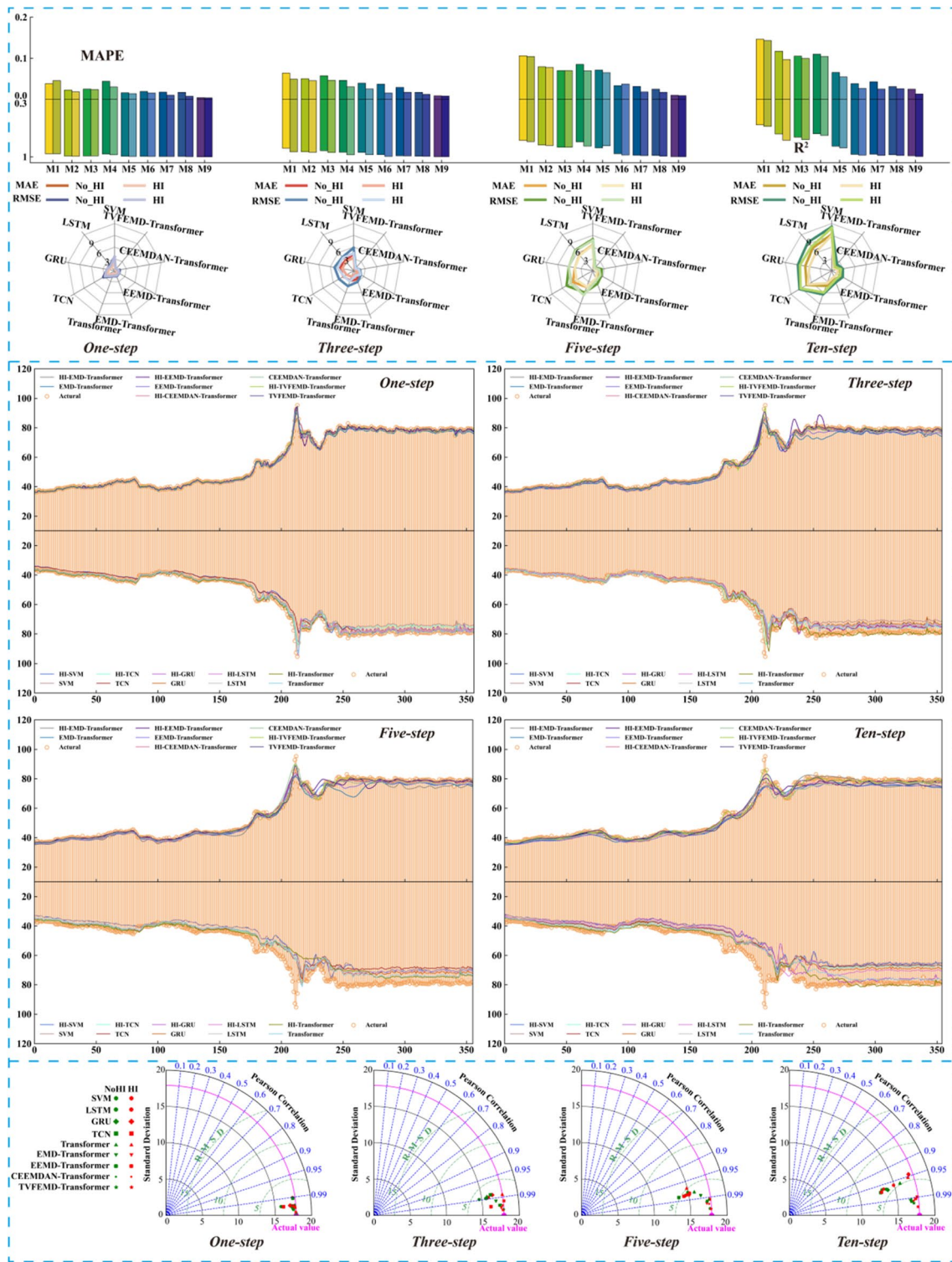
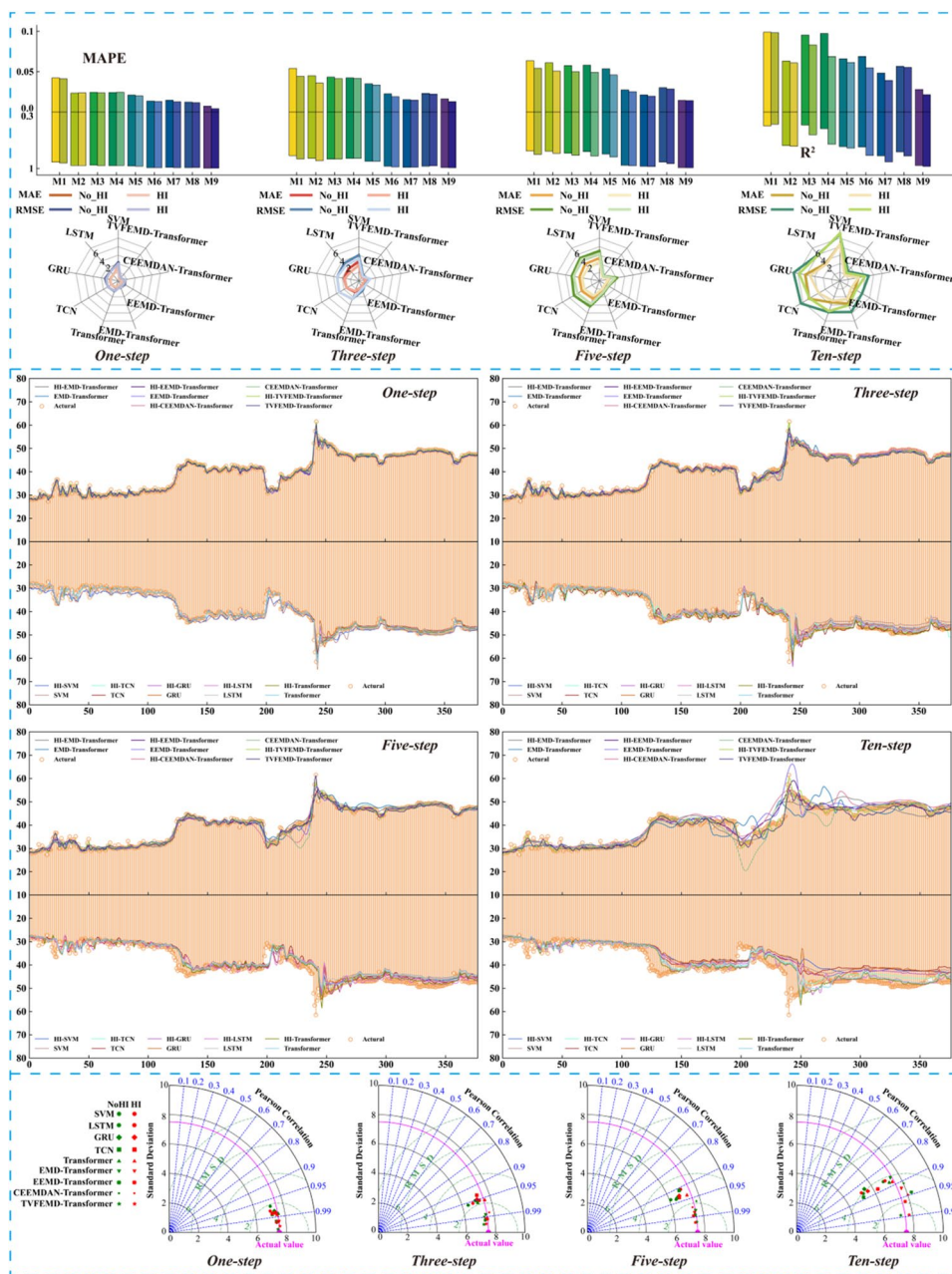


Fig. 5 Deterministic forecasting results of the Guangdong market. The upper part corresponds to the histogram and radar plot; the middle part of the histogram is the starting value of MAPE and R²; M1, M2, M3, M4, M5, M6, M7, M8, and M9 represent SVM, LSTM, GRU, TCN, transformer, EMD-transformer, EEMD-transformer, CEEMDAN-transformer, and TVFEMD-transformer, respectively;

CEEMDAN-transformer, and TVFEMD-transformer, respectively; the left side represents the plot without HI; and the right side represents the plot with HI. The middle part is the fitted curve. The part below it is the Taylor chart

Fig. 6 Deterministic forecasting results of the Hubei market



HI-TVFEMD-transformer has an average MAPE of 0.0341 in one-step-ahead forecasting, whereas HI-EEMD-transformer is 0.0818, HI-EEMD-transformer is 0.0775, and HI-CEEMDAN-transformer is 0.0802. The average MAPE of the transformer model is 0.1121. Comparing the different metrics in the predictions of various ahead steps, TVFEMD also exhibited the best performance, thereby proving the superiority of the TVFEMD method and improving the prediction accuracy of the proposed model.

Comparison IV: the comparison of different carbon markets. According to the Taylor diagram in Figs. 5, 6, 7, 8, and 9, the forecasting results of the Guangdong, Hubei, and

Shanghai carbon markets are more concentrated than those of Beijing and Shenzhen carbon markets, indicating that the carbon price fluctuations in these three carbon markets are more regular, thereby making the performance of various models less different. Consequently, the price series for the Beijing and Shenzhen carbon markets are more volatile, which better reflects the differences in the performance of various models. The HI-TVFEMD-transformer is closer to the actual point in the Taylor diagram in different carbon markets. Its performance is almost always optimal, indicating that it has strong robustness and can be adapted to different datasets.

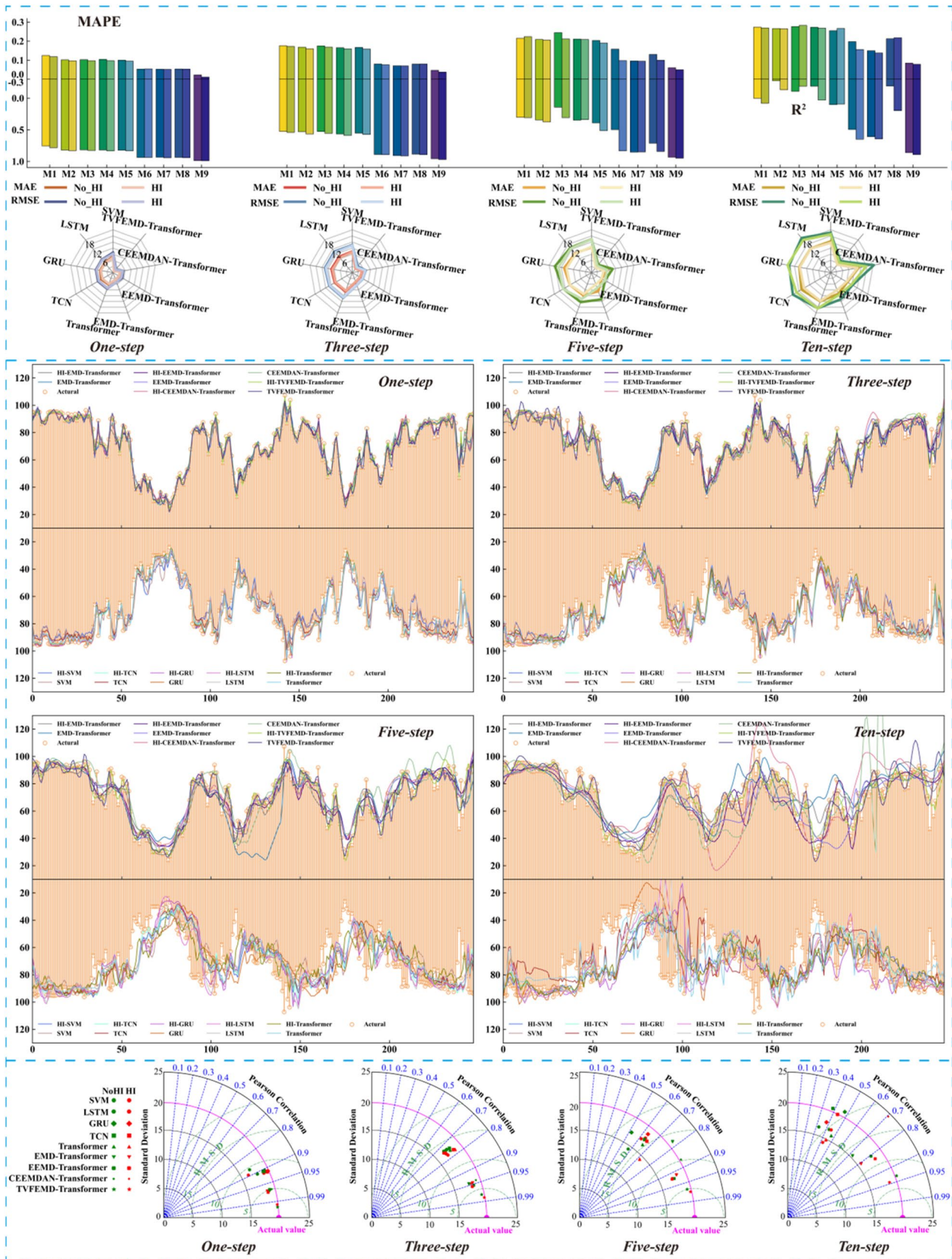


Fig. 7. Deterministic forecasting results of the Beijing market

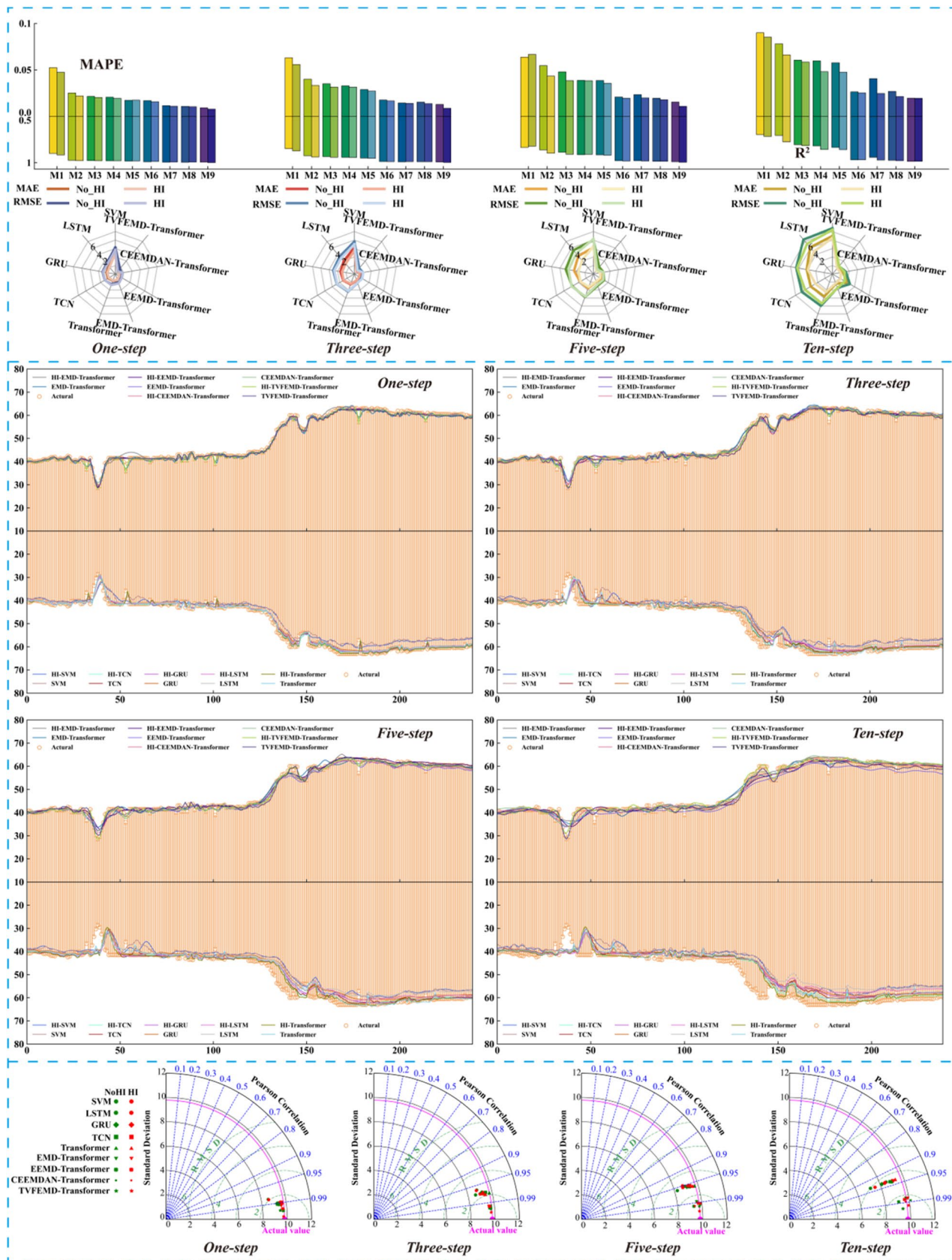


Fig. 8 Deterministic forecasting results of the Shanghai market

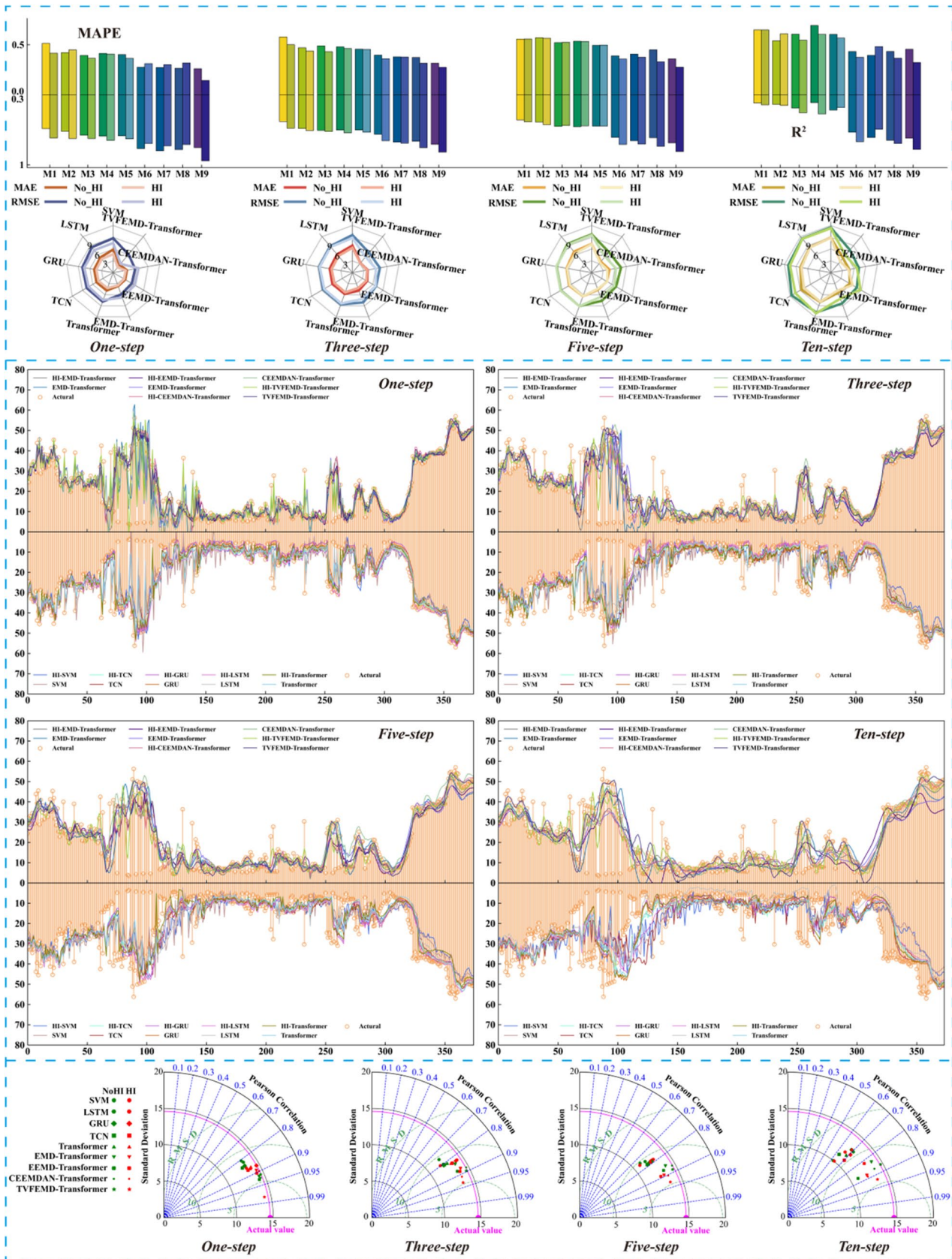


Fig. 9. Deterministic forecasting results of the Shenzhen market

Table 11 Indicator results of interval forecasting

		One-step			Three-step			Five-step			Ten-step		
		$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$
GD	AW	3.7517	2.9950	1.3256	7.4639	5.3737	3.4211	8.8214	5.7713	3.5696	12.4418	8.5105	5.1686
	PINAW	0.0636	0.0508	0.0225	0.1265	0.0911	0.0580	0.1495	0.0978	0.0605	0.2109	0.1443	0.0876
	PICP	0.9552	0.9356	0.8095	0.9859	0.9042	0.8310	0.9663	0.9494	0.7331	0.9268	0.8845	0.8169
HB	AW	2.7403	2.2124	1.5805	4.4639	2.9647	2.3842	5.8143	4.2485	3.2502	9.7962	6.1739	4.2321
	PINAW	0.0800	0.0646	0.0461	0.1303	0.0865	0.0696	0.1697	0.1240	0.0949	0.2859	0.1802	0.1235
	PICP	0.9472	0.9367	0.8575	0.9656	0.9206	0.8519	0.9894	0.9471	0.8968	0.9920	0.9072	0.7613
BJ	AW	16.2249	8.6393	8.5923	21.9262	15.4160	10.5396	27.6073	12.1699	11.5617	37.4047	26.1416	14.1968
	PINAW	0.1949	0.1038	0.1032	0.2633	0.1852	0.1266	0.3316	0.1462	0.1389	0.4493	0.3140	0.1705
	PICP	0.9799	0.9639	0.9719	0.9355	0.8508	0.8548	0.9435	0.8226	0.8105	0.9555	0.8583	0.7126
SH	AW	5.1689	3.3875	2.5800	7.0595	4.9152	3.4536	8.5376	6.0468	4.5218	11.7828	8.3761	6.2185
	PINAW	0.1503	0.0985	0.0750	0.2052	0.1429	0.1004	0.2482	0.1758	0.1314	0.3425	0.2435	0.1808
	PICP	0.9750	0.9667	0.9458	0.9833	0.9750	0.9625	0.9958	0.9833	0.9707	1.0000	0.9874	0.8361
SZ	AW	16.3022	11.1420	8.4882	21.3452	14.0983	10.1807	21.1311	14.9308	11.7690	25.8867	21.9949	14.1140
	PINAW	0.3049	0.2084	0.1588	0.3993	0.2637	0.1904	0.3953	0.2793	0.2201	0.4842	0.4114	0.2640
	PICP	0.9440	0.9120	0.8613	0.9467	0.9200	0.8693	0.9412	0.9037	0.8743	0.9517	0.9223	0.8445

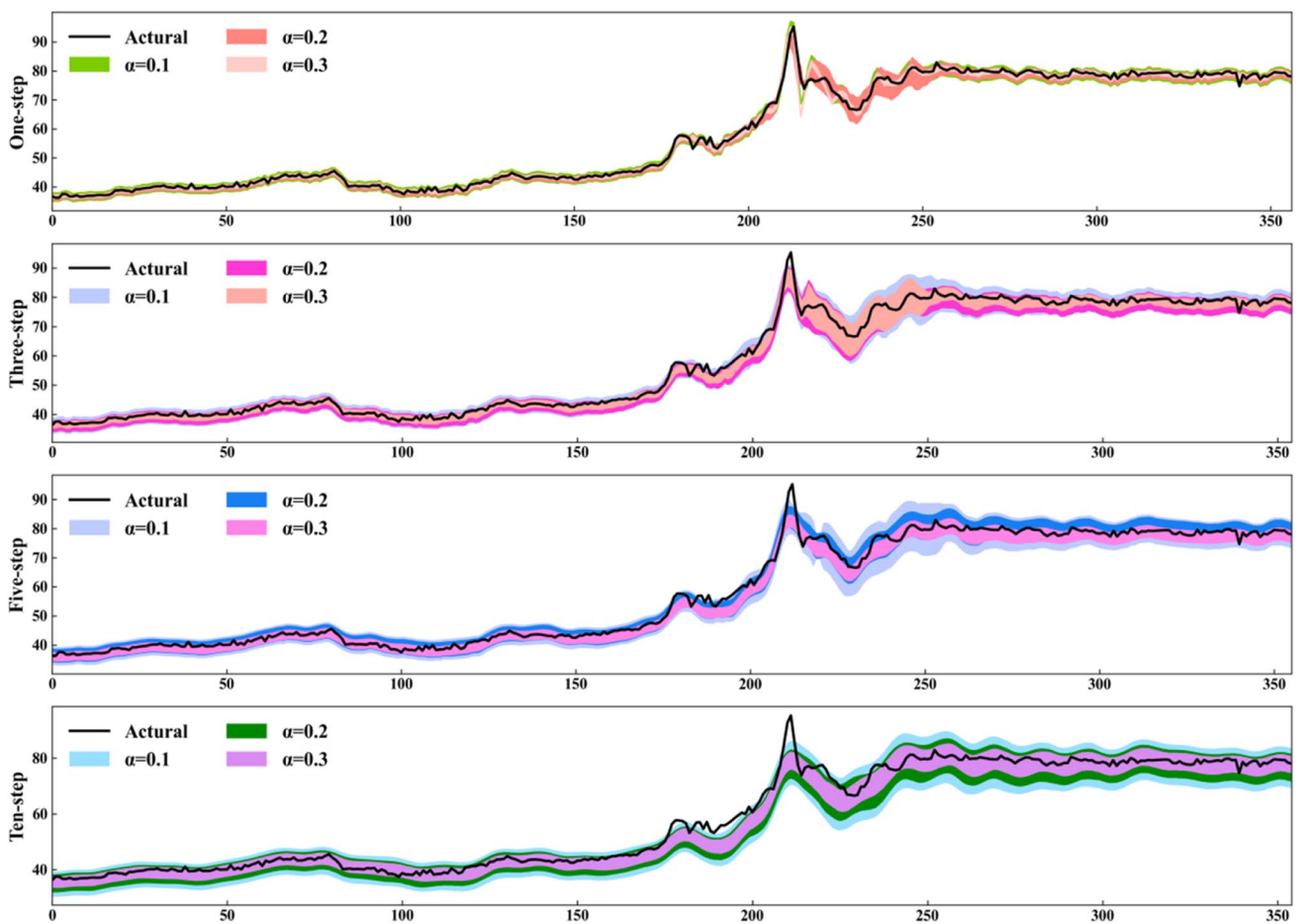


Fig. 10. Interval forecasting results of the Guangdong market

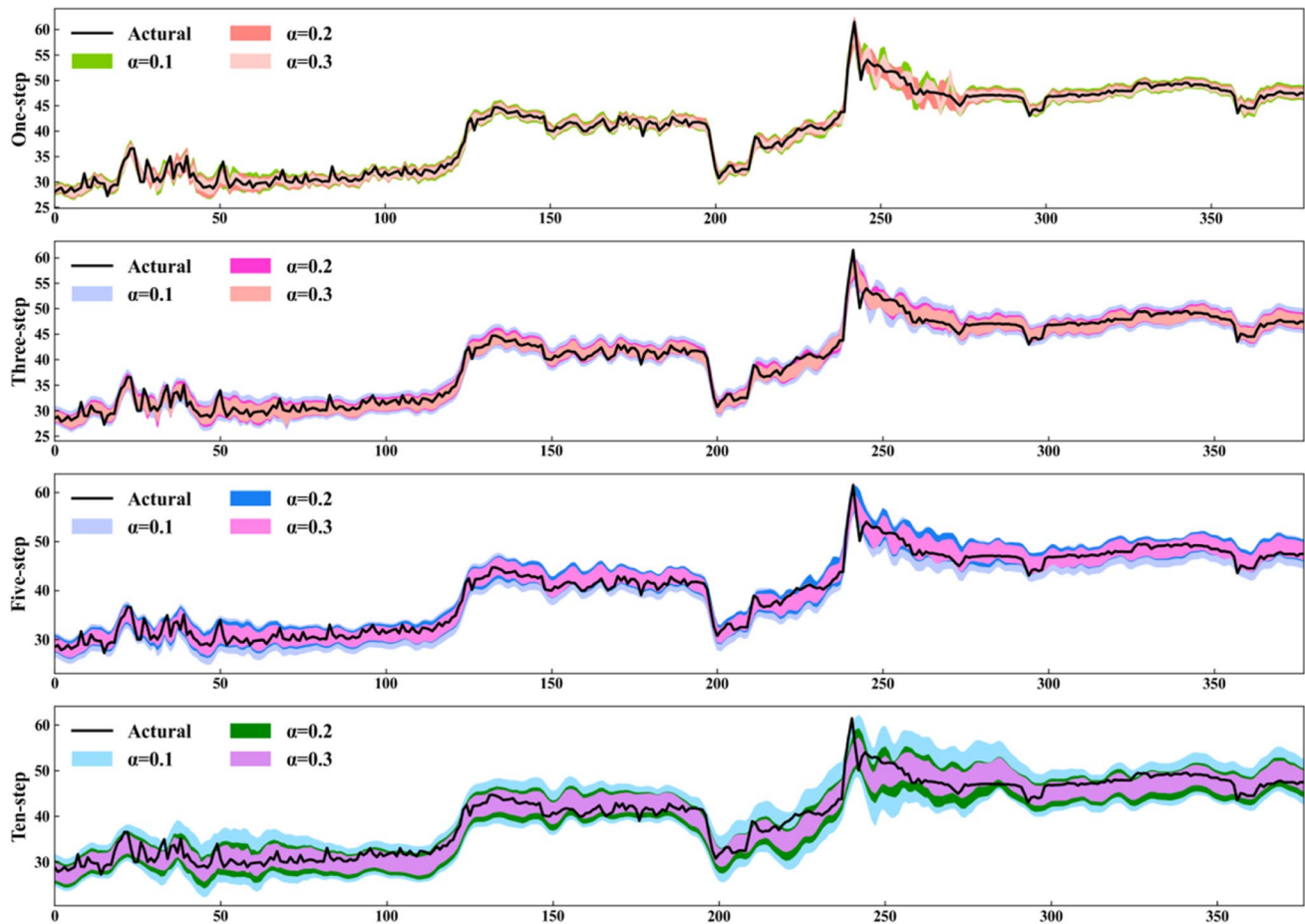


Fig. 11 Interval forecasting results of the Hubei market

Comparison V: the comparison of different step-ahead interval forecasting. The comparison results in Figs. 5, 6, 7, 8, and 9 and Tables 6, 7, 8, and 9 show that as the forecasting step increases, the forecasting accuracy of the model gradually decreases because of the lack of information required for forecasting. A single model often experiences varying degrees of forecasting lag. As the number of forecasting steps increases, a single model tends to consider the current price as the best forecast and learns only a simple mapping, thereby failing to capture the relevant features in the data. Conversely, the decomposition algorithm extracts the features of different carbon price time scales by decomposition. Thus, the model can learn smoother features, thereby effectively eliminating this phenomenon. As observed from the fitted curves in Figs. 5, 6, 7, 8, and 9, the prediction results of the hybrid forecasting model for the one-step and three-step-ahead forecasts will contain the fluctuation characteristics of the rapid oscillation in the sequence. In contrast, for the five-step and ten-step-ahead forecasts, the prediction results are smoother because learning the features in the high-frequency imfs for the larger span of steps is difficult

for the model. These characteristics are almost unpredictable in large time step gaps because high-frequency imfs represent the characteristics of the short-term fluctuations of carbon price with strong volatility and sharp fluctuations. In this situation, the model can only learn the characteristics of medium- and low-frequency imfs, which represent the low frequency of the medium- and long-term trends in carbon prices. These imfs are more periodic and regular, which leads to smoother prediction results in one-step and three-step-ahead forecasts compared with five-step and ten-step-ahead forecasts.

Interval forecasting

Interval forecasts of carbon prices hold significant relevance in capturing carbon price uncertainty. They serve as valuable tools for assessing the future volatility of carbon prices, thus providing decision-makers insights that deterministic forecasts alone cannot provide. This study is based on the HI-TVFMED-transformer and utilizes the quantile loss function

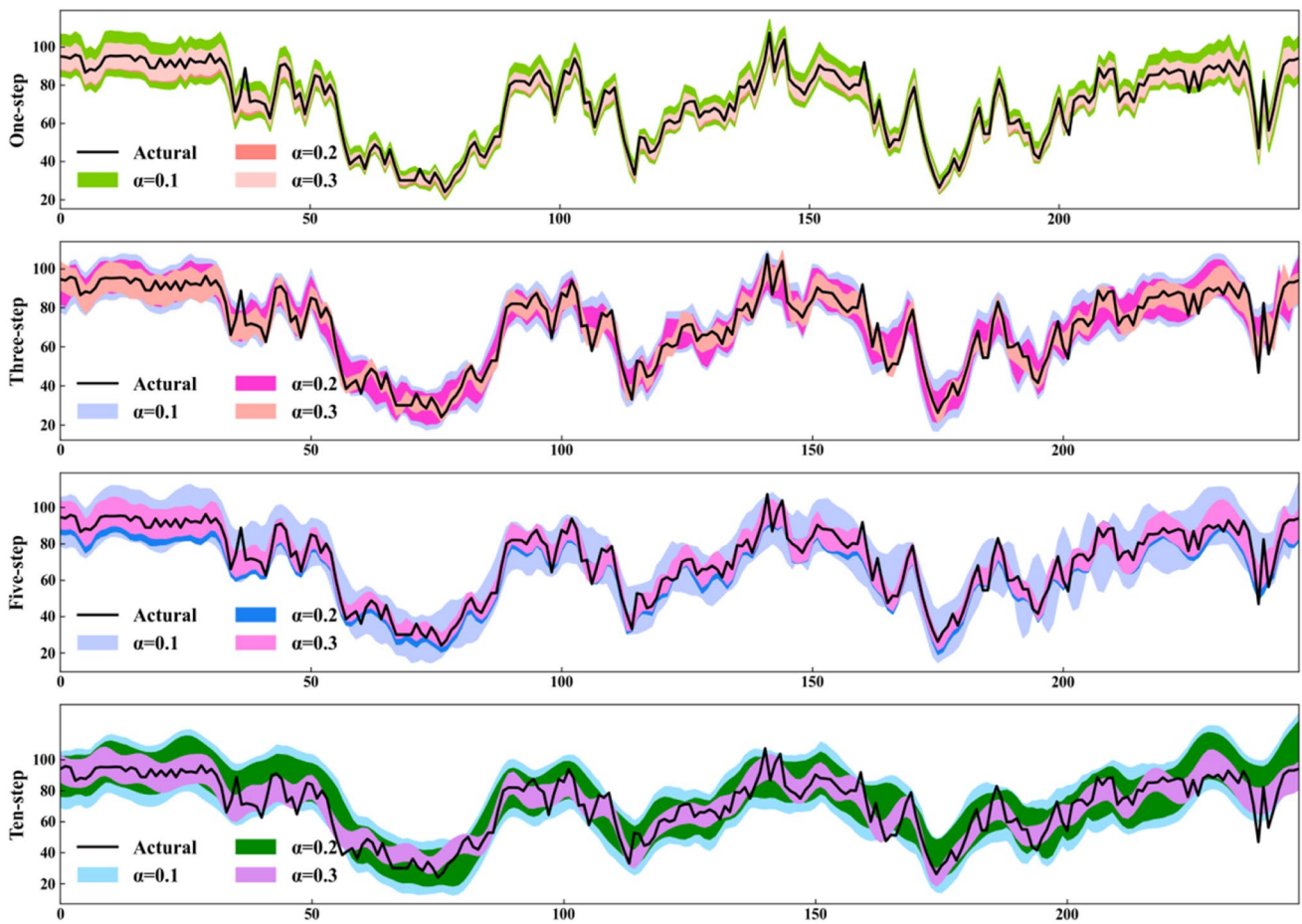


Fig. 12 Interval forecasting results of the Beijing market

for interval forecasting of carbon prices. The significance level α plays a crucial role in determining the width and coverage of the interval forecasting. To compare the performance of interval forecasting at various significance levels, this study considers different levels of α ($\alpha = 0.1, 0.2, 0.3$) for conducting interval forecasting. Table 11 and Figs. 10, 11, 12, 13, and 14 show the results of the interval forecasting for each carbon market.

We can draw the following conclusions from Table 11 and Figs. 10, 11, 12, 13, and 14. The prediction results will change with α . Guangdong is taken as an example, $\alpha = 0.1$, $AW = 3.7517$, $PINAW = 0.0636$, $PICP = 0.9552$; $\alpha = 0.2$, $AW = 2.9950$, $PINAW = 0.0508$, $PICP = 0.9356$; and $\alpha = 0.3$, $AW = 1.3256$, $PINAW = 0.0225$, $PICP = 0.8095$. With the increase in significance level, the interval width and coverage of the interval prediction gradually decreases, indicating that the trade-off between width and coverage is needed during interval forecasting, wherein high coverage but large width or small width but low coverage result in less information contained in the prediction interval. Further comparison

reveals that the AW of one-step-ahead interval forecast for the Beijing and Shenzhen carbon markets exceeds 15, whereas for other carbon markets, it hovers around 5. Additionally, when considering ten-step-ahead forecasts, the AW for the two aforementioned carbon markets exceeds 25, while for other carbon markets, it remains around 10. The interval forecast needs a larger width to cover the fluctuations of these two carbon markets, indicating that their volatility is significantly greater than those of other carbon markets. The AW of the Hubei carbon market is the smallest in almost all cases, thereby indicating a stable trend of low volatility. Moreover, the AW gradually increases with the same confidence interval as the forecasting step expands. For example, $AW = 5.1689$ for one-step-ahead and $AW = 11.7828$ for ten-step-ahead in Shanghai at $\alpha = 0.1$. As the forecasting steps increase, the carbon price will have more significant fluctuations relative to the current price, thereby resulting in a larger estimation interval. The interval prediction based on HI-TVFEMD-transformer can explain the volatility of the carbon price well, indicating that the result is good and credible.

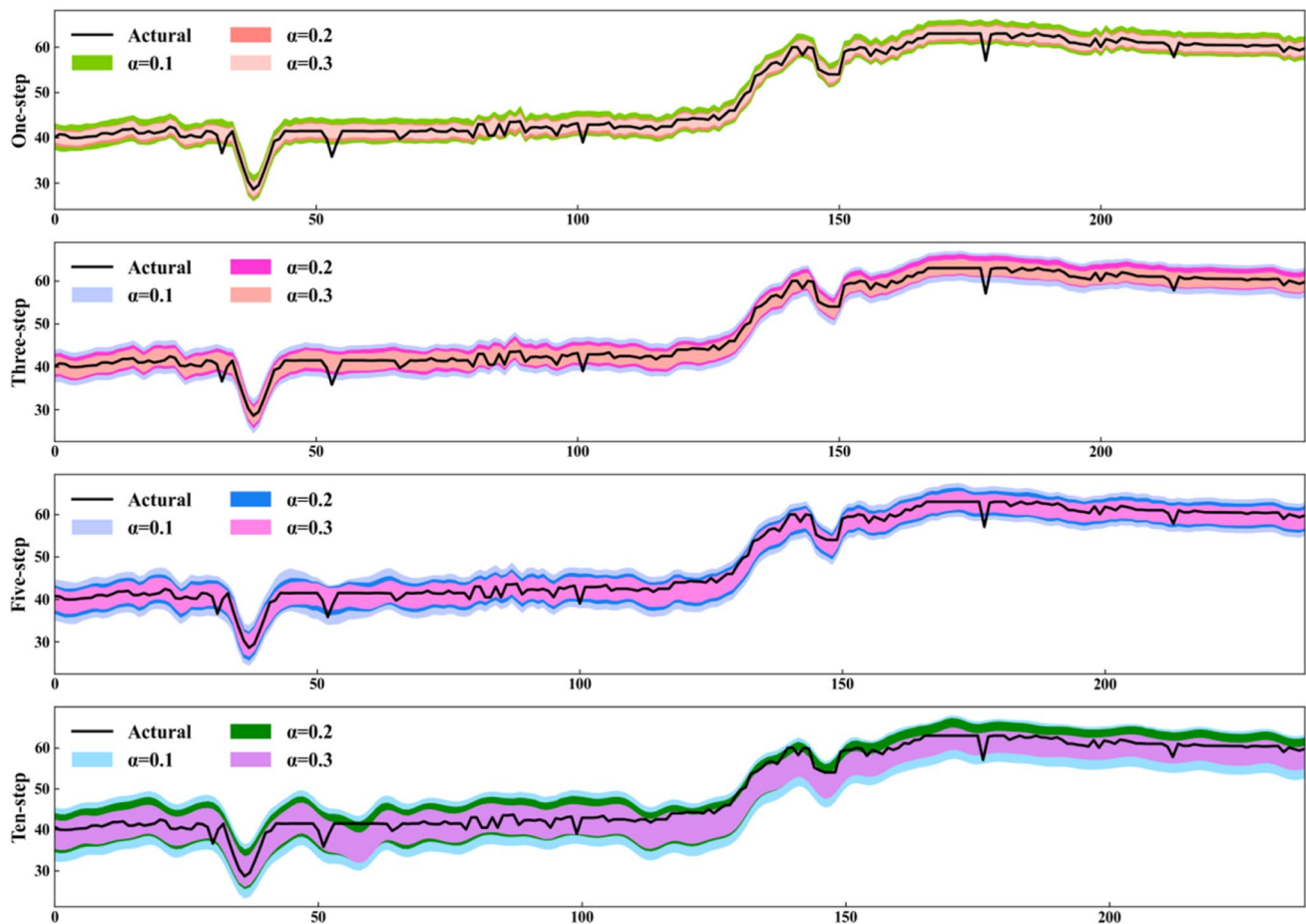


Fig. 13 Interval forecasting results of the Shanghai market

Conclusion

A hybrid forecasting model that combines HI, TVFEMD, and transformer is proposed to improve the accuracy of carbon price forecasting. First, HI is used to identify and correct any outliers present in the carbon price series. Then, the original carbon price series is decomposed into multiple imfs using TVFEMD to reduce non-stationarity, and the imfs are reconstructed by SE to reduce the time consumed by the model while retaining its complex features. Further, OATM optimized the hyperparameters of the transformer deep learning technique to select a reasonable hyperparameter configuration. Finally, Transformer and quantile loss functions are used for the multi-step-ahead and interval forecasting of carbon prices. To verify the validity of the proposed model from several aspects, five real carbon price data and four different forecasting steps were selected, and multiple models and decomposition methods were compared. Subsequently, interval forecasting was performed for each carbon market to measure carbon price uncertainty.

The main conclusions drawn from this study are presented as follows:

- (1) On the one hand, the experimental results indicate that transformer outperforms SVM, LSTM, GRU, and TCN because of its excellent ability to extract long-term dependence and global features in sequences. On the other hand, TVFEMD exhibited better performance in characterizing the multiscale time–frequency features of carbon price sequences than EMD, EEMD, and CEEM-DAN. Moreover, the prediction accuracy achieved by the proposed HI-TVFEMD-transformer hybrid forecasting model was better than most forecasting models, and its robustness and stability are demonstrated in multi-step-ahead forecasting. In most of the experiments, HI improves the prediction accuracy. In contrast, the effects of HI in a small proportion of the experiments are not evident, because although the outliers in the sequence can be eliminated to make the sequence more regular, they will make the sequence lose some fluctuation features.

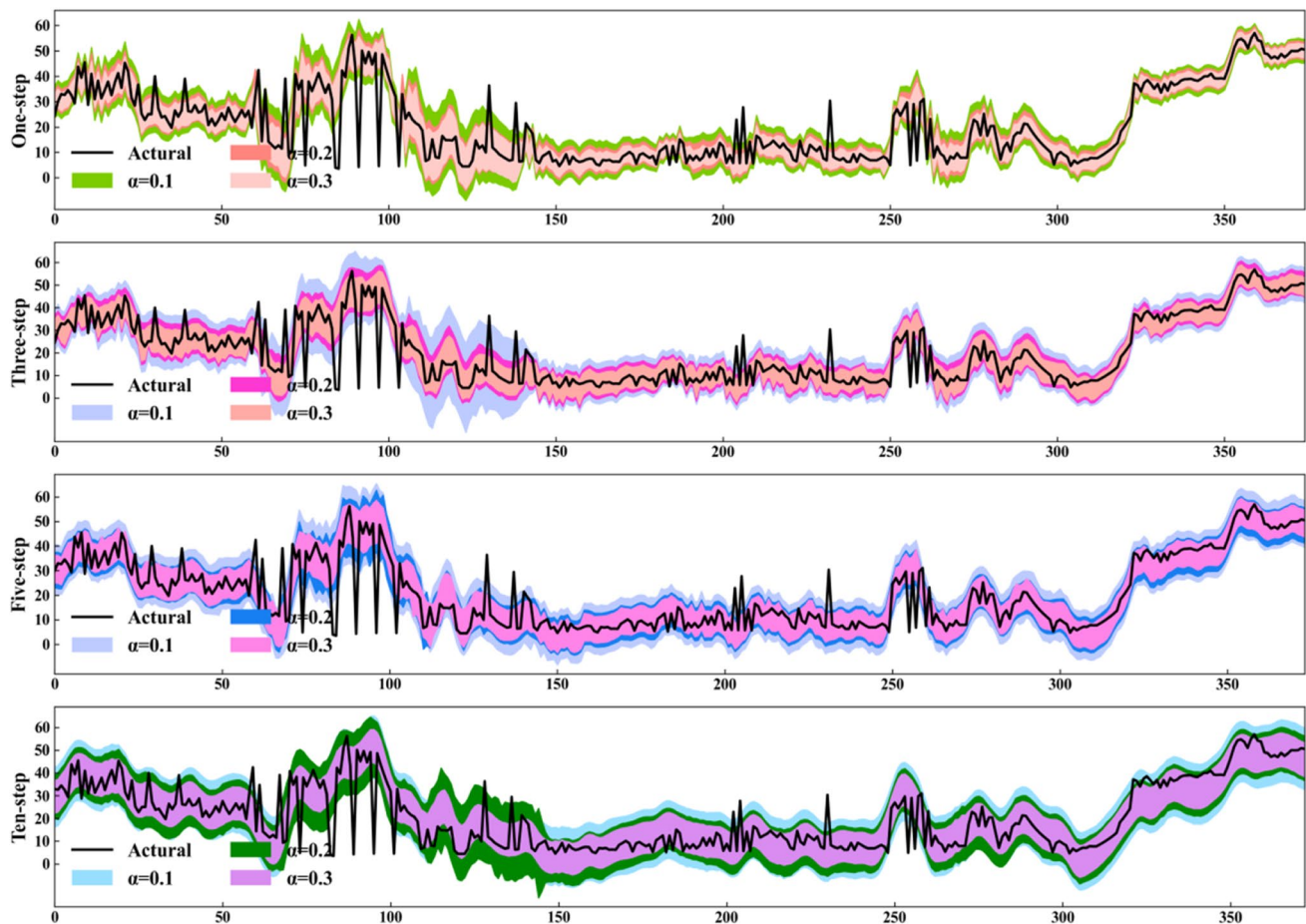


Fig. 14 Interval forecasting results of the Shenzhen market

- (2) The proposed hybrid model exhibits a high level of precision in forecasting carbon prices at multiple levels, encompassing deterministic and uncertain. Therefore, it can provide practical information for investors and policymakers. In the near future, investors can formulate investment strategies based on the results of multi-step-ahead forecasting and understand the possible risks through interval forecasting to avoid them. Meanwhile, policymakers can monitor the changes in carbon price fluctuations through the forecasting results in time and take corresponding measures to stabilize the prices when fluctuations are large.

Although the proposed hybrid forecasting model in this paper has shown excellent forecasting ability during the experiments, it still has room for improvement. First, the use of outlier identification and correction techniques can be considered. Also, more advanced outlier testing or denoising techniques can be used to process carbon prices in future research. Second, determining the SE threshold is based on the distribution and prior experience. However,

it is considering whether a more scientific method can be used to determine the threshold of SE. Finally, carbon prices are affected by several factors, such as fossil fuel prices and economic development levels. Hence, studying how to model these factors is also worthwhile.

Acknowledgements The authors express deep appreciation to the editors and reviewers for reading the manuscript tediously and for providing valuable suggestions and remarks.

Author contribution Wang Yue wrote and revised the original manuscript. Wang Xiaoyi Wang performed data collection and treatment. Wang Zhong contributed to the conception of the study. Kang Xinyu reviewed and revised the manuscript. All authors commented on previous versions of the manuscript and read and approved the final manuscript.

Funding This research was funded by the Project of Sichuan Oil and Natural Gas Development Research Center (Grant No. SKB20-06) and the Strategic Research and Consulting Project of the Chinese Academy of Engineering (2022-28-33, 2023-HZ-10).

Data availability All data analyzed in the duration of this study are included in the supplementary information files and available from the corresponding author upon reasonable request.

Declarations

Ethics approval This study follows all ethical practices during writing.

Consent to participate This is not applicable.

Consent for publication This is not applicable.

Competing interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- Benz E, Trück S (2009) Modeling the price dynamics of CO₂ emission allowances. *Energy Econ* 31:4–15
- Bommidi BS, Teeparthi K, Kosana V (2023) Hybrid wind speed forecasting using ICEEMDAN and transformer model with novel loss function. *Energy* 265:126383
- Byun SJ, Cho H (2013) Forecasting carbon futures volatility using GARCH models with energy volatilities. *Energy Econ* 40:207–221
- Du Y, Chen K, Chen S, Yin K (2022) Prediction of carbon emissions trading price in Fujian province: based on BP neural network model. *Front Energy Res* 10:939602
- Fan X, Lv X, Yin J, Tian L, Liang J (2019) Multifractality and market efficiency of carbon emission trading market: analysis using the multifractal detrended fluctuation technique. *Appl Energy* 251:113333
- Graves, A. (2013) Generating sequences with recurrent neural networks. arXiv preprint arXiv:1308.0850
- Han M, Ding L, Zhao X, Kang W (2019) Forecasting carbon prices in the Shenzhen market, China: the role of mixed-frequency factors. *Energy* 171:69–76
- Hao Y, Tian C, Wu C (2020) Modelling of carbon price in two real carbon trading markets. *J Clean Prod* 244:118556
- He Y, Li H (2018) Probability density forecasting of wind power using quantile regression neural network and kernel density estimation. *Energy Convers Manag* 164:374–384
- Huang Y, Dai X, Wang Q, Zhou D (2021) A hybrid model for carbon price forecasting using GARCH and long short-term memory network. *Appl Energy* 285:116485
- Li H, Li Z, Mo W (2017) A time varying filter approach for empirical mode decomposition. *Signal Process* 138:146–158
- Lim B, Arik SÖ, Loeff N, Pfister T (2021) Temporal fusion transformers for interpretable multi-horizon time series forecasting. *Int J Forecast* 37:1748–1764
- Liu H, Shen L (2019) Forecasting carbon price using empirical wavelet transform and gated recurrent unit neural network. *Carbon Manag* 11:25–37
- Lutz BJ, Pigorsch U, Rotfuss W (2013) Nonlinearity in cap-and-trade systems: The EUA price and its fundamentals. *Energy Econ* 40:222–232
- Ma B, Zhang T (2020) Single-channel blind source separation for vibration signals based on TVF-EMD and improved SCA. *IET Signal Proc* 14:259–268
- Martins, A. F. T. and Fernandez Astudillo, R. (2016) From Softmax to Sparsemax: a sparse model of attention and multi-label classification. arXiv preprint arXiv:1602.02068
- Meka R, Alaeddini A, Bhaganagar K (2021) A robust deep learning framework for short-term wind power forecast of a full-scale wind farm using atmospheric variables. *Energy* 221:119759
- Niu X, Wang J (2019) A combined model based on data preprocessing strategy and multi-objective optimization algorithm for short-term wind speed forecasting. *Appl Energy* 241:519–539
- Peters, B., Niculae, V. and Martins, A. F. T. (2019) Sparse sequence-to-sequence models. arXiv preprint arXiv:1905.05702
- Richman JS, Moorman JR (2000) Physiological time-series analysis using approximate entropy and sample entropy. *Am J Phys Heart Circ Phys* 278:H2039–H2049
- Sun G, Chen T, Wei Z, Sun Y, Zang H, Chen S (2016) A carbon price forecasting model based on variational mode decomposition and spiking neural networks. *Energies* 9:54
- Sun S, Jin F, Li H, Li Y (2021) A new hybrid optimization ensemble learning approach for carbon price forecasting. *Appl Math Model* 97:182–205
- Sun W, Huang C (2020a) A carbon price prediction model based on secondary decomposition algorithm and optimized back propagation neural network. *J Clean Prod* 243:118671
- Sun W, Huang C (2020b) A novel carbon price prediction model combines the secondary decomposition algorithm and the long short-term memory network. *Energy* 207:118294
- Sun W, Xu C (2021) Carbon price prediction based on modified wavelet least square support vector machine. *Sci Total Environ* 754:142052
- Sun W, Zhang C (2018) Analysis and forecasting of the carbon price using multi-resolution singular value decomposition and extreme learning machine optimized by adaptive whale optimization algorithm. *Appl Energy* 231:1354–1371
- Sun W, Zhang J (2020) Carbon price prediction based on ensemble empirical mode decomposition and extreme learning machine optimized by improved bat algorithm considering energy price factors. *Energies* 13:3471
- Taylor KE (2001) Summarizing multiple aspects of model performance in a single diagram. *J Geophys Res Atmos* 106:7183–7192
- Tian C, Hao Y (2020) Point and interval forecasting for carbon price based on an improved analysis-forecast system. *Appl Math Model* 79:126–144
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I. (2017) Attention is all you need. arXiv preprint arXiv:1706.03762
- Wang J, Du P, Hao Y, Ma X, Niu T, Yang W (2020a) An innovative hybrid model based on outlier detection and correction algorithm and heuristic intelligent optimization algorithm for daily air quality index forecasting. *J Environ Manag* 255:109855
- Wang J, Niu X, Zhang L, Lv M (2021) Point and interval prediction for non-ferrous metals based on a hybrid prediction framework. *Res Policy* 73:102222
- Wang J, Qiu S (2021) Improved multi-scale deep integration paradigm for point and interval carbon trading price forecasting. *Mathematics* 9:2595
- Wang K, Fu W, Chen T, Zhang B, Xiong D, Fang P (2020b) A compound framework for wind speed forecasting based on comprehensive feature selection, quantile regression incorporated into convolutional simplified long short-term memory network and residual error correction. *Energy Convers Manag* 222:113234
- Wang Y, Wang Z, Kang X, Luo Y (2022) A novel interpretable model ensemble multivariate fast iterative filtering and temporal fusion transform for carbon price forecasting. *Energy Sci Eng* 11:1148–1179
- Wen, Q., Zhou, T., Zhang, C., Chen, W., Ma, Z., Yan, J. and Sun, L. (2022) Transformers in time series: a survey. arXiv preprint arXiv:2202.07125
- Wen, R., Torkkola, K., Narayanaswamy, B. and Madeka, D. (2017) A multi-horizon quantile recurrent forecaster. arXiv preprint arXiv:1711.11053
- Wiegrefe, S. and Pinter, Y. (2019) Attention is not not explanation. arXiv preprint arXiv:1908.04626

- Wu, N., Green, B., Ben, X. and O'Banion, S. (2020) Deep transformer models for time series forecasting: the influenza prevalence case. arXiv preprint arXiv: 2001.08317v08311
- Wu Z, Huang NE (2004) Ensemble empirical mode decomposition: a noise-assisted data analysis method, vol 193. Centre for Ocean-Land-Atmosphere Studies Technical Report, p 51
- Xiong D, Fu W, Wang K, Fang P, Chen T, Zou F (2021) A blended approach incorporating TVFEMD, PSR, NNCT-based multi-model fusion and hierarchy-based merged optimization algorithm for multi-step wind speed prediction. *Energy Convers Manag* 230:113680
- Yang Y, Guo H, Jin Y, Song A (2021) An ensemble prediction system based on artificial neural networks and deep learning methods for deterministic and probabilistic carbon price forecasting. *Front Environ Sci* 9:740093
- Yao Z, Xie J, Tian Y, Huang Q (2019) Using Hampel identifier to eliminate profile-isolated outliers in laser vision measurement. *J Sens* 2019:1–12
- Yi L, Li Z, Yang L, Liu J (2017) The scenario simulation analysis of the EU ETS carbon price trend and the enlightenment to China. *J Environ Econ* 2017:22–35
- Yun P, Huang X, Wu Y, Yang X (2022) Forecasting carbon dioxide emission price using a novel mode decomposition machine learning hybrid model of CEEMDAN-LSTM. *Energy Science & Engineering*
- Zhang F, Wen N (2022) Carbon price forecasting: a novel deep learning approach. *Environ Sci Pollut Res* 29:54782–54795
- Zhang, X., Chen, X., Yao, L., Ge, C. and Dong, M. (2019a) Deep neural network hyperparameter optimization with orthogonal array tuning. arXiv preprint arXiv:1907.13359
- Zhang, X., Yao, L., Huang, C., Sheng, Q. Z. and Wang, X. (2017) Intent recognition in smart living through deep recurrent neural networks. arXiv preprint arXiv:1702.06830
- Zhang Y, Liu K, Qin L, An X (2016) Deterministic and probabilistic interval prediction for short-term wind power generation based on variational mode decomposition and machine learning methods. *Energy Convers Manag* 112:208–219
- Zhang Z, Qin H, Liu Y, Yao L, Yu X, Lu J, Jiang Z, Feng Z (2019b) Wind speed forecasting based on quantile regression minimal gated memory network and kernel density estimation. *Energy Convers Manag* 196:1395–1409
- Zhao Y, Zhao H, Li B, Wu B, Guo S (2022) Point and interval forecasting for carbon trading price: a case of 8 carbon trading markets in China. Research Square
- Zhou F, Huang Z, Zhang C (2022) Carbon price forecasting based on CEEMDAN and LSTM. *Appl Energy* 311:118601
- Zhu B, Han D, Wang P, Wu Z, Zhang T, Wei Y-M (2017) Forecasting carbon price using empirical mode decomposition and evolutionary least squares support vector regression. *Appl Energy* 191:521–530
- Zhu B, Wei Y (2013) Carbon price forecasting with a novel hybrid ARIMA and least squares support vector machines methodology. *Omega* 41:517–524
- Zhu BZ, Ye SX, Wang P, He KJ, Zhang T, Wei YM (2018) A novel multiscale nonlinear ensemble leaning paradigm for carbon price forecasting. *Energy Econ* 70:143–157

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.