Frontiers in Environmental Research
Print ISSN: 2960-0251
Online ISSN: 2960-026X
DOI: https://doi.org/10.61784/fer3045

AN AI-DRIVEN HYBRID LSTM-RF FRAMEWORK FOR DECOUPLING ENVIRONMENTAL AND ANTHROPOGENIC IMPACTS ON AIR QUALITY

HongYue Deng
School of Thermal Engineering, Shandong Jianzhu University, Jinan 250101, Shandong, China.

Abstract: Evaluating atmospheric emission policies requires accurately separating anthropogenic activities from meteorological influences. Traditional statistical approaches often fail to capture the non-linear, high-dimensional dynamics of atmospheric environments. This research introduces a hybrid machine learning framework integrating an LSTM neural network and Random Forest meteorological normalization to decouple these overlapping effects. A dynamic inversion model was constructed using 2021 hourly observations of $PM_{2.5}$, PM10, temperature, humidity, and wind speed from Jinan's Guodian station. The LSTM captured long-term temporal dependencies, achieving a coefficient of determination exceeding 0.80 across seasonal variations. To isolate meteorological impacts, the RF model employed a 14-day random resampling technique, calculating the conditional expectation of pollutant concentrations to marginalize weather randomness. Results revealed distinct seasonal contribution rates: meteorological factors contributed 47.36% to $PM_{2.5}$ concentrations during severe December pollution events, driven by stagnant conditions and high humidity. Conversely, this contribution dropped to 24.87% in February, indicating primarily anthropogenic origins. Feature analysis identified humidity and temperature as the primary meteorological drivers of particulate accumulation. This framework provides an interpretable methodology to evaluate the actual efficacy of emission interventions under varying weather scenarios.
Keywords: Meteorological normalization; Machine learning; Long short-term memory; Random forest; Decoupling analysis

1 INTRODUCTION

Precise urban air quality management intrinsically depends on differentiating anthropogenic emissions from meteorological conditions. Weather variations directly modulate the dispersion and accumulation of $PM_{2.5}$ and PM_{10}, frequently obscuring the actual effectiveness of localized emission reduction policies. Differentiating these entangled components—formally known as meteorological normalization—is a fundamental requirement for accurate environmental governance.

Conventional air quality forecasting methodologies predominantly rely on deterministic chemical transport models or traditional statistical techniques. Deterministic models require extensive emission inventories and incur substantial computational costs, constraining their utility for rapid, localized assessments. Meanwhile, parametric statistical models assume linear relationships and stationarity. However, atmospheric dynamics inherently violate these assumptions through complex, non-linear physical interactions, leading to suboptimal decoupling accuracy. To overcome these limitations, recent environmental governance paradigms have increasingly adopted data-driven modeling to isolate complex emission reduction effects. For instance, advanced machine learning architectures have been successfully deployed to quantify the synergistic effects of pollution control [1], while random forest models have demonstrated strong capabilities in mapping complex spatial patterns and drivers of regional carbon emissions [2].

Machine learning provides a structurally sound alternative by accommodating these complex non-linear approximations without strict parametric assumptions. In highly volatile environmental contexts, LSTM networks effectively model long-range temporal dependencies in sequential data through internal gating mechanisms. The structural superiority of LSTM has been broadly validated in meteorological forecasting, including precise temperature sequence predictions [3]. Concurrently, Random Forest (RF) algorithms demonstrate exceptional efficacy in handling high-dimensional environmental variables. Recent studies have confirmed the robustness of integrating LSTM with RF for complex ecological tracking, utilizing ensemble methods to extract features dynamically [4]. By randomly sampling historical weather spaces, RF models calculate expected pollutant concentrations under statistically average meteorological conditions, effectively isolating the anthropogenic baseline from transient weather anomalies.

This research implements a specialized hybrid LSTM-RF framework in Jinan, a central Yellow River Basin city facing dual $PM_{2.5}$ and PM_{10} pollution pressures. The study establishes a time-series dynamic inversion model utilizing LSTM to predict pollutant concentrations based on meteorological variables. Subsequently, an RF-based 14-day random sampling technique decouples the meteorological and anthropogenic contributions. The core objective is to quantitatively delineate seasonal variances in meteorological contribution rates (CRs) and mathematically identify primary environmental drivers of particulate accumulation, offering a purely data-driven statistical mechanism.

2 METHODOLOGY

2.1 Data Acquisition and Preprocessing Strategy

The empirical analysis targets Jinan, characterized by complex topographical features and a dense mixture of industrial and residential zones. To ensure the integrity and spatial consistency of the decoupling analysis, the study utilizes a single, reliable monitoring station: the Guodian monitoring station situated in the Licheng District. This specific location acts as a critical receptor, effectively capturing the intersection of urban anthropogenic emissions and regional atmospheric diffusion processes to yield a representative profile of the local atmospheric environment.

The foundational dataset comprises 8,760 hourly observations spanning the entirety of the calendar year from January 1, 2021, to December 31, 2021. Target predictive variables include the mass concentrations of $PM_{2.5}$ and PM_{10}, standardized to units of $\mu g/m^3$. Predictor variables encompass continuous surface meteorological observations: ambient temperature measured in ℃, relative humidity expressed as a percentage, and wind speed recorded in m/s. All meteorological monitoring instruments comply with precision standards dictated by the World Meteorological Organization (WMO).

Continuous environmental monitoring data frequently exhibit localized missing values due to sensor calibration cycles, telemetric transmission failures, or extreme weather events. The raw dataset exhibited minor missing values in humidity and wind speed, alongside 89 and 78 missing values for $PM_{2.5}$ and PM_{10}, respectively, as detailed in Table 1.

Table 1 Descriptive statistics and missing value imputation summary.

Variable	Valid Values (Before)	Missing Values (Before)	Missing Values (After Imputation)
$PM_{2.5}$ ($\mu g/m^3$)	8669	89	0
PM_{10} ($\mu g/m^3$)	8680	78	0
Humidity (%)	8754	4	0
Wind Speed (m/s)	8743	15	0
Temperature (C)	8754	4	0

A multiple imputation strategy utilizing median filling was applied, consistent with the data processing framework in our previous work [5]. This methodology preserves the central tendency and statistical distribution of the time series, avoiding synthetic extreme values that could perturb gradient updates during backpropagation. To accelerate neural network convergence and prevent independent features with larger numerical magnitudes from dominating the loss function topography, all continuous variables were standardized, following standard normalization protocols commonly utilized in multidimensional energy and environmental assessments [6]. The standardized input features, projected into the [0, 1] interval, are calculated by Eq. (1):

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

where x denotes the raw observation, $x_{\min}$ and $x_{\max}$ represent the temporal minimum and maximum bounds respectively. For specific data segments demonstrating significant localized outliers corresponding to acute pollution spikes. To maintain signal fidelity against acute pollution spikes, the robust scaling transformation is defined by Eq. (2):

$$x_{robust} = \frac{x - Median(x)}{IQR(x)} \tag{2}$$

where $Median(x)$ designates the 50th percentile of the temporal sequence, and $IQR(x)$ signifies the statistical dispersion between the 75th and 25th percentiles.

2.2 LSTM Formulation for Temporal Dynamics

The temporal evolution of atmospheric particulate matter exhibits strong autocorrelation and profound dependency on antecedent meteorological states.The LSTM network architecture was deployed to inherently mitigate the vanishing and exploding gradient problems encountered in standard recurrent frameworks, a structural robustness heavily validated by recent advancements in counterfactual air quality modeling and complex environmental sequence tracking [7-8].The LSTM computational unit controls temporal information flow via three specialized, mathematically defined gates: the forget gate f_t, the input gate i_t, and the output gate o_t, as illustrated in Figure 1.

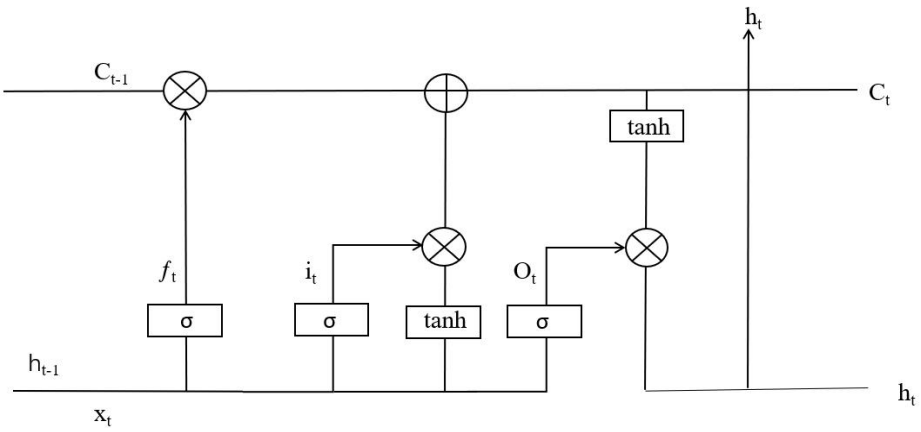

Figure 1 Internal Architecture and Gating Mechanisms of the LSTM cell

The temporal evolution and internal hidden state updates of the LSTM at time *t* are governed by Eqs. (3) - (7):

$$f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + b_f) \tag{3}$$

$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i) \tag{4}$$

$$g_t = \tanh(W_{gx}x_t + W_{gh}h_{t-1} + b_g) \tag{5}$$

$$s_t = g_t \odot i_t + s_{t-1} \odot f_t \tag{6}$$

$$h_t = o_t \odot \tanh(s_t) \tag{7}$$

where x_t denotes the multidimensional input vector encompassing meteorological parameters and historical particulate concentrations, h_{t-1} represents the prior hidden state mapped from the previous time step, s_t denotes the internal memory cell configuration, W and b signify the trainable weight matrices and bias vectors respectively, σ designates the standard logistic sigmoid activation function, and $\odot$ represents the element-wise Hadamard product.

The constructed deep learning architecture consists of three stacked LSTM layers, each instantiated with 32 hidden functional units. The initial layer captures short-term hourly meteorological fluctuations. Subsequent layers extract broader cross-diurnal cyclical patterns through cell state transmission mechanisms. To suppress network overfitting and prevent co-adaptation of hidden units, a Dropout layer with a retention probability complement of 0.2 follows each LSTM layer, randomly masking 20% of the neuronal connections during the forward training pass. The final multi-dimensional representation projects through a Dense fully connected layer utilizing a linear activation to output the unified predicted concentration vector mapping to a 1440×1 dimension.

The total number of trainable parameters for a single LSTM layer is determined by Eq. (8):

$$4 \times (n_{input} + n_{units} + 1) \times n_{units} \tag{8}$$

yielding a total parameter volume of 21,409 for the entire network. This configuration establishes an optimal equilibrium between feature learning capacity and computational efficiency. The objective function for model optimization, specifically the Mean Squared Error (MSE), is expressed by Eq. (9):

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{9}$$

where N denotes the total number of samples, y_i represents the actual observed particulate concentration, and $\hat{y}_i$ signifies the predicted value generated by the LSTM algorithm. Training was executed utilizing the Adam optimization algorithm initiating with a learning rate of 0.001. The model processed data in batches of 128 over 60 complete epochs.

2.3 Random Forest Normalization and Decoupling Mechanism

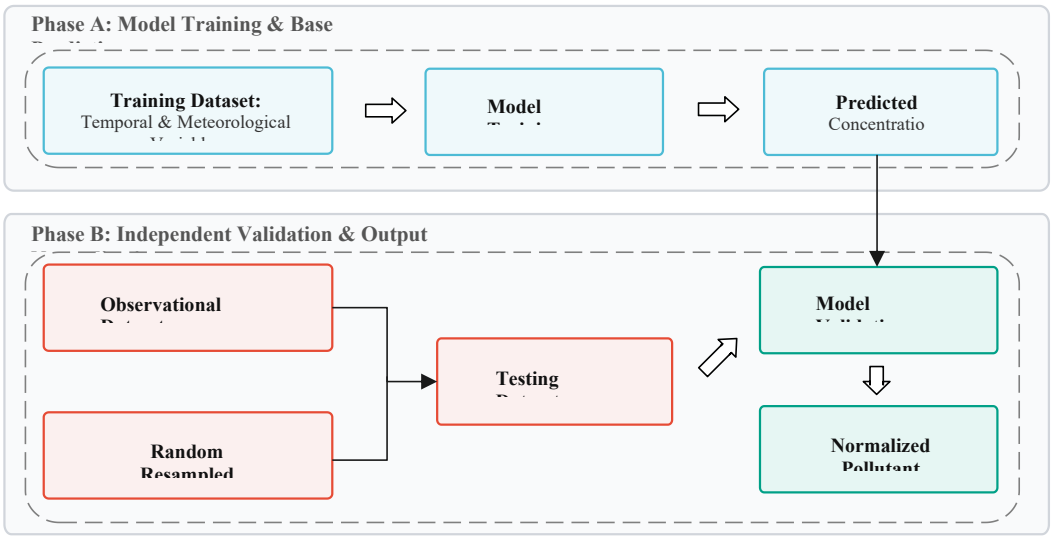

Figure 2 Conceptual Framework of the Random Forest Meteorological Decoupling Model

While the LSTM framework provides accurate predictive modeling of the observed state, isolating the precise quantitative impact of human activity requires mathematically removing the stochastic variance introduced by changing weather systems. A RF regressor executes this specific meteorological normalization, as depicted in Figure 2. The RF ensemble operates by predicting the target pollutant concentration $\hat{y}$ through averaging the output predictions of n_{tree} independently grown decision trees. During the training phase, each decision tree partitions the data by minimizing the variance at each internal node. The optimal feature split is determined by maximizing the reduction in Mean Squared Error, calculated by Eq. (10):

$$\Delta MSE = MSE_{parent} - \left(\frac{N_{left}}{N} MSE_{left} + \frac{N_{right}}{N} MSE_{right} \right) \tag{10}$$

where N designates the total number of samples at the parent node, while N_{left} and N_{right} represent the sample distributions in the subsequent child branches.

Central to this decoupling paradigm is the statistical reformulation of the expected pollutant concentration. In atmospheric modeling, isolating the emission-driven baseline requires marginalizing out the meteorological variables. The true meteorologically normalized concentration, representing the integral of the predictive model over the long-term probability density function of weather states, is mathematically defined by Eq. (11):

$$E[C \mid X_{anthro}] = \int \hat{f}(X_{meteo}, X_{anthro}) p(X_{meteo}) dX_{meteo} \tag{11}$$

where $E[C|X_{anthro}]$ denotes the expected concentration given fixed anthropogenic emissions, and $p(X_{meteo})$ represents the probability distribution of local meteorological conditions. Because this continuous integral is analytically intractable in high-dimensional weather spaces, the normalization technique resolves it via a Monte Carlo approximation. For a specified target date, relevant meteorological variables are randomly sampled from an established 14-day window (encompassing 7 days prior to and 7 days subsequent to the target date) across the historical observational record.

This random sampling procedure is computationally iterated 200 times per target vector. The sampled meteorological data arrays, combined with the target date's fixed temporal features (month, day of week, hour), are fed sequentially into the pre-trained RF model. By averaging the 200 distinct predictive outputs, the expected concentration, which effectively integrates out the random meteorological perturbations specific to that exact day, is calculated by Eq. (12):

$$C_{adjusted} \approx \frac{1}{M} \sum_{m=1}^{M} \hat{f}(X_{meteo}^{(m)}, X_{anthro}) \tag{12}$$

where $C_{adjusted}$ denotes the meteorologically normalized pollutant concentration, M represents the total number of resampling iterations (set to 200), $X_{meteo}^{(m)}$ signifies the m-th randomly sampled meteorological feature subset, and X_{anthro} designates the fixed temporal proxy variables representing anthropogenic activity.

The relative impact of weather, termed the Meteorological Contribution Rate (CR), is quantified by Eq. (13):

$$CR = \left(\frac{C_{adjusted} - C_{observed}}{C_{observed}} \right) \times 100\% \tag{13}$$

where $C_{observed}$ denotes the raw, unadjusted monitoring station data.

The RF architecture calculates feature importance by evaluating the reduction in impurity each feature contributes across all node splits in the ensemble:

$$Importance(X_j) = \frac{1}{T}\sum_{t=1}^{T}\sum_{v \in S_j(X_j)} \Delta I(v) \tag{14}$$

where T denotes the total number of trees, $S_j(X_j)$ designates the set of all nodes in tree t that split on feature X_j, and $\Delta I(v)$ represents the decrease in variance achieved by that specific node split.

3 RESULT AND DISCUSSION

3.1 Model Predictive Evaluation

The supervised LSTM network achieved stable convergence subsequent to 50 epochs of training. As shown in Figure 3, the training and validation loss curves descended smoothly and plateaued synchronously. This behavior demonstrates the absence of both underfitting and severe overfitting paradigms.

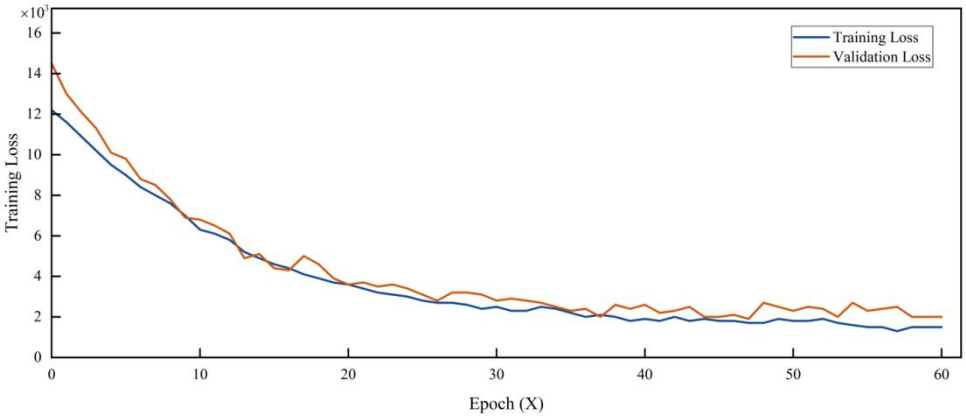

Figure 3 LSTM Training and Validation Loss Trajectories over 60 Epochs

The LSTM model replicated the non-linear coupling relationship between local weather patterns and particulate accumulation. The R^2 surpassed 0.80 across representative seasonal tracking months, as detailed in Table 2. This metric statistically validates the capability of the cell state gating mechanisms to handle long-range dependencies inherent in environmental time-series data.

Table 2 Seasonal R^2 for LSTM Predictions

Pollutant	February	May	September	December
$PM_{2.5}$	0.859	0.840	0.801	0.936
PM_{10}	0.878	0.871	0.864	0.910

Analytical review of the predictive vector outputs mapped against actual observation data reveals distinct seasonal amplitudes. During the winter phases of February and December, actual concentration values exhibited extreme outliers. For instance, in December, $PM_{2.5}$ concentrations peaked at approximately $200\mu g/m^3$ as illustrated in Figure 4. Concurrently, PM_{10} concentrations approached $300\mu g/m^3$ as visually confirmed , as shown in Figure 5. The prediction array for May captured an anomalous peak of $250\mu g/m^3$ for $PM_{2.5}$. This specific anomaly correlates mathematically with a period of exceptionally low wind speeds, reduced precipitation, and elevated regional agricultural biomass incineration. The September temporal matrix displayed the lowest systemic volatility and amplitude, reflecting stable baseline atmospheric dispersion and structurally lower anthropogenic heating emissions.

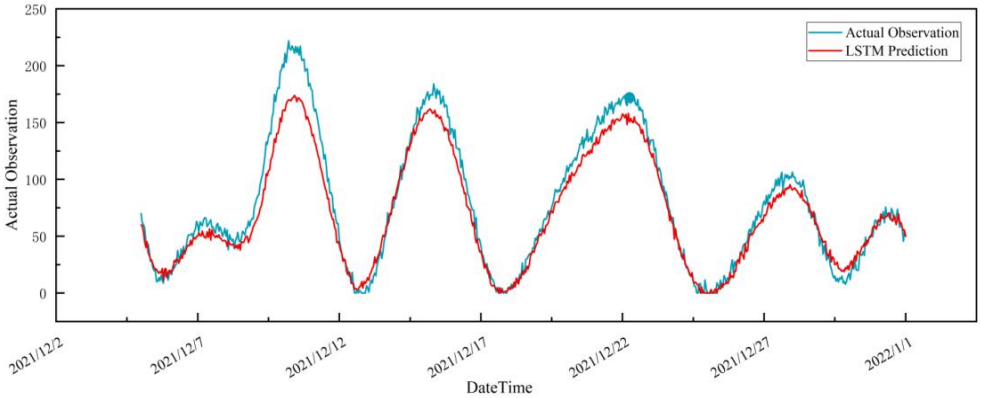

Figure 4 Comparison of LSTM Predictive Trajectories against Actual Observational Data for $PM_{2.5}$ in December

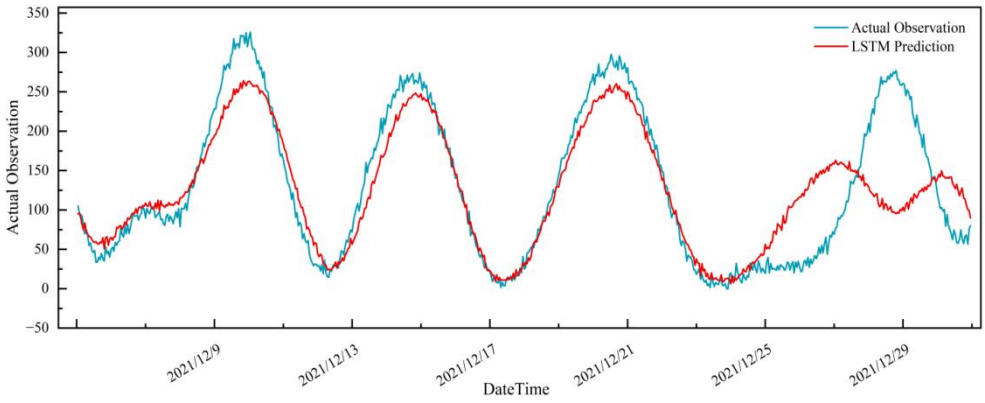

Figure 5 Comparison of LSTM Predictive Trajectories against Actual Observational Data for PM_{10} in December

3.2 Variable Importance and Meteorological Sensitivities

The RF regressor parameters were strictly optimized to contain 300 decision trees, restricted to a maximum depth of 20, establishing a minimum node split sample threshold of 5, and requiring a minimum leaf node size of 2. The ensemble model demonstrated stable generalization on the designated test set partition, yielding an R^2 of 0.87 alongside minimal Root Mean Square Error (RMSE) deviations, as verified in Table 3.

Table 3 Error Evaluation Metrics for the Random Forest Regressor on the Test Partition

Dataset	MSE	RMSE	MAE	R^2
Test Set	0.01	0.02	0.01	0.87

To quantitatively ascertain the hierarchy of environmental drivers dictating pollution severity, the internal RF feature importance metric was evaluated. The algorithmic output, illustrated in Figure 6, identifies the temporal "month" variable as having the most dominant statistical influence. This explicitly underscores the severe seasonal heterogeneity of localized pollution in Jinan.

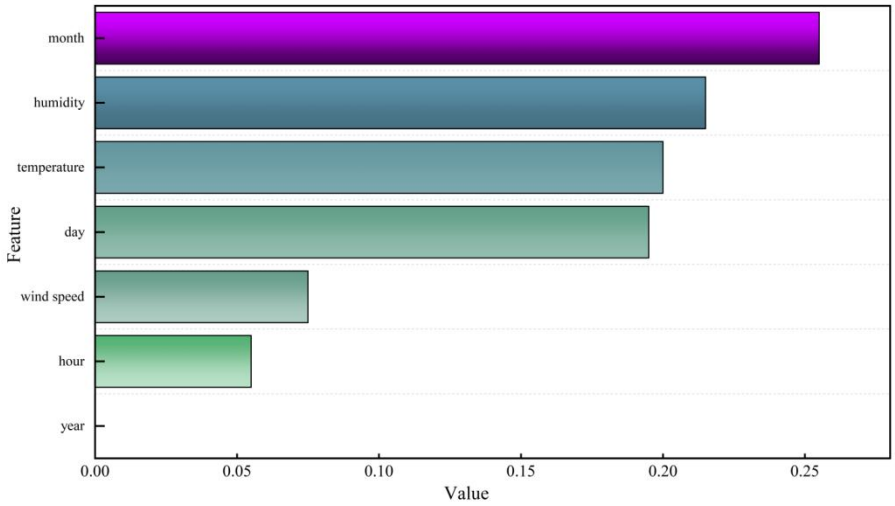

Figure 6 RF Feature Importance Matrix for Pollutant Concentration Drivers

Beyond sheer temporal variance, understanding these concentrations requires precise variable attribution. Among the continuous meteorological variables, relative humidity emerged as the most critical structural driver. Elevated moisture levels directly facilitate the hygroscopic growth of existing particulates. Concurrently, humid conditions accelerate the aqueous-phase oxidation of gaseous precursors,notably such as SO_2 and NO_x. This specific chemical process increases the secondary inorganic aerosol fraction within the total measured mass, driving up the $PM_{2.5}$ readings independently of primary emissions.

Temperature served as the secondary operational driver. Its influence primarily manifests through boundary layer dynamics rather than direct chemical catalysis. During winter months, localized surface cooling frequently induces temperature inversions. Such thermal stratification restricts the planetary boundary layer height. This physical limitation traps primary emissions near the surface, severely impeding vertical mixing and concentrating pollutants at ground-level monitoring stations.

Wind speed exhibited the lowest relative importance magnitude among the weather variables. Variance analysis indicates that Jinan's specific geographic topography restricts massive wind speed fluctuations. This results in a

distinctly low statistical variance that limits its proportional algorithmic influence on absolute concentration changes compared to the dominant hygroscopic effects of humidity.

3.3 Decoupled Contributions and Emission Dynamics

Two distinct, highly polluted temporal windows were isolated for the 14-day resampling normalization sequence: late winter (February 25–27) and early winter (December 21–23). These targeted periods represent varying baseline intensities of anthropogenic activity, contrasting Spring Festival socio-economic dynamics against centralized industrial winter heating outputs.

Subsequent to applying the RF conditional expectation normalization methodology, the R^2 values mapping the normalized concentrations against time approached absolute zero. This explicitly confirms mathematically that the stochastic meteorological interference was stripped from the observational time series. The residual variance remaining in the adjusted signal corresponds directly to localized emission-driven anthropogenic fluctuations.

The decoupled time-series data, summarized in Table 4, reveals a mathematically validated contrast in pollution formation mechanisms between the two distinct periods.

Table 4 Calculated Average Meteorological *CR*s during Target Analytical Windows

Target Analytical Window	$PM_{2.5}$ Meteorological *CR*	PM_{10} Meteorological *CR*
Early Winter (Dec 21 - Dec 23)	47.36%	41.14%
Late Winter (Feb 25 - Feb 27)	24.87%	27.26%

For the December 21–23 operational window, the actual observational $PM_{2.5}$ concentrations reached apexes of $168\mu g/m^3$. Post-normalization, the modeled expected concentration vector dropped sharply to $76\mu g/m^3$, yielding an average meteorological *CR* of 47.36% for $PM_{2.5}$ and 41.14% for PM_{10}. During this specific timeframe, micro-meteorological variables severely inhibited vertical and horizontal atmospheric diffusion. Specifically, wind speeds stabilized below 2 m/s combined with localized humidity exceeding 90%. The prevailing stagnation explicitly amplified the baseline anthropogenic emissions stemming from winter heating infrastructure, confirming that this severe pollution episode was primarily meteorologically modulated, as demonstrated in Figure 7.

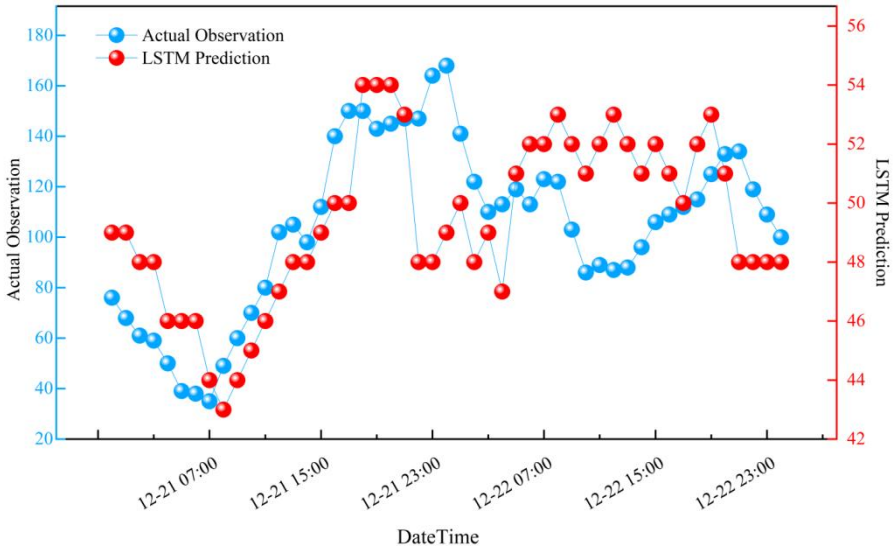

Figure 7 Dynamic Decoupling of PM2.5: Modeled Deweathered Baseline vs. Actual Observations during early Winter from December 21 to 23

Analyzing the February 25–27 normalization window presents a different thermodynamic profile. The average meteorological *CR* fell to 24.87% for $PM_{2.5}$ and 27.26% for PM_{10}. Analyzing February 25 specifically, the meteorological contribution was statistically negligible at merely 2.90%. The highly elevated particulate pollution levels observed during this specific multi-day sequence were driven predominantly by acute anthropogenic emissions—specifically, intensive fireworks displays associated with the Spring Festival combined with resuming localized industrial activity—rather than atmospheric stagnation, as shown in Figure 8.

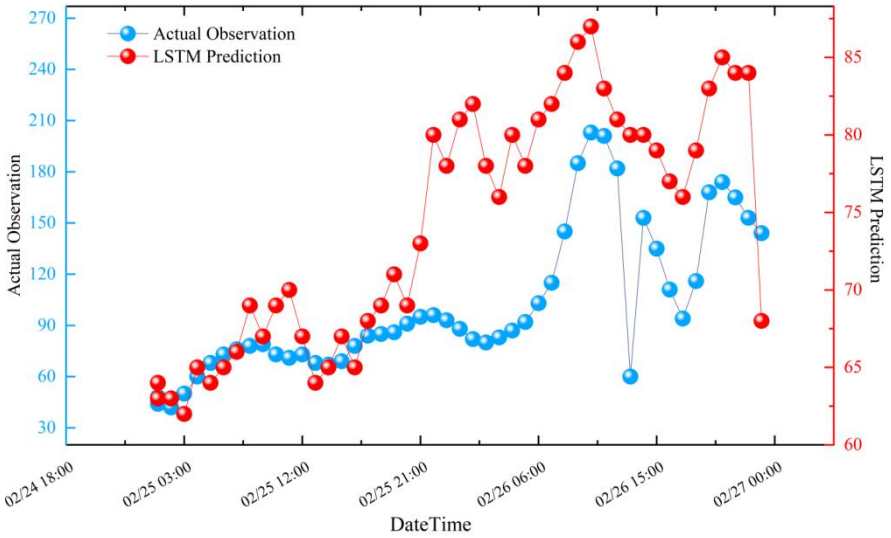

Figure 8 Dynamic Decoupling of PM2.5: Modeled Deweathered Baseline vs. Actual Observations during Late Winter from February 25 to 27

Detailed hourly tracking revealed that a sudden drop in regional wind speed on February 27 caused the meteorological *CR* to spike back to 48.05%. This demonstrates how rapidly shifting weather fronts dynamically interact with a high-emission anthropogenic baseline to trigger acute, localized pollution crises. This indicates that variations in meteorological conditions interact with baseline anthropogenic emissions to cause severe local pollution episodes. Therefore, decoupling these factors is necessary to reduce meteorological masking effects in environmental assessments [9-10].

4 CONCLUSION

This research employs a hybrid LSTM-RF framework to decouple anthropogenic emissions from meteorological variability. By capturing sequential dependencies, the LSTM architecture establishes a robust non-linear baseline, while randomized resampling isolates core emission signals. This dual-layered approach bypasses the rigid assumptions of deterministic modeling, offering an interpretable statistical mechanism for pollution attribution.

Empirical analysis in Jinan identified stark seasonal shifts. December pollution episodes were largely weather-driven, with a 47.36% contribution rate augmenting local heating effects. In contrast, February spikes remained primarily anthropogenic, as the weather-induced portion dropped to 24.87%. Feature evaluation confirmed that humidity and thermal inversions—rather than wind speed—serve as the primary drivers for particulate accumulation within this specific topography.

These insights support governance frameworks focused on humidity-sensitive thresholds to evaluate genuine policy impact. The framework's internal logic necessitates transitioning from single-station observations toward integrated spatial-temporal networks. Incorporating regional advection through graph architectures and assimilating satellite-derived boundary layer data represents the next objective trajectory. Bridging the gap between statistical inference and atmospheric physics remains essential for real-time emission calibration across wider geographical domains.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Zhao Y, Sun H, Ma D. Research on the impact of urban green finance reform on synergizing the reduction of pollution and carbon emissions. Industrial Economics Research, 2024(3): 15-28.

[2] Yu W, Zhang T, Shen D. Evolution of spatial pattern and influencing factors of county-level carbon emission intensity in China based on random forest model. China Environmental Science, 2022, 42(6): 2788-2798.

[3] Tao Y, Du J. Temperature prediction using long short-term memory network based on random forest. Computer Engineering and Design, 2019, 40(3): 737-743.

[4] Li C, Yin Y, Cui W. Study on carbon emission prediction and its interaction with green finance index. Statistical Theory and Practice, 2023(12): 18-26.

[5] Gao H, Wang H, Dong Z. Intelligent drive and enterprise green innovation: a quasi-natural experiment based on the national artificial intelligence pilot zone. Journal of Jiangnan University (Humanities and Social Sciences), 2024, 23(6): 55-69.

[6] Zhang H, Deng H. Photovoltaic evaluation in buildings integrally based on entropy weight and grey relational analysis// 2025 8th International Conference on Power and Energy Applications (ICPEA). 2025: 526-531.
[7] Liang W, Li Y, Liu X, et al. AI-based Bayesian structural time series modeling for assessing PM2.5 air quality improvements during the Beijing 2022 Winter Olympics. Atmospheric Environment, 2025, 358: 121328.
[8] Zhao J, Deng F, Cai Y, et al. Long short-term memory - fully connected (LSTM-FC) neural network for PM2.5 concentration prediction. Chemosphere, 2019, 220: 486-492.
[9] Chen M, Xu P, Liu Z, et al. Air pollution prediction based on optimized deep learning neural networks: PSO-LSTM. Atmospheric Pollution Research, 2025, 16(3): 102413.
[10] Huang G, Li X, Wu D. PM2.5 concentration forecasting at surface monitoring sites using GRU neural network based on empirical mode decomposition. Science of the Total Environment, 2021, 768: 144516.