Received: 7 May 2025 | Revised: 5 July 2025 | Accepted: 8 July 2025

DOI: 10.1002/cjce.70054

RESEARCH ARTICLE

THE CANADIAN JOURNAL OF CHEMICAL ENGINEERING

# Dynamic NOx emission prediction in coal-fired power plants based on joint multi-head attention CNN-GRU hybrid model

**Jianghong Chen**[1] | **Zhimin Lu**[1,2,3] | **Zhenghui Li**[2,3] | **Wenjing Li**[1] | **Anli Zhou**[1] | **Wenbo Pi**[1] | **Youxing Wei**[1] | **Shunchun Yao**[1,2,3]

[1]School of Electric Power, South China University of Technology, Guangzhou, China

[2]Guangdong Province Key Laboratory of Efficient and Clean Energy Utilization, Guangzhou, China

[3]Guangdong Engineering Technology Center for High-Efficiency and Low-Pollution Conversion of Energy, Guangzhou, China

**Correspondence**
Zhimin Lu and Shunchun Yao, School of Electric Power, South China University of Technology, Guangzhou, China.
Email: zhmlu@scut.edu.cn; epscyao@scut.edu.cn

**Funding information**
National Key Research and Development Program of China, Grant/Award Numbers: 2024YFC3909002, 2024YFC3909004-02; Guangdong Key Laboratory of Efficient and Clean Energy Utilization, South China University of Technology, Grant/Award Number: 2013A061401005; Guangzhou Science and Technology Elite, Grant/Award Number: 2024A04J4486

**Abstract**

The flexible operation of coal-fired power plants under deep peak-shaving conditions imposes significant challenges on accurate NOx prediction for SCR systems. Given the inherent complexities of the SCR denitrification process, characterized by dynamic nonlinearity, temporal variability, and multivariable coupling in nitrogen oxide emissions, this study proposes a convolutional neural network-gated recurrent unit (CNN-GRU) hybrid model integrated with multi-head attention (MA) mechanisms to address these system-specific characteristics for precise NOx prediction. The model combines the local feature extraction capability of CNNs, the long-term temporal dependency modelling strength of GRUs, and the adaptive feature weighting functionality of MA mechanisms, achieving dynamic weight allocation across feature channels and temporal scales to enhance robustness and feature representation. Furthermore, a sparrow search algorithm (SSA) is introduced to optimize model parameters adaptively, improving prediction accuracy and generalization performance. Experimental validation using real operational data demonstrates the model's superior performance, with mean absolute error (MAE) below 0.5 mg/m$^3$ and mean absolute percentage error (MAPE) below 2%. Ablation experiments confirm the effectiveness of the proposed architecture, showing over 28% prediction accuracy improvement compared to Transformer-based models while maintaining enhanced generalization capability.

**KEYWORDS**

CNN-GRU, coal-fired power plants, multi-head attention, NOx, selective catalytic reduction (SCR), sparrow search algorithm

## 1 | INTRODUCTION

Driven by China's "dual-carbon" policy objectives and the escalating integration of wind and solar power generation,[1,2] coal-fired power units with long-term flexible regulation capabilities have become essential for deep peak-shaving operations in modern power systems, thereby enhancing renewable energy accommodation.[3,4] Virtually all Chinese coal-fired plants utilize selective catalytic reduction (SCR) systems for NOx control, where injected ammonia converts NOx to nitrogen and water vapour. Under ultra-low emission standards, non-steady

  

operations cause excess ammonia injection, leading to secondary pollution, equipment corrosion, and operational issues, including air preheater blockage and induced draft fan damage.[5] Deep peak-shaving further intensifies NOx fluctuations in SCR flue gas, aggravating ammonia slip.[6,7] This necessitates real-time outlet NOx prediction for optimal ammonia dosing control.[8] Consequently, driven by the dual-carbon policy's requirements for flexible operation of coal-fired units, SCR system outlet NOx emission modelling has emerged as a critical research priority to reconcile deep peak-shaving demands with stringent emission constraints.

SCR outlet NOx emission modelling methodologies can be categorized into mechanism-driven and data-driven approaches. Mechanism-based methods face challenges in NOx online prediction, including unstable model accuracy and time-consuming modelling processes,[9,10] due to the complexity of SCR denitrification reaction mechanisms, which involve numerous influencing factors. To develop SCR models suitable for practical power plant applications, data-driven modelling approaches have been extensively studied owing to their powerful nonlinear modelling capacity and proficiency in learning complex feature relationships.

In data-driven methodologies, shallow learning and deep learning represent two predominant paradigms. Shallow machine learning relies on manual feature engineering. For instance, Yang et al.[11] employed mutual information (MI) for variable selection with least squares support vector machine (LSSVM) modelling, while Tang et al.[12] integrated principal component analysis (PCA) and extreme learning machine (ELM) for dynamic NOx prediction. Though achieving accuracy in trained scenarios, these models exhibit limited feature representation capacity and performance degradation under novel operating conditions. Lv et al.[13,14] mitigated the performance deterioration of shallow machine learning models by constructing representative operational databases to expand training sample coverage, though this approach requires additional data labeling and storage infrastructure. Collectively, shallow learning-based methods inadequately capture the multivariate coupled nonlinearities and temporal dependencies inherent in SCR systems, resulting in insufficient dynamic adaptability.

In contrast, deep learning employs multilayer nonlinear transformations to automatically extract features and model complex temporal dependencies, with its architectures providing novel pathways toward model generalization enhancement. Mohammadi et al.[15] developed an LSTM-based temporal NOx prediction model demonstrating stable performance across wide-load testing scenarios. Capitalizing on the inherent capability of (convolutional neural networks) CNNs to autonomously identify the multivariable coupling, Yin et al.[16] integrated CNN-LSTM architectures for SCR system modelling, achieving a 15.1% reduction in NOx prediction error compared to standalone LSTM implementations. Li et al.[17] incorporated a sparse-constrained stacked autoencoder (SAE) network as a feature extraction layer, combined with bidirectional LSTM, enhancing model accuracy in predicting NOx emissions under global operating conditions. However, while LSTM and CNN can improve the temporal feature expression of the model, they usually suffer from low efficiency in model training. The gated recurrent unit (GRU), a streamlined variant of LSTM, retains the capability to model long-term temporal dependencies while exhibiting structural simplicity and accelerated convergence—characteristics that have driven its progressive adoption in NOx prediction research.[18] Consequently, our study employs a CNN-GRU hybrid architecture for SCR system dynamic modelling, specifically engineered to simultaneously enhance global NOx prediction accuracy and computational efficiency.

Recent studies have also emphasized the importance of transient prediction capabilities in NOx emission control systems.[19] While temporal models (e.g., LSTM, CNN-GRU) achieve satisfactory global accuracy, their efficacy in capturing transient operational dynamics (critical for real-time NOx control) remains constrained by rigid temporal feature weighting mechanisms. This limitation originates from the temporal-scale dependency of feature saliency,[20] where load transient characteristics exhibit higher predictive relevance compared to steady-state operational features.[21] Empirical analyses reveal that abrupt load variations induce shifts in feature saliency, challenging existing models to dynamically prioritize mission-critical temporal dependencies. Consequently, multi-model ensemble approaches[22,23] have been developed. These methods capture feature saliency variations across system operational modes by establishing regime-specific local models. However, this approach significantly increases model lifecycle management costs. In contrast, the multi-head attention mechanism provides deep learning solutions.[24,25] Through adaptive weighting of critical temporal features, this mechanism enables the preservation of essential temporal information while preventing feature degradation. We integrate this mechanism into the NOx emission prediction framework, replacing conventional multi-model ensemble approaches. This machine learning paradigm achieves the dual optimization of the prediction robustness in mission-critical operational phases and the model management cost.

In summary, existing NOx emission prediction models face the following limitations: (1) shallow

machine learning architectures exhibit limited feature representation capability and insufficient generalization ability, while deep learning models suffer from poor efficiency in characterizing complex temporal dependencies; (2) current modelling frameworks fail to dynamically track feature saliency variations during operational transients such as power plant deep peak-shaving operations, resulting in severe performance degradation under abrupt load-changing scenarios. To address these challenges, this study proposes a joint multi-head attention CNN-GRU hybrid architecture. The principal contributions are threefold:

1. The hybrid CNN-GRU framework synergistically combines the local feature extraction capability of CNNs with the temporal modelling superiority of GRUs. This integrated architecture enables the dynamic characterization of SCR systems with nonlinear multivariate couplings.
2. Integration of a multi-head attention mechanism into the CNN-GRU framework enables adaptive weight adjustment across feature channels and temporal scales. The design dynamically adapts to feature saliency variations, significantly enhancing model responsiveness during operational transients.
3. An adaptive SSA-based parameter optimization strategy was developed for the proposed model. The joint optimization of architectural parameters and NOx-related variable temporal window sizes establishes structural-dataset synergy, thereby maximizing model performance.

Finally, the proposed model is evaluated and validated using historical operational datasets from a 330 MW subcritical coal-fired unit.

## 2 | OBJECT DESCRIPTION AND DATA PREPROCESSING

This study investigates an SCR flue gas denitrification system installed on a 330 MW subcritical coal-fired unit in Guangzhou. The SCR system adopts a parallel configuration with independent reactor chambers (A/B sides) positioned between the boiler economizer and air preheater, utilizing a dual-layer vanadium-based catalyst. Through reaction mechanism analysis, 10 variables influencing the outlet NOx concentration were preliminarily selected as model inputs. Given the operational symmetry between A/B-side reactors, this research focuses exclusively on Side A. Operational data were collected at 60-s intervals from the distributed control system (DCS), comprising a 10,000 sample dataset covering load ranges of 143 to 302 MW. Variable specifications and units are detailed in Table 1.

To mitigate inherent measurement noise in industrial environments, a Savitzky–Golay filter was systematically applied to all selected variables, with the denoising effects on NOx measurements illustrated in Figure 1A. The continuous emissions monitoring system (CEMS) occasionally generates missing/blank data during calibration/maintenance periods, manifesting as persistent constant values. Leveraging the high temporal correlation between inlet NOx concentration trends of Reactors A and B, a data imputation method was implemented to address the missing values. Specifically, during CEMS maintenance operations on Reactor A, the methodology utilizes the temporal gradient of Reactor B's inlet NOx concentration to calibrate Reactor A's missing data through the following formulation:

$$y(t) = y(t-1) + \frac{dy'(t)}{dt} * \nabla t \tag{1}$$

**TABLE 1** Input variables of the model.

| Variables | | Unit | Value range |
|---|---|---|---|
| SCR inlet NOx concentration | $C_{in}$ ($NO_x$) | $mg/m^3$ | 79.30–184.33 |
| SCR outlet NOx concentration | $C_{out}$ ($NO_x$) | $mg/m^3$ | 5.61–58.77 |
| Ammonia injection flow rate | $Q_v$ ($NH_3$) | $m^3/h$ | 1.31–32.95 |
| Ammonia slip | $C_{out}$ ($NH_3$) | ppm | 0.10–1.65 |
| SCR outlet $O_2$ content | $\eta_{in}$ ($O_2$) | % | 3.27–9.36 |
| SCR inlet temperature | $T$ | °C | 316.85–373.66 |
| SCR reactor pressure | $\Delta P$ | Pa | 136.54–401.71 |
| Unit load | $W$ | MW | 143.98–302.13 |
| Total coal feed rate | $Q_m$ (coal) | t/h | 64.93–133.34 |
| Total air flow rate | $Q_v$ (air) | $m^3/h$ | 1002.01–1826.41 |

Abbreviation: SCR, selective catalytic reduction.

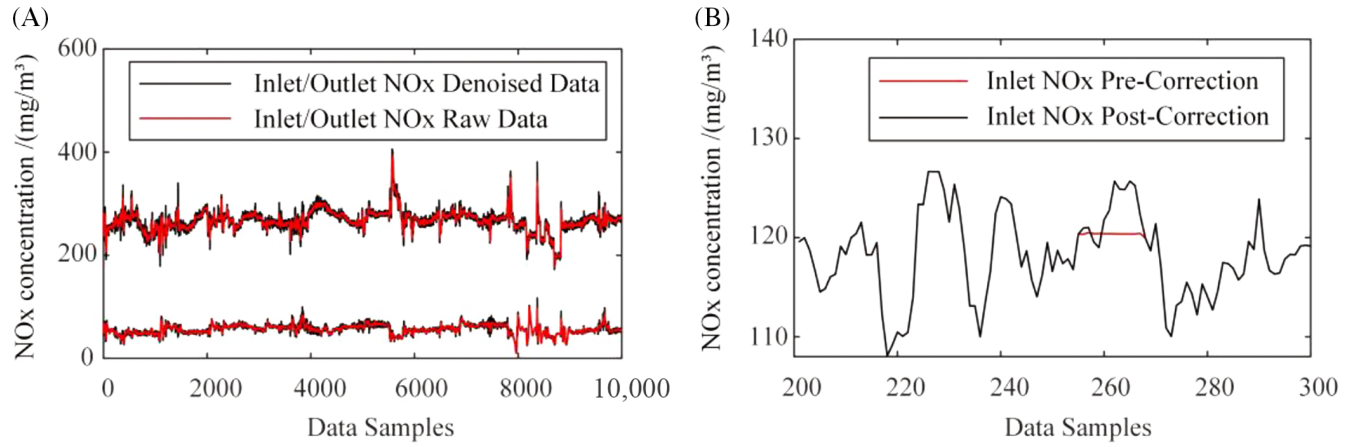


**FIGURE 1** (A) Savitzky–Golay filter noise reduction; (B) Correction processing for selective catalytic reduction (SCR) inlet and outlet variable NOx.

where $y(t)$ is the imputed NOx concentration at Reactor A (mg/m$^3$) and $y'(t)$ is the measured NOx concentration at Reactor B (mg/m$^3$). As shown in Figure 1B, the corrected results of inlet NOx are presented.

## 3 | METHODOLOGY

### 3.1 | CNN mechanism

Convolution fundamentally represents a weighted summation operation on input data. As a feedforward neural network architecture based on convolutional operations, CNN was initially developed for image recognition and has subsequently demonstrated significant potential in temporal feature extraction. The CNN comprises convolutional layers, pooling layers, and fully connected layers. Through convolutional kernels, the convolution operation extracts distinctive features from the input data to generate feature maps—comprehensive representations of the localized features that provide foundational insights for subsequent analytical processing. This operation can be mathematically expressed as follows:

$$W_{k,l} \otimes x_{i,j} = \sum_{m=0}^{\alpha-1} \sum_{n=0}^{\beta-1} w_{m,n} \times x_{i+m,j+n} \tag{2}$$

$$P = f(W_{k,l} \otimes x + b_k) \tag{3}$$

where $W_{k,l}$ is the $k \times l$ convolution kernel, $x_{i,j}$ is the input feature map pixel value, $\otimes$ is the convolution operation, $\alpha$ and $\beta$ are the kernel dimensions, $w_{m,n}$ is the positional weight of the kernel, $P$ is the output feature map, $b_k$ is the bias term, and $f(\cdot)$ is the ReLU activation function.

### 3.2 | GRU mechanism

The GRU, a streamlined variant of LSTM networks, employs gating mechanisms to selectively filter and dynamically regulate information flow, enabling more efficient sequential data modelling. As depicted in Figure 2, its core innovation lies in two gating units: the update gate $r_t$ and the reset gate $z_t$, along with direct utilization of the hidden state $h_t$ as integrated memory storage, eliminating explicit separate memory cells. This architectural simplification, relative to LSTM, enhances computational efficiency. The GRU operational workflow is mathematically expressed as follows:

$$r_t = \sigma(U_r \cdot [h_{t-1}, p_t] + b_r) \tag{4}$$

$$z_t = \sigma(U_z \cdot [h_{t-1}, p_t] + b_z) \tag{5}$$

$$\tilde{h}_t = \tanh(U_h \cdot [r_t \cdot h_{t-1}, p_t] + b_h) \tag{6}$$

$$h_t = z_t \cdot h_{t-1} + (1 - z_t) \cdot \tilde{h}_t \tag{7}$$

where σ is the sigmoid activation function; $U_r$, $U_z$, and $U_h$ are the weight matrices, respectively; $b_r$, $b_z$, and $b_h$ are the bias terms of the reset gate, update gate, and candidate hidden state, respectively; $p_t$ is the input vector at timestep $t$; and $\tilde{h}_t$ is the candidate hidden state integrating current input $p_t$ and previous hidden state $h_{t-1}$.

### 3.3 | MA mechanism

Effective identification of the most importance feature variations constitutes a critical challenge in sequential modelling tasks. The attention mechanism addresses this

challenge by dynamically weighting inter-position correlations across the sequences through query ($Q$), key ($K$), and value ($V$) computations. Its mathematical formalization via scaled dot-product operations is expressed as follows:

$$\text{Attention}(Q,K,V)=\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{8}$$

where $d_k$ denotes the dimensionality of key vectors, ensuring numerical stability. The architecture is illustrated in Figure 3A.

The MA mechanism implements concurrent computation of multiple independent attention heads. This design enables the model to learn manifold relational patterns across the distinct subspaces, as illustrated in Figure 3B. The mathematical formalization is expressed as follows:

$$\text{MultiHead}(Q,K,V)=\text{Concat}(\text{head}_1,\text{head}_2,...,\text{head}_h)W^O \tag{9}$$

$$\text{head}_i=\text{Attention}\left(QW_i^Q,KW_i^K,VW_i^V\right) \tag{10}$$

where $W_i^Q$, $W_i^K$, and $W_i^V$ are the projection matrices for the $i$-th attention head, and $W^O$ is the linear transformation matrix for the final output.

## 3.4 | SSA optimization

The SSA is a swarm intelligence optimization method that seeks equilibrium between global exploration and local exploitation by simulating sparrow foraging behaviours.[26] The algorithm comprises three agent types: producers (global search), scroungers (local search), and

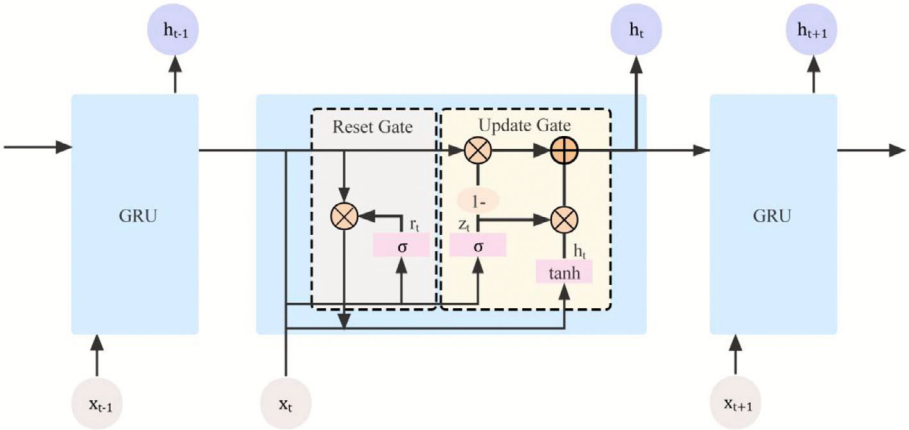


**FIGURE 2** Gated recurrent unit (GRU) network structure diagram.

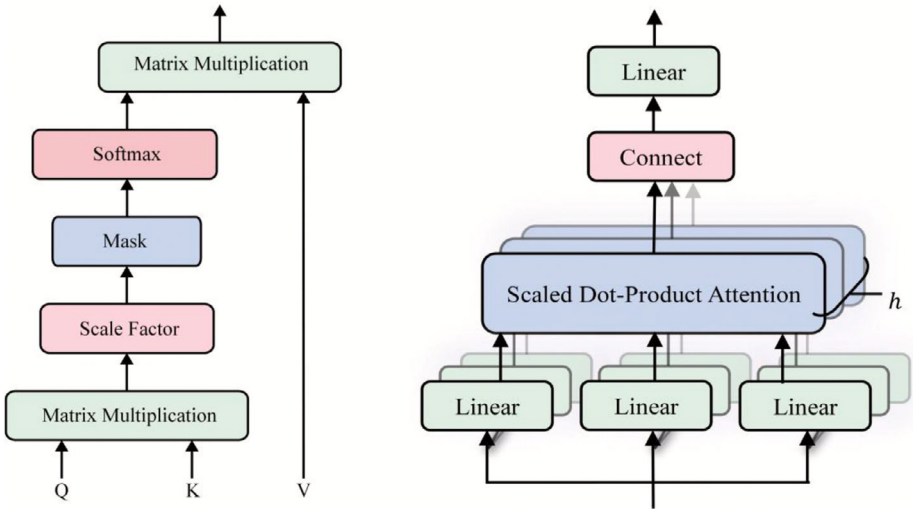


**FIGURE 3** Structure of the multiple attention mechanism: (A) Scaled dot-product attention; (B) Multi-head attention.

vigilantes (perturbation mechanism). Key operational formulae include the following:

1. Producer position update rule:

$$z_i^{t+1} = \begin{cases} z_i^t \cdot \exp\left(\frac{-i}{\alpha T}\right) & R < \mathrm{P} \\ z_i^t + QL & R \geq \mathrm{P} \end{cases} \tag{11}$$

where $z_i^t$ is the position vector of the $i$-th sparrow at generation $t$, $\alpha$ is a uniformly distributed random number in (0,1] for exponential decay control, $T$ is the maximum iteration count, $Q$ is a random scalar sampled from the standard normal distribution N (0,1), $\boldsymbol{L}$ is the Lévy flight directional vector for stochastic exploration. $R$, $P$ are the alert threshold and safety threshold, respectively.

2. Integrated academic formulation:

$$z_i^{t+1} = \begin{cases} \mathbf{z}_{\text{best}}^t + \boldsymbol{S}\left(z_i^t - \mathbf{z}_{\text{best}}^t\right) & f_i \neq f_{\text{g}} \\ z_i^t + \boldsymbol{K}\left(\frac{z_i^t - \mathbf{z}_{\text{worst}}^t}{|f_i - f_{\text{w}}| + e}\right) & f_i = f_{\text{g}} \end{cases} \tag{12}$$

where $\boldsymbol{S}$ is a normally distributed random scalar $N(0,1)$, $\boldsymbol{K}$ is a step-size control parameter uniformly sampled from $[-1,1]$, $e$ is an infinitesimal constant to prevent division by zero, $f_i$ denotes the fitness value of the $i$-th sparrow, and $f_g$ and $f_w$ represent the global best and worst fitness values in the current population, respectively.

## 3.5 | Gapped blocking cross-validation

To address the limitations of classical K-fold cross-validation in time-series forecasting, gapped blocking cross-validation (GB-CV) partitions data sequentially into K contiguous temporal blocks, eliminating future information leakage caused by random shuffling.[27] As illustrated in Figure 4, this method iteratively designates one block as the test set while using the remaining blocks for training, ensuring comprehensive evaluation across temporal phases and mitigating overfitting from spurious temporal patterns. Residual information leakage may persist due to the statistical dependency between training and test sets, degrading generalization fidelity. To resolve this issue, the gap mechanism is introduced by reserving the intermediate blocks between the training and test sets. This enforced temporal isolation enhances the dataset independence and strengthens the predictive robustness for future observations.[28]

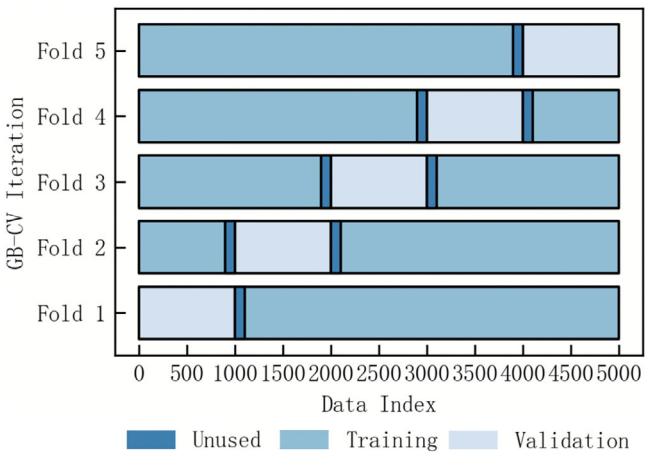


FIGURE 4 Cross-validation data block division method.

## 3.6 | Comprehensive evaluation

To holistically evaluate the performance of time-series forecasting models, we establish a comprehensive evaluation framework combining performance metrics and computational efficiency indicators. The performance evaluation employs the following metrics: mean absolute error (MAE), mean absolute percentage error (MAPE), mean squared error (MSE), root mean squared error (RMSE), coefficient of determination ($R^2$), and time cost (TC) for computational efficiency. The mathematical definitions are as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{13}$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \tag{14}$$

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{15}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{16}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{17}$$

$$TC \in R^+ \,(\text{training duration in seconds}) \tag{18}$$

where $n$ denotes the total number of samples, $y_i$ is the truth value of the $i$-th sample, and $\hat{y}_i$ represents the predicted value of the $i$-th sample.

## 4 | CNN-GRU-MA MODEL

### 4.1 | Forecasting model architecture

The proposed CNN-GRU-MA model for predicting the flue gas NOx concentration at the SCR outlet is illustrated in Figure 5. The overall architecture consists of an input layer, CNN layer, GRU-MA, and output layer. Multivariate time-series data are first restructured temporally through the input layer. The CNN layer extracts localized spatiotemporal features, followed by the GRU layer capturing global temporal dependencies. The MA layer subsequently applies adaptive weighting to the sequence features generated by the GRU, enhancing the recognition of critical temporal patterns. Finally, the weighted features are processed through a fully connected layer to produce prediction results.

1. Input layer. The input layer performs temporal restructuring on the selected variables X using a predefined sequence window length L. The restructured data $X^{(t)}_{\text{input}}$ is then transformed into a 4D tensor through a temporal folding operation to meet the input requirements of the CNN layers.

$$\boldsymbol{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,N} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1} & x_{m,2} & \cdots & x_{m,N} \end{bmatrix} \tag{19}$$

$$\boldsymbol{X}^{(t)}_{\text{input}} = \begin{bmatrix} x_{t,1} & x_{t,2} & \cdots & x_{t,N} \\ x_{t+1,1} & x_{t+1,2} & \cdots & x_{t+1,N} \\ \vdots & \vdots & \ddots & \vdots \\ x_{t+L-1,1} & x_{t+L-1,2} & \cdots & x_{t+L-1,N} \end{bmatrix} \tag{20}$$

The original data matrix $X \in R^{m \times N}$, where $m$ denotes the total number of samples, $N$ represents the feature dimensionality, and $t \in [1, m - L]$ is the index of temporally sliced samples.

2. CNN layer: The CNN layer performs robust feature extraction on each input $\boldsymbol{X}^{(t)}_{\text{input}}$ through a hierarchical architecture comprising two convolutional modules. Each module progressively extracts temporal features from low-level patterns (short-term correlations) to high-level abstractions (long-range dependencies) by 2D convolution followed by a maximum pooling layer. Detailed configurations of the convolutional kernel dimensions, stride

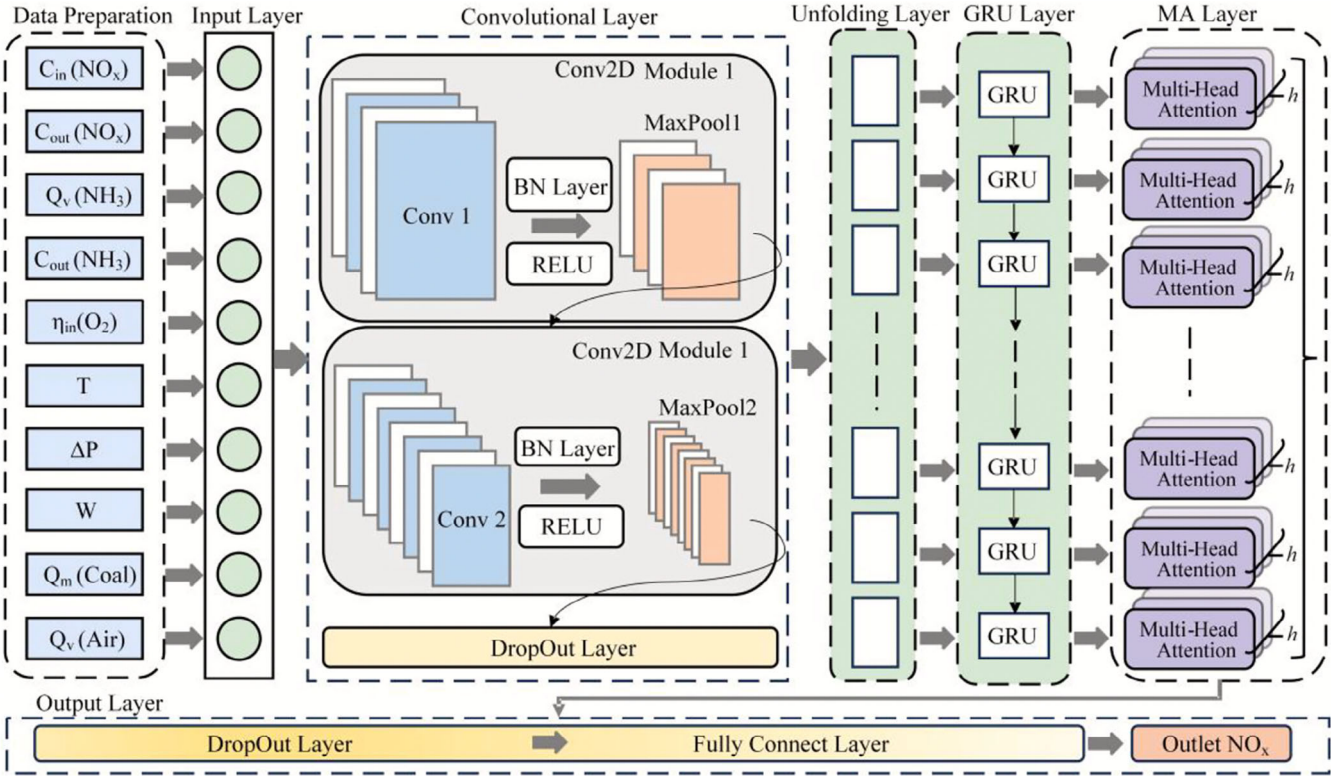


FIGURE 5 Architecture of multivariate time series forecasting model.

parameters, and pooling window sizes are provided in Table 2.

3. GRU-MA layer: The GRU-MA layer enhances the standard GRU architecture by integrating a multi-head attention mechanism. This hybrid design processes the feature vectors extracted by the CNN layer to capture long-term dependencies in power plant operational data. The multi-head attention module enables simultaneous focus on distinct feature subspaces, improving the model's capacity to interpret global temporal patterns through parallelized attention heads. The structural complexity of this layer necessitates systematic parameter optimization, with detailed hyperparameter configurations discussed in Section 5.
4. Output layer: The $N$-th feature is the target scalar, restructured via a sliding window to form $\boldsymbol{Y}$. The output from the GRU-MA layer passes through a memory forgetting layer and is then mapped to the target variable via a fully connected layer, generating the predicted value $\widehat{\boldsymbol{Y}}$.

$$\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{X}_{\boldsymbol{N}}^{(\boldsymbol{L}+\boldsymbol{1})} \\ \boldsymbol{X}_{\boldsymbol{N}}^{(\boldsymbol{L}+\boldsymbol{2})} \\ \vdots \\ \boldsymbol{X}_{\boldsymbol{N}}^{(\boldsymbol{m})} \end{bmatrix} \tag{21}$$

where each row of $\boldsymbol{Y}$ corresponds to the NOx outlet concentration at timestep $t+1$ for the input sequence $\boldsymbol{X}_{\mathbf{input}}^{(\boldsymbol{t})}$.

## 4.2 | SSA optimized computational framework

This study employs the sparrow search algorithm (SSA) to optimize seven critical hyperparameters of the CNN-GRU-MA model for enhanced NOx prediction performance, including learning rate, GRU units, attention heads, key dimensions, regularization parameters, forget rate, and window size (Table 3). The framework incorporates the window size into the optimization process mainly because of the complexity of the SCR process with high latency and high dynamics, and the need to consider the impact of the time series window length on the model's prediction capability and accuracy with respect to the sequential data. The computational architecture operates through three synergistic modules:

1. The SSA optimization module configures the sparrow search space $Z$ with initialized parameters (sparrow population size $N = 30$, iteration count $T = 5$). Through iterative updates of the objective function value $f(z) = \mathrm{MSE}_{\mathrm{train}}(z)$ and optimal position tracking, the algorithm dynamically adjusts the spatial positions of sparrow search agents. This mechanism minimizes the training loss to identify optimal hyperparameters for the neural network.
2. The data processing module executes preliminary feature selection and temporal window restructuring based on DCS data. After preprocessing, the dataset is partitioned into training and test sets at a 7:3 ratio, where the training set iteratively optimizes model weight parameters, and the test set evaluates the final model performance.
3. The model computation module incorporates the optimized parameters into the CNN-GRU-MA neural network, performs iterative computations to generate the objective value $f(z)$ and outputs the optimal model through comprehensive evaluation.

The framework achieves synergistic convergence of parameter optimization and model training through a bidirectional looping mechanism, and the termination condition under the constraint of the maximum number of iterations ensures computational efficiency. The overall computational flowchart is shown in Figure 6.

TABLE 2 Convolutional neural network (CNN) structure and parameters.

| Layer | Output channels | Kernel size | Stride | Padding |
|---|---|---|---|---|
| Conv1 | 8 | (4, 1) | (1, 1) | Valid |
| MaxPool1 | - | (2, 1) | (2, 1) | Same |
| Conv2 | 16 | (4, 1) | (1, 1) | Valid |
| MaxPool2 | - | (2, 1) | (2, 1) | Same |

# 5 | MODEL VALIDATION AND RESULT ANALYSIS

## 5.1 | Hyperparameter optimization

The model was trained on the first 5000 operational samples from the SCR system dataset, which includes multiple

TABLE 3 Hyperparameters and ranges.

| Hyperparameter | Data type | Range |
|---|---|---|
| Learning rate | Float (Continuous) | [0.001, 0.01] |
| GRU units | Integer (Discrete) | [8, 100] |
| Attention heads | Integer (Discrete) | [2, 16] |
| Key dimension | Integer (Discrete) | [4, 32] |
| Regularization | Float (Continuous) | [0.0001, 0.001] |
| Forget rate | Float (Continuous) | [0.0, 0.4] |
| Window size | Integer (Discrete) | [5, 30] |

rapid load-variation scenarios to rigorously evaluate its adaptability to dynamic industrial conditions. The CNN-GRU-MA time series model is used to predict the NOx emission, and the hyperparameters obtained after optimization by SSA algorithm are shown in Table 4, with the model input data structure of 23 × 10, the number of GRU hidden units of 24 × 1, the structure of multi-attention layer of 8 × 5, and the structure of output results of 1 × 1.

To validate the effectiveness of SSA and objectively evaluate model performance, this study compared the baseline CNN-GRU-MA architecture with its PSO-optimized and SSA-optimized counterparts under continuous wide-load operation scenarios (>80 h). Both PSO and SSA were configured with identical iteration counts for fair comparison. The prediction results and magnified partial views of the three models are shown in Figure 7.

The results demonstrated the following: (1) All CNN-GRU-MA-based models effectively captured the overall variation trends of NOx emissions; (2) In intervals with significant NOx fluctuations, the SSA- and PSO-optimized models exhibited superior prediction accuracy and stability across both training and test sets, whereas the unoptimized CNN-GRU-MA baseline showed relatively weaker performance in the training set and more pronounced local deviations in the test set.

Table 5 provides a quantitative comparison of evaluation metrics across all models on both training and test sets, systematically validating the architectural validity of the SSA-optimized model. Key conclusions are as follows:

**TABLE 4** Determination of hyperparameter values.

| Hyperparameter | Value |
|---|---|
| Learning rate | 0.0074 |
| GRU units | 24 |
| Attention heads | 8 |
| Key dimension | 5 |
| Regularization | 0.0002 |
| Forget rate | 0.2154 |
| Window size | 23 |

Abbreviation: GRU, gated recurrent unit.

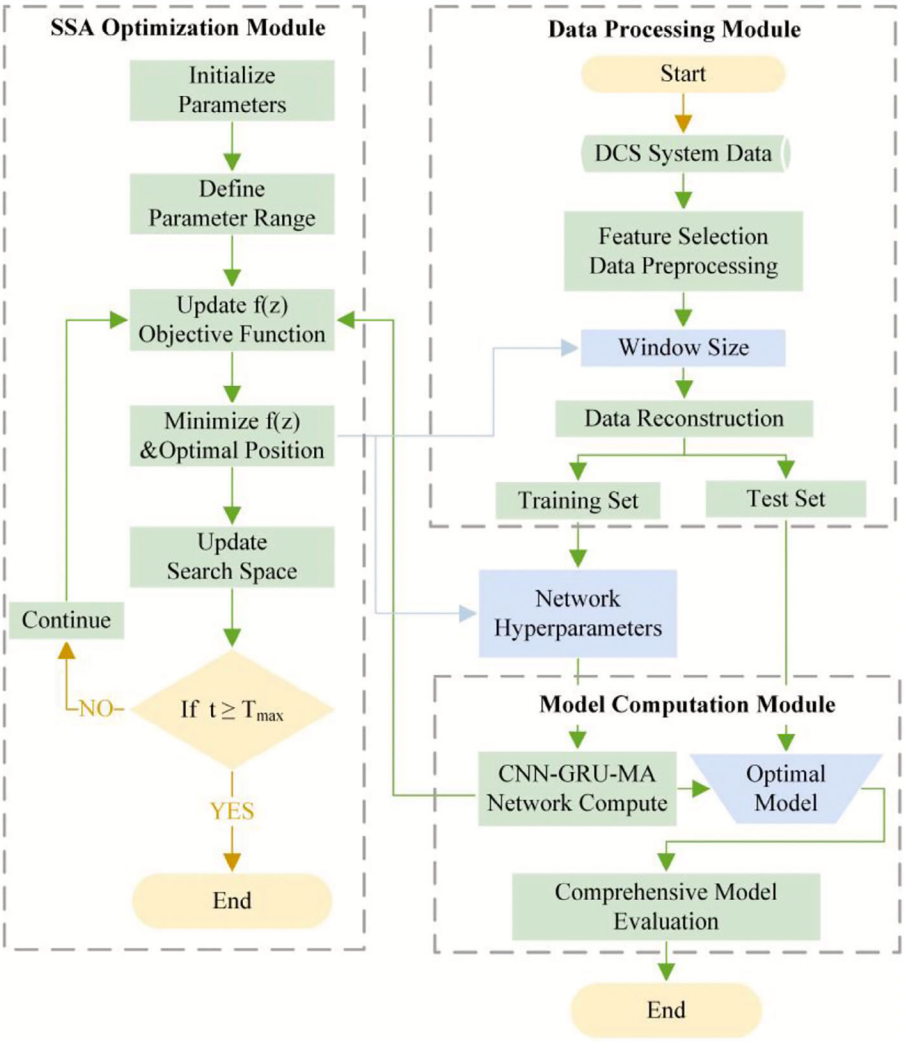


**FIGURE 6** Computational framework of sparrow search algorithm (SSA) optimized model.

### 5.1.1 | Generalization performance analysis

The baseline model exhibited strong performance on the training set but suffered significant performance degradation on the test set, with relative increases in MAE and MAPE reaching 160.19% and 140.01%, respectively, indicating notable generalization deficiencies. Optimization algorithms substantially mitigated this issue: after PSO optimization, the performance gaps between training and test sets were reduced to 91.43% (MAE) and 77.37% (MAPE), while SSA optimization further narrowed these gaps to 33.07% (MAE) and 20.47% (MAPE), demonstrating superior generalization robustness.

### 5.1.2 | Predictive performance comparison

The proposed SSA-CNN-GRU-MA model achieved the lowest prediction errors across both training and test sets (MAPE: 1.53% and 1.87%, respectively). Compared to the

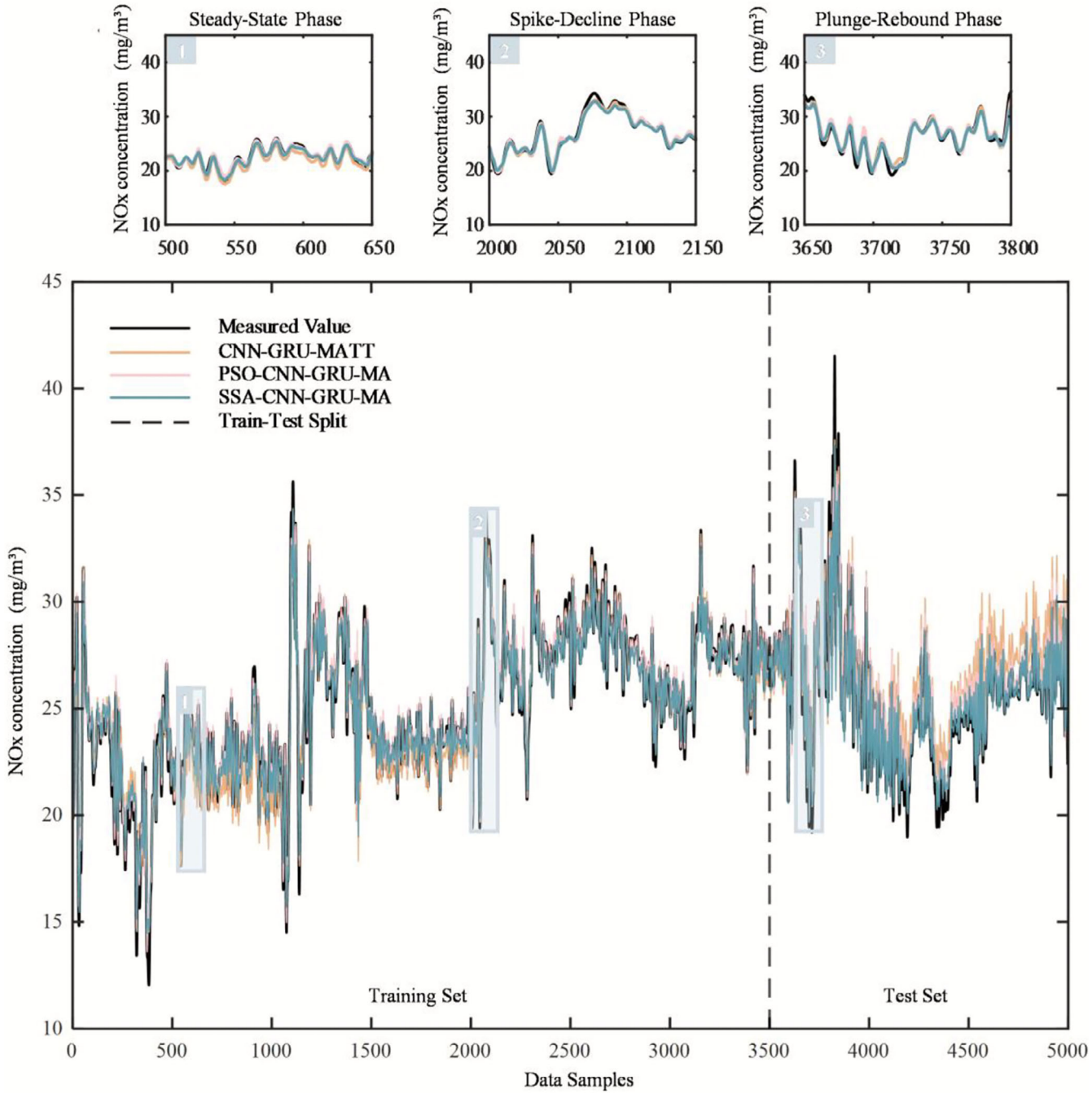


**FIGURE 7** Training-test exit NOx prediction curve of for different models (SSA-CNN-GRU-MA, PSO-CNN-GRU-MA, and CNN-GRU-MA). CNN, convolutional neural network; GRU, gated recurrent unit; MA, multi-head attention; PSO, particle swarm optimization; SSA, sparrow search algorithm.

**TABLE 5** Comprehensive evaluation indicators for different models (SSA-CNN-GRU-MA, PSO-CNN-GRU-MA, and CNN-GRU-MA).

| | SSA-CNN-GRU-MA | | PSO-CNN-GRU-MA | | CNN-GRU-MA | |
|---|---|---|---|---|---|---|
| | **Train** | **Test** | **Train** | **Test** | **Train** | **Test** |
| MAE ($mg/m^3$) | 0.37 | **0.49** | 0.38 | 0.73 | 0.43 | 1.12 |
| Relative gap | 33.07% | | 91.43% | | 160.19% | |
| MAPE (%) | **1.57%** | **1.89%** | 1.64% | 2.92% | 1.88% | 4.52% |
| Relative gap | **20.47%** | | 77.37% | | 140.01% | |
| MSE ($mg/m^3$) | 0.26 | 0.58 | 0.28 | 0.89 | 0.39 | 1.64 |
| Relative gap | 123.37% | | 218.93% | | 311.53% | |
| RMSE ($mg/m^3$) | 0.51 | 0.76 | 0.53 | 0.94 | 0.63 | 1.28 |
| Relative gap | 49.31% | | 78.30% | | 102.70% | |
| $R^2$ | 0.97 | 0.94 | 0.97 | 0.91 | 0.96 | 0.86 |
| Relative gap | 3.78% | | 6.76% | | 10.66% | |
| TC | **SSA < 300 s Train 18.53 s** | | OPS < 420 s Train 18.15 s | | Train 16.86 s | |

Abbreviations: CNN, convolutional neural network; GRU, gated recurrent unit; MA, multi-head attention; MAE, mean absolute error; MAPE, mean absolute percentage error; PSO, particle swarm optimization; RMSE, root mean squared error; SSA, sparrow search algorithm.

baseline and PSO-optimized models, the training set MAPE decreased by 16.61% and 4.44%, while the test set MAPE decreased by 58.14% and 35.10%, respectively.

#### 5.1.3 | Efficiency comparison

All three models showed comparable training time requirements (16–19 s), but significant differences in hyperparameter optimization efficiency: SSA completed global optimization 120 s faster than PSO, improving time efficiency by 86%.

In summary, the proposed SSA-CNN-GRU-MA model demonstrated optimal performance across both training and test sets, confirming that hyperparameter optimization effectively enhances model performance. SSA's superior search strategy significantly improved optimization efficiency while maintaining accuracy, outperforming traditional PSO in comprehensive performance.

### 5.2 | Model validity verification

To rigorously validate the effectiveness of the proposed model architecture, ablation experiments were conducted by systematically deconstructing the baseline model for component-level comparative analysis. The GB-CV method was implemented to evaluate each model adaptability to temporal data drift and generalization capability over extended operational periods. Utilizing 5000 newly collected power plant operational samples, we configured the training–test set block interval to 100 timesteps and performed five validation rounds to calculate the average metrics. As shown in Figure 8 and Table 6, the key findings are as follows:

#### 5.2.1 | Baseline model comparative analysis

The standalone CNN architecture demonstrated competent NOx emission trend capture under steady-state operating conditions but showed transient prediction inaccuracies during multiple load regulation phases (MAPE = 4.90%, $R^2 = 0.71$), indicating its limited capacity to discern feature variations across distinct operational stages. In contrast, the GRU model achieved relatively stable predictions by leveraging global temporal dependencies, yielding marginally superior overall performance (MAPE = 4.25%, $R^2 = 0.79$). However, this architecture remained constrained in prediction accuracy due to insufficient local feature extraction capability. The CNN-GRU hybrid framework synergistically enhanced both local feature resolution and global temporal modelling, reducing MAPE to 3.81% and elevating $R^2$ to 0.86.

#### 5.2.2 | Efficacy of the multi-head attention mechanism

The incorporation of the MA mechanism endowed the CNN-GRU-MA architecture with significantly enhanced robustness in regions of abrupt load variations. Across five GB-CV validation rounds, NOx predictions exhibited strong alignment with measured value (MAPE = 3.40%),

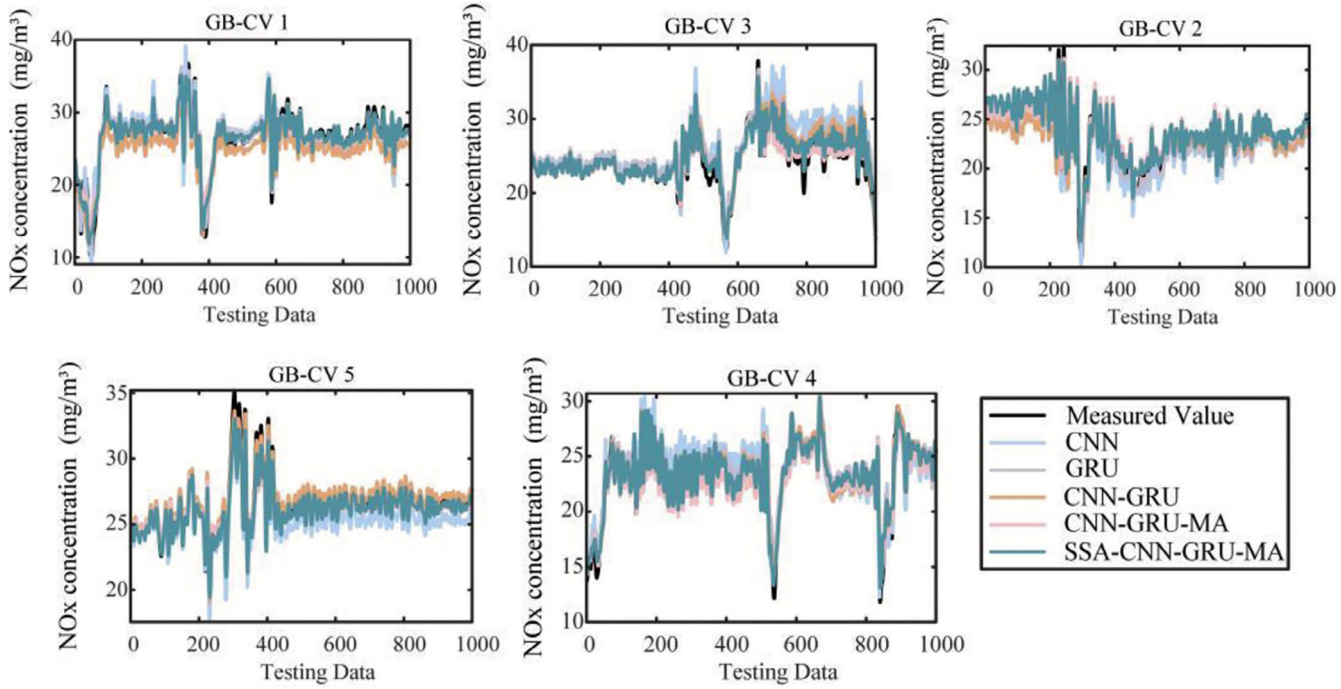


**FIGURE 8** Five rounds of gapped blocking cross-validation prediction results.

**TABLE 6** Comprehensive evaluation indicators of model structure (average of five rounds of cross-validation).

| Model | MAE | MAPE | MSE | RMSE | $R^2$ | TC |
|---|---|---|---|---|---|---|
| CNN | 1.18 | 4.90% | 2.48 | 1.48 | 0.71 | 13.11 s |
| GRU | 1.01 | 4.25% | 1.88 | 1.29 | 0.79 | 6.17 s |
| CNN-GRU | 0.91 | 3.81% | 1.61 | 1.14 | 0.86 | 16.15 s |
| CNN-GRU-MA | 0.77 | 3.40% | 1.13 | 1.03 | 0.88 | 17.01 s |
| SSA-CNN-GRU-MA | **0.60** | **2.63%** | **0.70** | **0.83** | **0.92** | 19.67 s |

Abbreviations: CNN, convolutional neural network; GRU, gated recurrent unit; MA, multi-head attention; MAE, mean absolute error; MAPE, mean absolute percentage error; PSO, particle swarm optimization; RMSE, root mean squared error; SSA, sparrow search algorithm.

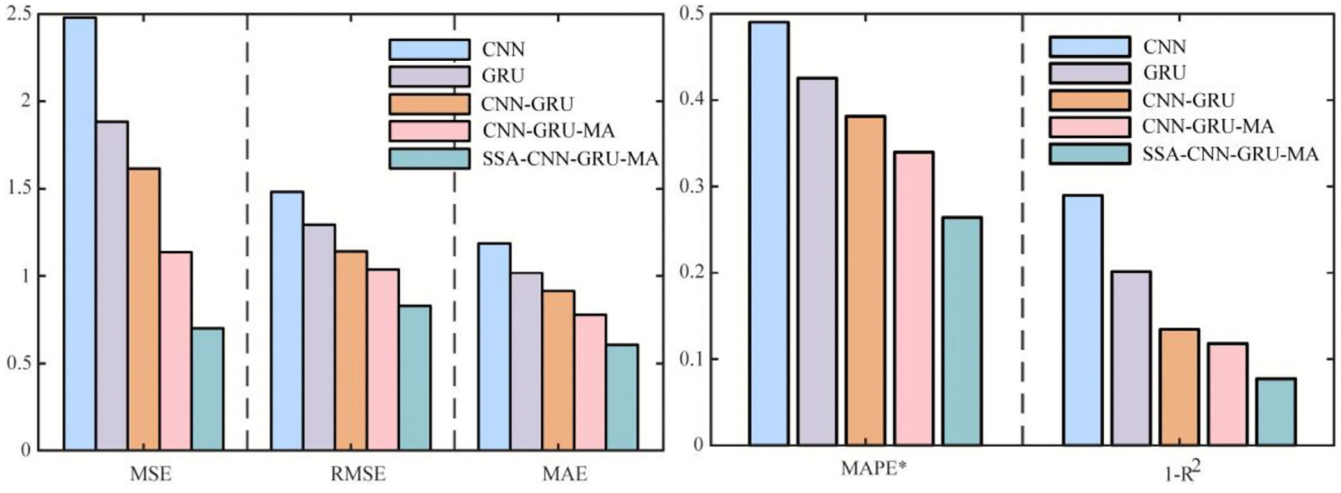


**FIGURE 9** Comparison of model performance metrics (where MAPE* represents the original MAPE values scaled by a factor of 10). MAPE, mean absolute percentage error.

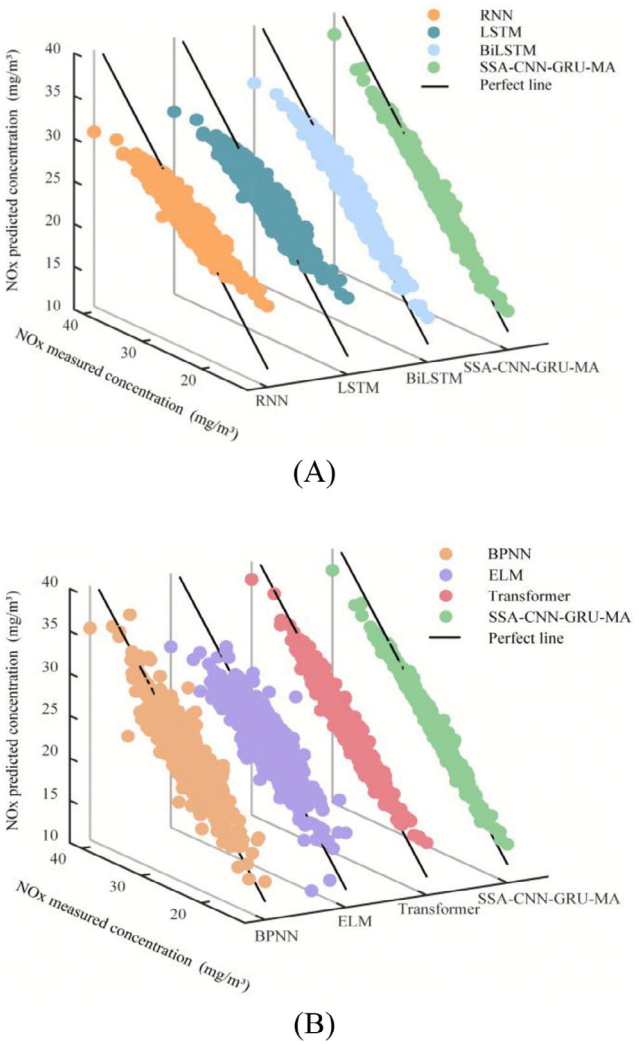


**FIGURE 10** Scatter plot of NOx prediction results of different models: (A) RNN, LSTM, BiLSTM, and SSA-CNN-GRU-MA; (B) BPNN, ELM, Transformer, and SSA-CNN-GRU-MA. BiLSTM, bidirectional long short-term memory network; BPNN, backpropagation neural network; CNN, convolutional neural network; ELM, extreme learning machine; GRU, gated recurrent unit; LSTM, long short-term memory network; MA, multi-head attention; RNN, recurrent neural network; SSA, sparrow search algorithm.

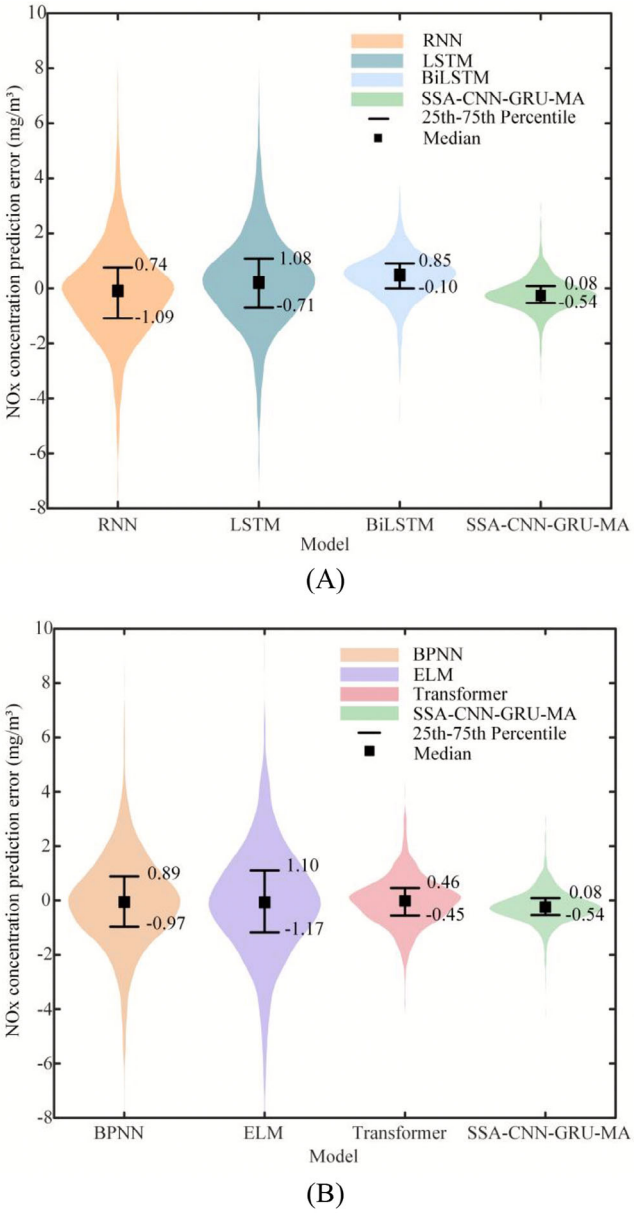


**FIGURE 11** Distribution of prediction error ranges for different models: (A) RNN, LSTM, BiLSTM and SSA-CNN-GRU-MA; (B) BPNN, ELM, Transformer and SSA-CNN-GRU-MA. BiLSTM, bidirectional long short-term memory network; BPNN, backpropagation neural network; CNN, convolutional neural network; ELM, extreme learning machine; GRU, gated recurrent unit; LSTM, long short-term memory network; MA, multi-head attention; RNN, recurrent neural network; SSA, sparrow search algorithm.

representing a 10.87% reduction compared to the CNN-GRU baseline. This improvement substantiates the MA mechanism's capacity to prioritize critical timestep features and strengthen discriminative modelling of peak-shaving phase characteristics in power plant operations.

### 5.2.3 | Performance of the SSA optimized baseline model

The SSA-CNN-GRU-MA model achieved optimal mean performance across all metrics (MAPE = 2.63%) and demonstrated superior stability and precision in responding to abrupt NOx fluctuations during multiple load regulation phases (e.g., GB-CV rounds 3, 4, and 5), maintaining MAPE below 3% (minimum 1.98%). These results substantiate the model enhanced resistance to temporal data drift and validate the synergistic interplay between hyperparameter optimization (via SSA) and the baseline architecture.

| Model | MAE | MAPE | MSE | RMSE | $R^2$ | TC |
|---|---|---|---|---|---|---|
| RNN | 1.25 | 5.10% | 2.99 | 1.73 | 0.71 | 8.71 s |
| LSTM | 1.20 | 4.99% | 2.71 | 1.64 | 0.73 | 9.79 s |
| BiLSTM | 0.75 | 3.07% | 0.98 | 0.99 | 0.90 | 10.36 s |
| BP | 1.24 | 4.96% | 2.83 | 1.68 | 0.73 | 3.90 s |
| ELM | 1.50 | 6.00% | 4.12 | 2.03 | 0.60 | 0.02 s |
| Transformer | 0.66 | 2.67% | 0.78 | 0.88 | 0.92 | 18.26 s |
| SSA-CNN-GRU-MA | **0.51** | **1.91%** | **0.60** | **0.79** | **0.94** | 18.87 s |

Abbreviations: BiLSTM, bidirectional long short-term memory network; BP, backpropagation; BPNN, backpropagation neural network; CNN, convolutional neural network; ELM, extreme learning machine; GRU, gated recurrent unit; LSTM, long short-term memory network; MA, multi-head attention; RNN, recurrent neural network; SSA, sparrow search algorithm; TC, time cost.

**TABLE 7** Comprehensive evaluation indicators for different models (BPNN, ELM, Transformer, and SSA-CNN-GRU-MA).

To enable comprehensive and intuitive comparison of the model performance metrics, the results from Table 6 were visualized as a multi-criteria bar chart (Figure 9), The SSA-CNN-GRU-MA model performs optimally in multiple performance evaluation metrics such as MAE and MAPE.

In conclusion, the proposed architecture effectively integrates the localized feature extraction capabilities of CNN, the global temporal modelling capacity of GRU, and the MA mechanism's ability to identify the most critical time-step features. Through SSA-based hyperparameter optimization, the model achieves significant enhancements in prediction accuracy and generalization capability.

## 5.3 | Comparative analysis of predictive performance

To validate the superiority of the SSA-CNN-GRU-MA model, this study compared it with the traditional machine models (backpropagation neural network [BPNN] and ELM) and deep learning models (RNN, LSTM, BiLSTM, Transformer), using an additional 5000 new highly dynamic data samples (train–test split: 7:3). Figure 10 illustrates through its scatterplot of predicted versus actual NOx concentrations that the SSA-CNN-GRU-MA predictions exhibit the closest alignment with the perfect line. Figure 11 illustrates the standardized error frequency distributions of all models in the test set. The SSA-CNN-GRU-MA model demonstrates superior performance, with its smallest interquartile range (IQR = 0.62), surpassing all comparable baseline models. Compared to traditional machine learning models BPNN (IQR = 1.86) and ELM (IQR = 2.27), the SSA-CNN-GRU-MA model reduces the IQR by 66.67% and 72.69%, respectively, and exhibits a significantly lower frequency in high-error regions. Similarly, other deep learning models such as RNN (IQR = 1.83), LSTM (IQR = 1.79), BiLSTM (IQR = 0.95) and Transformer (IQR = 0.91) demonstrate reduced frequencies in high-error regions compared to traditional counterparts.

Table 7 presents the comprehensive evaluation results of all models on the test set, with SSA-CNN-GRU-MA demonstrating superior performance across all metrics. Its MAPE of 1.91% significantly outperforms BPNN (4.96%) and ELM (6.00%), achieving reductions of 61.6% and 68.3%, respectively. Compared to the Transformer (2.67%), the proposed model performances exhibit a 28.7% improvement, confirming its enhanced adaptability to wide-load fluctuation scenarios. As an enhanced architecture evolving from RNN and LSTM frameworks, the BiLSTM model achieves 3.07% MAPE. While demonstrating greater accuracy than its RNN and LSTM counterparts, it exhibits marginally inferior performance compared to the Transformer model.

While BPNN and ELM exhibit faster training speeds (3.90 and 0.02 s, respectively), their insufficient accuracy renders them unsuitable for precision-critical deep peak-shaving operations of plant. The SSA-CNN-GRU-MA model achieves a TC of 18.87 s, maintaining computational efficiency comparable to the Transformer while delivering superior accuracy, thereby demonstrating enhanced engineering practicality within the power plant's 60-s sampling cycle constraint.

# 6 | CONCLUSIONS

To address the dynamic nonlinearity, temporal correlations, and multivariate coupling characteristics of SCR operational data in coal-fired power plants, this study proposes an SSA-CNN-GRU-MA time series model for predicting NOx concentrations in the flue gas at the SCR outlet, being validated by the operational data from a power plant in Guangzhou. Key conclusions are as follows:

1. The SSA-CNN-GRU-MA model effectively leverages multi-module synergies to improve NOx prediction accuracy and generalization during deep peak-shaving operations. Compared to the baseline model (CNN-GRU-MA), it demonstrates a 54% reduction in test-set mean absolute percentage error (MAPE = 1.89%), with a relative gap of only 20% between training and test MAPE, demonstrating robust adaptability to dynamic load variations.
2. Ablation experiments confirm the architectural effectiveness of the proposed model, demonstrating the model's capability to predict highly dynamic NOx fluctuations during deep peak-shaving. Across five operational scenarios, the model maintains MAPE below 3% (averaging 2.63%), which conclusively validates its robustness under abrupt load variations.
3. Under equivalent TC, the SSA-CNN-GRU-MA model reduces MAPE by 28% compared to the Transformer model.

Although the proposed model demonstrates the capability of capturing dynamic load variations, quantitative metrics for load variations are lacking (currently dependent on empirical observations). Future work will focus on developing quantitative load change metrics and validating applicability to SCR system loads dynamics. In addition, adaptive model updating mechanisms will be implemented to further improve engineering applicability.

## AUTHOR CONTRIBUTIONS

**Jianghong Chen:** Methodology; writing – original draft; visualization; validation; investigation; formal analysis. **Zhimin Lu:** Writing – review and editing; resources; conceptualization. **Zhenghui Li:** Supervision; writing – review and editing. **Wenjing Li:** Investigation; validation. **Anli Zhou:** Investigation; data curation. **Wenbo Pi:** Investigation. **Youxing Wei:** Supervision. **Shunchun Yao:** Supervision; resources.

## ACKNOWLEDGEMENTS

This work was supported by the National Key Research and Development Program of China [2024YFC3909002], the Sub-project of National Key Research and Development Program of China [2024YFC3909004-02], Guangdong Key Laboratory of Efficient and Clean Energy Utilization, South China University of Technology [2013A061401005], and the Guangzhou Science and Technology Elite "Leading" Project [2024A04J4486].

## CONFLICT OF INTEREST STATEMENT

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.

## PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1002/cjce.70054.

## DATA AVAILABILITY STATEMENT

Research data are not shared.

## ORCID

*Shunchun Yao* https://orcid.org/0000-0002-3287-9609

## REFERENCES

[1] J. Li, M. S. Ho, C. Xie, N. Stern, *Renewable Sustainable Energy Rev.* **2022**, *158*, 112112.
[2] M. Shahbaz, C. Raghutla, K. R. Chittedi, Z. Jiao, X. V. Vo, *Energy* **2020**, *207*, 118162.
[3] G. Zhang, P. Xie, S. Huang, Z. Chen, M. Du, N. Tang, Y. Niu, F. Hong, *Front. Energy Res.* **2021**, *9*, 767277.
[4] J. Li, C. Zhang, M. R. Davidson, X. Lu, *Appl. Energy* **2025**, *377*, 124459.
[5] L. Cao, J. Zhang, *Clean Energy* **2024**, *8*, 37.
[6] Z. Zhuang, B. Guan, J. Chen, C. Zheng, J. Zhou, T. Su, Y. Chen, C. Zhu, X. Hu, S. Zhao, J. Guo, H. Dang, Y. Zhang, Y. Yuan, C. Yi, C. Xu, B. Xu, W. Zeng, Y. Li, K. Shi, Y. He, Z. Wei, Z. Huang, *Chem. Eng. J.* **2024**, *486*, 150374.
[7] S. M. Wang, *Engineering* **2020**, *6*, 167.
[8] J. Zhu, Z. Ye, D. Gong, Q. Wang, Q. Luo, *iScience* **2025**, *28*, 111588.
[9] G. Wang, O. I. Awad, S. Liu, S. Shuai, Z. Wang, *Energy* **2020**, *198*, 117286.
[10] K. K. Jain, *Degree Thesis*, Purdue University (West Lafayette, IN) **2017**.
[11] T. Yang, K. Ma, Y. Lv, Y. Bai, *Fuel* **2020**, *274*, 117811.
[12] Z. Tang, S. Wang, S. Cao, T. Shen, in *Chinese Automation Congress* (*CAC*), IEEE, Shanghai, China. IEEE, **2020**, 3226-3229.
[13] Y. Lv, C. E. Romero, T. Yang, F. Fang, J. Liu, *Appl. Therm. Eng.* **2018**, *143*, 160.
[14] Y. Lv, X. Lv, F. Fang, T. Yang, C. E. Romero, *Energy* **2020**, *192*, 116589.
[15] M. Mohammadi, D. Saloglu, H. Dertli, M. Ghaffari-Moghaddam, *Water, Air, Soil Pollut.* **2024**, *235*, 297.
[16] Z. Yin, C. Yang, X. Yuan, F. Jin, B. Wu, *Front. Energy Res.* **2023**, *10*, 1054427.
[17] Z. Li, S. Yao, D. Chen, L. Li, Z. Lu, W. Liu, Z. Yu, *Energy* **2024**, *306*, 132477.
[18] J. Liao, J. Hu, P. Chen, H. Wu, M. Wang, Y. Shao, Z. Li, *Energy Sources, Part A* **2024**, *46*, 1800.
[19] A. García, J. Monsalve-Serrano, J. Marco-Gimeno, E. Iñiguez, *Fuel* **2025**, *394*, 135150.
[20] G. Sundaram, T. Gehra, J. Ulmen, M. Heubaum, D. Görges, M. Günthner, *Modeling of Transient Gasoline Engine Emissions using Data-Driven Modeling Techniques*. (SAE Technical Paper 2023-01-0374), **2023**, https://doi.org/10.4271/2023-01-0374

[21] C. Liu, Z. Wei, L. Zhou, Y. Shao, *Complex Intell. Syst.* **2025**, *11*, 1.
[22] B. Hu, C. Liu, Y. Yang, B. Wang, D. Cai, W. Xu, *IEEE Access* **2022**, *10*, 24769.
[23] Y. Zhu, C. Yu, W. Fan, H. Yu, W. Jin, S. Chen, X. Liu, *Energy* **2023**, *280*, 128128.
[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, *Adv. Neural Inf. Process Syst.* **2017**, *30*, 2.
[25] Z. Niu, G. Zhong, H. Yu, *Neurocomputing* **2021**, *452*, 48.
[26] J. Xue, B. Shen, *Syst. Sci. Control Eng.* **2020**, *8*, 22.
[27] C. Bergmeir, R. J. Hyndman, B. Koo, *Comput. Stat. Data Anal.* **2018**, *120*, 70.
[28] V. Cerqueira, L. Torgo, I. Mozetič, *Mach. Learn.* **2020**, *109*, 1997.

**How to cite this article:** J. Chen, Z. Lu, Z. Li, W. Li, A. Zhou, W. Pi, Y. Wei, S. Yao, *Can. J. Chem. Eng.* **2025**, 1. https://doi.org/10.1002/cjce.70054