



# Prediction of NO<sub>x</sub> emission concentration from coal-fired power plant based on joint knowledge and data driven

Zheng Wu<sup>a,b</sup>, Yue Zhang<sup>a,b</sup>, Ze Dong<sup>a,b,\*</sup>

<sup>a</sup> School of Control and Computer Engineering, North China Electric Power University, Beijing, 102206, China

<sup>b</sup> Hebei Technology Innovation Center of Simulation & Optimized Control for Power Generation, Baoding, 071003, China

## ARTICLE INFO

Handling Editor: Isabel Soares

### Keywords:

NO<sub>x</sub> emission concentration prediction

Knowledge driven

Data driven

Informer

Modal decomposition

## ABSTRACT

Accurate NO<sub>x</sub> concentration prediction is of great significance for the pollutant emission control and safe operation of coal-fired power plants. The global properties of the research object cannot be adequately described by a single data driven model, which hinders generalization performance. We propose a NO<sub>x</sub> emission concentration prediction method based on joint knowledge and data driven. First, we introduce a knowledge driven combined feature selection method to provide a global feature basis for data driven modeling. Second, we enable adaptive decomposition of the variational modal decomposition (VMD) using the modal energy difference and sample entropy. The method can extract deep time-frequency information in nonlinear and non-smooth features. Finally, we use the Informer combined with an adaptive time series segmentation method to predict NO<sub>x</sub> concentration. The experimental results indicate that the proposed method predicts the NO<sub>x</sub> concentration better than several comparative models.

## 1. Introduction

One of the main sources of atmospheric pollutants, nitrogen oxides (NO<sub>x</sub>), is produced in huge quantities during the operation of coal-fired power plants [1]. During the unit denitrification reaction, accurate NO<sub>x</sub> concentration measurement is the foundation for determining an appropriate amount of ammonia injection [2]. An insufficient amount of ammonia injection will cause ammonia or NO<sub>x</sub> to escape and harm the environment [3]. Continuous emission monitoring system (CEMS) [4] is widely used in thermal power units to track and monitor NO<sub>x</sub> concentrations in real time. However, CEMS measurement is hampered by issues such as a harsh working environment, signal interference, and its own failure [5]. Therefore, it is critical to develop an accurate NO<sub>x</sub> emission concentration prediction model from coal-fired units for denitrification and environmental protection [6].

A variety of methods for predicting NO<sub>x</sub> concentrations have been proposed. These models can be classified into knowledge driven [7–10] and data driven models [12–14] according to their principles.

Knowledge driven models primarily use NO<sub>x</sub> generation and fuel combustion mechanisms in boilers to simulate the combustion process and predict NO<sub>x</sub> emissions. Zhao et al. [7] established a mathematical model for real-time thermal efficiency and NO<sub>x</sub> concentration in boilers

based on combustion mechanisms. Chang et al. [8] proposed an integrated CFD model that included coal combustion and NO<sub>x</sub> emissions. Devarakonda et al. [9] developed a kinetic model based on the SCR system's reaction mechanism and investigated the effects of various factors on denitrification efficiency. However, the knowledge driven model requires a number of assumptions that are challenging to meet in actual production, and the model parameters are numerous and complicated to calculate [10]. Therefore, the knowledge driven model is difficult to use for fast and accurate NO<sub>x</sub> concentration prediction.

Coal-fired power plants with Distributed Control Systems (DCS) and Supervision Information Systems (SIS) provide a wealth of field operation data for data driven modeling [11]. Tuttle et al. [12] compared 10 data driven methods for NO<sub>x</sub> emissions from thermal power units. The results showed that the GRU was the best performing prediction model. Wang [13] and Tang [14] et al. recently proposed two hybrid models (CEEMDAN-AM-LSTM and AE-ELM), which also showed superior performance in NO<sub>x</sub> concentration prediction. However, data driven models are constrained by the manner, size, and quality of the sample data chosen. The model frequently performs inadequately in terms of generalization because it can typically only describe local features of the research object [15].

The proposed joint knowledge and data driven method has recently

\* Corresponding author. School of Control and Computer Engineering, North China Electric Power University, Beijing, 102206, China.

E-mail addresses: [wuzhengncepu@ncepu.edu.cn](mailto:wuzhengncepu@ncepu.edu.cn) (Z. Wu), [zhyhot\\_zhyhot@sina.com](mailto:zhyhot_zhyhot@sina.com) (Y. Zhang), [dongze@ncepu.edu.cn](mailto:dongze@ncepu.edu.cn) (Z. Dong).

realized the organic combination of global and local features, rules, and object experiences. Based on this concept, many joint driven models with improved performance have been proposed [16,17]. The joint driven models have achieved excellent results in research areas such as group intelligent decision making [18], fault diagnosis [19], traffic prediction [20], and smart manufacturing [21]. However, the application of joint driven modeling for NOx concentration prediction is still in the exploratory stage. Li et al. [15] reviewed the main applications of the current joint driven models in power systems. The joint driven method's superior performance and future prospects are illustrated. Also, several researchers, including Li et al. [22], Yi et al. [23], and Wang et al. [24], confirmed that the joint driven method produces excellent results in a variety of power system application scenarios. Therefore, this research attempts to apply the concepts of joint knowledge and data driven modeling to predict the NOx emission concentration.

However, the operational data of SCR systems exhibits nonlinearity and randomness. These traits make it more difficult for the model to capture the time-frequency features of NOx concentration prediction, lowering forecast accuracy. To handle such challenges, a hybrid model incorporating signal decomposition and deep learning is efficient [25].

Variational mode decomposition (VMD) is a popular method for signal decomposition [26]. It can decompose a signal into multiple Band-limited Intrinsic Mode Function (BIMF) with limited bandwidth in order to extract different frequency features and reduce nonlinearity and randomness interference [27]. Sun et al. [28] used VMD to describe the unstable and stochastic characteristics as steady-state BIMF, which improved the prediction performance of long and short-term memory neural network (LSTM). To deal with the strong nonlinear characteristics of wind power prediction, Xiong et al. [29] used VMD with nonlinear weighted combined learning. Experiments revealed the addition of VMD led to a significant improvement in prediction accuracy. However, one of the major challenges is the inability of VMD to decompose adaptively.

Deep learning excels at uncovering hidden quantitative relationships in time-series data. Transformer [30] provides a new deep network architecture for time series prediction. However, the problems of high time complexity and memory utilization increase the difficulty of its application [31]. Zhou et al. [32] proposed an efficient informer model. It solves the problem by using the ProbSparse self-attention mechanism and self-attention distilling operation. The model achieved superior prediction performance on four large-scale datasets. Gong et al. [33] developed a load prediction method for centralized heating systems using the informer. The experiments proved that the method is generalizable and robust. These researches demonstrate that informer offers a novel solution for improving the accuracy of time series prediction.

This paper proposes a hybrid model based on joint knowledge and data driven. First, we integrate the NOx generation mechanism and utilize a knowledge driven combined feature selection method to calculate the variable correlation and offer the model's feature base. Second, we introduce modal energy difference and sample entropy to enhance VMD and extract deep time-frequency information from nonlinear feature parameters. Finally, we employ informer to predict NOx emission concentrations following adaptive segmentation.

The contributions of this paper are as follows.

- A hybrid model based on joint knowledge and data driven is proposed. The NOx emission concentration prediction challenge is accomplished by integrating NOx generating mechanism knowledge, combined feature selection, signal decomposition, and deep learning.
- A knowledge driven combined feature selection is established to mine the best collection of features extensively and avoid big disparities in describing the correlation of variables by a single method.
- Nonlinear time-series features are decomposed using a novel signal decomposition method (MEVMD). The decomposed residuals are filtered based on the permutation entropy measurements.

**Table 1**  
Main parameters of boiler system.

No.	Parameter description	Unit	Design	Check
1	Superheated steam flow	t/h	2955	2955
2	Superheated steam pressure	MPa(a)	27.56	27.56
3	Superheated steam temperature	°C	605	605
4	Reheat steam flow	t/h	2448	2448
5	Reheater inlet pressure	MPa(g)	6.2	6.2
6	Reheater outlet pressure	MPa(g)	6.0	6.0
7	Reheater inlet temperature	°C	377	377
8	Reheater outlet temperature	°C	603	603
9	Economizer inlet temperature	°C	295	295
10	Preheater inlet primary air temperature	°C	23	23
11	Preheater inlet secondary air temperature	°C	27	27
12	Preheater outlet primary air temperature	°C	330	334
13	Preheater outlet secondary air temperature	°C	340	343
14	Exhaust gas temperature	°C	131	134
15	Guaranteed efficiency BRL condition	%	125	128
16	Minimum stable load	%	93.72	–
17	Air preheater leakage rate	%	30	30
18	NOx emission	mg/m <sup>3</sup>	6	6

- The time series are adaptively segmented by the slicing module in the input layer of the encoder stack. The relationship between historical NOx concentration and future NOx concentration is described.
- Informer is applied to the prediction of NOx emission concentration. Deep learning is utilized to capture time-series data dependent coupling.

The rest of this paper is organized as follows. Section 2 describes the boiler in this research and NOx generation mechanism. Section 3 discusses the VMD, Informer, and Methodology. Section 4 presents the proposed prediction model in detail. Section 5 conducts a comparison experiment to validate the model performance. The final section concludes this paper.

## 2. The boiler and mechanistic knowledge

### 2.1. Description of the boiler object

In this research, the DC boiler of Shanghai Caojing Power Plant's 1000 MW ultra-supercritical coal-fired generating unit was used. Shanghai Boiler Plant Co. manufactured the boiler, which is a variable pressure operating spiral tube coil water-cooled wall DC furnace. It is a coal-fired boiler with a single chamber, primary intermediate reheat, four-corner cut circle combustion method, balanced ventilation, solid slag discharge, all-steel suspension structure, tower type, and open-air arrangement. The furnace height is 320.8 m, with a cross section of 156.7 m × 156.7 m. The low NOx rotary burner is arranged in the wall-type boiler. The boiler tail flue is equipped with a silicon controlled device. Ammonia is used as the reducing agent to control NOx concentration. The main parameters of the boiler are shown in Table 1.

### 2.2. Mechanistic analysis of NOx generation

Selective Catalyst Reduction (SCR) is a complex chemical reaction system. It mainly consists of a denitrification reaction system, an ammonia storage and supply system, and an ammonia/air spray system. The SCR reactor usually adopts a high ash arrangement, i.e., it is arranged in the flue between the coal saver and the air preheater.

SCR denitrification technology is characterized by no secondary pollution, simple device structure, reliable operation, and easy maintenance. Under reasonable conditions and suitable temperature conditions, denitrification efficiency can reach 80–90%. Fig. 1 shows the overall structure and production flow chart of the SCR denitrification plant for the 1000 MW ultra-supercritical coal-fired unit used in this paper.

Multiphase catalytic reactions generally follow the Eley-Rideal

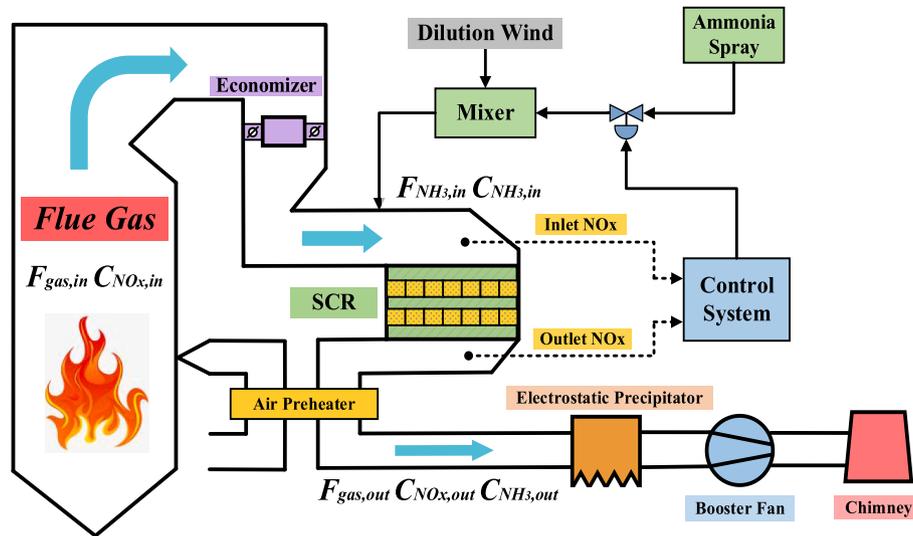


Fig. 1. The overall structure and production process of SCR denitration plant.

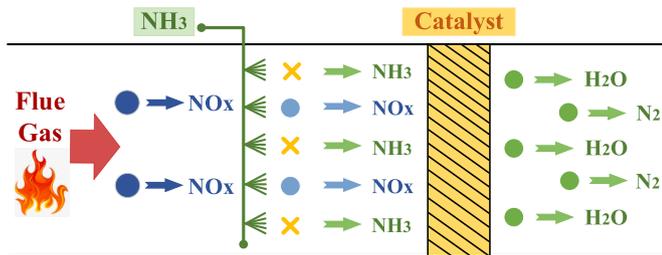
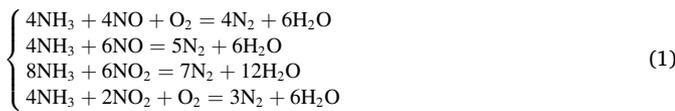
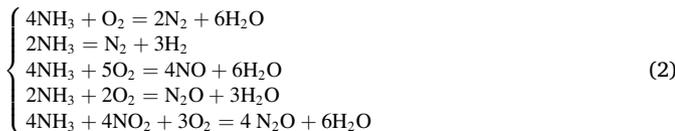


Fig. 2. SCR denitrification reaction schematic.

mechanism [34] and the Langmuir-Hinshelwood mechanism [35]. Most scholars currently believe that SCR denitrification follows the reaction Eley-Rideal mechanism [36]. Ammonia (NH<sub>3</sub>) is used as the reducing agent and V<sub>2</sub>O<sub>5</sub>/TiO<sub>2</sub> catalyst is filled in the SCR reactor. NH<sub>3</sub> is mixed with flue gas and a series of chemical reactions take place under the flue gas temperature conditions of 280–420 °C. After the denitrification reaction, the NO<sub>x</sub> in the flue gas is reduced to non-polluting nitrogen (N<sub>2</sub>) and water (H<sub>2</sub>O) [37]. The schematic diagram of the SCR flue gas denitrification reaction is given in Fig. 2. Among them, the main reactions are:



In addition to the main reactions, the following secondary reactions may be present:



The flue gas flow path in the SCR denitrification system is not unique depending on how the equipment is configured. However, the reaction process of gaseous NO<sub>x</sub> in flue gas is generally similar. The flue gas is mixed with ammonia by jets and deflectors before entering the reactor to begin the denitrification process. Denitrification is completed by diffusion from the micropores to the outer surface into the flue gas stream after a series of complex chemical reactions and product desorption.

Through the thermodynamic analysis of the reaction mechanism of

Table 2  
Candidate input variables.

No.	Parameter description	Unit	Notation	Scope
1	Unit load	MW	$UL$	[759.32–885.23]
2	Main steam temperature	°C	$ST$	[571.71–605.35]
3	Main steam pressure	MPa	$SP$	[16.77–31.38]
4	Inlet NO <sub>x</sub> concentration	mg/m <sup>3</sup>	$N_{in}$	[25.20–595.36]
5	Net flue gas NO <sub>x</sub> content	mg/m <sup>3</sup>	$N_f$	[11.66–189.81]
6	Inlet CO concentration	mg/m <sup>3</sup>	CO	[4.28–864.53]
7	Inlet oxygen content	%	$O_{in}$	[2.73–21.28]
8	Inlet flue gas pressure	kPa	$P_{in}$	[-1.16–0.55]
9	Inlet flue gas flow rate	m <sup>3</sup> /h	$F_{in}$	[128.33–159.72]
10	Inlet flue gas temperature	°C	$T_{in}$	[355.19–366.97]
11	Outlet flue gas pressure	kPa	$P_{out}$	[-1.61–0.98]
12	Outlet flue gas temperature	°C	$T_{out}$	[350.88–371.05]
13	Outlet oxygen content	%	$O_{out}$	[2.77–12.05]
14–19	Instantaneous coal volume of coal feeder (A-F)	°C	$C_{A-F}$	[38.86–76.96]
20	Dilution fan current	A	D	[0.22–0.71]
21	Total coal volume	t/h	TC	[301.04–378.82]
22	Total air volume	t/h	TA	[2895.9–3433.7]
23	Superheat degree	°C	SH	[34.05–77.29]
24	Ammonia opening degree	m <sup>3</sup> /h	$A_i$	[36.31–49.68]
25–29	Inlet air temperature of coal mill (A-F)	°C	$I_{A-F}$	[29.45–218.08]
30–34	Mill outlet air temperature (A-F)	°C	$U_{A-F}$	[21.92–86.58]
35	Exhaust temperature	°C	$T$	[130.07–140.15]
36	Water to coal ratio	%	R	[6.62–8.94]
37	Ammonia spray flow rate	m <sup>3</sup> /h	Q	[51.18–107.68]
38	Ammonia escape	ppm	$NH_3$	[0.86–3.42]
39	Denitrification efficiency	%	DE	[11.65–100.00]

SCR system and the emission characteristics of NO<sub>x</sub> generated by combustion, we initially selected 39 feature variables related to the denitrification reaction as the candidate inputs for NO<sub>x</sub> concentration prediction. Table 2 lists the candidate input variables.

### 3. The VMD and informer

#### 3.1. Description of the VMD model

Assume that the original time series is  $x(t)$  and the mode  $u_k(t)$  is a finite bandwidth of  $k$  with center frequencies decomposed by the input signal. Each of these sequences has a finite bandwidth of center fre-

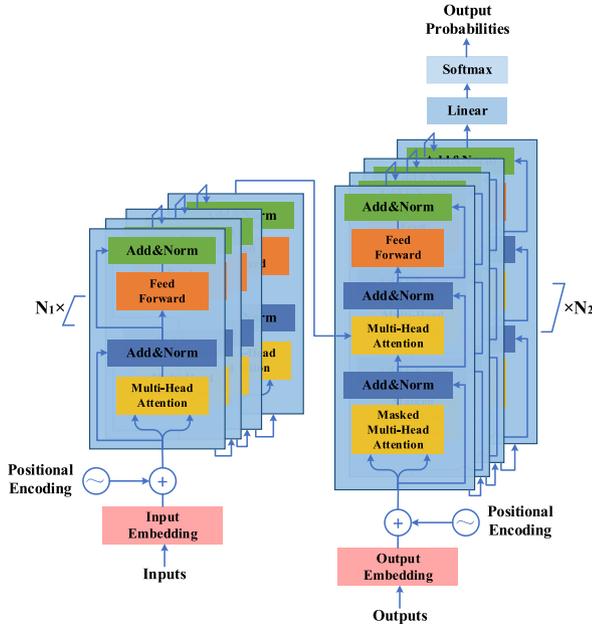


Fig. 3. The structure of the Transformer.

quency, and the sum of the bandwidths is guaranteed to be minimal. The VMD with constraints is calculated as follows:

$$\begin{cases} \min_{\{u_k\}, \{\omega_k\}} \left\{ \sum_k \left\| \partial_t \left[ \left( \delta(t) + \frac{j}{\pi t} \right) \times u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \right\} \\ \text{s.t. } \sum_k u_k = x(t) \end{cases} \quad (3)$$

where  $\{u_k\} = \{u_1, \dots, u_k\}$  is the decomposition of each modal component.  $\{\omega_k\} = \{\omega_1, \dots, \omega_k\}$  is the center frequency of each modal component.  $\delta(t)$  is the impulse function.  $\partial_t$  is the partial derivative of  $t$ .

To find the optimal solution to the constrained variational problem, we introduce penalty factors and Lagrange operators to convert the constrained problem into an unconstrained problem.

$$\begin{aligned} L(\{u_k\}, \{\omega_k\}, \lambda) = & \alpha \sum_k \left\| \partial_t \left[ \left( \delta(t) + \frac{j}{\pi t} \right) \times u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \\ & + \left\| x(t) - \sum_k u_k(t) \right\|_2^2 + \left[ \lambda(t), x(t) - \sum_k u_k(t) \right] \end{aligned} \quad (4)$$

where  $\alpha$  is the penalty factor, and it ensures the reconstruction accuracy of the signal in the presence of gaussian noise.  $\lambda(t)$  is the Lagrangian multiplication operator, and it maintains the stringency of the constraints.

We calculate the optimal solution of the variational modal model by alternately updating  $u_k^{n+1}$ ,  $\omega_k^{n+1}$ , and  $\lambda_k^{n+1}$ . Thus, the time series  $x(t)$  is decomposed into  $k$  narrowband IMF components. The calculation formula is given in Eq. (5)(6) (7).

$$\hat{u}_k^{n+1}(\omega) = \frac{\hat{x}(\omega) - \sum_{i \neq k} \hat{u}_i(\omega) + \frac{\hat{\lambda}(\omega)}{2}}{1 + 2\alpha(\omega - \omega_k)^2} \quad (5)$$

$$w_k^{n+1} = \frac{\int_0^\infty \omega |\hat{u}_k^{n+1}(\omega)|^2 d\omega}{\int_0^\infty |\hat{u}_k^{n+1}(\omega)|^2 d\omega} \quad (6)$$

$$\hat{\lambda}^{n+1}(\omega) = \hat{\lambda}^n(\omega) + \tau \left[ \hat{x}(\omega) - \sum_k \hat{u}_k^{n+1} \right] \quad (7)$$

$$\sum_k \left\| \hat{u}_k^{n+1} - \hat{u}_k^n \right\|_2^2 \Big/ \left\| \hat{u}_k^n \right\|_2^2 < \varepsilon \quad (8)$$

where  $\tau$  is the update factor.  $\hat{u}_k^{n+1}(\omega)$ ,  $\hat{x}(\omega)$ , and  $\hat{\lambda}^{n+1}(\omega)$  represent the Fourier transforms of  $u_k^{n+1}$ ,  $x(t)$ , and  $\lambda_k^{n+1}$ , respectively. The discriminant accuracy is set to  $\varepsilon > 0$ , and the iteration is stopped when Eq. (8) is satisfied.

### 3.2. Description of the informer model

Time series prediction tasks require models that can efficiently and accurately capture the long-term dependent coupling between outputs and inputs. The Transformer model [30], based on a self-attentive mechanism, uses a classical encoder-decoder structure and has been successfully applied to a variety of sequence-to-sequence tasks.

Fig. 3 illustrates the basic structure of the Transformer model. In recent years, the Transformer model has performed superiorly in time series prediction. However, the problems with Transformer's higher memory resource usage and computing time prevent it from being directly applied to time series prediction.

Zhou et al. proposed an efficient Informer model to solve these problems [32]. The Informer greatly reduces the running cost and memory resource usage while ensuring the model's performance remains unchanged. Improvements include: 1) Proposing the ProbSparse self-attention mechanism to reduce the time complexity of memory occupation to  $O(L \log L)$  ( $L$  is the length of time series). 2) Halving the cascading layer inputs by the self-attention distilling operation to highlight the dominance of attention and efficiently process the input time series data. 3) Using a generative decoder to prevent the accumulation of cumulative errors during the inference phase.

### 3.3. Methodology

The weaknesses of single knowledge driven or data driven models in complex NOx concentration prediction are demonstrated. The joint knowledge and data driven model can improve prediction accuracy and generalization ability. For the various needs of power system application scenarios, the joint driven models primarily include: parallel mode, serial mode, bootstrap mode, and feedback mode [15].

In this research, the bootstrap mode is used to establish the NOx concentration prediction model. We take the knowledge driven combined feature selection method as the basis to guide the construction of data-driven section model. Because the nonlinear properties of unit operating data affect prediction accuracy, a combination of MEVMD and Informer is employed to build the data driven section model. The newly proposed sub-models are detailed in the following section.

## 4. The proposed models

### 4.1. MEVMD

Data driven models are deficient in predicting data with nonlinear characteristics [15]. Because feature parameters such as ammonia spray flow are nonlinear and nonstationary, the time-frequency information in the variables is minimal, making prediction more difficult.

We use VMD to decompose the nonlinear time series features into simple and smooth IMF and residual pairs. The decomposed subsequence is relatively stable and has local time-frequency feature information. In addition, we use the permutation entropy to filter the randomness of the residuals generated after decomposition [38].

The VMD requires a manual determination to set the decomposition layers and penalty factors. Due to the engineering characteristics of the thermal process data, the empirical method is usually used to take the values at present. We propose a modal energy and sample entropy based

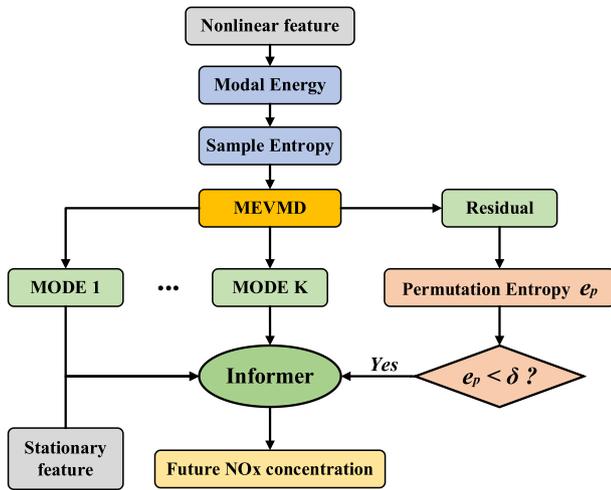


Fig. 4. Overall flow of MEVMD method.

variational modal decomposition (MEVMD). The method can determine the parameters adaptively to improve the predict performance. Fig. 4 shows the overall flow of the MEVMD.

#### 4.1.1. Modal energy difference

We use the modal energy difference to measure the decomposition quality of the eigenmodes and adaptively determine the number of layers of the VMD decomposition. The calculation steps are as follows.

**Step 1.** Define the signal sequence as  $u(t)$ . The total length of the signal sequence is  $N$ . Calculate the signal energy of each modal component:

$$E = \sqrt{\frac{\sum_{t=1}^N u^2(t)}{N}} \quad (9)$$

where  $E$  is the signal energy of each modal component.

**Step 2.** The energy difference between the sum of the energy of each modal component and the energy of the original signal is calculated. The calculation formula is given in Eq. (10).

$$\xi = \frac{\left| \sum_{k=1}^K E_k - E_t \right|}{E_t} \quad (10)$$

where  $E_k$  is the energy of the  $k$  modal component.  $E_t$  is the energy of the original signal.  $\xi$  is the signal energy difference.

**Step 3.** When  $|\xi_{K+1} - \xi_K| > 0.1$  is satisfied, the modal number  $K$  is determined as the optimal number of decomposition layers [39].

#### 4.1.2. Sample entropy

The value of the penalty factor also has a large impact on the VMD decomposition results. Sample entropy is an improved method based on approximate entropy. It proves that the self-similarity of a sequence shows an inverse ratio to its entropy value [40]. We consider that the thermal process data has the property of high self-similarity. Therefore, according to the sample entropy theory, we can choose the penalty factor that minimizes the sum of the sample entropy values of each mode as the optimal value.

**Step 1.** For a signal of sequence length  $u(t)$ , construct the vector  $u_m(i)$  of  $N - m + 1$ . The defining equation is as follows:

$$u_m(i) = [u(i), u(i+1), \dots, u(i+m-1)] \quad (11)$$

$$i = 1, \dots, N - m + 1$$

where  $m$  is the number of embedding dimensions.

**Step 2.** Define the vector spacing as  $d_m$ . Calculate the probability of matching two sequences at point  $m$  and calculate the mean value.

$$d_m[u_m(i), u_m(j)] = \max[u_m(i+k) - u_m(j+k)], 0 \leq k \leq m-1 \quad (12)$$

$$B_{m,i}(r) = \frac{1}{N - m + 1} v_m(i) \quad i = 1, \dots, N - m + 1 \quad (13)$$

$$B_m(r) = \frac{1}{N - m} \sum_{i=1}^{N-m} B_{m,i}(r) \quad (14)$$

Where  $r$  is the tolerance for accepting matrices.  $v_m$  is the number of  $d_m \leq r$ .  $B_{m,i}(r)$  is the probability of matching two points.  $B_m(r)$  is the mean value of the matching probability.

**Step 3.** Calculating the sample entropy.

$$SampleEn(m, r) = -\ln \frac{B_{m+1}(r)}{B_m(r)} \quad (15)$$

Where  $SampleEn(m, r)$  is the sample entropy value.

**Step 4.** The step size is set to 100, and the traversal is performed in the interval [1000, 10,000]. Calculate the sample entropy values for each mode of the time series data. The penalty factor that minimizes the sum of each entropy value is selected as the optimal value.

#### 4.2. Knowledge driven combined feature selection method

Feature selection is an important step before building a prediction model. To avoid the problem that a single feature selection method makes it difficult to comprehensively explore the best set of features for NOx concentration, we propose a knowledge driven combined feature selection method. We use MIC, CART, RF, and XGBoost algorithms to determine the input feature variables in combination with mechanistic knowledge.

**Step 1.** Based on the reaction mechanism analysis of NOx generation in Section 2.2, we initially select 39 feature variables as candidates for NOx concentration prediction.

**Step 2.** We use MIC, CART, RF, and XGBoost algorithms to calculate the correlations between 39 candidate variables and NOx concentrations.

**Step 3.** For each candidate variable, we calculate the mean value by adding the correlation results of the four different algorithms and then dividing by the total number of algorithms. We regard the mean value as the variables' importance. All variables are calculated in turn and ranked.

**Step 4.** Based on the ranking results, variables with importance greater than 2% are selected as model inputs. According to the mechanistic knowledge of NOx generation, direct correlation features such as inlet NOx concentration and ammonia spray flow rate are retained as default input feature variables.

We use the uniform input representation module of Informer to extract time features [32]. The time features consist of three parts, such as a scalar projection, the local time stamp, and the global time stamp. We set the global time stamp to be embedded by minutes, days, weeks, and months. The time features of the input  $x_i^t$  can be expressed as:

$$x_i^t = u_i^t + PE_{(L_x \times (t-1) + i)} + \sum_p [SE_{(L_x \times (t-1) + i)}]_p \quad (16)$$

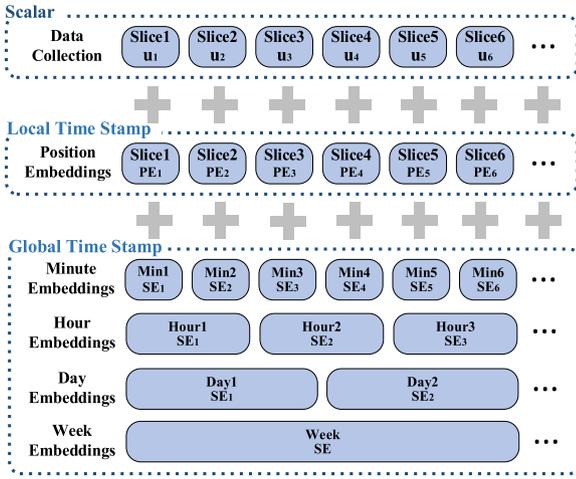


Fig. 5. Time feature extraction.

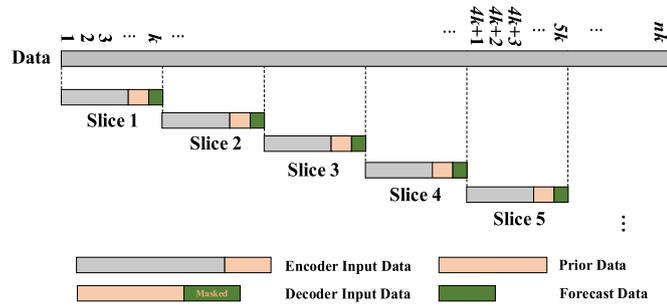


Fig. 6. Time series segmentation process.

where  $p \in \{\text{minute, time, week, month}\}$ .  $u_i$  denotes a scalar projection.  $SE(pos)$  is the global time stamp that can be learned. The local time position embedding can be expressed as:

$$PE_{(pos, 2j)} = \sin\left(\frac{pos}{(2L_x)^{2j/d}}\right) \quad (17)$$

$$PE_{(pos, 2j+1)} = \cos\left(\frac{pos}{(2L_x)^{2j/d}}\right) \quad (18)$$

where  $j \in \{1, \dots, d/2\}$ .  $pos$  denotes location information.  $PE(pos)$  denotes the local time stamp.  $d$  denotes the input feature dimension.  $2j + 1$  and  $2j$  denote the odd and even dimensions, respectively. Fig. 5 presents a depiction of time feature extraction.

#### 4.3. Adaptive timing series segmentation

In Informer, the slicing module of the encoder stack's input layer will slice the historical time series segments. The input time series length is set to  $L$ . Each slice contains  $ClnL$  consecutive measurements [32].  $C$  is a constant sampling factor. The sampling factor is set to 12 in this research.

The ProbSparse self-attentive mechanism extracts sparsity relations in adaptive temporal slices and also reduces the time and space complexity of the method. The output matrix of the ProbSparse self-attentive mechanism is:

$$\mathcal{A}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\overline{\mathbf{Q}\mathbf{K}^\top}}{\sqrt{d}}\right)\mathbf{V} \quad (19)$$

where  $\mathbf{Q}$ ,  $\mathbf{K}$ , and  $\mathbf{V}$  are the query matrix, key matrix, and value matrix, respectively.  $d$  is the input dimension.  $\mathbf{Q} \in \mathbb{R}^{L_Q \times d}$ ,  $\mathbf{K} \in \mathbb{R}^{L_K \times d}$ ,  $\mathbf{V} \in \mathbb{R}^{L_V \times d}$ . The sparse metric matrix is defined as:

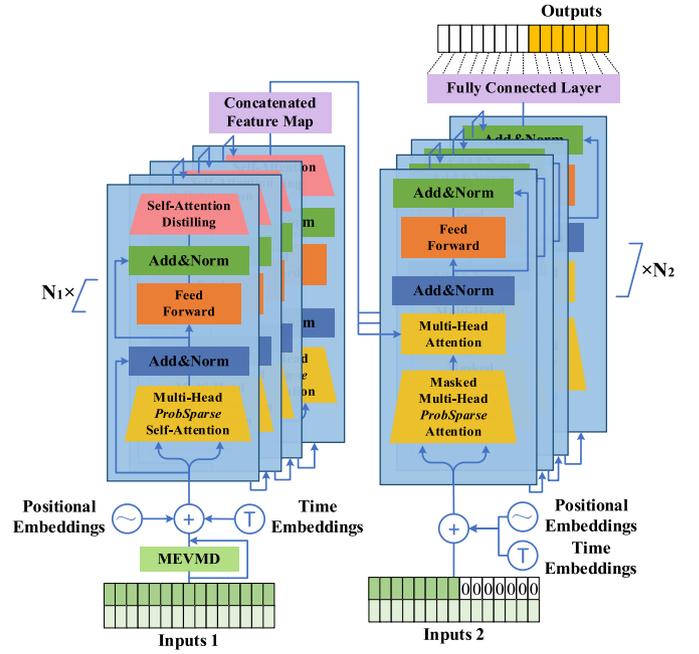


Fig. 7. The structure of the ME-INF.

$$\overline{M}(\mathbf{q}_i, \mathbf{K}) = \max_j \left\{ \frac{\mathbf{q}_i \mathbf{k}_j^\top}{\sqrt{d}} \right\} - \frac{1}{L_K} \sum_{j=1}^{L_K} \frac{\mathbf{q}_i \mathbf{k}_j^\top}{\sqrt{d}} \quad (20)$$

where  $L_K = L_Q = L$ .  $\overline{\mathbf{Q}}$  is the sparse matrix containing only the query data under the sparse metric  $\overline{M}(\mathbf{Q}, \mathbf{K})$ , and zero padding is used for the other dot product pairs.

In this research, based on the data length, we let each historical time series slice include 120 feature data points and set 66 of them as prior knowledge. The slices are used to predict the next 36 data points until the entire sequence is searched. It should be noted that, due to the fundamental properties of time series data, no further filling is performed when the amount of data to be sliced at the end is insufficient. Fig. 6 illustrates the adaptive time series segmentation for predicting NOx concentration.

#### 4.4. Overall prediction model

We propose a joint knowledge and data driven NOx concentration prediction model. For convenience, the proposed model is abbreviated as ME-INF in the following sections. Fig. 7 shows the overall framework of ME-INF. The overall steps are as follows.

- Data preparation. Input the original dataset. Perform preprocessing operations such as removing outliers, filling in exceptions and empty data, and data normalization. Data with a difference between time series greater than twice the mean value of the series is defined as invalid data points and filled with the mean value of the 25 points before and after.
- Feature variable selection. Candidate feature variables are initially chosen based on knowledge of the SCR reaction mechanism. Calculate the relevance of the variables and rank them using a knowledge driven combined feature selection algorithm. Determine the input feature variables for the model.
- Nonlinear sequence decomposition. The MEVMD method is used to decompose the nonlinear and unstable feature series into simple and smooth IMF and its residual pairs. The deep time-frequency information of the variables is extracted through the subsequence. And

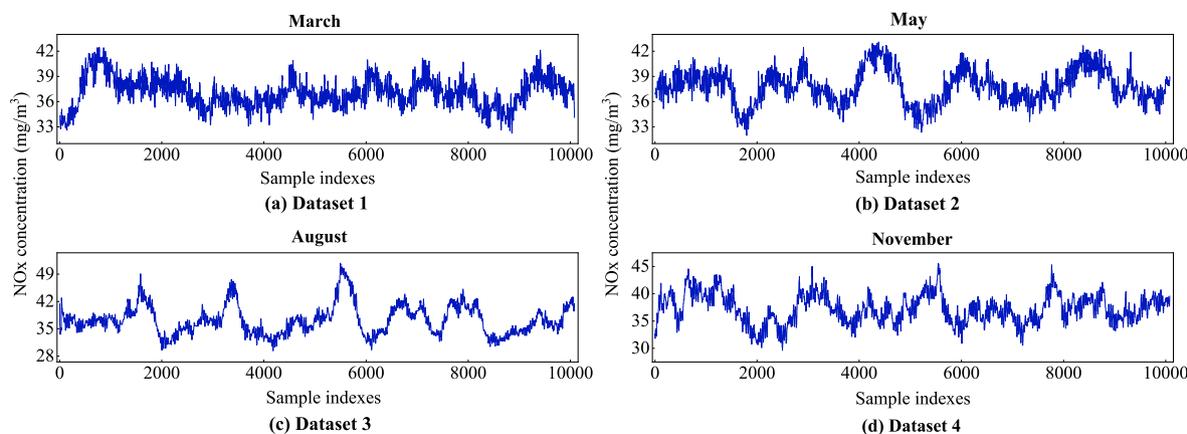


Fig. 8. The images of the experimental datasets.

the randomness of the decomposed residuals is filtered according to the permutation entropy.

- Model training. Input the mechanistic feature variables, time feature variables, and decomposed subsequences into the model. Initialize the model parameters. Set parameters such as learning rate and number of iterations to train the prediction model. Select the optional hyperparameters in the model using the grid search method.
- Model testing. Input test dataset. Output the prediction results of NOx concentration at the outlet of the SCR system by the ME-INF model.

### 5. Forecast results and analysis

#### 5.1. Experimental datasets

In the research, historical operational data of 1000 MW coal-fired generating units at Shanghai Caojing Power Plant are used for NOx concentration prediction comparison experiments. In order to verify the generalization of the prediction model, we selected a dataset consisting of measured data from four consecutive weeks in March, May, August, and November. The four datasets cover four seasons: spring, summer, autumn, and winter, which fully reflect the operating characteristics of the plant in different periods.

The data sampling interval is 1 min, and each dataset contains 10,080 sample points. Each sample includes 13-dimensional input features after filtering. The division ratio of the training set, validation set, and test set is 7:1:2. For convenience, we note the four datasets as: D1, D2, D3, and D4. Fig. 8 provides the images of the four experimental datasets.

It should be noted that the CEMS performs a 3-min blowback every 4 h or so. The control logic maintains the NOx concentration measurements during the blowdown. NOx concentration data during the blowdown period can have significant distortions. During the blowdown period, we mark this part of the data when the blowdown signal is issued and do not use it for the data.

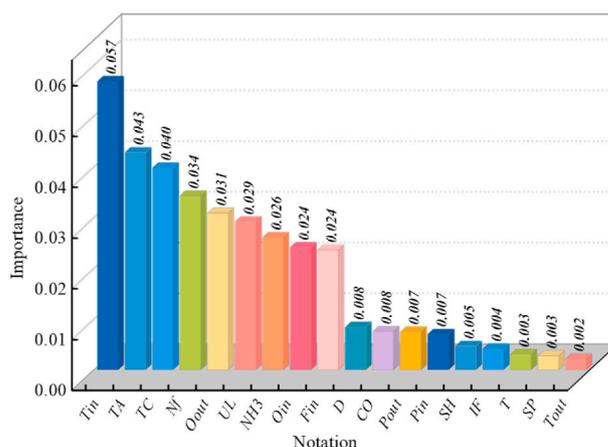


Fig. 9. Importance ranking of candidate feature variables.

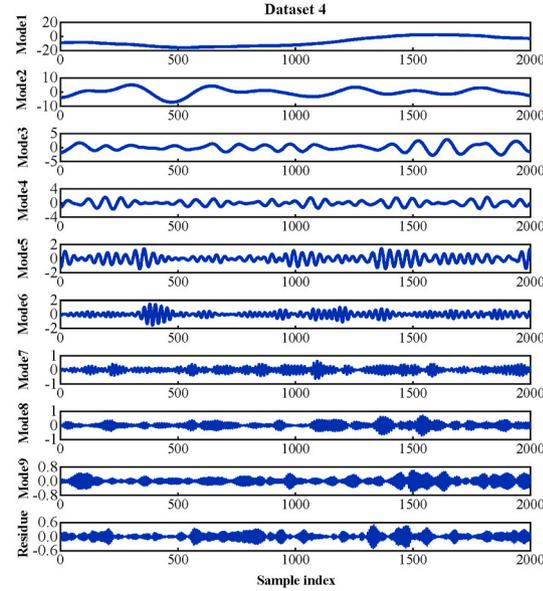
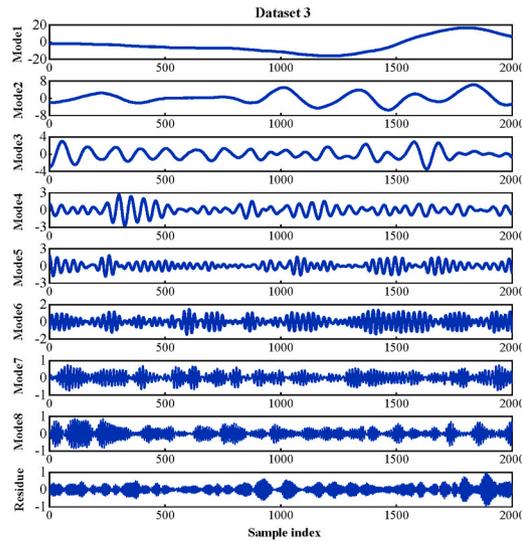
Table 3 Results of correlation analysis of input feature variables.<sup>a</sup>

Notation	MIC	CART	RF	XGBoost
UL	0.0740	0.0014	0.0016	0.0397
N <sub>f</sub>	0.0046	0.0357	0.0376	0.0592
T <sub>in</sub>	0.1026	0.0418	0.0276	0.0553
O <sub>in</sub>	0.0227	0.0074	0.0078	0.0587
O <sub>out</sub>	0.0587	0.0048	0.0041	0.0561
NH <sub>3</sub>	0.0370	0.0025	0.0029	0.0622
TC	0.0647	0.0051	0.0055	0.0840
TA	0.0772	0.0031	0.0033	0.0879
F <sub>in</sub>	0.0601	0.0023	0.0025	0.0298
A <sub>i</sub>	0.1530	0.1166	0.0971	0.0517
DE	0.1553	0.3963	0.3928	0.0866
Q	–	–	–	–
N <sub>in</sub>	–	–	–	–

<sup>a</sup> The results of the correlation analysis are normalized using the L1 parametric.

**Table 4**  
Optimal number of decomposition layers and penalty factors.

Dataset	Decomposition number	Penalty factor
D1	8	2500
D2	9	3100
D3	9	3900
D4	10	4500



**Fig. 10.** Decomposition results for ammonia spray flow rate feature of D3 and D4.

5.2. Assessment indicators

We choose three commonly used assessment indicators: the mean absolute error (MAE), the root mean square error (RMSE) and the mean absolute percentage error (MAPE) to reflect the predictive effect of the model. The formulas are as follows:

$$MAE = \frac{1}{N} \sum_{t=1}^N |x(t) - \tilde{x}(t)| \tag{21}$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{t=1}^N (x(t) - \tilde{x}(t))^2} \tag{22}$$

$$MAPE = \frac{1}{N} \sum_{t=1}^N \left| \frac{x(t) - \tilde{x}(t)}{x(t)} \right| \tag{23}$$

**Table 5**  
The central frequency and permutation entropy of each mode.<sup>a</sup>

Mode	D3		D4	
	$u_k$	$e_p$	$u_k$	$e_p$
Mode1	0.84	1.71	0.52	1.48
Mode2	5.00	1.81	5.22	1.82
Mode3	17.40	2.04	16.14	2.01
Mode4	37.10	2.22	30.06	2.14
Mode5	60.96	2.31	53.46	2.26
Mode6	93.00	2.36	85.42	2.32
Mode7	149.62	2.37	143.56	2.36
Mode8	211.22	2.39	192.52	2.37
Mode9	-	-	239.76	2.36
Residue	276.88	3.12	306.39	2.98

<sup>a</sup> The results of D1 and D2 are given in the supplementary materials (Table 14).

where  $x(t)$  is the real NOx concentration value.  $\tilde{x}(t)$  is the forecasted NOx concentration value, and  $N$  is the total number of data.

5.3. The combined feature selection

Fig. 9 ranks the importance of the candidate variables after the combined feature selection. Table 3 shows the correlation analysis re-

sults of the finalized 13 input features.

From Table 3, it can be seen that each algorithm has different importance ranking results for the candidate variables. For example, the importance of total air volume (TA) under MIC and XGBoost algorithms is significantly higher than other two algorithms. The importance of net flue gas NOx content ( $N_f$ ) under CART and RF algorithms is also higher than the MIC algorithms. The ranking results of the inlet flue gas temperature ( $T_{in}$ ) were inconsistent for all four algorithms.

Therefore, single feature selection methods differ greatly in explaining the importance of candidate features due to differences in principle. The single feature selection method is not sufficient to comprehensively reflect the influence of each candidate feature on the NOx concentration. The combined feature selection approach can provide a more comprehensive analysis of the correlation between the feature variables and predicted data.<sup>12</sup>

5.4. Modal decomposition

We use modal energy difference and sample entropy to adaptively determine the optimal values of decomposition layers and penalty factors for VMD. Table 4 shows the optimal number of decomposition layers and penalty factors for each dataset.

After that, we decompose the nonlinear time series features, such as ammonia spray flow rate, using the MEVMD method. The method can extract the deep local time-frequency information of the features, thus reducing the difficulty of fitting the nonlinear relationships.

<sup>1</sup> For clarity of presentation, we list only 18 of all the feature variables.

<sup>2</sup> To shorten the text, the results of the decomposition of D1 and D2 are shown in the supplemental materials (Fig. 14).

**Table 6**  
Reconstructed signal performance comparison.<sup>a</sup>

Model	D3		D4	
	Coef	RMSE	Coef	RMSE
EMD	0.9873	4.0195	0.6536	6.5386
VMD	0.9903	1.1744	0.9805	1.6477
CEEMD	0.9892	2.0193	0.8209	1.9812
HVMD	0.9936	1.0372	0.9863	1.3219
MEVMD	<b>0.9989</b>	<b>0.4125</b>	<b>0.9988</b>	<b>0.4008</b>

<sup>a</sup> The results of D1 and D2 are given in the supplementary materials (Table 15).

**Table 7**  
Prediction results for each mode of the D3 dataset.<sup>a</sup>

Mode	ARIAM			LSTM			INFORMER		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
Mode1	1.51	1.73	<b>1.19</b>	0.15	0.22	1.37	<b>0.13</b>	<b>0.17</b>	1.70
Mode2	1.49	1.86	8.69	<b>0.12</b>	<b>0.16</b>	<b>2.68</b>	0.25	0.32	4.46
Mode3	1.07	1.26	9.98	<b>0.24</b>	<b>0.29</b>	5.50	0.58	0.74	<b>2.29</b>
Mode4	0.34	0.48	14.13	<b>0.28</b>	<b>0.36</b>	10.88	0.79	1.00	<b>6.94</b>
Mode5	0.92	1.11	21.52	0.37	0.44	12.34	<b>0.32</b>	<b>0.36</b>	<b>11.36</b>
Mode6	1.69	2.03	34.75	0.51	0.57	14.73	<b>0.45</b>	<b>0.52</b>	<b>12.62</b>
Mode7	1.71	2.12	39.57	<b>0.55</b>	<b>0.59</b>	24.07	0.60	0.76	<b>21.33</b>
Mode8	1.83	2.27	40.41	0.88	0.98	33.37	<b>0.78</b>	<b>0.97</b>	<b>31.20</b>
Residue	2.30	2.97	64.12	1.03	1.24	50.71	<b>0.94</b>	<b>1.21</b>	<b>41.21</b>

<sup>a</sup> The results of D1, D2, and D4 are presented in the supplementary materials (Tables 11–13).

**Table 8**  
Comparison of NOx concentration prediction results.<sup>a,b</sup>

Model	D1			D2			D3			D4		
	MAE	RMSE	MAPE									
GRU	2.13	2.65	5.98	2.33	3.11	5.93	3.20	3.88	9.21	3.50	4.32	9.14
BP	3.08	3.57	8.46	2.56	3.38	6.75	3.43	4.58	9.26	5.81	6.48	15.30
AE-ELM	1.42	1.77	3.94	1.61	1.92	4.13	2.57	3.44	7.10	2.97	3.49	7.75
CA-LSTM	1.23	1.72	3.37	1.84	2.91	4.68	2.21	4.03	5.99	3.45	4.20	8.98
Informer <sup>†</sup>	1.24	1.58	3.46	1.64	2.01	4.19	2.76	3.69	7.40	3.15	5.17	8.18
Informer	1.15	1.46	3.21	1.47	1.80	3.93	1.83	2.27	4.98	1.60	1.96	4.18
ME-INF <sup>†</sup>	1.11	1.40	3.08	1.50	1.81	3.96	2.38	2.99	6.44	2.23	2.63	6.04
ME-INF	<b>0.94</b>	<b>1.17</b>	<b>2.60</b>	<b>1.08</b>	<b>1.33</b>	<b>2.83</b>	<b>1.07</b>	<b>1.41</b>	<b>2.92</b>	<b>1.27</b>	<b>1.55</b>	<b>3.39</b>

<sup>a</sup> † indicates that the model has not undergone knowledge driven combined feature selection.

<sup>b</sup> The experimental results are taken as the mean of the best 5 out of 20 experiments.

The decomposition results of ammonia spray flow rates for D3 and D4 are taken as examples<sup>3</sup> (Fig. 10). The central frequency  $u_k$  and the permutation entropy  $e_p$  of each mode are shown in Table 5. It can be concluded that the central frequency of each mode is basically proportional to the permutation entropy. At the same time, the residual pairs have relatively large values of the permutation entropy. Therefore, the permutation entropy [38] is used as the randomness measure. Set the threshold  $h$  to filter the invalid residuals. In this paper, the threshold value  $h$  is set to 3.0.

To verify the effectiveness of the method improvement, we used VMD [26], EMD [41], CEEMD [13], and HVMD [42] for the reconstruction comparison experiments after modal decomposition.

The correlation coefficient  $Coef$  is usually used to evaluate the match between the reconstructed and original signal. The more  $|Coef|$  approaches to 1, the stronger the correlation. The calculation formula is:

$$Coef = \frac{\sum_{t=1}^N x(t)y(t) - \left[ \sum_{t=1}^N x(t) \sum_{t=1}^N y(t) \right] / N}{\sqrt{\left( \sum_{t=1}^N x^2(t) - \left[ \left( \sum_{t=1}^N x(t) \right)^2 / N \right] \right) \left( \sum_{t=1}^N y^2(t) - \left[ \left( \sum_{t=1}^N y(t) \right)^2 / N \right] \right)}} \quad (24)$$

where  $x(t)$  is the reconstructed signal.  $y(t)$  is the original signal.  $N$  is the signal length.

From the results in Table 6, the reconstructed signal processed by the

MEVMD method outperforms other comparison methods in terms of correlation coefficient and RMSE of the original signal. The effectiveness of the MEVMD improvement is proved.

Meanwhile, we compare the prediction results of INFORMER with ARIMA and LSTM for each mode. The results are summarized in Table 7. It can be seen that the prediction accuracy of LSTM is better than ARIMA for low-frequency and high-frequency modes overall. INFORMER shows comparable performance to LSTM on low-frequency modes. And the

**Table 9**  
Model parameter setting.

Parameter	Numerical setting
Adaptive sequence length	120
Number of encoder layers	2
Number of decoder layers	2
Muti-head number	8
Prob attention factor	5
Model dimension	64
FCN dimension	128
Train epochs	12
Batch size	32
Initial learning rate	0.0001
Dropout	0.05
Loss function	MSE

<sup>3</sup> † indicates that the model has not undergone knowledge driven combination feature selection.

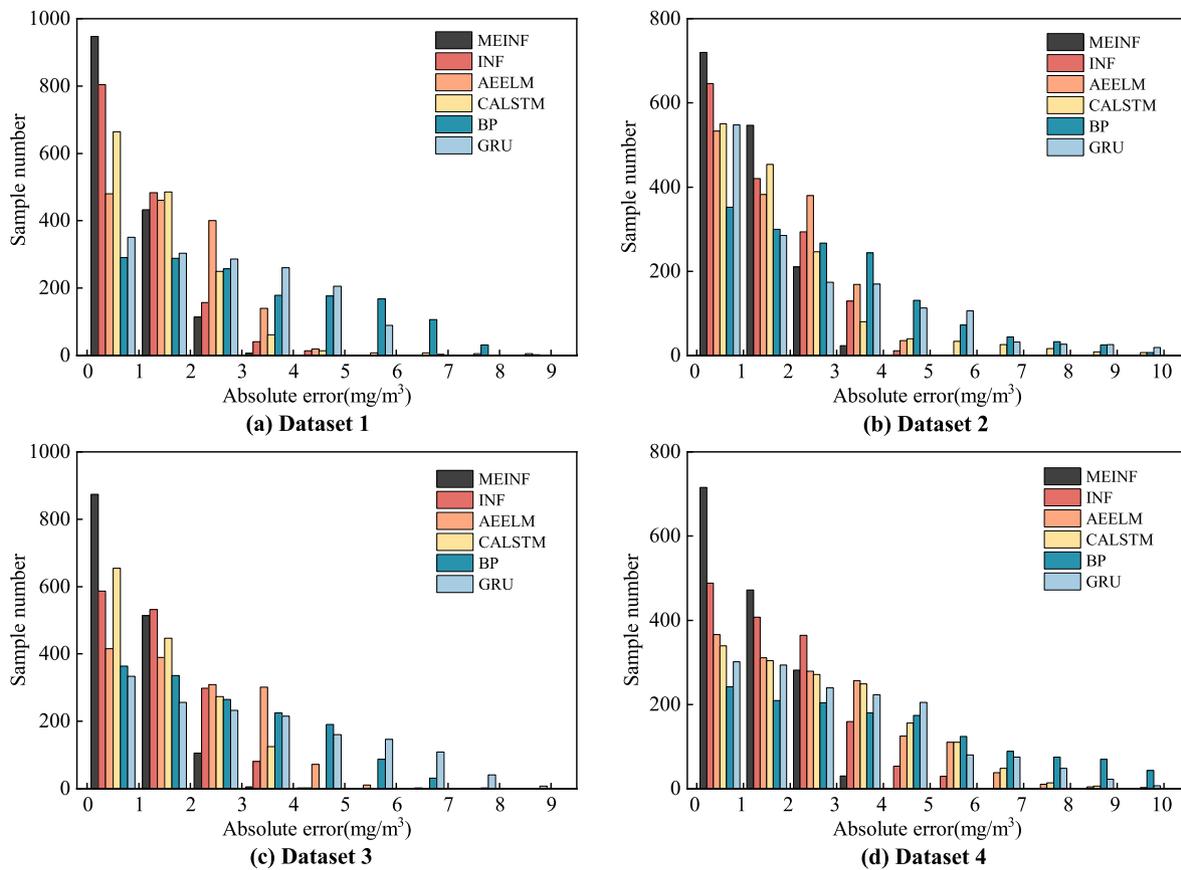


Fig. 11. Absolute error distribution of different models.

**Table 10**  
The quantitative improvement effect of model prediction performance.

Model	D1			D2			D3			D4		
	MAE	RMSE	MAPE									
<b>Single Model</b>												
GRU	-55.87	-55.85	-56.52	-53.65	-57.23	-52.28	-66.56	-63.66	-68.30	-63.71	-64.12	-62.91
BP	-69.48	-67.23	-69.27	-57.81	-60.65	-58.07	-68.80	-69.21	-68.47	-78.14	-76.08	-77.84
Informer <sup>†</sup>	-24.19	-25.95	-24.86	-34.15	-33.83	-32.46	-61.23	-61.79	-60.54	-59.68	-70.02	-58.56
Informer	-18.26	-19.86	-19.00	-26.53	-26.11	-27.99	-41.53	-37.89	-41.37	-20.63	-20.92	-18.90
<b>Hybrid Model</b>												
AE-ELM	-33.80	-33.90	-34.01	-32.92	-30.73	-31.48	-58.37	-59.01	-58.87	-57.24	-55.59	-56.26
CA-LSTM	-23.58	-31.98	-22.85	-41.30	-54.30	-39.53	-51.58	-65.01	-51.25	-63.19	-63.10	-62.25
ME-INF <sup>†</sup>	-15.32	-16.43	-15.58	-28.00	-26.52	-28.54	-55.04	-52.84	-54.66	-43.05	-41.06	-43.87

prediction performance is better in high frequency and residual modes. Therefore, INFORMER is expected to show its excellent performance in the prediction of combinatorial models based on signal decomposition.

### 5.5. Prediction comparison experiment

We conduct experiments comparing ME-INF with a variety of methods. The comparison methods include two single models, GRU [12] and BP, and two recently proposed hybrid models, CA-LSTM<sup>4</sup> [13] and AE-ELM [14]. D1, D2, D3 and D4 are chosen as the experimental datasets.

The parameter settings in the ME-INF model are given in Table 9. Feature parameters are input by the results of mechanistic knowledge and combined feature selection. Time features are extracted by minutes, days, weeks, and months. The parameter settings of the comparison

methods were kept the same in the original research. Both training and testing procedures are implemented on python 3.9 platform.

Table 8 summarizes the prediction results of each model. From the data, it can be obtained that: 1) The hybrid model outperforms the single model in terms of prediction performance. This suggests that the single model is difficult to handle thermal process data with strong coupling, nonlinearity, and perturbation. 2) The prediction performance of the ME-INF is better compared to the Informer. This is because the MEVMD method is used to deal with the feature variables with nonlinearity and non-smoothness, which can enhance the model's performance in resolving time-frequency information. 3) The prediction performance of the data driven models without incorporating knowledge driven combined feature selection (Informer<sup>†</sup> and ME-INF<sup>†</sup>) varies widely across datasets. However, the prediction performance of the ME-INF model is generally consistent across the four datasets. It also outperforms the other comparison models in terms of prediction accuracy. The reason is the introduction of mechanistic knowledge provides an important global feature basis for data driven modeling. It compensates for the drawback

<sup>4</sup> We abbreviate CEEMDAN-AM-LSTM as CA-LSTM in Ref. [13].

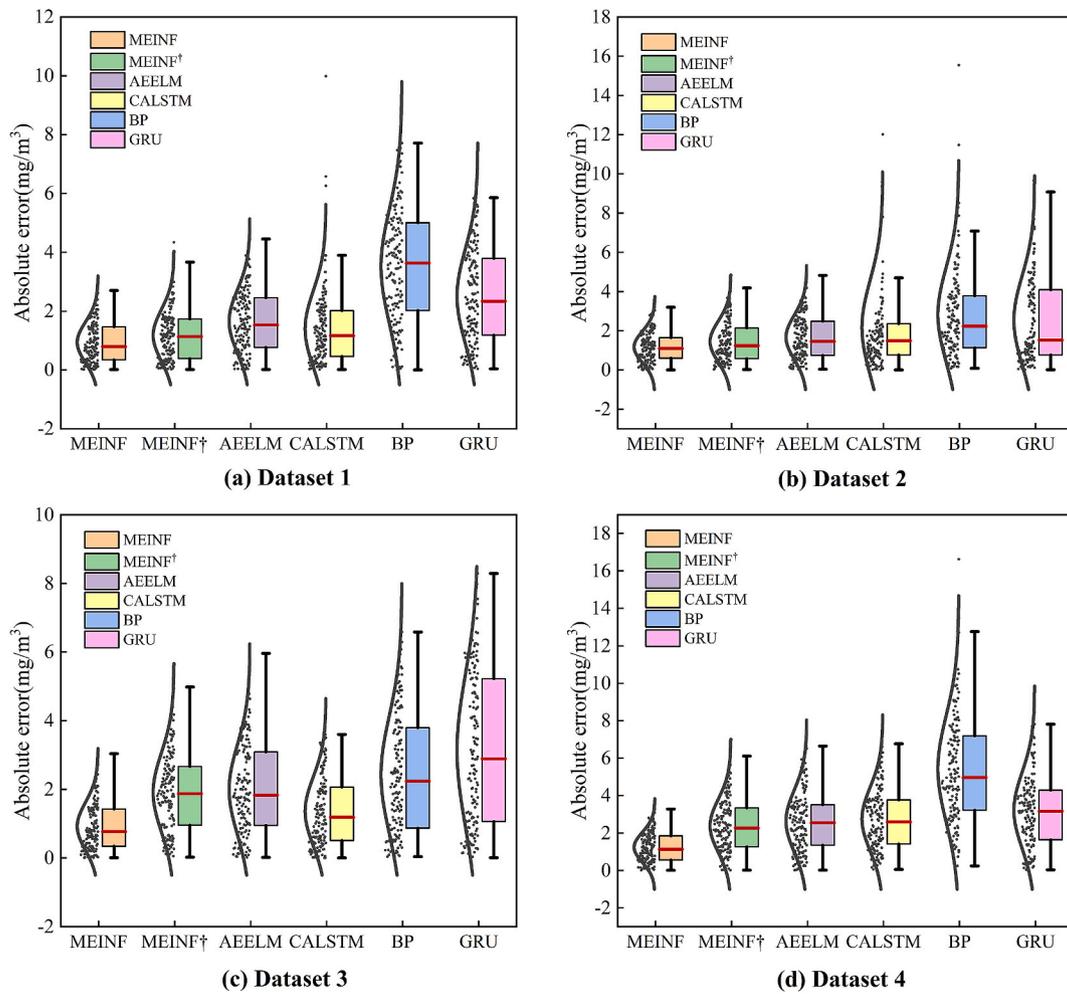


Fig. 12. The box line plot of model prediction errors.

that the data driven model cannot comprehensively characterize the research object. The joint knowledge and data driven model greatly improves accuracy and generalization ability of NOx concentration prediction.

Fig. 11 presents the absolute error distributions of the different models. It can be seen that the absolute error distribution of the ME-INF model prediction results is in the smallest interval [0,4] compared to the comparison method. Meanwhile, the vast majority of the error values are in the interval [0,1], and there is no distribution in the high error interval. This demonstrates the higher prediction accuracy of the ME-INF model than other comparison methods.

To more clearly illustrate the prediction performance improvement effect of the proposed model, we introduced a quantitative indicator proposed in Ref. [43].

$$I_m = \frac{EI_p - EI_o}{EI_o} \quad (25)$$

where  $I_m$  denotes the effect of model prediction performance improvement.  $EI_p$  is the value taken for the assessment indicator of ME-INF.  $EI_o$  is the value taken by the comparison model under the same assessment indicator.

Table 10 provides the prediction performance improvement effect of the ME-INF model. It can be found that each assessment indicator of the ME-INF is greatly reduced. Taking the D3 dataset as an example, compared with GRU, BP, AE-ELM, CA-LSTM and Informer, RMSE is reduced by 63.66%, 69.21%, 59.01%, 65.01% and 61.79%, and MAPE is reduced by 68.30%, 68.47%, 58.87%, 51.25% and 60.54%, respectively.

Therefore, the proposed model improves the prediction accuracy of NOx concentration significantly.

#### 5.6. The effectiveness of the knowledge driven combined feature selection method

Fig. 12 presents the box line plots of the absolute errors between predicted and true values for ME-INF, ME-INF<sup>†</sup>, and other comparison models.<sup>5</sup> The red line indicates the median error of the models. The curves show the distribution of the errors. The greater the concentration of the absolute error distribution, the smaller the median error, indicating greater prediction accuracy. According to Fig. 12, when the knowledge driven combined feature selection method is not embedded, the method has a wide range of error variation. The ME-INF model reduces the median error, makes the error distribution more concentrated, and improves prediction accuracy. Meanwhile, ME-INF performs essentially the same across four datasets with minimal error compared with other models. The effectiveness of the proposed knowledge driven combined feature selection method is demonstrated.

#### 5.7. The effectiveness of the MEVMD method

The prediction results of the model with nonlinear decomposition

<sup>5</sup> † indicates that the method has not undergone knowledge driven combined feature selection.

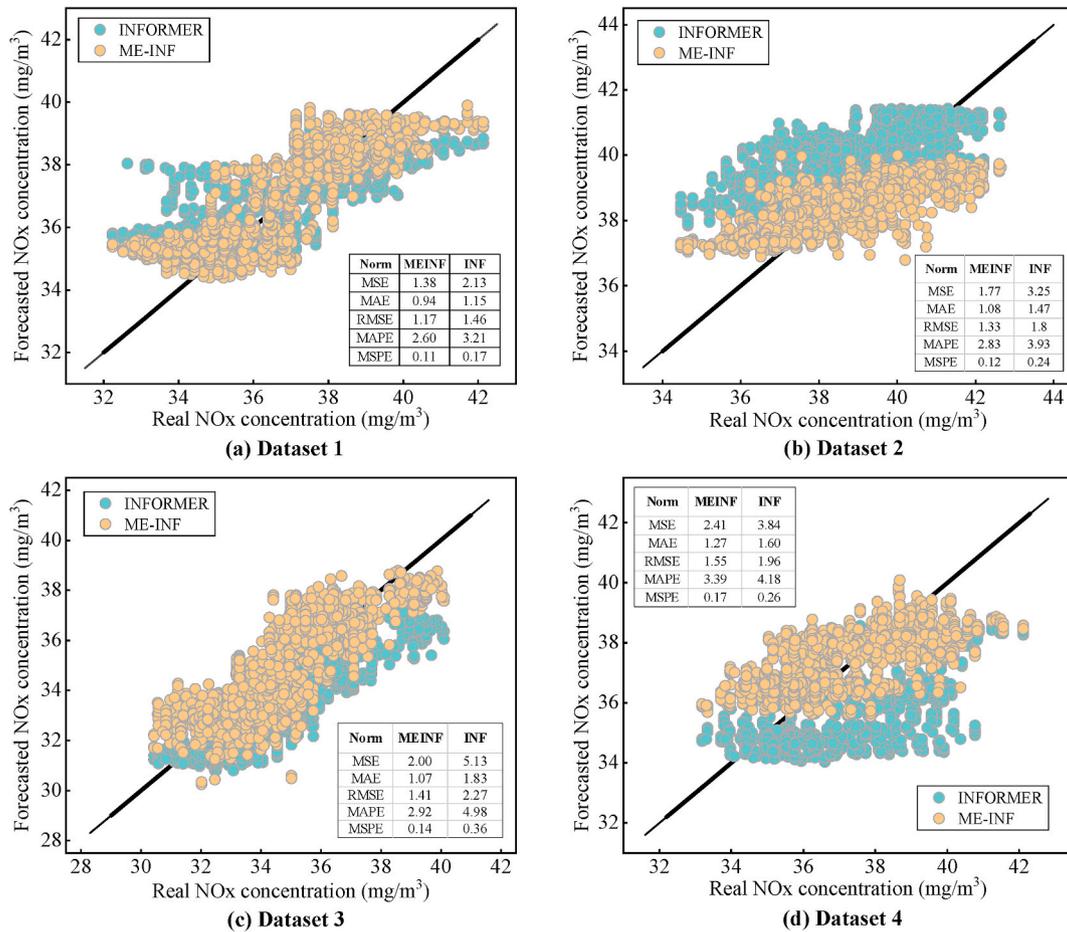


Fig. 13. Comparison of NOx concentration prediction results.

using MEVMD (ME-INF) and without decomposition (INF) are compared in Fig. 13. The black line segment represents the ideal straight line where the predicted and true values are equal. The higher the prediction accuracy, the closer the circular blob is to the line distribution. From Fig. 13, there is a significant difference between the forecasted and real values of the model without nonlinear feature decomposition. The prediction accuracy has improved since the introduction of the MEVMD method. This is because the MEVMD method eliminates the non-smoothness of signals such as ammonia spray flow rate and enhances the model's ability to extract information from nonlinear features. It is advantageous to fit the NOx emission concentration and the feature variables' deep time-frequency connection.<sup>6</sup>

## 6. Discussion

Through a series of experimental comparisons and analyses, we can obtain the following conclusions.

- The ME-INF solves the problem of data driven model being unable to describe the global characteristics of the research object, as well as improving prediction accuracy and generalization performance. Compared with GRU, BP, Informer, AE-ELM, and CA-LSTM, the prediction performance of ME-INF is basically consistent across the four datasets (Table 8). The prediction results' absolute errors were distributed within the minimum interval [0,4] (Fig. 11). The

prediction accuracy under several assessment indicators was higher than that of other comparison methods (Table 10).

- Different feature selection methods differ significantly in their principle of resolving variable correlation. In comparison to the single method, the knowledge driven combined feature selection method completely resolves the correlation between candidate variables and predicted data (Table 3, Fig. 9). The introduction of this method reduces the median error of ME-INF and concentrates the error distribution. The generalization performance of ME-INF outperforms many comparable models (Fig. 12).
- The modal energy difference and sample entropy can determine the optimal number of decomposition layers and penalty factor of VMD. Compared with VMD, EMD, CEEMD, and HVMD, the reconstructed signal following MEVMD has a stronger correlation with the original signal (Table 6, 15). MEVMD effectively eliminates the effects of nonlinear features such as ammonia injection flow. The introduction of this method improves the prediction accuracy of ME-INF (Fig. 13). The permutation entropy filters the random residuals and avoids the negative impact of noise (Table 5, 14).
- Informer performs well in hybrid prediction models involving signal decomposition and deep learning. In high-frequency and residual modes, informer outperforms ARIMA and LSTM, and in low-frequency modes, it outperforms ARIMA and is comparable to LSTM (Table 7, 12-14).

The limitations of ME-INF are as follows.

- ME-INF does not take into account feature factors such as different operating conditions and coal quality variations. These factors may

<sup>6</sup> We resample the data points in units of 10 to make a clearer presentation of the sample points in Fig. 12.

be included in future research to check whether the prediction will be improved.

- ME-INF is unable to achieve satisfactory NOx emission concentration prediction results on distorted data of CEMS blowdown events. We may try to develop other algorithms for predicting distorted data in future research.

## 7. Conclusions

Establishing an accurate and stable NOx emission concentration prediction model is the foundation for realizing denitrification and environmental protection in coal-fired power plants. A hybrid prediction model based on joint knowledge and data driven is proposed for the NOx emission concentration prediction. First, the knowledge driven combined feature selection is utilized to provide a global feature base for data driven modeling. Second, the VMD is improved by modal energy difference and sample entropy to enhance the model's capacity to extract nonlinear information. Finally, informer and adaptive segmentation are employed to forecast NOx emission concentrations. In comparison to other methods, the experimental results demonstrate the model's strong generalization and prediction accuracy.

The concept of joint knowledge and data driven modeling offers a novel solution for complex time-series prediction tasks. However, its application to NOx concentration prediction is still in the exploratory stage. The goal of future research is to expand the model's dynamic prediction capabilities by including additional feature parameters, such as operating conditions, coal quality changes, etc.

## Credit author statement

**Zheng Wu:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision. **Yue Zhang:** Investigation, Validation. **Ze Dong:** Resources, Conceptualization, Data curation.

## Declaration of competing interest

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.

## Data availability

The data that has been used is confidential.

## Acknowledgements

This work was supported by the S&T Program of Hebei (No. 22567643H) and the National Natural Science Foundation of China (No. 71471060) and the Fundamental Research Funds for Central Universities (No. 2018QN096) and the Hebei Provincial Natural Science Foundation (No. E2018502111).

## Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.energy.2023.127044>.

## References

- [1] Jin Y, Scherer L, Sutanudjaja E, et al. Climate change and CCS increase the water vulnerability of China's thermoelectric power fleet. *Energy* 2022;245:123339.
- [2] Jiang Y, Lee B, Oh D, et al. Influence of various air-staging on combustion and NOx emission characteristics in a tangentially fired boiler under the 50% load condition. *Energy* 2022;244:123167.
- [3] Du X, Yang G, Chen Y, et al. The different poisoning behaviors of various alkali metal containing compounds on SCR catalyst. *Appl Surf Sci* 2017;392:162–8.
- [4] Gu X, Li B, Sun C, et al. An improved hourly-resolved NOx emission inventory for power plants based on continuous emission monitoring system (CEMS) database: a case in Jiangsu, China. *J Clean Prod* 2022;369:133176.
- [5] Yang T, Ma K, Lv Y, et al. Real-time dynamic prediction model of NOx emission of coal-fired boilers under variable load conditions. *Fuel* 2020;274:117811.
- [6] Huang D, Tang S, Zhou D, Hao J. NOx emission estimation in gas turbines via interpretable neural network observer with adjustable intermediate layer considering ambient and boundary conditions. *Measurement* 2022;189:110429.
- [7] Zhao H, Shen J, Li Y, Bentsman J. Coal-fired utility boiler modelling for advanced economical low-NOx combustion controller design. *Control Eng Pract* 2017;58:127–41.
- [8] Chang J, Wang X, Zhou Z, Chen H, Niu Y. CFD modeling of hydrodynamics combustion and NOx emission in a tangentially fired pulverized-coal boiler at low load operating conditions. *Adv Powder Technol* 2021;32:290–303.
- [9] Devarakonda M, Tonkyn R, Tran D, et al. Modeling species inhibition of NO oxidation in urea-SCR catalysts for diesel engine NOx control. *J Eng Gas Turbines Power* 2011;133:491–8.
- [10] Liu H, Wang Y, Li X, Yang G. NOx emission prediction of coal-fired power plants based on mutual information-graph convolutional neural network. *Chin. J. Electr. Eng.* 2022;42:1052–60.
- [11] Yuan Z, Meng L, Gu X, et al. Prediction of NOx emissions for coal-fired power plants with stacked-generalization ensemble method. *Fuel* 2021;289:119748.
- [12] Tuttle JF, Blackburn LD, Andersson K, Powell KM. A systematic comparison of machine learning methods for modeling of dynamic processes applied to combustion emission rate modeling. *Appl Energy* 2021;292:116886.
- [13] Wang X, Liu W, Wang Y, Yang G. A hybrid NOx emission prediction model based on CEEMDAN and AM-LSTM. *Fuel* 2022;310:122486.
- [14] Tang Z, Wang S, Chai X, et al. Auto-encoder-extreme learning machine model for boiler NOx emission concentration prediction. *Energy* 2022;256:124552.
- [15] Li F, Wang Q, Hu J, Tang Y. Research progress of joint data and knowledge driven method and its application prospect in power system. *Chin. J. Electr. Eng.* 2021;41:4377–90.
- [16] Yin T, Lu N, Guo G, et al. Knowledge and data dual-driven transfer network for industrial robot fault diagnosis. *Mech Syst Signal Process* 2023;182:109597.
- [17] Yin J, Ren X, Liu R, et al. Quantitative analysis for resilience-based urban rail systems: a hybrid knowledge-based and data-driven approach. *Reliab Eng Syst Saf* 2022;219:108183.
- [18] Pu C, Yi J, Liu Z, et al. A review of research on collaborative knowledge and data driven group intelligence decision making methods. *J Automation* 2022;48:627–43.
- [19] Liu J, Li G, Liu B, et al. Knowledge discovery of data-driven-based fault diagnostics for building energy systems: a case study of the building variable refrigerant flow system. *Energy* 2019;174:873–85.
- [20] Xu X, Liu W, Yu L. Trajectory prediction for heterogeneous traffic-agents using knowledge correction data-driven model. *Inf Sci* 2022;608:375–91.
- [21] Ahmed F, Kim K. Recursive approach to combine expert knowledge and data-driven RSW weldability certification decision making process. *Robot Comput Integrated Manuf* 2023;79:102428.
- [22] Li F, Wang Q, Tang Y, et al. An integrated method for critical clearing time prediction based on a model-driven and ensemble cost-sensitive data-driven scheme. *Int J Electr Power Energy Syst* 2021;125:106513.
- [23] Yi J, Lin W, Hu J, et al. An integrated model-driven and data-driven method for on-line prediction of transient stability of power system with wind power generation. *IEEE Access* 2020;8:2991534.
- [24] Wang Q, Li F, Tang Y, Xue Y. Online prediction method of grid transient frequency characteristics based on physical-data fusion model. *Power Syst Automation* 2018;42:1–9.
- [25] Bokde N, Feijóo A, Villanueva D, et al. A review on hybrid empirical mode decomposition models for wind speed and wind power prediction[J]. *Energies* 2019;12:254.
- [26] Dragomiretskiy K, Zosso D. Variational mode decomposition. *IEEE Trans Signal Process* 2014;62:531–44.
- [27] Yao X, Yi T, Qu C. Autoregressive spectrum-guided variational mode decomposition for time-varying modal identification under nonstationary conditions. *Eng Struct* 2022;251:113543.
- [28] Sun Z, Zhao M, Zhao G. Hybrid model based on VMD decomposition, clustering analysis, long short memory network, ensemble learning and error complementation for short-term wind speed forecasting assisted by Flink platform. *Energy* 2022;261:125248.
- [29] Xiong B, Meng X, Xiong G, et al. Multi-branch wind power prediction based on optimized variational mode decomposition. *Energy Rep* 2022;8:11181–91.
- [30] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: *Proceedings of the 31st International Conference on neural information processing systems*; 2017. p. 6000–10.
- [31] Kitaev N, Kaiser Ł, Levskaya A. Reformer: the efficient transformer. 2020, 04451. arXiv:2001.
- [32] Zhou H, Zhang S, Peng J, et al. Informer: beyond efficient transformer for long sequence time-series forecasting. 2020, 07436. arXiv:2012.
- [33] Gong M, Zhao Y, Sun J, et al. Load forecasting of district heating system based on Informer. *Energy* 2022;253:124179.
- [34] Koebel M, Strutz E. Thermal and hydrolytic decomposition of urea for automotive selective catalytic reduction systems: thermochemical and practical aspects. *Ind Eng Chem Res* 2003;42:2093–100.
- [35] Kawai T. Activities of V<sub>2</sub>O<sub>5</sub>/TiO<sub>2</sub> and V<sub>2</sub>O<sub>5</sub>/Al<sub>2</sub>O<sub>3</sub> catalysts for the reduction of NO and NH<sub>3</sub> in the presence of O<sub>2</sub>. *Ind Eng Chem Prod Res Dev* 1989;21:424–8.

- [36] Beeckman J, Hegedus L. Design of monolith catalysts for power plant nitrogen oxide emission control. *Ind Eng Chem Res* 1991;30:969–78.
- [37] Zhou H, Zhang Z, Zhang J, et al. Hybrid structure-radial basis function neural network optimal control of flue gas denitrification ammonia injection in supercritical boilers. *Chin J Electr Eng* 2011;31:108–13.
- [38] Cao H, Tung W, Gao J, et al. Detecting dynamical changes in time series using the permutation entropy. *Phys Rev* 2004;70:046217.
- [39] Lian J, Liu Z, Wang H, Dong X. Adaptive variational mode decomposition method for signal processing based on mode characteristic. *Mech Syst Signal Process* 2018; 107:53–77.
- [40] Widodo A, Shim M, Caesarendra W, Yang B. Intelligent prognostics for battery health monitoring based on sample entropy. *Expert Syst Appl* 2011;38:11763–9.
- [41] Huang N, Shen Z, Long S, et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London Series A*; 1998. p. 454.
- [42] Gai J, Shen J, Hu Y, Wang H. An integrated method based on hybrid grey wolf optimizer improved variational mode decomposition and deep neural network for fault diagnosis of rolling bearing. *Measurement* 2020;162:107901.
- [43] Zhang J, Wei Y, Tan Z. An adaptive hybrid model for short term wind speed forecasting. *Energy* 2020;190:115615.