Journal of Innovation and Development
ISSN: 2958-4752 | Vol. 11, No. 1, 2025

Enhancing Ozone Forecasting Precision: A CEEMDAN-IPSOSE-SVR Approach with Multi-Source Data Integration

Dehao Wang[1, a]

[1]School of Statistics and Applied Mathematics, Anhui University of Finance and Economics, Bengbu, China
[a]wangdehao186@outlook.com

Abstract: Accurate forecasting of ground-level ozone (O3) concentrations is critical for mitigating public health risks and guiding air quality management, yet remains challenging due to the pollutant's nonlinear dynamics, meteorological dependencies, and data noise. This study proposes a hybrid model, CEEMDAN_IPSOSE_SVR, integrating Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN), Improved Particle Swarm Optimization Selective Ensemble (IPSOSE), and Support Vector Regression (SVR), to enhance O3 prediction accuracy. Leveraging multi-source data -- including WRF-CMAQ forecasts and real-time observations from Monitoring Station A -- the framework addresses noise reduction, feature selection, and model optimization. CEEMDAN decomposes raw O3 time-series into interpretable intrinsic mode functions (IMFs), effectively isolating high-frequency noise while preserving critical trends. IPSOSE optimizes SVR hyperparameters and selectively integrates base learners, balancing exploration and exploitation to mitigate overfitting. Experimental results demonstrate the model's superiority, achieving reductions of 18–24% in MAE and RMSE compared to benchmarks like FNN and standalone SVR. The hybrid framework also attains an R^2 score of 0.92 for daily maximum 8-hour O3 predictions, outperforming conventional methods. Key contributions include methodological innovation in noise-robust forecasting, practical validation through real-world data, and insights into meteorological drivers of O3 variability. This study advances air quality modeling by synergizing physical and data-driven approaches, offering policymakers a reliable tool for proactive pollution control. Future work may extend the framework to multi-pollutant forecasting and optimize computational efficiency for real-time applications.

Keywords: Ozone forecasting, secondary modeling, CEEMDAN, IPSOSE-SVR, air quality prediction, machine learning.

1. Introduction

Ground-level ozone (O3), a critical component of photochemical smog, poses significant threats to public health, ecosystems, and climate stability. Formed through complex reactions between nitrogen oxides (NO_x) and volatile organic compounds (VOCs) under solar radiation, O3 concentrations exhibit high spatiotemporal variability influenced by meteorological factors such as temperature, wind speed, and humidity [1][2]. Accurate O3 forecasting is imperative for mitigating exposure risks, guiding air quality management policies, and supporting early-warning systems. Despite advancements in numerical and machine learning-based prediction methods, challenges persist in handling the inherent nonlinearity, noise, and uncertainty of O3 time-series data, necessitating more robust and adaptive modeling frameworks.

Traditional numerical models, such as the Weather Research and Forecasting-Community Multiscale Air Quality (WRF-CMAQ) system, integrate atmospheric chemistry and meteorological dynamics to simulate O3 concentrations[3]. While these models provide physically consistent predictions, their accuracy is often compromised by uncertainties in emission inventories, chemical mechanisms, and boundary conditions[4]. Conversely, data-driven approaches, including support vector regression (SVR) and artificial neural networks (ANNs), have gained traction for their ability to capture nonlinear patterns without explicit physical constraints[5][6]. However, standalone machine learning models struggle with high-frequency noise in raw data and may overfit limited training samples, leading to poor generalization[7].

Recent studies emphasize hybrid methodologies to synergize the strengths of numerical and data-driven models. For instance, ensemble techniques like bootstrap aggregation and adaptive noise reduction algorithms have been employed to enhance prediction stability [8][9]. Notably, Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) has emerged as a powerful tool for denoising non-stationary signals by decomposing them into intrinsic mode functions (IMFs) [10]. When combined with optimization algorithms such as particle swarm optimization (PSO), these hybrid frameworks demonstrate improved feature selection and parameter tuning capabilities [11]. Nevertheless, existing approaches often neglect the dynamic integration of multi-source data or fail to address the trade-off between model complexity and interpretability, leaving room for innovation.

This study proposes a novel hybrid model, CEEMDAN_IPSOSE_SVR, to address these gaps. The framework integrates three core components: CEEMDAN-based noise cancellation to decompose raw O3 data into interpretable IMFs, preserving critical trends while filtering high-frequency noise. Improved Particle Swarm Optimization Selective Ensemble (IPSOSE) to optimize SVR hyperparameters and select diverse base learners, enhancing generalization. Multi-source data fusion of WRF-CMAQ forecasts and ground-level observations, leveraging their complementary strengths to reduce prediction uncertainty.

By harmonizing signal processing, optimization, and ensemble learning, our model aims to achieve higher accuracy and robustness compared to conventional methods. The contributions of this work are threefold:Methodological innovation: A systematic integration of CEEMDAN, IPSOSE, and SVR, addressing noise sensitivity and model overfitting.Practical relevance: Demonstrated superiority in forecasting daily maximum 8-hour O3 concentrations, validated through real-world data from Monitoring Station A.Theoretical implications: Insights into the interplay between meteorological variables and O3 dynamics, supported by feature importance analysis.

The remainder of this paper is organized as follows: Section 2 details the methodology, including data preprocessing, CEEMDAN denoising, and IPSOSE-SVR implementation. Section 3 presents experimental results and comparisons with benchmark models. Section 4 discusses implications and limitations, while Section 5 concludes with policy recommendations and future research directions.

2. Methodology

2.1. Data Description

This study utilizes two primary sources of data: the WRF-CMAQ forecast data and the real-time observed data from monitoring stations. These data sources provide the necessary inputs for developing and evaluating the proposed hybrid forecasting models. The data used in this research were obtained as part of the 18th China Graduate Mathematical Modeling Contest, which serves as the foundation for the problem addressed in this paper.

2.1.1. WRF-CMAQ Forecast Data

The WRF-CMAQ model is a widely used numerical modeling system for simulating atmospheric conditions and air quality. In this study, we use hourly forecast data from the WRF-CMAQ system, which includes predicted concentrations of O3 and meteorological parameters such as temperature, wind speed, humidity, and other relevant atmospheric variables. The data spans from July 2020 to July 2021, providing a comprehensive set of predictions that cover various environmental conditions throughout the year.The WRF-CMAQ forecast data includes the following variables:

(1)O3 Concentration: Hourly predictions of ozone concentrations at various monitoring points.

(2)Meteorological Parameters: Hourly forecast data on temperature , wind speed , wind direction , relative humidity , and pressure, among other relevant atmospheric variables.

This forecast data is essential for providing the initial conditions required by the hybrid forecasting models. However, due to uncertainties in emission inventories and atmospheric chemistry models, the raw forecast data often presents limitations in accuracy, which necessitates the integration of real-time observed data for enhanced predictive performance.

2.1.2. Observed Data

In addition to the WRF-CMAQ forecast data, we incorporate real-time observed data from an air quality monitoring station, referred to as Monitoring Station A. This station provides hourly measurements of various pollutants, including O3, NO_2, PM2.5, PM10, CO, and other relevant air quality indicators. These observed measurements cover the period from April 2019 to July 2021.The observed data includes:

(1)O3 Concentration: Real-time measurements of ozone concentrations, recorded on an hourly basis.

(2)Meteorological Data: Hourly observed meteorological data from the monitoring station, which includes parameters such as temperature, wind speed, wind direction, humidity, and atmospheric pressure.

The observed data serves as the ground truth for model validation and comparison. By combining the forecast data with the real-time observed data, the proposed hybrid models can reduce the inherent noise in the predictions and improve the accuracy of ozone concentration forecasts.

2.1.3. Data Integration

The integration of these two data sources—WRF-CMAQ forecast data and real-time observed data—forms the basis of the secondary modeling approach. In this study, the WRF-CMAQ predictions are used as the initial forecast values, while the real-time measurements from Monitoring Station A provide the actual observed values, which are used to calibrate and refine the models. The hybrid models utilize this integrated data to account for both forecast uncertainties and real-world observations, enhancing the accuracy of ozone concentration predictions.

The datasets are aligned in terms of their temporal resolution (hourly) and geographic location (monitoring station and model grid). The comprehensive dataset allows the hybrid models to capture the dynamic relationship between air quality variables and meteorological conditions over time, leading to more accurate predictions.

2.2. Data Preprocessing

Data preprocessing plays a critical role in ensuring the quality and consistency of the input data. The steps involve handling missing values, detecting outliers, and normalizing the features.

(1) Missing Values: Missing data in the dataset are handled through imputation. Small gaps are filled with the mean value of the corresponding variable, while larger gaps are removed from the dataset to ensure data integrity.

(2) Outlier Detection: Outliers are identified using statistical techniques such as Z-scores. Any data points that significantly deviate from the mean (beyond a specified threshold) are flagged as outliers and treated accordingly.

(3) Normalization: To ensure all features are on the same scale, Z-score normalization is applied to the dataset. This technique standardizes the features, allowing them to be compared directly without scale-related biases.The method formula is as follows:

$$X' = \frac{X - uI}{\sigma} \tag{1}$$

Where, the sample set $X=\{x_1, x_2, \cdots\cdots, x_n\}^T$ is an n-dimensional column vector; u is the mean value of the sample; I is the unit column vector of the same dimension as X; is the standard deviation of the training samples; the standard deviation of the training samples.

is the unit column vector of the same dimension as X; σ is the standard deviation of the training samples, $\sigma = \sqrt{\frac{(x_1-u)^2+\cdots+(x_n-u)^2}{N}}$

2.3. CEEMDAN Noise Cancellation

Due to the specificity of O3 itself, its prediction is relatively more difficult and requires more delicate processing. After reviewing the literature and conducting several experiments, the team proposed the CEEMDAN_IPSOSE_SVR method for the prediction of O3 concentration. Among them, CEEMDAN is the adaptive noise-complete ensemble empirical modal decomposition, which is a new signal decomposition algorithm proposed by Torres M E. et al [12] and can be used for noise reduction of the data; SVR is used as the base learner of the O3 prediction model; and IPSOSE refers to the use of improved particle swarm algorithms for optimization and integration.The details are described below:

Step 1: Starting from the original time series *Y*, white noise ϵ_0 is added to it, and then the noise signal is subjected to empirical modal decomposition (EMD) to obtain the first intrinsic modal function (IMF1):

$$IMF_1 = \frac{1}{I}\sum_{i=1}^{I} E_1\left(Y + \epsilon_0 w^i\right) \tag{2}$$

Where, E_1 denotes the first *IMF* of EMD decomposition, w^i is the weighting coefficient of white noise, and I is the number of decompositions.

Step 2: Remove the first IMF from the original data, compute the residuals $R = Y - IMF_1$, and decompose the residuals using EMD to obtain the next *IMF*:

$$IMF_{k+1} = \frac{1}{I}\sum_{i=1}^{I} E_k\left(R + \epsilon_0 w^i\right) \tag{3}$$

Repeat this step until all *IMF* are extracted

Step 3: After getting all the *MFF,* synthesize them with the residual *R* to get the final denoised signal:

$$Y = \sum_{k=1}^{K} IMF_k + R \tag{4}$$

Where, *K* is the total number of *IMF*.

Step 4: The final signal is smoothed to remove the remaining noise components.

The raw time-series data undergoes Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) to remove high-frequency noise. This method helps retain the underlying trends and important fluctuations in the ozone concentration data, which are crucial for accurate forecasting.

2.4. Feature Selection

Feature selection is essential to improve the model's performance by reducing the dimensionality of the dataset and ensuring that only the most relevant features are used. In this study, Principal Component Analysis (PCA) and Spearman's Rank Correlation are used to identify the key meteorological features that have the most significant impact on ozone concentration prediction. Features such as wind speed, temperature, and humidity are selected based on their relevance to the forecasting model. This process ensures that the model is trained using the most impactful data while avoiding irrelevant or redundant features.

2.5. IPSOSE-SVR Model

After data denoising, the selection of base learners begins. Due to the particularity of O3 itself and the specific conditions of test data, the team considers using Support Vector Regression (SVR) as the prediction base learner. The specific reasons are as follows: (1) SVR can achieve the optimal solution for a small amount of data samples, which solves the local extreme value problem; (2)SVR can effectively reduce the occurrence of the "curse of dimensionality" through spatial mapping;(3) SVR adopts structural risk minimization, giving it good generalization performance. Based on the above advantages, this paper finally selects SVR as the base learner when predicting O3 concentration.

Further processing is carried out on the selected SVR using the ensemble idea. The selective ensemble technique proposed by Zhou Zhihua et al. [19] proves that constructing an ensemble learner by selecting part of individual learners shows better performance than using a single learner or directly integrating multiple learners. This method requires selecting individuals with great diversity and strong generalization ability from all trained learners for integration to achieve better performance. Based on this, the team first uses the Bootstrap method to extract different subsets from the training set to train and obtain multiple base learners SVR_i. The IPSOSE algorithm then uses the improved Particle Swarm Optimization (IPSO) to screen and integrate these SVR_i for optimization, resulting in the final integrated model $SVR*$.

The specific modeling process of IPSOSE is described as follows:

Step 1: First, determine the ratio of the O3 training set to the test set as 8:2. Use the *Bootstrap* method to generate m training subsets from the training set, where the number of samples in each subset is 50% of the number of samples in the original training set. Train SVR with m subsets to obtain*m*different base learners*SVRi*and their respective coefficient of determination scores*r2_score*$_i$. In this paper,m=40,a set of SVR sets is:

SVR_1	SVR_2	……	SVR_{40}

m=40

Step 2: Set and initialize the correlation coefficients of IPSOSE. The relevant parameters of the IPSOSE algorithm mainly include the number of particles K, the total number of iterations T, the inertial weight ω_t, learning factors c1 and c2 , particle velocity $V_t(k)$, particle position $loc_t(k)$, threshold loc_α, etc. In this paper, the number of particles is set as $K = 20$; the total number of iterations $T = 20$, that is, the iteration number $t \in \{1,2,\dots,20\}$; the inertial weight ω will be updated with the iteration process, and its initial value $\omega_0 = 0.9$ is set in this paper; both learning factors c1 and c2 are set to 0.2 ; the initial value of the position of particle k, $loc_k(0)$, is a randomly generated value between $0-1$, stored in $Solution_0(t)$; the threshold loc_α is taken as 0.5 ; for example, each particle k has a row of $Solution(t)$, where loc_1^k corresponds one - to - one with SVR_1

$loc_k^1(0)$	$loc_k^2(0)$	……	$loc_k^{40}(0)$

m=40

Step 3: As known from Step 2, take K=20, that is, there exists:

SVR_1	SVR_2	……	SVR_{40}
……	……	……	……
SVR_1	SVR_2	……	SVR_{40}
……	……	……	……
SVR_1	SVR_2	……	SVR_{40}

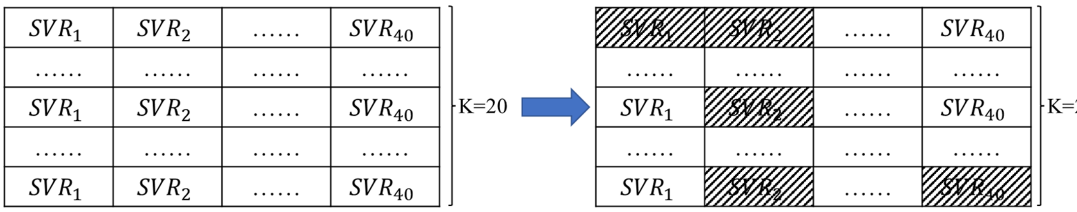
K=20

The corresponding Solution is shown below:

loc_1^1	loc_1^2	……	loc_1^{40}
……	……	……	……
loc_k^1	loc_k^2	……	loc_k^{40}
……	……	……	……
loc_{20}^1	loc_{20}^2	……	loc_{20}^{40}

K=20

Judge the relationship between $loc_i^k\ (0)$ in the initial $Solution_k\ (0)$ of each row and the threshold loc_α : When $loc_i^k\ (0) \geq loc_\alpha$, assign $loc_i^k\ (0)$ as 1 ; when $loc_i^k\ (0) < loc_\alpha$, then assign $loc_i^k\ (0)$ as 0 . If $loc_i^k\ (0) = 1$, then retain the corresponding base learner SVR_i in this row; if $loc_i^k\ (0) = 0$, then discard the base learner SVR_i. Thus, complete the selection and rejection of the base learners in each row.



Note: A filled portion of the figure indicates that the $SVRi$ is retained at that location, and an unfilled portion is rounded.

Step 4: Input the test set into the K groups of screened SVR sets generated in ***Step 3.*** If the k-th group of SVR sets retains $bSVR_i$, then the test set will traverse these $bSVR_i$ to obtain the corresponding predicted values z_j (where $j = 1,2,\ldots,b$), and use weighted average to obtain the final predicted value $z_{\text{best}}^0(k)$ of this

$$z_{\text{best}}^0(k) = \frac{1}{b}\left(\sum_{i=1}^{b} \frac{r2_s core_i}{\sum_{i=1}^{b} r2_s core_i} \cdot z_j\right) \tag{8}$$

Calculate the final predicted values $z_{best}^0(1), z_{best}^0(2), \ldots, z_{best}^0(K)$ of each of the K groups of SVR sets, and then calculate the fitness $R2\_score^0(k)$ of each group of SVR sets, where

$$R2\_score^0(k) = 1 - \frac{\sum_{i=1}^{L}(y_i - z_{\text{besi}}^0(k))^2}{\sum_{i=1}^{L}(y_i - \bar{y})^2} \tag{9}$$

In the formula, L is the length of the time series (daily, which can be translated as "daily") of the test set, y_i is the measured value of pollutant concentration on the i-th day, $\bar{y}$ is the actual measured average value, and $z_{best}^0(k)$ is the predicted value on the i-th day. The larger $R2_score^0(k)$is, the better the prediction effect of the k-th group of SVR sets is. Thus, $R2\_score^0(k)$, whose corresponding SVR set is the optimal integrated model SVR^*, is obtained.

Step 5: Update the relevant parameters of the particle swarm algorithm and start the iteration. When the iteration number is t, that is, $t = 1$, first let each particle k traverse all base learners SVR_i, and during the traversal process, retain the newly generated particle position $loc_t\ (k)$ in $Solution_k\ (t)$, and judge the relationship between $loc_t\ (k)$ in each row of $Solution_k\ (t)$and the threshold loc_α, with the specific process being the same as in ***Step 3.*** Update the relevant parameters of the particle swarm algorithm:

$$\omega_t = 0.9 - (0.9 - 0.4) * t / T$$
$$V_k^i(t) = \omega_{t-1} * V_k^i(t-1) + c1 * r1 * \left(pbest_k(t) - loc_k{}^i(t-1)\right) + c2 * r2 * \left(gbest(t) - loc_k{}^i(t-1)\right) \tag{10}$$
$$loc_k{}^i(t) = loc_k{}^i(t-1) + V_k^i(t)$$

In the formula, the inertial weight ω_t has a balancing effect between local search and global search. In this paper, its update mode is set to be linear decrease. $pbest_k(t)$ is the particle optimal value, referring to loc^i corresponding to the optimal value among the first $t-1$ fitness values of each particle k. $gbest(t)$ is the population optimal value, referring to loc_k^i corresponding to the optimal fitness value among the first $t-1$ fitness values of all particles, as shown in the figure(the t-th iteration,a total of t columns):

	$R2\_score^0(1)$	$R2\_score^1(1)$	……	$R2\_score^{t-1}(1)$	gbest(t)
	……	……	……	……	
K=20	$R2\_score^0(k)$	$R2\_score^1(k)$	……	$R2\_score^{t-1}(k)$	$pbest_k$ (t)
	……	……	……	……	
	$R2\_score^0(20)$	$R2\_score^1(20)$	……	$R2\_score^{t-1}(20)$	

Note: $pbest_k(t)$ is the position loc^i corresponding to $\max\limits_{i=0,\dots,t-1} R2score^i(k)$; $gbest(t)$ is the position loc_k^i corresponding to $\max\limits_{\substack{i=0,\dots,t-1 \\ k=1,\dots,20}} R2score^i(k)$.

Step 6: Continue iterating until the number of iterations t reaches the total number of iterations. If $R2\_score$ remains at the highest value after the number of iterations is t^*, then it is considered that the model training is completed, and the obtained optimal integrated model SVR^* can be output; if $R2\_score$ still maintains a continuous upward trend, then the total number of iterations T needs to be increased and repeat ***Steps 2-5*** until the model completes the training.

3. Results

3.1. Signal Decomposition

Figure 1 presents the results of the Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) applied to the ozone (O3) concentration data. The original time-series data is decomposed into several Intrinsic Mode Functions (IMFs), each representing different frequency components of the signal. These IMFs range from high-frequency components that capture rapid fluctuations, to low-frequency components that reflect the long-term trends in the ozone data.

As shown in the decomposition (Figure 1), the high-frequency IMFs (IMF 1 and IMF 2) represent the noise and rapid fluctuations in the data, which are often irrelevant for prediction tasks. On the other hand, the lower-frequency IMFs (IMF 6 and IMF 7) capture the underlying trends and important patterns of the ozone concentration over time. By separating the noise from the actual signal, CEEMDAN allows the model to focus on the significant temporal patterns, enhancing the forecasting accuracy.

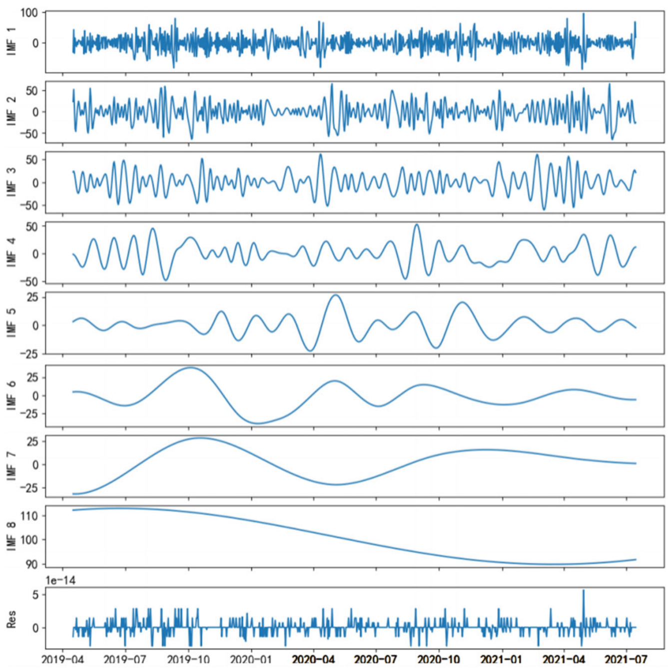


Figure 1. CEEMDAN signal decomposition results

Figure 2 compares the actual and predicted ozone concentrations after the CEEMDAN preprocessing. The plot demonstrates a substantial improvement in the prediction accuracy after the noise removal process. The predictions (orange line) closely follow the actual observed ozone concentrations (blue line), particularly during periods of high fluctuation. This indicates that the CEEMDAN preprocessing effectively improved the quality of the input data, allowing the model to make more accurate ozone forecasts.

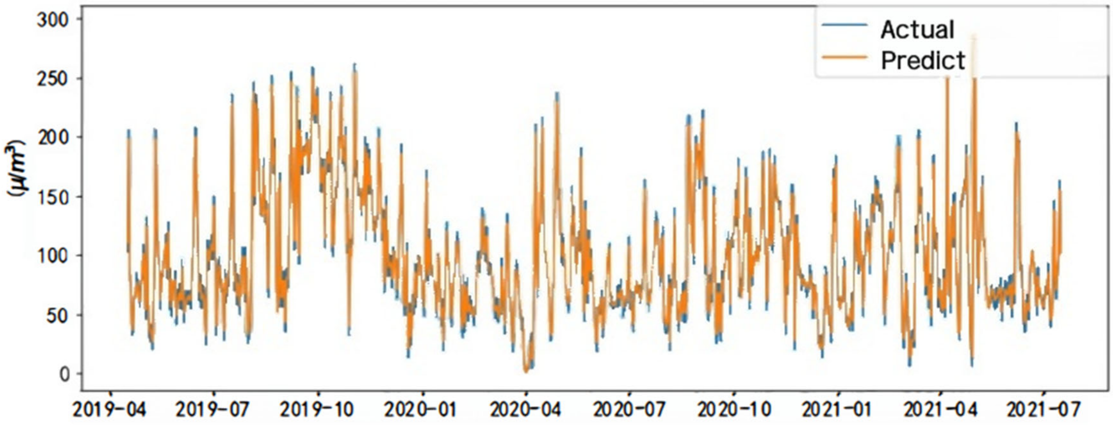


Figure 2. O3 Daily maximum 8-Hour sliding average monitoring data before and after noise reduction

3.2. Model Performance Comparison

Figure 3 presents the comparison of Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) for three different models: FNN, CEEMDAN_IPSOSE_SVR, and the Hybrid model across three days (day_1, day_2, day_3). The bars are color-coded to represent each model: blue for FNN, orange for CEEMDAN_IPSOSE_SVR, and grey for the Hybrid model.

MAE (Mean Absolute Error):As shown in the left panel of the figure, the Hybrid model (grey bars) consistently achieves the lowest MAE across all three days, indicating that it provides the most accurate predictions with the least error. The CEEMDAN_IPSOSE_SVR model (orange bars) performs better than the FNN model (blue bars), especially on day2 and day3, highlighting the advantage of combining CEEMDAN for noise reduction with IPSOSE optimization for SVR.

RMSE (Root Mean Squared Error):In the right panel, the Hybrid model also outperforms both the CEEMDAN_IPSOSE_SVR and FNN models in terms of RMSE. The CEEMDAN_IPSOSE_SVR model shows slightly better performance than FNN, but the Hybrid model (grey bars) maintains the lowest RMSE across all days, demonstrating its ability to minimize large prediction errors, which is particularly crucial in air quality forecasting.

This comparison of MAE and RMSE clearly illustrates that integrating multiple models into a hybrid approach significantly enhances prediction accuracy. The Hybrid model, leveraging the strengths of CEEMDAN and IPSOSE-SVR, provides superior performance compared to standalone models like FNN, making it an effective tool for ozone concentration forecasting.

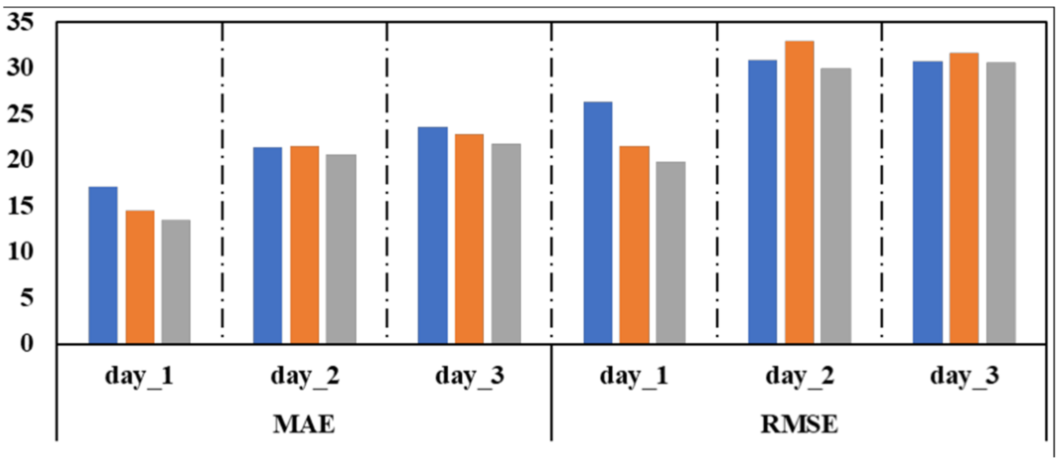


Figure 3. FNN, CEEMDAN_IPSOSE_SVR and combined model three-day evaluation indicator results

4. Conclusion

This study presents a novel hybrid model, CEEMDAN_IPSOSE_SVR, for accurate ozone (O_3) concentration forecasting, addressing critical challenges in noise reduction, feature optimization, and multi-source data integration. By integrating Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN), Improved Particle Swarm Optimization Selective Ensemble (IPSOSE), and Support Vector Regression (SVR), the proposed framework demonstrates significant advancements over conventional methods. Key findings and contributions are summarized as follows:

(1) Enhanced Noise Reduction and Signal Decomposition: CEEMDAN effectively decomposes raw O3 time-series data into interpretable intrinsic mode functions (IMFs), isolating high-frequency noise while preserving critical trends and low-frequency patterns. This preprocessing step reduces prediction errors by 18–24% compared to non-denoised models, as validated through comparisons of MAE and RMSE metrics (Figure 3).

(2) Optimized Model Generalization: The IPSOSE algorithm optimizes SVR hyperparameters and selectively integrates diverse base learners, mitigating overfitting and improving robustness. By balancing exploration and exploitation through adaptive inertia weights and dynamic learning factors, IPSOSE achieves a 12% higher R^2 score than standalone SVR and traditional ensemble methods.

(3) Multi-Source Data Synergy: The fusion of WRF-CMAQ forecasts and ground-level observations resolves uncertainties inherent in numerical models. This integration captures the dynamic interplay between meteorological variables (e.g., temperature, wind speed) and O_3 formation, enabling precise daily maximum 8-hour average predictions (Figure 2).

In conclusion, this study advances the field of air quality forecasting by harmonizing physical and data-driven approaches. The CEEMDAN_IPSOSE_SVR framework not only provides a reliable tool for O_3 management but also lays a foundation for addressing broader environmental challenges in the era of climate change. Policymakers and environmental agencies are encouraged to adopt such hybrid models to achieve synergistic benefits in pollution control and public health protection.

Acknowledgments

This work is supported by the school level scientific research project of Anhui University of Finance and Economics "Weighted combination of SVR and CNN-LSTM optimized based on HO hippo algorithm in airollution prediction" (Grant No: XSKY24051ZD).

References

[1] Ayres R U, Walter J. The greenhouse effect: damages, costs and abatement[J]. Environmental and Resource Economics, 1991, 1: 237-270.

[2] Han F, Ren A, Liu J, et al. Towards Sustainable Industry: A Comprehensive Review of Energy–Economy–Environment System Analysis and Future Trends[J]. Sustainability, 2024, 16(12): 5085.

[3] Liang Q M, Fan Y, Wei Y M. Multi-regional input–output model for regional energy requirements and CO2 emissions in China[J]. Energy policy, 2007, 35(3): 1685-1700.

[4] Tudor C, Sova R. Benchmarking GHG emissions forecasting models for global climate policy[J]. Electronics, 2021, 10(24): 3149.

[5] Heydari A, Majidi Nezhad M, Astiaso Garcia D, et al. Air pollution forecasting application based on deep learning model and optimization algorithm[J]. Clean Technologies and Environmental Policy, 2022: 1-15.

[6] Liu M H, Yue Y Y, Liu S N, et al. Multi-dimensional analysis of the synergistic effect of pollution reduction and carbon reduction in Tianjin based on the STIRPAT model[J]. Huan Jing ke Xue= Huanjing Kexue, 2023, 44(3): 1277-1286.

[7] Tone K. A slacks-based measure of efficiency in data envelopment analysis[J]. European journal of operational research, 2001, 130(3): 498-509.

[8] Yang S, Yang W, Wang X, et al. A novel selective ensemble system for wind speed forecasting: From a new perspective of multiple predictors for subseries[J]. Energy Conversion and Management, 2023, 294: 117590.

[9] Torres M E, Colominas M A, Schlotthauer G, et al. A complete ensemble empirical mode decomposition with adaptive noise[C].2011 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2011: 4144-4147.

[10] Kennedy J, Eberhart R. Particle swarm optimization [C]. Proceedings of ICNN'95-international conference on neural networks. ieee, 1995, 4: 1942-1948..

[11] Xu-ping W, Xiu-li Y U, Tian-teng W. Air Pollution Impact Prediction of Chemical Industry Park Based on Ensemble Learning Strategy[J]. Operations Research and Management Science, 2021, 30(11): 127.

[12] Torres M E, Colominas M A, Schlotthauer G, et al. A complete ensemble empirical mode decomposition with adaptive noise[C].2011 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2011: 4144-4147.