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 Yang2 · Ting Chen1 · Yang Jinxing3 · Ke Li2 · Zhou Qinglong2

Abstract
Combined thermal power (CHP) production mode plays a more important role in energy production, but the impact of its pollutant emission on the natural environment is still difficult to eradicate. Traditional pollutant control adopts post-treatment process to degrade the generated pollutants, but there is little research on controlling the generation of pollutants from the source. Therefore, starting from the source, this paper predicts the pollutants through the prediction model, so as to provide countermeasures for production regulation and avoiding excessive emission. In this paper, 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 · Machine learning

Nomenclature

| Symbol | Description |
|-------|-------------|
| 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 [-] |
| t     | Current time [h] |
| h     | Prediction horizon [h] |

Introduction

Among all environmental pollution sources in China, power industry pollution is one of the most dominant pollution sources (Pan et al. 2021; Xiao and Liao 2021). Human production activities, such as automobile exhaust emission, power generation in thermal power plants, and fuel combustion in other industrial production, highly increase the content of NOX and SO2 in the air. According to the preliminary estimation of the United Nations Environmental Planning (UNEP) deployment project “Comprehensive Consideration of Environmental Factors in Energy Planning,” 70% of atmospheric emissions of NOX in China originates from the direct combustion of
coal (Zhang 2016), which has an immeasurable impact on nature.

In recent years, numerous effective measures have been taken in many fields to control anthropogenic pollutant emissions (Meng 2021). For instance, in the field of thermal power generation, the application of combined heat and power (CHP) production mode (Hyeunguk et al. 2021) can improve resource utilization and reduce environmental impact under the same energy consumption. The Emission Standard of Air Pollutants for Thermal Power Plants (2004), which was implemented in China in 2012, has been strictly stipulated that the NOx and SO2 emissions of the new circulating fluidized bed coal-fired boilers are 100 mg/m^3, and there have been even more stringent requirements for certain regions in China. The traditional power plant adopts the post-treatment process to degrade the generated pollutants. But, the power generation capacity of power plants far exceeds the power consumption of users and produces too many pollutants, which is difficult to guarantee only after the post-treatment. From the source of pollutants, if the future production law and pollutant emission can be predicted in advance according to the historical operation of the power plant and the law of regional power supply, the current equipment operation conditions can be adjusted to avoid excess coal consumption. Combined with post-treatment, the impact of pollutants on the environment can be reduced more effectively.

In pollutant emission prediction, a method of physical mechanism analysis has been commonly used to establish the mathematical model between the parameters of various power plant components and pollutants. However, it is a great challenge to establish a physical model due to the complex structure of CHP equipment and cumbersome production process (Nakaishi et al. 2021). Considering the nonlinearity and heterogeneity of pollutant emission data in industrial production activities, the data-driven statistical method can capture the complexity of mechanism modeling and shows good performance (Lusis et al. 2017). Pollution emission prediction is a multi-dimensional time series prediction problem. For dealing with this problem, the neural network model in the field of machine learning has excellent results. The long short-term memory network (LSTM) model can well describe the data feature relationship contained in time series (Zou et al. 2020; Chang et al. 2020), and it has been widely used in time series prediction (Zhang and Guo 2020). Yang et al. (2020) used mutual information variable selection and an LSTM to predict the NOx emission from power plant boilers dynamically, which improved the accuracy and robustness of the prediction model. Weng et al. (2020) proposed a NOx emission prediction model based on deep bidirectional LSTM, which reduces the prediction error about 5%.

In the field of big industrial data, it has become one of the mainstream trends to combine different methods with LSTM to establish hybrid models to extract the time features of data (Lin and Zhang 2021). For instance, convolutional neural network (CNN) can extract the deep features of multi-dimensional time series and significantly improve the performance of feature extraction. Attention mechanism (AM) (Zhang et al. 2021a, b, c; Guo and Zhu, 2021; Ghaffarian et al. 2021) can optimize the model weight distribution and improve the model training efficiency by increasing the weight that has a stronger impact on the results. Peng et al. (2021) combined CNN with LSTM and used their advantages to predict the coal storage of power plants. Zhang et al. (2021a, b, c) combined the AM with LSTM network and proposed a short-term multi-energy load forecasting method for power plants, which was verified by the actual collected data.

However, the applications of hybrid CNN, LSTM, and AM models in pollutant emission prediction of power plants have been very few. Therefore, combining the respective advantages of the above three models, this paper proposes a hybrid model for CHP emission prediction. The main contributions of this study can be summarized as follows:

1) CNN is used to extract the features of input data through its powerful feature extraction function. The temporal features of multi-dimensional time series are extracted by LSTM.

2) The attention mechanism is used to distribute the weight of time series features, which improves the accuracy of emission prediction of CHP.

3) The working and environmental variables of equipment are filtering and refactoring by feature engineering, and the pollutant emission law caused by seasonality is analyzed, which improve prediction accuracy and efficiency.

4) The real data of combined cycle power plant are used to verify the proposed model, which provides a case for the application of artificial intelligence technology in the field of environmental protection and energy utilization.

**Data overview**

Zhejiang Jiaxing Tongxiang Tayes Environmental Protection Energy Limited Company is a CHP power plant (Fig. 1). It provides the power and heat demand of the
surrounding industries and residents. The pulverized coal in the boiler burns to heat water to produce the superheated steam. The steam is gathered in the main pipe and distributed to each steam turbine to drive the steam turbine to generate electricity and supply electricity to users. The waste heat transmitted by the gas condenser can also produce heat to the users so as to make full use of resources. The NOX and SO2 produced by the power plant are controlled by the post-treatment process, such as desulfurization, and then discharged into the atmosphere.

The hourly operation data of the CHP units collected from July 2, 2020, to July 9, 2021, are used to extract the NOX and SO2 data from the flue gas emission data of the power plant, which are then used as a prediction target. Historical operation data have typically been used as input data of a pollution emission prediction model because of the high dependency on the prediction objectives. The pollutant emission curve of a thermal power plant reflects daily and weekly production affected by the production equipment of the power plant and daily activities in the power supply area. As the power generation of a power plant shows a certain periodicity, the number of discharged pollutants changes regularly. Moreover, owing to the influence of seasonal factors, there are large differences in regional power supply between summer and winter periods, and their distribution ranges of emission values are completely different, which requires using a high-accuracy prediction model. The emission curves of NOX and SO2 from June 4 to June 9 in summer and from November 4 to November 9 in winter periods are presented in Fig. 2. As shown in Fig. 2, in the summer, the NOX emission is in the range of 10–30 mg/m³, but in the winter, it is less than 3 mg/m³; meanwhile, the SO2 emission shows smaller differences between the summer and winter periods, namely, it is higher than 20 mg/m³ in the summer and lower than 20 mg/m³ in the winter. Analyzing the daily emission results, it can be seen that
the NOX emission results show high periodicity during daytime but low periodicity at night in both summer and winter periods, while the SO2 emission results are chaotic and irregular. Therefore, it is more challenging to predict SO2 compared to NOX.

**Proposed method for NOX and SO2 emission prediction**

**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 data, determine the factors highly correlated with the prediction target, and reduce data dimension. The data preprocessing includes the Pearson correlation analysis (Zhang et al. 2021a, b, c) and mutual information analysis (Liu et al. 2021; Ghosh et al. 2021; Miyashita and Yonezawa 2021), which are respectively used to extract the factors that are linearly and nonlinearly related to NOX and SO2 emission.

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 to achieve better prediction results. The summer data (collected from April to September) are divided into the training and test sets, which are used to train and

---

**Fig. 2** Emission curves of NOX and SO2

**Fig. 3** Process of a pollutant prediction method development
test the depth neural network model, respectively; the methods of gradient descent and error backpropagation are used for model training.

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

Feature engineering

In order to achieve better follow-up neural network feature mining, it is necessary to select and reduce the dimension of the original data 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 and the mutual information.

Pearson correlation analysis

First, the physical mechanism analysis and variance test are performed to remove irrelevant variables and less volatile variables. Then, the Pearson correlation coefficient ($\rho_{X,Y}$) shown in Eq. (1) is used to measure the correlation between various parameters and pollutant emission. Pearson correlation coefficient is a measure of the linear correlation degree between two variables. Its value is distributed in the $[-1,1]$ interval. The greater its absolute value is, the greater the correlation degree of the two variables involved in the calculation is.

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

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

The calculation results of Pearson coefficient between each variable and the prediction target are shown in Table 1, which are less than 0.3, indicating that the linear correlation between single performance parameters of CHP Equipment and pollution emission is very weak (Dancey and Reidy 2017); this shows that in power production, the complex equipment structure makes the coupling relationship between various parameters not a simple linear relationship, but a more complex nonlinear relationship. Therefore, it is necessary to further measure the nonlinear correlation between parameters and targets through other indicators.

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, which is calculated by (Peng et al. 2005):

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

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.

As shown in Table 1, the mutual information between each variable and the prediction target is basically higher than 0.9, and the maximum value reaches 0.970, indicating that the variable contains a large amount of nonlinear information of the objective function, and its change has a strong impact on the prediction results.

Through the above two steps, the ten most relevant indicators screened are shown in Table 1, where the maximum five values of each column are denoted in bold.

| Tag | Factors | Abbreviation | $\rho_{X,Y}$ | $I(X,Y)$ |
|-----|---------|--------------|-------------|----------|
| 1   | Electricity consumption | $C_E$ | 0.224 | 0.256 | 0.939 | 0.951 |
| 2   | Coal consumption | $C_C$ | 0.286 | 0.231 | 0.943 | 0.949 |
| 3   | Total power generation load | $L_{PG}$ | 0.217 | 0.289 | 0.898 | 0.702 |
| 4   | Total load of power supply | $L_{PS}$ | 0.211 | 0.131 | 0.921 | 0.941 |
| 5   | Oxygen content | $O_2$ | 0.239 | 0.292 | 0.937 | 0.925 |
| 6   | Feed water flow | $F_W$ | 0.237 | 0.214 | 0.926 | 0.970 |
| 7   | Outlet steam temperature | $T_{OS}$ | -0.234 | -0.165 | 0.923 | 0.938 |
| 8   | Boiler feed water temperature | $T_{BW}$ | -0.233 | -0.263 | 0.920 | 0.801 |
| 9   | Boiler feed water pressure | $P_{BW}$ | 0.207 | 0.109 | 0.892 | 0.920 |
| 10  | Deaerator temperature | $T_D$ | -0.211 | -0.228 | 0.917 | 0.934 |
Because the linear correlation of variables to the target is very low, it is not feasible to adopt the traditional linear regression model to predict pollutant emission, but the nonlinear correlation is very high, which poses a higher challenge to the selection and design of the model. The deep learning neural network has strong nonlinear feature extraction ability, so this study builds the pollutant emission prediction model based on the neural network.

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 correlation with the target pollutants as the input data, which are $C_E$, $C_C$, $O_2$, $F_W$, $T_{OS}$. The $O_2$ here represents the sum of the oxygen content at the left and right inlets of the high-temperature economizer of three boilers.

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

$$L_i^t = [l_{i-h+1}^t, l_{i-h+2}^t, ..., l_{i-h}^t]^T,$$

(3)

where $l_m^t$ denotes the $i$th constituent element at time $m$, $\text{lag}$ 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

$$L^i_{t,j} = [l_{i-h+1}^t, l_{i-h+1}^{t+1}, l_{i-h+1}^{t+2}, ..., l_{i-h}^{t+n-1}]^T, j = 1, 2, ..., T,$$

(4)

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

$$T = \text{lag} + 1 - n.$$

(5)

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

$$L_{i,1}^{\text{res}} = [L_{i,1}^1, L_{i,2}^1, ..., L_{i,T-1}^1, L_{i,T}^1]^T,$$

(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.

### Data normalization

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

To improve the neural network training efficiency and mine the potential features of data, it is necessary to normalize the original data to the interval of $[0,1]$ and remove the dominant features of the macro dimension. In this study, the Max–min normalization (Ge et al. 2021) is used, which is shown in Eq. (7):

$$x_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}},$$

(7)

where $x$ denotes the original data, $x_{\max}$ is the maximum data, $x_{\min}$ is the minimum data.

### Table 2: Statistical characteristics of data

| Variable | Summer | Winter |
|----------|--------|--------|
|          | Max    | Min    | Mean  | Std   | Max    | Min    | Mean  | Std   |
| $C_E$(MW)   | 47.02  | 0      | 11.32 | 2.11  | 13.69  | 7.19   | 10.04 | 1.20  |
| $C_C$(t/h)  | 72.78  | 31.88  | 50.97 | 10.87 | 61.33  | 32.26  | 46.26 | 8.59  |
| $O_2$(vol%) | 36.87  | 10.34  | 17.40 | 4.37  | 50.08  | 12.51  | 21.87 | 8.64  |
| $F_W$(t/h)  | 693.56 | 278.51 | 450.64 | 98.56 | 622.87 | 318.27 | 461.31 | 87.64 |
| $T_{OS}$(℃) | 541.67 | 518.74 | 530.17 | 3.04  | 541.03 | 455.72 | 529.97 | 9.87  |
| $SO_2$(mg/m³) | 46.59  | 1.56   | 20.35 | 5.14  | 48.29  | 0      | 13.02 | 7.16  |
| $NO_X$(mg/m³) | 45.06  | 20.00  | 26.37 | 2.92  | 6.11   | 0.33   | 3.16  | 1.18  |
CNN-LSTM and AM hybrid model

1D-CNN

The 1D-CNN is a deep feedforward neural network with the characteristics of local connection and weight sharing (Wei and Liu 2021). Its operating principle is to extract local spatial information through the sliding of the convolution kernel on the data matrix, reduce the number of features and parameters by the pooling layer, and improve the training efficiency so as to replace the traditional single fully connected layer. A typical CNN is composed of a convolution layer, a activation layer, a pooling layer, and a fully connected layer (Ghorbani and Behzadan 2021).

LSTM

The LSTM is used in this study because it can capture long-term dependencies in time series. The LSTM represents an extension of the recurrent neural network (RNN) architecture (Abbasimehr et al. 2020), and its structure is shown in Fig. 4. Each LSTM unit has a storage unit and three gates (Han et al. 2019), namely, forgetting, input, and output gates. These internal structures control the flow of information related to the state of memory cells and enable the LSTM to learn the time relationships of long-term sequences. The operation of the three gates at the time step $t$ is given by the following formula (Gers et al. 2000; Hochreiter et al. 1997):

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

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

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

where $V_f, V_i,$ and $V_o$ refer to the recurrent weights of the LSTM; $W_f, W_i,$ and $W_o$ indicate the weight matrixes of the forget gate, input gate, and output gate, respectively; $b_f, b_i,$ and $b_o$ represent the biases of the three gates, respectively; “sig” is the sigmoid activation function.

AM

The AM is highly suitable for pollutant emission prediction models involving an LSTM network. The internal correlation between hidden features and the effects of the hidden features in different 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. 5, and the data conversion process of the AM is presented in Fig. 6.

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

$$s_{j,d} = f_{scg}(W_{j,d}(h_{1,d}, h_{2,d}, ..., h_{T,d})), d = 1, ..., n, j = 1, ..., T,$$

where $W_{j,d}$ is the weight matrix of the AM, $h_{t,d}$ is the hidden feature of the $d$th dimension at the $t$th time step, and $f_{scg}$ is the function that calculates the score.
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 $f_{\text{scn}}(\cdot)$ is implemented using the dense layer (Choi et al. 2018) and neurons number is equal to the time step; $n$ 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.

$$s_{j,d}^\text{nor} = \frac{\exp(s_{j,d})}{\sum_{d=1}^{T} \exp(s_{k,d})}, \quad (12)$$

Finally, the normalized scores of each dimension are 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.

$$M_j = \frac{1}{T} \sum_{d=1}^{T} s_{j,d}^\text{nor}, \quad (13)$$

### Table 3 Model structural parameters

| Layer         | Output shape |
|---------------|--------------|
| Input         | 40 × 6       |
| Conv1D        | 40 × 64 × 2  |
| Max-pooling   | 40 × 8 × 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        |

Fig. 6 Data conversion flow of the AM

Fig. 7 Structure of the proposed pollution emission prediction model
where $M_j$ denotes the weight of the $j$th time step, and $h_j^{\text{At}}$ is the $j$th weight vector of the hidden feature.

$$h_j^{\text{At}} = M_j h_j^j,$$  \hspace{1cm} (14)

The structure of the proposed hybrid CNN-LSTM-AM model is shown in Fig. 7. It 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 emis-
sions are fed to the input layer, the potential local fea-
tures 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 flat-
ten layer; the flatten layer transforms the output results
into a one-dimensional vector. Finally, the output layer
is used to obtain the prediction value corresponding to
each superposition component so as to obtain the final
prediction result of each combination.

Fig. 9 The NO\textsubscript{X} and SO\textsubscript{2} emission prediction results in winter
Proposed model verification

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 NOX 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. 7, 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.

Performance indicators

Three evaluation indexes were used to evaluate the performance of the proposed pollution emission prediction model, including the mean absolute percentage error (MAPE), root mean square error (RMSE), and mean square error (MSE) (Chen et al. 2019), which are respectively given by Eqs. (15)–(17):

$$e_{\text{MAPE}} = \frac{|\hat{L}_i - L_i|}{L_i} \times 100\%,$$  \hspace{1cm} \hspace{1cm} \hspace{1cm} \hspace{1cm} (15)

$$e_{\text{RMSE}} = \frac{1}{N} \sqrt{\sum_{i=1}^{N} (\hat{L}_i - L_i)^2},$$  \hspace{1cm} \hspace{1cm} \hspace{1cm} \hspace{1cm} (16)

$$e_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^{N} (\hat{L}_i - L_i)^2,$$  \hspace{1cm} \hspace{1cm} \hspace{1cm} \hspace{1cm} (17)

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

For the same batch of datasets, the closer the above three indexes are to zero, the better the prediction effect is.

Table 4 Comparison of prediction errors of each model

|          | Summer NOX | Summer SO2 | Winter NOX | Winter SO2 |
|----------|------------|------------|------------|------------|
|          | $e_{\text{RMSE}}$ | $e_{\text{MAPE}}$ | $e_{\text{RMSE}}$ | $e_{\text{MAPE}}$ |
| KNN      | 2.108      | 6.36%      | 3.473      | 13.73%     |
| LS-SVM   | 2.342      | 8.35%      | 2.163      | 9.03%      |
| LSTM     | 0.788      | 4.29%      | 1.84       | 6.74%      |
| CNN-LSTM | 1.242      | 2.22%      | 1.851      | 5.963%     |
| LSTM-Attention | 0.749    | 2.43%      | 1.475      | 6.23%      |
| Proposed method | 0.739  | 1.91%      | 1.752      | 5.67%      |

Fig. 10 Loss function comparison results

© Springer
Experimental parameters

The proposed prediction model was trained using the error BP algorithm based on time expansion; namely, the neural network was expanded into a deep network in the time order, and then the expanded network was trained by the error BP algorithm (Li and Zhu 2021). To control the learning rate of the network and to prevent the problems of gradient disappearance and slow convergence speed, the Adam optimization algorithm (Tang et al. 2021) was used to update the network parameters, and the initial learning rate $lr$ was set to 0.001. The dropout regularization method was introduced before the LSTM layer to avoid overfitting, and the retention rate was set to 0.8. Through experimental observation, when the number of epochs exceeds 50, the loss function has almost converged and does not decline. Therefore, the training parameters of the neural network were set as follows: the maximum number of epochs was 50, and the small batch size was 512.

Results and discussion

Performance comparison of different prediction models

The proposed pollution emission prediction model was compared with the traditional prediction models including the K-nearest neighbor (KNN), least squares support vector machine (LS-SVM), LSTM, CNN-LSTM, and LSTM-Attention models. Their network parameter settings are consistent with the method proposed in this paper. The number of LSTM cyclic cores was 40, and the number of CNN convolution cores was 64. The optimal model parameters were selected in the other models according to the prediction effect. The prediction timestep was 1 h, and the continuous data of 40 previous hours were used to predict the NO$_X$ and SO$_2$ emission levels in the next hour. The results of the above-mentioned models on June 2, 2021 are shown in Fig. 8.

As shown in Fig. 8, the traditional model was not effective in predicting the emission of the thermal power plant. In the second figure in Fig. 8, 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 SO$_2$ emission, which had a weak regularity, the effect of the traditional 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 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 (Fig. 9) was further verified using the winter data as a verification set.

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. 9, 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 Fig. 7. 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 model are shown in Table 4, where the minimum error values in each column are boldfaced. The $e_{RMSE}$ 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 other models. The eMAPE of the proposed model was lower in predicting NOX than in predicting SO2, but both values were lower than the traditional models.

**Loss function change experiment**

To prove the superiority of the proposed prediction model over the traditional models, the proposed model was compared under the changes in the loss function (eMSE) during the training with the LSTM, CNN-LSTM, and LSTM-Attention models, and the comparison results are shown in Fig. 10. The proposed method converged to the minimum value of the loss function in the early stage of model training, after only several iteration rounds. Compared with the proposed model, although the convergence speed of LSTM with AM is slower, the curve is still smooth, while the single LSTM and CNN-LSTM models showed an upward trend of the loss function in the early stage of training, which proved that there was an overfitting phenomenon; the convergence speed of the follow-up model was also slow. The LSTM did not converge even after 40 rounds, and its training effect was poor. This is enough to prove that adding the AM to the time series prediction model can effectively mitigate the overfitting phenomenon and improve model training efficiency and final prediction effect.

**Attention layer weight visualization**

The weight of the attention layer of the power plant pollution emission prediction model was used to draw a thermodynamic diagram (Fig. 11). In the thermodynamic diagram, the brighter and darker parts indicate higher and lower attention of the AM during weight allocation, respectively. As shown in Fig. 11, the difference between brighter and darker parts increases in the areas close to the prediction time point for both NOX and SO2 data; thus, the data characteristics closer to the prediction point have a more obvious impact on it, which is consistent with common sense.

**Conclusions and future work**

To solve the problem of accurate prediction of CHP pollutant emission, this study proposes a CHP emission prediction model named the hybrid CNN-LSTM and AM. The hybrid CNN-LSTM and AM is trained with the NOX and SO2 data collected in summer and tested on the data collected in winter; the results are compared with the traditional prediction models. Based on the comparison results, the following conclusion can be drawn:

1) The hybrid CNN-LSTM and AM can accurately predict the future NOX and SO2 emission levels, and its eRMSE is less than 1 mg/m^3.
2) Compared with the traditional prediction models based on machine learning and deep learning, the proposed model has higher prediction accuracy and better robustness.
3) The introduction of attention mechanism and the selection of key variables by feature engineering improve the effect of the prediction model.

This study gives the measures of artificial intelligence application in clean energy usage in the power production industry. And in the follow-up work, the working condition regulation of power plant equipment will be studied according to the prediction results, so as to reduce the emission of pollutants.

**References**

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

Chang Y, Abimannan S, Chiao H, Lin C, Huang Y (2020) An ensemble learning based hybrid model and framework for air pollution forecasting. Environ Sci Pollut Res Int 27:1–14. https://doi.org/10.1007/s11356-020-09855-1

Chen H, Zhang J, Tao Y, Tan F (2019) Asymmetric GARCH type models for asymmetric volatility characteristics analysis and wind power forecasting. Protect Control Mod Power Syst 4:356–366. https://doi.org/10.1186/s41601-019-0146-0

**Author contribution** AW: writing—original draft. JY: methodology. TC: supervision. YJ: formal analysis. KL: writing—introduction. ZQ: writing—introduction.

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

**Data availability** Not applicable.

**Declarations**

**Ethics approval** Not applicable.

**Consent to participate** All authors have agreed to submit the manuscript.

**Consent for publication** Not applicable.

**Competing interests** The authors declare no competing interests.
Choi H, Cho K, Bengio Y (2018) Fine-grained attention mechanism for neural machine translation. Neurocomputing 284:171–176. https://doi.org/10.1016/j.neucom.2018.01.007

Dancey C-P, Reidy J (2017) Statistics without maths for psychology (7th edition), Prentice-Hall, Inc.

Ge Z, Bai X-Q, Wang Y-X (2021) Short-time multi-energy load forecasting method based on CNN-Seq2Seq model with attention mechanism. Mach Learn Appl 5:100064. https://doi.org/10.1016/j.mlawa.2021.100064

Gers FA, Schmidhuber J, Cummins F (2000) Learning to forget: continual prediction with LSTM. Neural Comput 12:2451–2471. https://doi.org/10.1162/089976600300015015

Ghaffarian S, Valente J, Tekinerdogan B (2021) Effect of attention mechanism in deep learning-based remote sensing image processing: a systematic literature review. Remote Sens 13:2965–2965. https://doi.org/10.3390/rs13152965

Ghosht-S, Mandyopadhyay S, Mitra D (2021) Nonlinear classification of emotion from EEG signal based on maximized mutual information. Expert Syst Appl 185:115605. https://doi.org/10.1016/j.eswa.2021.115605

Ghorbani Z, Behzadan A-H (2021) Monitoring offshore oil pollution using multi-class convolutional neural networks. Environ Pollut 289:117884. https://doi.org/10.1016/j.envpol.2021.117884

Guo H, Zhu K-P (2021) Attention-based dual-scale hierarchical LSTM for tool wear monitoring. Manuf Lett 29:99–103. https://doi.org/10.1016/j.mfglet.2021.07.015

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

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 US. Energy Policy 157:112485. https://doi.org/10.1016/J.EPEJ.2021.10.009

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

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.psee.202540

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.APENER.2017.07.114

Meng X (2021) Research on technological innovation effect of environmental regulation from perspective of industrial transfer: Evidence in China’s thermal power industry. Clean Eng Technol 4:100178. https://doi.org/10.1016/J.CLET.2021.100178

Ministry of ecology and environment, China (2004). Emission standard of air pollutants for thermal power plants. GB 13223–2003, China Environmental Science Press. http://www.mee.gov.cn/ywgz/fbbz/bz/bzwb/dqghbjb/dqghdwywwpbzu/200401/W0201 10127387986915138.pdf

Miyashita N, Yonezawa Y (2021) Mutual information analysis of the dynamic correlation between side chains in proteins. J Chem Phys 155:044107. https://doi.org/10.1063/5.0055662

Nakaishi T, Kagawa S, Takayabu H, Lin C (2021) Determinants of technical inefficiency in China’s coal-fired power plants and policy recommendations for CO2 mitigation. Environ Sci Pollut Res Int 28:52064–52081. https://doi.org/10.1007/s11356-021-14394-4

Pan J, Hou G, Wang J (2021) Research on environmental efficiency of power industry in china and its influencing factors: based on provincial panel data of 2008–2017. J Technol Econ 40:20–30

Peng H, Long F, Ding C (2005) Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27:1226–1238. https://doi.org/10.1109/TPAMI.2005.159

Peng D, Zhu Q, Che Q, Zhao H (2021) Short-term coal storage forecasting of thermal power plant for power grid dispatching based on CNN-LSTM neural network. Electr Power Autom Equip 47:127–132. https://doi.org/10.16081/epae.202102025

Tang S, Teng Z, Stil B, Hu Q, Pan X (2021) Improved BP neural network with ADAM optimizer and the application of dynamic weigh. Journal of Electronic Measurement and Instrumentation 35:127–135. https://doi.org/10.13382/j.emi.B2003558

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

Weng X, Qian Y, Peng D, Ma H (2020) NOx emission prediction model of SCR system based on deep bidirectional LSTM. J Eng Thermal Energy Power 35:57–64. https://doi.org/10.16146/j.cnki.rndgc.2020.10.009

Wei Z, Liu Y (2021) Construction of super-resolution model of remote sensing image based on deep convolutional neural network. Comput Commun 178:191–200. https://doi.org/10.1016/J.COMCOM.2021.06.022

Xiao C, Liao Y (2021) Transformer order demand forecasting based on grey forecasting model. IOP Conf Ser Earth Environ Sci 831:012004. https://doi.org/10.1088/1755-1315/831/1/012004

Yang G, Wang Y, Li X, Liu K (2020) Dynamic prediction of boiler NOx emission based on mutual information variable selection and LSTM. Journal of North China Electric Power University 47:66–74. CNKI:SUN:HBLD.0.2020–03–009

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

Zhang J (2016) Study on China’s strategic measures to achieve peak carbon dioxide emissions in 2030. Energy research and utilization: 18–21. https://doi.org/10.16404/j.cnki.ins1001-5523.2016.06.007

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

Zhang G, Yang J, Zheng Y, Wang Y, Wu Y, Chen S (2021a) Hybrid-attention guided network with multiple resolution features for person re-identification. Inf Sci 578:525–538. https://doi.org/10.1016/J.INFSCI.2021.07.058

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

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

Zou Q, Xiong Q, Li Q, Yi H, Yu Y, Wu C (2020) A water quality prediction method based on the multi-time scale bidirectional long short-term memory network. Environ Sci Pollut Res Int 27:16853–16864. https://doi.org/10.1007/s11356-020-08087-7

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