Dynamic Pollution Emission Prediction Method of a Combined Heat and Power System Based on the Hybrid CNN-LSTM Model and Attention Mechanism

Anping Wan

ZUCC: Zhejiang University City College

Jie Yang

zhejiang university city college

Ting Chen (✉ chenting@zucc.edu.cn)

Zhejiang university City College

Yang Jinxing

Bansgan Power plant

Ke Li

Zhejiang university

Zhou Qinglong

Zhejiang University

Research Article

Keywords: Pollution emission prediction, CHP system, CNN-LSTM structure, attention mechanism

Posted Date: November 30th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-988024/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Dynamic pollution emission prediction method of a combined heat and power system based on the hybrid CNN-LSTM model and attention mechanism

Anping Wan¹, Jie Yang^{1,2}, Ting Chen^{1*}, Yang Jinxing³, Ke Li², Zhou Qinglong²

¹ Department of Mechatronics Engineering, Zhejiang University City College, Hangzhou, 310015, China

² School of Mechanical Engineering, Zhejiang University, Hangzhou, 310015, China

³ Banshan Power Plant, Hangzhou, 310015, China

*Corresponding author. E-mail: chenting@zucc.edu.cn

Abstract: The prediction of pollution emission from a combined heat and power (CHP) system is very important for the production regulation and emergency response of a power system. The composition and structure of the CHP equipment are complex, and the production process is cumbersome. The fuel chemical reaction of the pulverized coal in the boiler represents a highly nonlinear and strongly interrelated process that is strongly affected by external environmental factors, which causes a certain level of volatility and uncertainty. In this study, a pollution emission prediction method of CHP systems based on feature engineering and a hybrid deep learning model is proposed. Feature engineering performs multi-step preprocessing on the original data, refines the correlation factors, and removes redundant variables. The hybrid deep learning model has a multi-variable input and is established by combining the convolutional neural network-long short-term memory network with the attention mechanism. The case study is conducted on the collected actual dataset. The influence of the prediction target periodicity on the prediction results is analyzed seasonally to verify the effectiveness of the hybrid model. The results show that the root mean square error of the proposed method is less than one, and the error is reduced compared

to the other basic methods, which proves the superiority of the proposed pollution emission prediction method over the existing methods.

Keywords: Pollution emission prediction, CHP system, CNN-LSTM structure, attention mechanism

Nomenclature

L	constituent element [-]
lag	the number of continuous lag prediction targets [-]
T	total number of sliding windows [-]
n	dimension of input features [-]
f	forgetting gate [-]
i	input gate [-]
o	output gate [-]
C	storage unit [-]
H	external state [-]
W	weight [-]
b	bias [-]

Subscripts

t	current time [h]
h	prediction horizon [h]

1. Introduction

Among all environmental pollution sources in China, power industry pollution is one of the most

30 dominant pollution sources (Pan et al. 2021; Xiao et al. 2021). Nitrogen oxides (NO_x) and sulfur
31 dioxide (SO₂) are the most conventional air pollutants. Excessive NO_x and SO₂ can cause
32 photochemical smog and acid rain and have a serious impact on the ozone content in the
33 atmosphere. In nature, NO_x is mainly produced under thunderstorm weather conditions; however,
34 human production activities, such as automobile exhaust emission, power generation in thermal
35 power plants, and fuel combustion in other industrial production, highly increase the content of
36 NO_x and SO₂ in the air. According to the preliminary estimation of the United Nations
37 Environmental Planning (UNEP) deployment project “Comprehensive Consideration of
38 Environmental Factors in Energy Planning,” 70% of atmospheric emissions of NO_x in China
39 originates from the direct combustion of coal (Zhang et al. 2016), which has an immeasurable
40 impact on nature. Therefore, it is particularly important to control the emission of NO_x and SO₂.

41 In recent years, numerous effective measures have been taken in many fields to control
42 anthropogenic pollutant emissions (Meng et al. 2021). For instance, in the field of thermal power
43 generation, the application of CHP production mode (Hyeunguk et al. 2021) can improve resource
44 utilization and reduce environmental impact under the same energy consumption. In the Emission
45 Standard of Air Pollutants for Thermal Power Plants, which was promulgated and implemented in
46 China in 2012, it has been strictly stipulated that the NO_x and SO₂ emissions of the new circulating
47 fluidized bed coal-fired boilers are 100 mg/m³, and there have been even more stringent
48 requirements for certain regions in China. Zhejiang Jiaxing Tongxiang Teyes Environmental
49 Protection Energy Limited Company is a CHP power plant. Its power generation equipment adopts
50 the method where multiple boilers and steam turbines transmit steam through the main pipe to
51 complete simultaneous regional heating and power supply. The structural diagram of a CHP system

52 is shown in Fig. 1, where it can be seen that it includes a single boiler and steam turbine involved
53 through the main pipe. The pulverized coal in the boiler burns to heat water to produce the
54 superheated steam. The steam is gathered in the main pipe and distributed to each steam turbine to
55 drive the steam turbine to generate electricity and supply electricity to users. The waste heat
56 transmitted by the gas condenser can also produce heat to the users so as to make full use of
57 resources. The NO_x and SO₂ produced by the power plant are controlled by the post-treatment
58 process, such as desulfurization, and then discharged from the chimney into the atmosphere.
59 However, before a pollutant discharge, the power plant needs to predict the emission amount at a
60 certain time in the future and whether the post-treatment effect reaches the expectation. If the
61 pollutant discharge is excessive, the power plant shall take necessary measures in advance to adjust
62 the production equipment parameters so that to reduce the emission.

63 In pollutant emission prediction, a method of physical mechanism analysis has been commonly
64 used to establish the mathematical model between the parameters of various power plant
65 components and pollutants, and the values of pollutants have been calculated through the mapping
66 relationship. For instance, Zhang et al. (2008) analyzed the generation mechanism of NO_x and SO₂
67 in thermal power plants according to the organic chemical structure of coal and concluded that the
68 furnace temperature, the uniformity of furnace temperature, the residence time of flue gas in the
69 high-temperature zone, excess air coefficient, and coal quality characteristics had a dominant
70 impact on pollutant generation; they also established a dynamic model to predict the pollutant
71 emissions.

72 However, the composition and structure of the CHP equipment are complex and the production
73 process is cumbersome. The establishment of a physical model reflecting equipment operation has

74 always been challenging (Nakaishi et al. 2021). Chemical reaction of the fuel in the boiler is highly
75 nonlinear. In addition, coal combustion properties, catalyst properties, reaction operation conditions,
76 and other factors can affect energy combustion and emission. Thus, it is very difficult to establish a
77 kinetic reaction model based on physical mechanisms.

78 Considering at the nonlinearity and heterogeneity of pollutant emission data in industrial
79 production activities, the data-driven statistical method can capture the complexity of mechanism
80 modeling and shows good performance (Lusis et al. 2017). Traditional methods for pollutant
81 emission prediction use statistical models, such as the ARIMA (Wang et al. 2021) and linear
82 regression. Muhammad et al. (2013) combined the genetic algorithm (GA) and linear regression
83 (LR) model to predict NO_x emission of a power plant and achieved satisfactory results. With the
84 beginning of the third upsurge of artificial intelligence, machine learning (ML) has been widely
85 used in the prediction of pollutant emission of power plants. For example, Zhou et al. (2020) used
86 the genetic algorithm to optimize the support vector machine (SVM) to predict NO_x emission of
87 thermal power plants, and the relative error of the final model was less than 3%. Thompson et al.
88 (2000) established a multi-layer perceptron (MLP) model for the long-term prediction of nitrogen
89 emission from the coal-fired power plants and analyzed the impact of different prediction time
90 lengths on the prediction result. Owing to the continuous upgrading of computer hardware, the
91 application of deep learning neural networks with a deeper number and more parameters in the field
92 of industrial production has gradually increased, and the field of power plants has not been an
93 exception. Belén et al. (2003) used an artificial neural network (ANN) to predict pollutant SO₂ near
94 a thermal power plant in Colomba Province, Spain; the accuracy of the artificial neural network was
95 improved compared with the semiparametric model. Sriram et al. (2006) used a multi-variable-input

three-layer feedforward neural network (FNN) to predict the amount of the suspended particulate matter in the emission of Neyveli Thermal Power Plant in India and achieved excellent results.

The long short-term memory network (LSTM) (Zou et al. 2020; Chang et al. 2020) model can well describe the data feature relationship contained in time series, and it has been widely used in text analysis (Zhang et al. 2020), time series prediction (Pappas et al. 2008; Vu et al. 2017), and speech recognition (Lee et al. 2011). Yang et al. (2020) used mutual information variable selection and an LSTM to predict the NO_x emission from power plant boilers dynamically, which improved the accuracy and robustness of the prediction model. Weng et al. (2020) stated that the traditional backpropagation (BP) neural network prediction model could consider only a one-way time relationship, resulting in a poor prediction effect and an inability to take into account the NO_x emission index and proposed a NO_x emission prediction model based on deep two-way LSTM, thus reducing the prediction error by about 5%.

In the field of big industrial data, performing multi-dimensional spatiotemporal data processing by combining different methods with an LSTM to establish hybrid models and extract data features from the space and time domains has become one of the dominant trends (Lin et al. 2021). For instance, a one-dimensional convolutional neural network (1D-CNN) can extract deep features of time series. Also, when used as the previous layer of an LSTM, it can significantly improve the feature extraction performance for input multi-variable time series. The attention mechanism (AM) (Zhang et al. 2021; Guo et al. 2021; Ghaffarian et al. 2021) can optimize the LSTM weight distribution and improve model training efficiency and recognition results by increasing the weight that has a stronger impact on the results and reducing the weight that has a weaker impact on the results. In the power industry, many neural network models have been used for load forecasting,

118 including CNN, LSTM, and mixed models. Peng et al. (2021) designed a CNN-LSTM neural
119 network forecasting model by combining a convolutional neural network (CNN) and an LSTM
120 neural network to predict the coal storage of power plants. This method uses the advantages of both
121 models, namely, a good feature extraction ability of the CNN and the special memory prediction
122 function of the LSTM neural network. Zhang et al. (2021) combined the attention mechanism with
123 a CNN seq2seq network and proposed a short-term multi-energy load forecasting method for power
124 plants, which was verified by the actual collected data.

125 However, the applications of hybrid CNN, LSTM, and AM models in pollutant emission
126 prediction of power plants have been very few. Therefore, this study proposes a pollutant emission
127 prediction method of a thermal power plant that combines feature engineering and a hybrid
128 CNN-LSTM-AM model to improve the prediction effect and achieve accurate prediction on an
129 hourly basis.

130 To integrate the most comprehensive information related to the pollutant prediction target, this
131 study proposes a strategy to determine the data dimension based on the Pearson correlation and
132 mutual information. The data used for the proposed model development are obtained from actual
133 production equipment, and their diversity and quantity are large, which addresses the overfitting
134 challenge of the deep learning-based methods.

135 The main contributions of this study can be summarized as follows:

136 (1) A multi-step feature engineering method, which includes the variance test, physical
137 mechanism filtering, Pearson correlation analysis, and mutual information calculation, is proposed
138 to process the target data. The collected data are screened and dimensionally reduced to obtain an
139 appropriate input for the prediction model.

(2) By combining CNN, LSTM, and AM, a hybrid model with multi-variable input is proposed.

The AM is introduced into the adaptive weight allocation of time features to identify more important inputs for the LSTM layer. The hybrid model compensates for the deficiency of a single LSTM and can ensure better depth feature extraction.

(3) An improved strategy for pollutant emission control of CHP, which makes decisions in advance by predicting changes in the future emissions, which are further used to adjust production equipment and reduce the damage to the environment, is proposed. In addition, a scheme for the application of the artificial intelligence-based method in the field of power production is introduced.

Fig. 1. Structure diagram of a CHP system

2. Data overview

The hourly operation data of the CHP units collected by Zhejiang Jiaxing Tongxiang Tayes Environmental Protection Energy Limited Company from July 2, 2020, to July 9, 2021, are used to

154 extract the NO_x and SO₂ data from the flue gas emission data of the power plant, which are then
155 used as a prediction target.

156 Historical operation data have typically been used as input data of a pollution emission
157 prediction model because of the high dependency on the prediction objectives. The pollutant
158 emission curve of a thermal power plant reflects daily and weekly production affected by the
159 production equipment of the power plant and daily activities in the power supply area. As the power
160 generation of a power plant shows a certain periodicity, the number of discharged pollutants
161 changes regularly. The collected operation parameters include both linear and nonlinear parameters
162 related to NO_x and SO₂ emission. Moreover, owing to the influence of seasonal factors, there are
163 large differences in regional power supply between summer and winter periods, and their
164 distribution ranges of emission values are completely different, which requires using a
165 high-accuracy prediction model. The emission curves of NO_x and SO₂ from June 4 to June 9 in
166 summer and from November 4 to November 9 in winter periods are presented in Fig. 2. As shown
167 in Fig. 2, in the summer, the NO_x emission is in the range of 10–30 mg/m³, but in the winter, it is
168 less than 3 mg/m³; meanwhile, the SO₂ emission shows smaller differences between the summer
169 and winter periods, namely, it is higher than 20 mg/m³ in the summer and lower than 20 mg/m³ in
170 the winter. Analyzing the daily emission results, it can be seen that the NO_x emission results show
171 high periodicity during daytime but low periodicity at night in both summer and winter periods,
172 while the SO₂ emission results are chaotic and irregular. Therefore, it is more challenging to predict
173 SO₂ compared to NO_x.

Fig. 2. Emission curves of NO_x and SO₂

3. Proposed method for NO_x and SO₂ emission prediction

3.1 Proposed method operation flow

The proposed pollutant emission prediction method mainly includes feature engineering, model establishment, and prediction, as shown in Fig. 3.

Feature engineering uses two methods to preprocess the original dataset, determine the factors highly correlated with the prediction target, reduce data dimension, and prevent the data dimension-related problem during model training. The pretreatment stage includes the Pearson correlation analysis (Zhang et al. 2021) and mutual information analysis (Liu et al. 2021; Ghosh et al. 2021; Miyashita et al. 2021), which are respectively used to extract the factors in the dataset that are linearly and nonlinearly related to NO_x and SO₂, which are then used as the model input.

The hybrid prediction model consists of CNN and LSTM, which are combined with the AM. Among them, the CNN is used to extract the potential spatial features of data and simplify network parameters, whereas the LSTM is used to extract the time distribution relationship between data; the AM is used to optimize the network weights so that to achieve better prediction results. The

191 summer data (collected from April to September) of the preprocessed historical operation data are
 192 divided into the training and test sets, which are used to train and test the depth neural network
 193 model, respectively; the methods of gradient descent and error backpropagation are used for model
 194 training.

195 After the model has converged, the model parameters are saved; the winter data (collected from
 196 October to March) are used as a verification set, which is used for model verification and to obtain
 197 the final prediction results.

198

199

200 **Fig. 3.** Process of a pollutant prediction method development

201 **3.2 Feature engineering**

202 To achieve better subsequent neural network feature mining, feature selection and

dimensionality reduction of the original data are needed to reduce the error caused by irrelevant factors. In this study, two methods are used to reduce the dimension of the collected data; namely, the Pearson correlation coefficient is used to measure the linear correlation between the data, and the mutual information method is used to extract the variables with high nonlinear correlation with the prediction target.

208

3.2.1 Influencing factor extraction based on Pearson correlation coefficient

First, the physical mechanism analysis and variance test are performed to remove irrelevant variables and variables with a little change. Then, the Pearson correlation coefficient is used to measure the correlation between various parameters and pollutant emission.

The Pearson correlation coefficient $\rho_{X,Y}$ is calculated as follows:

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}, \quad (1)$$

where $Cov(X,Y)$ denotes the covariance of X and Y , and σ_X is the variance of X .

The variables with a correlation coefficient greater than 0.2 are considered to have a certain linear correlation with the prediction target.

3.2.2 Data dimensionality reduction based on mutual information

Mutual information is used to evaluate the amount of information contributed by the occurrence of one event to another event. It has been often used to test the nonlinear correlation between multiple variables in feature engineering, and it is calculated by

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}, \quad (2)$$

where $P(x)$ and $P(y)$ are the marginal probability distributions of random variables X and y , respectively, and $p(x, y)$ denotes their joint probability distribution.

The characteristics of NO_x and SO₂ are selected. The analysis results obtained using the Pearson correlation coefficient and mutual information are shown in Table 1, where the maximum five values of each column are denoted in bold.

Table 1

Feature selection results obtained using the Pearson correlation coefficient and mutual information

Tag	Parameter	Pearson correlation coefficient		Mutual information	
		NO _x	SO ₂	NO _x	SO ₂
1	Power consumption rate	0.224	0.256	0.939	0.951
2	Coal consumption	0.286	0.231	0.943	0.949
3	Total power generation load	0.217	0.289	0.898	0.702
4	Total load of power supply	0.211	0.131	0.921	0.941
5	Oxygen content of flue gas	0.239	0.292	0.937	0.925
6	Feed water flow	0.237	0.214	0.926	0.970
7	Outlet steam temperature	-0.234	-0.165	0.923	0.938
8	Boiler feed water temperature	-0.233	-0.263	0.920	0.801
9	Boiler feed water pressure	0.207	0.109	0.892	0.920
10	Deaerator temperature	-0.211	-0.228	0.917	0.934

The input dimension of a deep learning-based model should not be too large; otherwise, the dimension problem will arise, resulting in a too-long model training time and difficulty in model fitting. This study selects five variables with the highest linear and nonlinear correlation with the target pollutant as the input data, and they are as follows: tag 1, tag 2, tag 5, tag 6, and tag 7 in Table 1.

In this study, the pollutant emission of the power plant is predicted on an hourly basis; namely, the prediction horizon h takes 1. Therefore, the i th continuous lag prediction target is given by

$$L_t^i = [l_{t-h-lag+1}^i, l_{t-h-lag+2}^i, \dots, l_{t-h}^i]^T, \quad (3)$$

where l_m^i denotes the i th constituent element at time m , lag represents the number of continuous lag prediction targets, and t is the prediction time.

The vector of the j th sliding window is given by

$$L_{t,j}^i = [l_{t-h-lag+j}^i, l_{t-h-lag+j+1}^i, \dots, l_{t-h-lag+n-1}^i]^T, j=1, 2, \dots, T, \quad (4)$$

where n is the sliding window width, and T is the total number of sliding windows, which is calculated by

$$T = lag + 1 - n. \quad (5)$$

Multiple input variables and continuous lag targets are reconstructed into two-dimensional (2D) features as follows:

$$L_{t,1}^{i \text{ res}} = [L_{t,1}^i, L_{t,2}^i, \dots, L_{t,T-1}^i, L_{t,T}^i]^T, \quad (6)$$

where $T \times n$ denotes the size of a 2D feature map, and the two dimensions correspond to the time step number and dimension of input features in the LSTM, respectively. In this study, n is set to six, including five pollutant emission influencing factors and prediction objectives.

3.3 Data normalization

The statistical characteristics of each feature of the dataset after the feature selection are calculated. As shown in Table 2, the data distribution intervals of the features are different.

Table 2

Statistical characteristics of data

Variable	Summer				Winter			
	Max	Min	Mean	Std	Max	Min	Mean	Std
Tag 1	47.02	-1.67	11.32	2.11	13.69	7.19	10.04	1.20
Tag 2	72.78	31.88	50.97	10.87	61.33	32.26	46.26	8.59
Tag 5	36.87	10.34	17.40	4.37	50.08	12.51	21.87	8.64
Tag 6	693.56	278.51	450.64	98.56	622.87	318.27	461.31	87.64
Tag 7	541.67	518.74	530.17	3.04	541.03	455.72	529.97	9.87
SO ₂	46.59	1.56	20.35	5.14	48.29	0	13.02	7.16
NO _x	45.06	20.00	26.37	2.92	6.11	0.33	3.16	1.18

258

259 To improve the neural network training efficiency and mine the potential features of data, it is
 260 necessary to normalize the original data to the interval of [0,1] and remove the dominant features of
 261 the macro dimension.

262 In this study, the maximum-minimum normalization is used, and the normalization formula is as
 263 follows:

$$264 \quad X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}, \quad (7)$$

265 where X denotes the original data value, X_{max} is the maximum data value, X_{min} is the minimum data
 266 value, and X_{norm} is the normalized data value.

267

268 3.4 CHP pollutant emission prediction by CNN-LSTM-AM model

269 3.4.1 CNN

270 The CNN is a deep feedforward neural network with the characteristics of local connection and
 271 weight sharing (Wei et al. 2021). Its operating principle is to extract local spatial information

272 through the sliding of the convolution kernel on the data matrix, reduce the number of features and
 273 parameters by the pooling layer, and improve the training efficiency so as to replace the traditional
 274 single fully connected layer. A typical CNN is composed of a convolution layer, a pooling layer, and
 275 a fully connected layer (Ghorbani et al.2021), as shown in Fig. 4.

276
 277 **Fig. 4.** CNN network structure diagram

278 3.4.2 LSTM

279 The LSTM is used in this study because it can capture long-term dependencies in time series.
 280 The LSTM represents an extension of the recurrent neural network (RNN) architecture
 281 (Abbasimehr et al. 2020), and its structure is shown in Fig. 5. Each LSTM unit has a storage unit
 282 and three gates (Han et al. 2019), namely, forgetting, input, and output gates. These internal
 283 structures control the flow of information related to the state of memory cells. The forgetting gate f_t
 284 controls how much information needs to be forgotten in the internal storage unit state at the
 285 previous time, the input gate i_t updates the storage unit state according to the information from the
 286 current input and the previously hidden state, and the output gate o_t controls how much information
 287 needs to be output in the internal state at the current time. These functions enable the LSTM to learn
 288 the time relationships of long-term sequences.

289 The operation of the three gates and storage unit at the time step t is given by the following

290 formula:

$$291 \quad f_t = \text{sig}(W_f x_t + V_f h_{t-1} + b_f), \quad (7)$$

$$292 \quad i_t = \text{sig}(W_i x_t + V_i h_{t-1} + b_i), \quad (8)$$

$$293 \quad o_t = \text{sig}(W_o x_t + V_o h_{t-1} + b_o), \quad (9)$$

$$294 \quad C'_t = \tanh(W_c x_t + V_c h_{t-1} + b_c), \quad (10)$$

$$295 \quad C_t = C_{t-1} \odot f_t + i_t \odot C'_t, \quad (11)$$

$$296 \quad h_t = o_t \odot \tanh(C_t), \quad (12)$$

297 where V_f , V_c , V_i , and V_o refer to the recurrent weights of the LSTM; W_f , W_c , W_i , and W_o indicate the
 298 weight matrixes of the forget gate, memory cell, input gate, and output gate, respectively; b_f , b_c , b_i ,
 299 and b_o represent the biases of the forget gate, memory cell, input gate, and output gate, respectively;
 300 sig is the sigmoid activation function, whose output ranges from zero to one, where zero indicates
 301 that information is totally forgotten, and one indicates that complete information that is reserved;
 302 $\odot$ stands for the point-wise multiplication operation.

303
304 **Fig. 5.** Structure diagram of the LSTM memory unit

305 3.4.3 AM

306 The AM is highly suitable for pollutant emission prediction models involving an LSTM network.
 307 The internal correlation between hidden features and the effects of the hidden features in different
 308 time steps on the prediction results are difficult to identify using an LSTM network (Wang et al.

2019). The method proposed in this study uses adaptive weighted hidden features to enhance key information and weaken trivial information. The introduction of the AM helps to identify the importance of the LSTM input so as to extract high-level features better.

The structure block diagram of the AM inspired by the work of reference (Zang et al. 2021) is shown in Fig. 6, and the data conversion process of the AM is presented in Fig. 7.

Fig. 6. Structure block diagram of the AM

First, the correlation between hidden features of different time steps is discussed from various dimensions. The score of the j th time step of the d th dimension of the hidden feature is calculated by

$$s_{j,d} = f_{sco}(W_{j,d}[h_{1,d}, h_{2,d}, \dots, h_{T,d}]), d = 1, \dots, n, j = 1, \dots, T, \quad (13)$$

where $h_{j,d}$ denotes the d th dimension of the hidden feature at the j th time step; $W_{j,d}$ is the weight vector, which shall be adjusted during training; function $f_{sco}(\cdot)$ is implemented using the full connection layer (Choi et al. 2018), where the number of neurons is equal to the number of time steps; n and T are the dimension and time step of hidden features, respectively.

The softmax function, which is defined by Eq. (14), is used to normalize the dimension score,

and the sum is equal to one.

$$s_{j,d}^{nor} = \frac{\exp(s_{j,d})}{\sum_{k=1}^T \exp(s_{k,d})}. \quad (14)$$

Finally, the normalized scores of each dimension is normalized by Eq. (15) to obtain the weight of the corresponding time step. Using Eq. (16), the hidden feature vector is multiplied by the weight and used as the input of the next layer.

$$M_j = \frac{1}{i} \sum_{d=1}^i s_{j,d}^{nor}, \quad (15)$$

$$h_j^{Att} = M_j h_j, \quad (16)$$

where M_j denotes the weight of the j th time step, and h_j^{Att} is the j th weight vector of the hidden feature.

Fig. 7. Data conversion flow of the AM

3.4.4 Proposed model structure

The structure of the proposed hybrid CNN-LSTM-AM model is shown in Fig. 8, where it can be seen that the hybrid model has one input layer, one CNN layer, one LSTM layer, one attention layer, one flatten layer, and one output layer; the AM is located in the attention layer.

The multi-variable time series data of pollution emissions are fed to the input layer, the potential local features are extracted by convolution of the CNN layer, and the time correlation of target components is extracted by the LSTM layer. The extracted hidden features are transmitted to the

attention layer, where they are weighted by the AM before being transmitted to the flatten layer; the flatten layer transforms the output results into a one-dimensional vector. Finally, the fully connected layer is used to obtain the prediction value corresponding to each superposition component so as to obtain the final prediction result of each combination, which is output by the output layer.

Fig. 8. Structure of the proposed CNN-LSTM-AM prediction model

4. Proposed model verification

4.1 Model structural parameters

To evaluate the performance of the proposed pollution emission prediction model, a case study was conducted using the real power plant data. The NO_x and SO₂ emission data collected in summer were used as the input data of the two prediction models, and the data were divided into training and test sets according to the ratio of 4:1, and the data collected in winter were used as a verification set intended for model testing. The pollution emission prediction model was constructed based on the structure displayed in Fig. 8, and through many experiments, the model super parameters were finally determined according to the optimal value of the model prediction effect and are shown in Table 3.

360

361 **Table 3**

362 Model structural parameters

Layer	Output shape
Input	40×6
Conv1D	$40 \times 64 \times 2$
Max-pooling	$40 \times 8 \times 2$
Dropout	40×8
Bi-LSTM	40×40
Permute_1	40×40
Dense	40×40
Activation	40×40
Permute_2	40×40
Multiply	40×40
Flatten	1600×1
Output	1×1

363

364 **4.2 Performance indicators**

365 Three evaluation indexes were used to evaluate the performance of the proposed pollution
 366 emission prediction model, including the mean absolute percentage error (MAPE), root mean
 367 square error (RMSE), and mean square error (MSE) (Chen et al. 2019), which are respectively
 368 given by Eqs.(18)–(20):

369

$$e_{MAPE} = \frac{|\hat{L}_i - L_i|}{L_i} \times 100\%, \quad (18)$$

370

$$e_{RMSE} = \frac{1}{N} \sqrt{\sum_{i=1}^N (\hat{L}_i - L_i)^2}, \quad (19)$$

371

$$e_{MSE} = \frac{1}{N} \sum_{i=1}^N (\hat{L}_i - L_i)^2, \quad (20)$$

372

where n is the total number of training samples, $\hat{L}_i$ is the predicted value, and L_i is the true value.

373

For the same batch of datasets, the closer the above three indexes are to zero, the better the

374

prediction effect is.

375

376

4.3 Experimental parameters

377

The proposed prediction model was trained using the error BP algorithm based on time

378

expansion; namely, the neural network was expanded into a deep network in the time order, and

379

then the expanded network was trained by the error BP algorithm (Li et al. 2021; Tang et al. 2021).

380

To control the learning rate of the network and to prevent the problems of gradient disappearance

381

and slow convergence speed, the Adam optimization algorithm (Tang et al. 2021) was used to

382

update the network parameters, and the initial learning rate lr was set to 0.001. The dropout

383

regularization method was introduced before the LSTM layer to avoid overfitting, and the retention

384

rate was set to 0.8. Through experimental observation, when the number of epochs exceeds 50, the

385

loss function has almost converged and does not decline. Therefore, the training parameters of the

386

neural network were set as follows: the maximum number of epochs was 50, and the small batch

387

size was 512.

388

389

4.4 Experimental hardware and software

390

This study was implemented on a personal computer having an NVIDIA GeForce GTX1650 and

391

16-GB RAM. The programming language was Python, and the integrated development environment

392 was Spyder. TensorFlow 2.1.1 and Keras 2.3.1, which denote the open-source deep learning
393 platform and software library, respectively, were used to develop the proposed model.

394 **5 Results and discussion**

395 **5.1 Performance comparison of different prediction models**

396 The proposed pollution emission prediction model was compared with the traditional prediction
397 models based on machine learning and deep learning. The comparison models included the
398 *K*-nearest neighbor (KNN), least squares support vector machine (LS-SVM), LSTM, CNN-LSTM,
399 and LSTM-Attention models. Their network parameter settings are consistent with the method
400 proposed in this paper. The number of LSTM cyclic cores was 40, and the number of CNN
401 convolution cores was 64. The optimal model parameters were selected in the other models
402 according to the prediction effect. The prediction time unit was one hour, and the continuous data of
403 40 previous hours were used to predict the NO_x and SO₂ emission levels in the next hour. The
404 predicted and actual results of the above-mentioned models on June 2, 2021 are shown in Fig. 9.

Fig. 9. The NO_x and SO₂ emission prediction results in summer

As shown in Fig. 9, the traditional machine learning-based model was not effective in predicting the pollution emission of the thermal power plant. In the second figure in Fig. 9, it can be seen that the NO_x emission showed a certain periodicity and regularity during the day. It should be noted that the machine learning model takes periodicity as a surface characteristic when capturing data characteristics to predict the final result. However, since there was no such regularity in SO₂ emission, which had a weak regularity, the effect of the traditional machine learning-based model was unsatisfactory. In contrast, the proposed CNN-LSTM-AM model could not only learn the shallow regularity characteristics of data but also mine the depth characteristics by the CNN. Therefore, compared to the traditional machine learning-based model, the prediction effect of the proposed model was better and its prediction results fitted closely with the real curve, and also, its changing trend was consistent with the real one. The generalization performance of the proposed model was further verified using the winter data as a verification set, and the results are shown in Fig. 10.

Fig. 10. The NO_x and SO₂ emission prediction results in winter

Owing to the obvious difference in the distribution characteristics between winter and summer data, a model with a poor training effect will perform unsatisfactorily on the verification set. As shown in Fig. 10, the performances of the KNN and LS-SVM models on the winter data were poor, and they differed significantly from the performance on the summer data, which are shown in Figure 8. Thus, the robustness of these models was low, and they could difficultly adapt to changes in data. The prediction effects of the LSTM and CNN-LSTM models declined slowly, and these

models had certain generalization performances. After the introduction of the AM, the characteristics learned in the same period were more accurate, and the prediction effect was gradually improved. This was because the weight distribution paid more attention to the factors that have a strong impact on the results and ignored irrelevant factors, so the model training efficiency was higher.

The errors of each of the models are shown in Table 4, where the minimum error values in each column are boldfaced. The RMSE value of the proposed CNN-LSTM-AM in predicting the pollution emissions of the thermal power plant was lower than one, and it was lower than the RMSE values of the traditional models. The MAPE value of the proposed model was lower in predicting NO_x than in predicting SO₂, but both values were lower than those of the traditional models.

Table 4
Comparison of prediction errors of each model

	Summer				Winter			
	NO _x		SO ₂		NO _x		SO ₂	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
KNN	2.108	6.36%	3.473	13.73%	3.214	8.15%	5.112	15.90%
LS-SVM	2.342	8.35%	2.163	9.03%	3.981	11.29%	4.375	8.07%
LSTM	0.788	4.29%	1.84	6.74%	1.431	4.75%	3.144	8.17%
CNN-LSTM	1.242	2.22%	1.851	5.963	2.726	2.22%	2.019	5.41%
LSTM-Attention	0.749	2.43%	1.475	6.23%	1.037	2.13%	1.779	8.93%
Proposed method	0.739	1.91%	1.752	5.67%	0.741	1.94%	0.171	5.36%

443

444 5.2 Loss function change experiment

445 To prove the superiority of the proposed CNN-LSTM-AM model over the traditional models,
 446 the proposed model was compared under the changes in the loss function (MSE) during the training
 447 with the LSTM, CNN-LSTM, and LSTM-Attention models, and the comparison results are shown
 448 in Fig. 11. The proposed method converged to the minimum value of the loss function in the early
 449 stage of model training, after only several iteration rounds. Compared with the proposed model,
 450 although the convergence speed of LSTM with AM is slower, the curve is still smooth, while the
 451 single LSTM and CNN-LSTM models showed an upward trend of the loss function in the early
 452 stage of training, which proved that there was an overfitting phenomenon; the convergence speed of
 453 the follow-up model was also slow. The LSTM did not converge even after 40 rounds, and its
 454 training effect was poor. This is enough to prove that adding the AM to the time series prediction
 455 model can effectively mitigate the overfitting phenomenon and improve model training efficiency
 456 and final prediction effect.

457
 458 **Fig. 11.** Loss function comparison results

459 5.3 Attention layer weight visualization

460

461 The weight of the attention layer of the power plant pollution emission prediction model was

462 used to draw a thermodynamic diagram, as shown in Fig. 12. In the thermodynamic diagram, the
463 brighter and darker parts indicate higher and lower attention of the AM during weight allocation,
464 respectively. As shown in Fig. 12, the difference between brighter and darker parts increases in the
465 areas close to the prediction time point for both NO_x and SO₂ data; thus, the data characteristics
466 closer to the prediction point have a more obvious impact on it, which is consistent with common
467 sense.

468
469 **Fig. 12.** Visualization of the attention layer weights
470 (a) NO_x; (b) SO₂

471 **6. Conclusions and future work**

472 To solve the problem of accurate prediction of CHP pollutant emission, this study proposes a
473 CHP pollution emission prediction model named the CNN-LSTM-AM. The proposed model uses
474 the deep mining ability of the CNN, the time series data processing ability of the LSTM, and the
475 weight optimization ability of the AM to improve prediction accuracy so as to provide technical
476 support for pollutant emission control.

477 The proposed CNN-LSTM-AM prediction model is trained with the NO_x and SO₂ data collected
478 in summer and tested on the data collected in winter; the proposed model's prediction results are

479 compared with the real data. Also, the proposed model is compared with the traditional prediction
480 models. Based on the comparison results, the following conclusion can be drawn:

481 1) The proposed CNN-LSTM-AM prediction model can accurately predict the future NO_x and
482 SO₂ emission levels, and its RMSE is less than 1 mg/m³.

483 2) Compared with the traditional prediction models based on machine learning and deep learning,
484 the proposed model has higher prediction accuracy and better robustness.

485 From feature engineering and data processing to neural network prediction, this study gives the
486 measures of artificial intelligence application in clean energy usage in the power production
487 industry. Although the proposed model can accurately predict the pollutant emission level at a
488 single time point, it is difficult to predict the changing trend of pollutant emission in a longer period.
489 To address this challenge, in our future research, the multi-step time-series prediction over a long
490 period will be studied, and other variables such as pollutant emission of a power plant will be
491 considered to solve the actual demand.

492

493 **Acknowledgements**

494 This research is financially supported by the National Natural Science Foundation of China
495 (51705455), Scientific Research Foundation of Zhejiang University City College (No.J-202113).

496

497 **Authors Contributions**

498 Anping Wan: Writing – Original draft. Jie Yang: Methodology. Ting Chen: Supervision. Yang
499 Jinxing: Formal analysis. Ke Li: Writing – Introduction. Zhou Qinglong: Writing – Introduction.

500
501

502
503
504

505
506

507
508
509

510
511
512
513

514

515

516
517

518
519
520
521

Ethical Approval

Not applicable.

Consent to Participate

All authors have agreed to submit the manuscript.

Consent to Publish

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.

Availability of data and materials

Not applicable.

Reference

Abbasimehr H, Shabani M, Yousefi M (2020) An optimized model using LSTM network for demand forecasting. Comput Ind Eng 143: 106435.
<https://doi.org/10.1016/j.cie.2020.106435>
Bunyamin M-A, Yap K-S, Aziz N, Sheih K-T; Shen Y-W; Md F-K (2013) A hybrid genetic

522 algorithm and linear regression for prediction of NO_x emission in power generation plant.

523 Iop Conference 16: 012101. <https://doi.org/10.1088/1755-1315/16/1/012101>

524 Castro B, Sanchez J, Manteiga W, Bande M , Cela J, Fernandez J (2003) Prediction of SO₂

525 levels using neural networks. JOURNAL OF THE AIR & WASTE MANAGEMENT

526 ASSOCIATION 53: 532-539. <https://doi.org/10.1080/10473289.2003.10466186>

527 Chang Y, Abimannan S, Chiao H, Lin C, Huang Y (2020) An ensemble learning based hybrid

528 model and framework for air pollution forecasting. Environmental science and pollution

529 research international 27: 1-14. <https://doi.org/10.1007/s11356-020-09855-1>

530 Choi H, Cho K, Bengio Y (2018) Fine-grained attention mechanism for neural machine

531 translation. Neurocomputing 284: 171-176. <https://doi.org/10.1016/j.neucom.2018.01.007>

532 Chen H, Zhang J, Tao Y, Tan F, (2019) Asymmetric GARCH type models for asymmetric

533 volatility characteristics analysis and wind power forecasting. Protect Control Mod Power

534 Syst 4: 356-366. <https://doi.org/10.1186/s41601-019-0146-0>

535 Guo H, Zhu K-P (2021) Attention-based dual-scale hierarchical LSTM for tool wear monitoring.

536 Manufacturing Letters 29: 99-103. <https://doi.org/10.1016/J.MFGLET.2021.07.015>

537 Ghaffarian S, Valente J, Tekinerdogan B (2021) Effect of Attention Mechanism in Deep

538 Learning-Based Remote Sensing Image Processing: A Systematic Literature Review.

539 Remote Sensing 13: 2965-2965. <https://doi.org/10.3390/rs13152965>

540 Ghosh S-M, Bandyopadhyay S, Mitra D (2021) Nonlinear classification of emotion from EEG

541 signal based on maximized mutual information. Expert Systems With Applications 185:

542 115605. <https://doi.org/10.1016/J.ESWA.2021.115605>

543 Ghorbani Z, Behzadan A-H (2021) Monitoring offshore oil pollution using multi-class

convolutional neural networks. Environmental Pollution 289: 117884.
<https://doi.org/10.1016/J.ENVPOL.2021.117884>

Hyeunguk A, Miller W, Sheaffer P, Tutterow V, Rapp V (2021) Opportunities for installed combined heat and power (CHP) to increase grid flexibility in the U.S.. Energy Policy 157: 112485. <https://doi.org/10.1016/J.ENERPOL.2021.112485>

Han L, Jing H, Zhang R, Gao Z (2019) Wind power forecast based on improved Long Short Term Memory network. Energy 189: 116300. <https://doi.org/10.1016/j.energy.2019.116300>

Lusis P, Khalilpour K-R, Andrew L, Liebman A (2017) Short-term residential load forecasting: impact of calendar effects and forecast granularity. Appl Energy. 205: 654-669. <https://doi.org/10.1016/j.apenergy.2017.07.114>

Lee C-M, Ko C-N (2011) Short-term load forecasting using lifting scheme and ARIMA models. Expert Syst Appl 38: 5902-5911. <https://doi.org/10.1016/j.eswa.2010.11.033>

Lin B, Zhang C (2021) A novel hybrid machine learning model for short-term wind speed prediction in inner Mongolia, China. Renewable Energy 179: 1565-1577. <https://doi.org/10.1016/J.RENENE.2021.07.126>

Liu H, Wang Y, Li X, Yang G (2021) Prediction of NOx Emissions of Coal-Fired Power Plants Based on Mutual Information-Graph Convolutional Neural Network. Proceedings of the CSEE: 1-9. <https://doi.org/10.13334/j.0258-8013.pcsee.202540>

Li Y, Zhu T (2021) Application of the collision mathematical model based on a BP neural network in railway vehicles. Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit 235: 713-725. <https://doi.org/10.1177/0954409720959869>

Meng X (2021) Research on technological innovation effect of environmental regulation from

566 perspective of industrial transfer: Evidence in China's thermal power industry. *Cleaner*
 567 *Engineering and Technology* 4: 100178. <https://doi.org/10.1016/J.CLET.2021.100178>
 568 Miyashita N, Yonezawa Y (2021) Mutual information analysis of the dynamic correlation
 569 between side chains in proteins. *The Journal of chemical physics* 155: 044107.
 570 <https://doi.org/10.1063/5.0055662>
 571 Nakaishi T, Kagawa S, Takayabu H, Lin C (2021) Determinants of technical inefficiency in
 572 China's coal-fired power plants and policy recommendations for CO2 mitigation.
 573 *Environmental science and pollution research international* 28: 52064-52081.
 574 <https://doi.org/10.1007/s11356-021-14394-4>
 575 Pan J, Hou G, Wang J (2021) Research on Environmental Efficiency of Power Industry in China
 576 and Its Influencing Factors: Based on Provincial Panel Data of 2008-2017. *Journal of*
 577 *Technology Economics* 40: 20-30.
 578 Pappas S-S, Ekonomou L, Karamousantas D-Ch, Chatzarakis G-E, Katsikas S-K, Liatsis P (2008)
 579 Electricity demand loads modeling using AutoRegressive Moving Average (ARMA) models.
 580 *Energy* 33: 1353-1360. <https://doi.org/10.1016/j.energy.2008.05.008>
 581 Peng D, Zhu Q, Che Q, Zhao H (2021) Short-term coal storage forecasting of thermal power
 582 plant for power grid dispatching based on CNN-LSTM neural network. *Electric Power*
 583 *Automation Equipment* 47: 127-132. <https://doi.org/10.16081/j.epae.202102025>
 584 Sriram G, Mohan N-K, Gopalasamy V (2006) Neural network for short-term predictions of
 585 ambient particulate matter around thermal power plant. *Journal of Scientific & Industrial*
 586 *Research* 65: 639-645.
 587 Thompson S, Li K (2000) A MLP prediction model for power plant NOx emission. *IFAC*

588 Proceedings Volumes 33: 117-122. [https://doi.org/10.1016/S1474-6670\(17\)35457-5](https://doi.org/10.1016/S1474-6670(17)35457-5)

589 Tang S, Teng Z, Sun B, Hu Q, Pan X (2021) Improved BP neural network with ADAM optimizer
590 and the application of dynamic weigh. Journal of Electronic Measurement and
591 Instrumentation 35: 127-135. <https://doi.org/10.13382/j.jemi.B2003558>

592 Vu D-H, Bouzerdoum A (2017) Short-term electricity demand forecasting using autoregressive
593 based time varying model incorporating representative data adjustment. Appl Energy 205:
594 790-801. <https://doi.org/10.1016/j.apenergy.2017.08.135>

595 Wang X, Tian W, Liao Z (2021) Statistical comparison between SARIMA and ANN's
596 performance for surface water quality time series prediction. Environmental science and
597 pollution research international 28:33531-33544.
598 <https://doi.org/10.1007/s11356-021-15574-y>

599 Weng X, Qian Y, Peng D, Ma H (2020) NOx Emission Prediction Model of SCR System based
600 on Deep Bidirectional LSTM. JOURNAL OF ENGINEERING FOR THERMAL ENERGY
601 AND POWER 35: 57-64. <https://doi.org/10.16146/j.cnki.rndlgc.2020.10.009>

602 Wang S, Wang X, Wang S, Wang D (2019) Bi-directional long short-term memory method based
603 on attention mechanism and rolling update for short-term load forecasting. Int J Electr
604 Power Energy Syst 109: 470-479. <https://doi.org/10.1016/j.ijepes.2019.02.022>

605 Wei Z, Liu Y (2021) Construction of super-resolution model of remote sensing image based on
606 deep convolutional neural network. Computer Communications 178: 191-200.
607 <https://doi.org/10.1016/J.COMCOM.2021.06.022>

608 Xiao C, Liao Y (2021) Transformer Order Demand Forecasting Based on Grey Forecasting
609 Model. IOP Conference Series: Earth and Environmental Science 831: 012004.

610 <https://doi.org/10.1088/1755-1315/831/1/012004>

611 Yang G, Wang Y, Li X, Liu K (2020) Dynamic Prediction of Boiler NO_x Emission Based on
612 Mutual Information Variable Selection and LSTM. Journal of North China Electric Power
613 University 47: 66-74. <https://doi.org/CNKI:SUN:HBDL.0.2020-03-009>

614 Zhang J (2016) Study on China's strategic measures to achieve peak carbon dioxide emissions in
615 2030. Energy research and utilization: 18-21.
616 <https://doi.org/10.16404/j.cnki.issn1001-5523.2016.06.007>

617 Zhang L, Mai Y, Zhang Y, Huang Q (2008) Prediction model of SO₂ and NO_x pollutants in large
618 coal-fired power plant based on Neural Network. Thermal Power Generation: 9-13.
619 <https://doi.org/10.19666/j.rlfed.2008.10.003>

620 Zhou J, Liang H, Zhao Y (2020) Prediction of NO_x Emissions from Thermal Power Plants in
621 China Based on a Grey Support Vector Machine Combination Model. Research of
622 Environmental Sciences 24: 489-496. <https://doi.org/10.13198/j.res.2011.05.17.zhaoy.004>

623 Zou Q, Xiong Q, Li Q, Yi H, Yu Y, Wu C (2020) A water quality prediction method based on the
624 multi-time scale bidirectional long short-term memory network. Environmental science and
625 pollution research international 27:16853-16864.
626 <https://doi.org/10.1007/s11356-020-08087-7>

627 Zhang G, Guo J (2020) A novel method for hourly electricity demand forecasting. IEEE Trans
628 Power Syst 35: 1351-1363. <https://doi.org/10.1109/TPWRS.2019.2941277>

629 Zhang G, Yang J, Zheng Y, Wang Y, Wu Y, Chen S (2021) Hybrid-attention guided network with
630 multiple resolution features for person re-identification. Information Sciences 578:
631 525-538. <https://doi.org/10.1016/J.INS.2021.07.058>

632 Zhang G, Bai X, Wang Y (2021) Short-time multi-energy load forecasting method based on
633 CNN-Seq2Seq model with attention mechanism. Machine Learning with Applications 5:
634 100064. <https://doi.org/10.1016/J.MLWA.2021.100064>

635 Zhang B, Yan L, Lu L, Huang H, Feng J, Zhang E (2021) Single-pole-to-ground fault
636 identification of photovoltaic power station collection system based on Pearson correlation
637 coefficient. Proceedings of the CSU-EPSA: 1-8.
638 <https://doi.org/10.19635/j.cnki.csu-epsa.000798>

639 Zang H, Xu R, Cheng L, Ding T, Liu L, Wei Z, Sun G (2021) Residential load forecasting based
640 on LSTM fusing self-attention mechanism with pooling. Energy 229: 120682.
641 <https://doi.org/10.1016/J.ENERGY.2021.120682>