https://doi.org/10.5194/egusphere-2026-1103
Preprint. Discussion started: 27 April 2026

A Hybrid STL-Ensemble Framework for Multivariate Time-Series Forecasting of Source-Specific PM2.5 Emissions

Jintu Borah[1], Deepali Kushwaha[2], Sachchida Nand Tripathi[3], and Rajesh M. Hegde[2, 3]

[1]Airawat Research Foundation, IIT Kanpur, India
[2]Department of Electrical Engineering, IIT Kanpur, India
[3]Kotak School of Sustainability, IIT Kanpur, India

Correspondence: Jintu Borah (jintub64@gmail.com) and Deepali Kushwaha (deepkush@iitk.ac.in)

Abstract. Forecasting the evolution of source-specific particulate emissions is central to modern air quality management strategies. Existing source identification methods result in non-uniqueness and instability in source profiles, leading to uncertainties in source identification and quantification. In this work, we present an approach that integrates receptor modeling with supervised machine learning to overcome this limitation. The hybrid model integrates statistical decomposition, feature-engineered multivariate learning, and ensemble regression techniques to predict the temporal trajectory of PM2.5 source contributions. The concentrations of elemental and organic species from high-resolution measurement systems were processed through source apportionment to identify the target sources. A time-series pipeline was developed, including temporal imputation, autocorrelation-guided feature engineering, Seasonal-Trend Decomposition using LOESS (STL), and multi-output ensemble regression. The proposed method demonstrated improved predictive performance across diverse emission categories, highlighting the importance of decomposition for interpretability and providing a robust foundation for the operational forecasting of air quality dynamics. Compared to the source-specific PM2.5 emission forecasting without STL, the proposed method is able to improve the R2 score from 0.22 to 0.95 in aggregate. The proposed comprehensive modeling framework is robust and can be adapted to various multi-source environmental datasets.

1 Introduction

Fine particulate matter (PM2.5) is one of the most critical pollutants affecting human exposure, public health, and urban environmental management. Owing to its small aerodynamic diameter ($\leq 2.5\ \mu$m), PM2.5 can penetrate deep into the alveolar region of the respiratory tract and translocate into the bloodstream, thereby contributing to cardiopulmonary morbidity and premature mortality. In addition to the total mass concentration, the physicochemical composition of PM2.5 governs its toxicity, oxidative potential, and atmospheric lifetime. Its composition varies significantly over space and time under the influence of heterogeneous emission sources, including vehicular exhaust, industrial combustion, biomass burning, crustal re-suspension, and secondary aerosol formation through gas-to-particle conversion mechanisms (Bhandari et al., 2020; Bukhari et al., 2022; Alfoldy et al., 2023). Meteorological drivers, such as boundary layer height, relative humidity, temperature, and wind dynamics, further modulate dispersion and chemical transformation processes, resulting in complex spatiotemporal patterns.

https://doi.org/10.5194/egusphere-2026-1103
Preprint. Discussion started: 27 April 2026

Source-attribution techniques, including receptor models and factor analysis approaches (Dinh et al., 2025; Zografou et al., 2022), have been widely employed to deconvolve observed pollutant species into interpretable source factors. These methods effectively reconstruct historical contributions from sources and provide actionable insights into regulatory interventions. However, the predictive dimension, specifically forecasting future source-specific contributions, remains relatively unexplored. Anticipating the evolution of source-specific emissions is essential for proactive air quality management, exposure mitigation strategies, and scenario-based policy evaluation.

Traditional forecasting approaches typically treat the aggregated PM2.5 mass as a univariate time series, implicitly assuming homogeneous emission dynamics (Antad et al., 2025). Such formulations neglect inter-species correlations, cross-source interactions, and nonlinear dependencies among precursor gases, secondary products, and environmental variables (Cohen et al., 2010; Bousiotis et al., 2022). This simplification limits the ability to interpret and attribute forecast changes to specific emission categories. In contrast, multivariate forecasting frameworks explicitly model the relationships between correlated chemical species and derived source factors (Zhao et al., 2025a; Mei et al., 2025). Using cross-covariance structures and joint temporal dependencies, these approaches enhance predictive robustness and provide greater explanatory value (Bousiotis et al., 2021).

Motivated by these considerations, this study proposes a hybrid multivariate forecasting pipeline that captures temporal dynamics, nonlinear interactions, and multi-source dependencies. The framework integrates statistical signal decomposition to isolate temporal modes with ensemble learners to predict individual components. By combining decomposition-driven feature extraction with ensemble-based regression, the system aims to improve generalization under non-stationary atmospheric conditions. The proposed methodology ultimately seeks to generate reliable, source-resolved emission forecasts from observed pollutant species and their apportionment-derived factors, thereby supporting data-driven environmental management and planned policy design.

2 Literature Review

Existing research on PM2.5 modeling typically spans three major domains: (i) source apportionment, (ii) univariate or multivariate air quality forecasting, and (iii) statistical decomposition and hybrid machine learning methodologies. Each domain addresses a distinct component of the atmospheric modeling pipeline, ranging from source identification and quantification to temporal prediction and structural signal characterization. Despite substantial progress in each stream, their integration into source-resolved predictive modeling remains limited.

Source apportionment traditionally relies on receptor-oriented frameworks that separate mixed atmospheric samples into constituent source categories based on their measured chemical compositions. These approaches assume that ambient particulate matter is a linear combination of multiple source profiles, with contributions that vary temporally but maintain chemically consistent fingerprint patterns. Among the most widely adopted techniques is Positive Matrix Factorization (PMF) (Chauhan et al., 2025), a multivariate factor-analytic method that constrains solutions to non-negative source contributions and profiles. PMF is particularly advantageous for high-dimensional datasets because it accommodates measurement uncertainties and missing values while minimizing residual errors through weighted least squares optimization. Additional receptor-based strategies

https://doi.org/10.5194/egusphere-2026-1103
Preprint. Discussion started: 27 April 2026

include exploratory factor analysis (Madukpe et al., 2025; Das, 2025), principal component regression (Rabhi et al., 2025), and constrained chemical mass balance formulations (Ryoo et al., 2025; Zhao et al., 2025b; Edwards, 2025). These methods have demonstrated effectiveness in handling multi-species datasets derived from advanced instrumentation platforms.

Empirical studies have consistently shown that chemically speciated measurements obtained using instruments such as aerosol mass spectrometers (AMS) and Xact elemental analyzers provide sufficient chemical resolution to disentangle primary and secondary sources (Chakraborty et al., 2025). Organic aerosol fractions, elemental tracers, and ion fragments facilitate the identification of traffic-related emissions, biomass burning contributions, industrial combustion, crustal dust, and various secondary aerosol formation pathways. Secondary processes, including photochemical oxidation and heterogeneous reactions, introduce additional complexity by generating regionally transported oxidized components (Manousakas et al., 2025). Although receptor models (Zhu et al., 2018) effectively reconstruct historical source contributions and reveal dominant emission categories, they are inherently retrospective. Their primary objective is explanatory, rather than predictive, limiting their direct application in forward-looking environmental management scenarios.

Parallel to source apportionment, air quality forecasting has evolved from deterministic and linear statistical formulations to sophisticated data-driven architectures (Morain et al., 2025). Early approaches primarily relied on autoregressive integrated moving average models (Song et al., 2025), state-space representations (Muhammad et al., 2026), and Kalman filtering frameworks (Lawrence and Bhathmanabhan, 2026). These techniques assume linear temporal dependence and stationarity, offering interpretability but often failing to capture nonlinear atmospheric interactions. With the growth of computational resources, nonlinear regression models and artificial neural networks have been increasingly adopted to model complex pollutant dynamics (Jin et al., 2026; Tongkhow and Kantanantha, 2012). In particular, feedforward neural networks, recurrent neural networks, and long short-term memory architectures have shown improved performance in modeling lagged dependencies, seasonal patterns, and abrupt episodic variations in time series data.

Tree-based ensemble learners, including Random Forests and gradient boosting machines (Shukla and Pal, 2011; Cheng et al., 2025), have further enhanced predictive capability by aggregating multiple weak learners to reduce variance and bias. These algorithms are particularly robust in high-dimensional contexts, accommodating correlated predictors and nonlinear feature interactions without making strict parametric assumptions. However, a substantial portion of the forecasting literature focuses on predicting aggregated PM2.5 mass concentrations rather than decomposed source contributions. This aggregation obscures the heterogeneity of the emission processes and diminishes the interpretability from a regulatory perspective. Studies have indicated that multivariate forecasting frameworks outperform univariate approaches when pollutant species exhibit strong correlation structures, cross-dependencies, and shared meteorological drivers (Gao et al., 2022; Mohapatra et al., 2026). By jointly modeling multiple variables, multivariate systems exploit covariance structures and reduce the information loss inherent in scalar aggregation.

Another domain of PM2.5 modeling methods encompasses statistical decomposition and hybrid machine learning methodologies designed to address nonstationarity and multi-scale temporal variability (Yuan et al., 2025; Diapouli et al., 2017; Liu et al., 2025). Atmospheric time series data typically exhibit seasonal cycles, long-term trends, episodic spikes, and stochastic noise components. Directly training predictive models on raw data may obscure the underlying structures and degrade

https://doi.org/10.5194/egusphere-2026-1103
Preprint. Discussion started: 27 April 2026

generalization. Decomposition techniques aim to isolate interpretable temporal modes before forecasting. Seasonal-trend decomposition using locally estimated scatter-plot smoothing (STL) has emerged as a flexible nonparametric approach for separating trend, seasonal, and remainder components (Zhou et al., 2025; Li et al., 2025). STL is particularly advantageous because it accommodates nonlinear seasonal variations and is robust to outliers. Alternative decomposition strategies include wavelet transforms (Qiao et al., 2021), empirical mode decomposition (Lahmiri, 2017), and variational mode decompositions (Dragomiretskiy and Zosso, 2014). These methods operate in either the time-frequency or intrinsic mode domains, enabling the extraction of oscillatory components at different temporal scales. Once decomposed, individual components can be modeled using tailored predictive algorithms, thereby reducing the complexity and enhancing the interpretability (Sharma et al., 2025). Hybrid frameworks combine decomposition with machine learning predictors, capitalizing on their complementary strengths. For example, low-frequency trends can be captured by linear or ensemble regressors, whereas high-frequency residuals can benefit from nonlinear learners. Again, ensemble learning techniques have demonstrated strong performance in handling nonlinearity, high-dimensional feature spaces, and complex variable interactions (Zhou et al., 2024). Random Forests mitigate overfitting through bootstrap aggregation and randomized feature selection. Gradient boosting sequentially refines the residual errors, thereby improving the predictive accuracy. Multi-output regression architectures extend these benefits to simultaneously forecast multiple correlated targets, preserving inter-source relationships. These ensemble systems also provide variable-importance metrics, supporting interpretability and feature relevance assessment.

Despite these advances, few studies have explicitly integrated source apportionment outputs with decomposition-based multivariate forecasting in a unified pipeline. Most research efforts treat these domains independently, either focusing on retrospective source identification or aggregate-concentration prediction. The integration of STL-based temporal decomposition with ensemble-driven multi-output regression offers a promising pathway to bridge this gap. By decomposing source-resolved time series data into structured components and leveraging ensemble predictors for each component, the framework can address nonstationarity, preserve cross-source dependencies, and enhance predictive reliability. Therefore, this study adopted an STL-ensemble architecture tailored for source-specific PM2.5 forecasting. This approach synthesizes receptor-derived source factors, multivariate temporal modeling, and ensemble-based regression, within a coherent predictive framework. The objective is not merely to forecast the total particulate mass but to anticipate the temporal evolution of individual emission categories, thereby supporting targeted intervention strategies, exposure risk assessment, and evidence-based environmental policy design.

3 Dataset

This section describes the AMS and Xact datasets used for source-specific PM2.5 emission forecasting. It also outlines the methodology employed to detect missing values and prepare the data for training machine learning models.

3.1 Data Collection and Cleaning

The dataset comprised high-resolution aerosol measurements from three monitoring sites representing urban, peri-urban, and institutional settings across two observation cycles. The CSIR (urban), Talkatora (peri-urban), and BBAU (institutional) mon-

https://doi.org/10.5194/egusphere-2026-1103
Preprint. Discussion started: 27 April 2026

Table 1. Description of the dataset and missing window percentages.

Sensor	Cycle	Site	Sampling Period	Samples	Missing data (%)
AMS	C-I	CSIR	29 Sep. - 9 Oct. 2023	402	10
		BBAU	12 - 23 Oct. 2023	519	0
		Talkatora	8 - 19 Nov. 2023	499	0
	C-II	CSIR	19 - 29 Feb. 2024	341	40
		BBAU	18 - 30 Mar. 2024	406	7
		Talkatora	2 - 20 Apr. 2024	404	34
Xact	C-I	CSIR	30 Sep. - 8 Oct. 2023	373	2
		BBAU	12 - 23 Oct. 2023	507	3
		Talkatora	5 - 19 Nov. 2023	633	0
	C-II	CSIR	18 - 29 Feb. 2024	500	0
		BBAU	18 - 31 Mar. 2024	467	0
		Talkatora	29 Mar. - 20 Apr. 2024	581	0

itoring sites are located within Lucknow (a city in the state of Uttar Pradesh, India). The first-cycle data were recorded in September to November of 2023, and the second cycle was recorded in February to April of 2024. The data captured a variety of meteorological and anthropogenic conditions at approximately 30-minute intervals, as real-world data are subject to irregular sampling due to operational interruptions.

Each cycle contains organic and inorganic species measured by the AMS systems, and trace elemental concentrations recorded by Xact analyzers. The combined dataset provided a multi-species representation of PM2.5 composition. Using the speciated AMS and Xact datasets, source apportionment was performed to identify characteristic sources, including traffic emissions, dust, biomass burning, e-waste incineration, and regional oxidized aerosol. Each factor consists of a profile vector and a time-varying contribution signal. The time series data served as the prediction targets for the forecasting models.

To ensure temporal consistency, all species and source-apportioned data were aligned with t_i, the timestamp of the i-th sample on a fixed grid $\Delta t = 30$ minutes:

$$t_i = i\Delta t. \tag{1}$$

Missing timestamps were inserted, and the concentration gaps were linearly interpolated. Before modeling, irregularities such as abrupt spikes, negative values, and sensor dropouts were removed using the local sliding-window logic. This ensured that the final dataset exhibited stationary characteristics consistent with multivariate time-series modeling. This resulted in missing windows in the input sequences. Missing periods were distributed throughout the dataset. The aggregated percentage is presented in Table 1.

https://doi.org/10.5194/egusphere-2026-1103
Preprint. Discussion started: 27 April 2026

Table 2. Dominant sectors identified from the sensing unit across the dataset.

Sensor	Species	Emmission Sectors	Sector code
AMS	Organics, SO_4, NH_4, NO_3, Cl^-	Urban Oxidized	A-1
		Traffic	A-2
		Regional Oxidized	A-3
		Oxidized Biomass Burning	A-4
		Primary Biomass Burning	A-5
Xact	Al, Si, P, K, Ca, Sc, Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, Ge, As, Se, Br, Rb, Sr, Y, Zr, Mo, Ag, Cd, In, Sn, Sb, Te, Cs, Ba, La, Ce, Au, Hg, Tl, Pb, Bi	Lead Recycling	X-1
		Informal Waste Burning	X-2
		E-Waste incineration	X-3
		Solid-fuel combustion	X-4
		Power-plants	X-5
		Dust	X-6
		Tyre wear	X-7
		Refuse Burning	X-8
		Industrial	X-9
		Ferrous smelting	X-10
		Fireworks	X-11
		Coal combustion	X-12

3.2 Training Data Generation

Source apportionment was performed using receptor-based models to identify and quantify the major sources of particulate matter using high-time-resolution chemical composition data. Positive Matrix Factorization (PMF) and Chemical Mass Balance (CMB) techniques were applied to speciated organic compound data from the AMS and elemental concentration time series measured by the Xact multi-metals monitor. PMF decomposes the measured concentration matrix into factor profiles and their temporal contributions, subject to non-negativity constraints, enabling the identification of physically interpretable sources without prior assumptions. CMB is used selectively where reliable source profiles were available to complement and validate the PMF results.

PMF analysis of AMS-derived compounds resolved five organic aerosol source factors: traffic, oxidized Biomass Burning, regional oxidized, urban oxidized, and primary biomass burning, representing both primary emissions and secondary formation processes. Elemental source apportionment using Xact data identified multiple industrial, combustion-related, and crustal sources, including power plants, lead recycling, ferrous smelting, e-waste incineration, informal and refuse waste burning,

https://doi.org/10.5194/egusphere-2026-1103
Preprint. Discussion started: 27 April 2026

solid-fuel combustion, dust resuspension, and firework emissions. The combined use of AMS- and Xact-based source apportionment provided a comprehensive characterization of the dominant anthropogenic and natural sources influencing ambient particulate matter. The sectors identified are shown in Table 2.

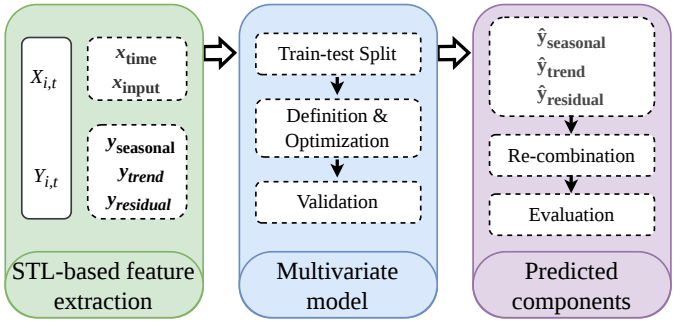

Figure 1. Flow diagram for the proposed STL-ensemble framework.

4 Feature extraction using STL

This section presents the input feature enhancement method, its credibility, and the model implementation. The flow of the proposed work is depicted in Algorithm 1 and Fig. 1.

4.1 Autocorrelation in Individual Sequence

Temporal feature engineering plays a critical role in capturing interdependencies and long-range dynamics. Autocorrelation is a key property of time-series data, measuring the degree of linear dependence between observations separated by a given time lag. For a discrete sequence $\{x_t\}_{t=1}^{T}$ of length T, the autocorrelation (ρ) at lag τ is defined as

$$\rho(\tau) = \frac{\sum_{t=1}^{T-\tau}(x_t - \mu)(x_{t+\tau} - \mu)}{\sum_{t=1}^{T}(x_t - \mu)^2}, \tag{2}$$

where μ denotes the mean of the series.

In this study, significant autocorrelation was consistently observed across all analyzed sequences. Figure 2 illustrates the autocorrelation functions computed for four representative source time series, and similar structures were identified for the remaining sequences. In these plots, two confidence ranges are typically shown. The inner range is bounded by solid lines, while the outer range is indicated by dashed lines. The solid-line bounds lie entirely within the dashed-line limits, representing a stricter confidence interval compared to the wider dashed-line boundary. The PM2.5 data exhibit pronounced and regularly spaced crests at approximately the 48th lag, followed by secondary peaks at the 96th lag and subsequent multiples. Given

https://doi.org/10.5194/egusphere-2026-1103
Preprint. Discussion started: 27 April 2026

Algorithm 1 Expanded STL-Ensemble Forecasting Procedure

Input: Multivariate input $\mathbf{x}_{1:T}$, target $\mathbf{y}_{1:T}$, lag k, window W, period P, forecast horizon Δ
Output: Forecasted source contributions $\hat{\mathbf{y}}_{T+1:T+\Delta}$
Step 1: Data Preprocessing
Remove outliers using threshold filtering
Fill missing timestamps via linear interpolation
Normalize each source using min-max scaling
Step 2: STL Decomposition
for each source $s = 1$ to S **do**
    Decompose $y_t^{(s)}$ into $\left(T_t^{(s)}, S_t^{(s)}, R_t^{(s)}\right)$
end for
Step 3: Feature Construction
for $t = k$ to $T - \Delta$ **do**
    Construct lagged vector $\mathbf{x}_{t-k:t}$
    Encode seasonality using $\sin\left(\frac{2\pi t}{P}\right)$ and $\cos\left(\frac{2\pi t}{P}\right)$
    Form feature vector $\mathbf{z}_t = \left[\mathbf{x}_{t-k:t}, \sin\left(\frac{2\pi t}{P}\right), \cos\left(\frac{2\pi t}{P}\right)\right]$
end for
Split dataset into training and validation sets using final window of length W for validation
Step 4: Ensemble Model Training
Define candidate models $\{\mathcal{F}_1, \mathcal{F}_2, ..., \mathcal{F}_M\}$
for each model $\mathcal{F}_m$ **do**
    Train $\mathcal{F}_m$ to predict $(T_{t+\Delta}, S_{t+\Delta}, R_{t+\Delta})$ from $\mathbf{z}_t$
    Compute validation error (RMSE)
end for
Select best-performing model $\mathcal{F}^*$ with minimum validation error
Step 5: Multi-step Forecasting
for $h = 1$ to Δ **do**
    Construct feature vector $\mathbf{z}_{T+h-\Delta}$
    Predict STL components: $\hat{T}_{T+h}, \hat{S}_{T+h}, \hat{R}_{T+h} = \mathcal{F}^*\left(\mathbf{z}_{T+h-\Delta}\right)$
    Reconstruct forecast: $\hat{y}_{T+h} = \hat{T}_{T+h} + \hat{S}_{T+h} + \hat{R}_{T+h}$
end for
Return $\hat{\mathbf{y}}_{T+1:T+\Delta}$

the sampling resolution of two observations per hour, these lags correspond to periodicities of 24 hours and its harmonics, clearly indicating strong diurnal cycles embedded in the data. The persistence of these peaks demonstrates that daily activity patterns, such as traffic intensity, residential energy use, and boundary-layer dynamics, exert a dominant influence on the temporal evolution of source contributions. Moreover, the slow decay of autocorrelation between successive peaks suggests

https://doi.org/10.5194/egusphere-2026-1103
Preprint. Discussion started: 27 April 2026

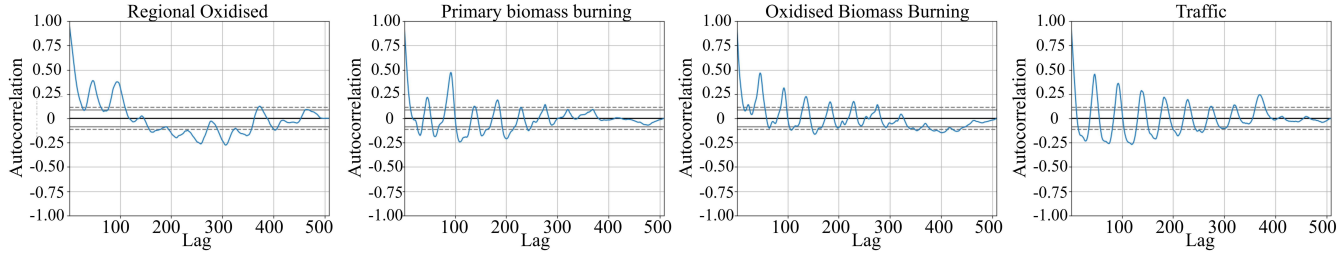

Figure 2. Autocorrelation plot observed from the sequences in both cycles.

that these cycles are not isolated events but are reinforced over time, leading to structured temporal dependence rather than purely stochastic behavior.

The identification of such regular periodicity has important modeling implications. Standard regression approaches that do not explicitly account for temporal cycles may fail to capture this recurring structure, resulting in systematic bias and reduced predictive accuracy. Consequently, explicit time-cycle encoding becomes a necessary component of the forecasting framework. By representing time using sinusoidal functions with periods aligned with observed diurnal cycles, the model can internalize daily repetition patterns and improve its ability to anticipate recurring emission behavior.

Given strong autocorrelation within several source categories, lagged features were incorporated:

$$\boldsymbol{X}_{\text{lag}}(t) = [x(t-1), x(t-2), \ldots, x(t-p)], \tag{3}$$

where p was chosen based on autocorrelation decay and partial autocorrelation structure. This allows the models to leverage historical data for future prediction. Following this, rolling-window averages and standard deviations were computed to capture short-term variability:

$$\mu_t = \frac{1}{W} \sum_{i=t-W}^{t} x_i, \quad \sigma_t = \sqrt{\frac{1}{W} \sum_{i=t-W}^{t} (x_i - \mu_t)^2}, \tag{4}$$

where W is the window length.

These features represent local smoothness and volatility, which are important for emission sources that fluctuate with distinct temporal patterns. Temporal cycles were incorporated using trigonometric encodings for daily and weekly periodicities:

$$\sin\left(\frac{2\pi t}{P}\right), \quad \cos\left(\frac{2\pi t}{P}\right), \tag{5}$$

where P represents the number of steps per cycle, such encodings are effective at modeling recurring diurnal traffic peaks and seasonal biomass-burning cycles.

4.2 STL decomposition

Each target source contribution time series is decomposed into trend, seasonal component, and residual components:

$$y_t = T_t + S_t + R_t, \tag{6}$$

https://doi.org/10.5194/egusphere-2026-1103
Preprint. Discussion started: 27 April 2026

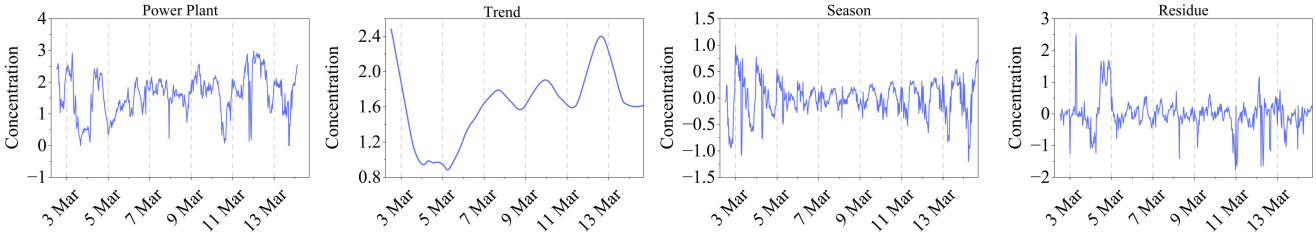

Figure 3. Decomposed components of Power Plant source sequence.

Table 3. Hyperparameters space for finding optimum values using grid search.

Parameter	Random Forest	XGBoost	LightGBM	Step
n_estimator	100 - 800	100 - 800	100 - 800	50
max_depth	3 - 30	3 - 10	-	1
min_sample_split	2 - 10	-	-	1
min_sample_leaf	1 - 6	-	-	1
learning_rate	-	0.01 - 0.3	0.01 - 0.3	0.02
subsample	-	0.6 - 0.4	0.6 - 0.4	0.1
colsample_bytree	-	0.6 - 0.4	0.6 - 0.4	0.1
num_leaves	-	-	16 - 128	8

where T_t is the trend, S_t is the seasonal component, and R_t is the residual. STL decomposition enables the forecasting models to focus on structure-specific patterns rather than a general-purpose signal. Trends typically capture long-term drift, seasonal components represent periodic variations, and residuals capture short-term irregularities. The prediction model learns a mapping:

$$\hat{Y}(t+\Delta) = f(X(t-k:t)), \tag{7}$$

where $f(\cdot)$ denotes a nonlinear regression function. Multiple ensemble regressor were evaluated, including tree-based ensembles such as Random Forest, gradient-boosting variants, regularized linear models, and multi-output regression wrappers. Each model was tuned using a predefined hyperparameter search to optimize performance. All components (trend, seasonal, and residual) across all target sources are predicted jointly:

$$\hat{\mathbf{y}}_\mathbf{i} = [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_q]. \tag{8}$$

where $\hat{y}_i$ represents output sequence at the i-th decomposition vector and q represents evaluation length. This formulation accounts for cross-source correlations arising from shared human activities and meteorological influences.

https://doi.org/10.5194/egusphere-2026-1103
Preprint. Discussion started: 27 April 2026

5 Performance Evaluation

This section presents the experimental setup for the proposed method, followed by the experimental results on comparison of the proposed forecasting with the ground truth.

5.1 Experimental Setup

The experimental evaluation is designed to reflect realistic forecasting conditions for multivariate air quality time series. All models were trained and tested using temporally ordered data to preserve causal structure and prevent information leakage. For each source-apportioned sequence, the dataset was divided into non-overlapping training and testing segments using a chronological split. The initial portion of the time series (80%) was used for model training and hyperparameter optimization, while the remaining portion was reserved exclusively for out-of-sample testing. A total length of No random shuffling was applied at any stage of the data partitioning, ensuring that predictions were always generated using only past information.

Model training and evaluation were conducted in a controlled simulation environment implemented in Python. The pipeline incorporated standard scientific computing and machine learning libraries for numerical processing, statistical decomposition, and ensemble regression. STL decomposition was applied independently to each target sequence prior to model training, and all feature engineering steps were executed using only historical observations within the training window. Hyperparameter tuning was performed within the training set using time-aware cross-validation, after which the selected model configuration was retrained on the full training data and evaluated on the held-out test set. The selection grid is shown in Table 3. All experiments were executed on a workstation-class system, and identical preprocessing, feature construction, and evaluation protocols were maintained across models to ensure a fair and reproducible comparison.

5.2 Experimental Results

5.2.1 Experiments on Prediction using STL Decomposition

The prediction of the PM2.5 over time using STL decomposition is compared with direct prediction and is shown in Fig. 4. The decomposition-based approach showed high predictive accuracy for trend and seasonal components, which generally varied smoothly over time. Residual components were more unpredictable due to their stochastic nature, yet still benefited from ensemble learning.

Table 4 shows the prediction performance metrics for each sector. The table shows how well the model predicted outcomes in the A and X sectors using two different configurations (C-I and C-II) and three different locations (L-I, L-II, and L-III). The results were measured using R^2 score, MAE, and RMSE. In general, the models are very good at making predictions, with R^2 values mostly between 0.97 and 0.99 for most combinations of sectors. This shows that the fit is good and the learning behavior is reliable. In most cases, the error magnitudes stay pretty low, especially in the X-sectors. Sources such as regional oxidized, biomass burning, dust, and e-waste incineration demonstrated strong forecast ability. Source categories dominated by anthropogenic cycles were particularly well modeled due to the effectiveness of cyclic features and decomposition. The

https://doi.org/10.5194/egusphere-2026-1103
Preprint. Discussion started: 27 April 2026

Table 4. Cross-sensor prediction performance over location and cycle variations.

Sector	C-I									C-II								
	L-I (BBAU)			L-II (CSIR-CIMAP)			L-III (Talkatora)			L-I (BBAU)			L-II (CSIR-CIMAP)			L-III (Talkatora)		
	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE
A-1	0.99	0.46	0.69	0.99	0.15	0.27	0.98	0.92	1.25	0.99	0.23	0.41	0.99	0.40	0.68	–	–	–
A-2	0.99	0.29	0.47	0.99	0.34	0.59	0.99	1.22	1.71	0.97	0.78	1.96	0.98	0.39	0.66	0.99	0.49	0.78
A-3	0.99	0.84	1.11	0.98	0.57	1.17	0.99	1.58	2.17	0.99	0.67	0.91	0.98	0.31	0.49	0.99	0.49	0.72
A-4	0.99	0.33	0.57	0.98	0.29	0.55	0.99	1.21	1.61	0.98	0.46	0.65	0.98	0.37	0.62	0.98	0.28	0.47
A-5	0.99	0.44	0.64	–	–	–	0.99	1.37	1.94	0.98	0.21	0.40	0.98	0.18	0.29	0.98	0.84	1.32
X-1	0.98	0.03	0.06	0.99	0.01	0.03	0.99	0.14	0.33	0.99	0.02	0.04	0.98	0.01	0.01	0.99	0.01	0.03
X-2	0.97	0.02	0.04	–	–	–	0.98	0.04	0.10	0.98	0.01	0.01	0.99	0.01	0.02	0.96	0.01	0.03
X-3	0.98	0.05	0.10	0.99	0.03	0.05	0.97	0.02	0.05	0.95	0.02	0.04	0.98	0.02	0.03	0.98	0.01	0.04
X-4	0.98	0.08	0.13	0.99	0.04	0.08	0.99	0.20	0.33	0.98	0.04	0.09	0.98	0.03	0.05	0.99	0.05	0.09
X-5	0.99	0.05	0.07	0.98	0.06	0.09	0.99	0.10	0.19	0.98	0.08	0.11	0.98	0.05	0.08	0.98	0.04	0.05
X-6	0.99	0.06	0.09	0.99	0.02	0.04	0.98	0.04	0.07	0.98	0.04	0.06	0.99	0.04	0.07	0.98	0.11	0.27
X-7	0.96	0.01	0.08	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
X-8	0.99	0.04	0.06	0.99	0.03	0.08	0.98	0.15	0.30	0.98	0.04	0.07	0.98	0.02	0.06	0.98	0.04	0.06
X-9	–	–	–	0.98	0.02	0.05	–	–	–	–	–	–	–	–	–	0.99	0.14	0.26
X-10	–	–	–	–	–	–	0.99	0.06	0.11	0.99	0.02	0.04	–	–	–	–	–	–
X-11	–	–	–	–	–	–	0.99	3.04	6.47	–	–	–	–	–	–	–	–	–
X-12	–	–	–	–	–	–	–			0.99	0.03	0.06	0.98	0.03	0.04	0.99	0.03	0.06

Xact sectors always have lower MAE and RMSE values than the A-sectors. This means they have smoother temporal behavior and stronger feature relevance. The Xact sector predictions have high R^2 values and smaller errors, which shows that these sectors are easier to predict. Some sectors, on the other hand, have missing results, which implies that there were not enough observations. An interesting case is X-11, where R^2 stays very high, but RMSE increases. This shows that the model is good at capturing the trend but has bigger absolute deviations, due to scaling.

Comparison with observed values indicated strong alignment and stable temporal evolution across test periods.

https://doi.org/10.5194/egusphere-2026-1103
Preprint. Discussion started: 27 April 2026

Table 5. Comparison of sectoral prediction performance metrics across sites aggregated across sources.

Model	Site	Sensor	With STL			Without STL		
			MAE	RMSE	R^2 Score	MAE	RMSE	R^2 Score
Random Forest	CSIR-CIMAP	AMS	1.24	1.67	0.94	6.83	10.11	0.22
		Xact	0.17	0.32	0.95	7.39	10.32	0.11
	BBAU	AMS	0.94	1.38	0.97	6.44	9.91	0.21
		Xact	0.11	0.16	0.96	0.20	1.94	0.18
	Talkatora	AMS	1.11	1.35	0.97	19.94	28.6	0.41
		Xact	0.59	1.26	0.96	2.50	4.11	0.31
XGBoost	CSIR-CIMAP	AMS	0.11	0.17	0.97	5.62	7.44	0.33
		Xact	0.17	0.33	0.94	2.78	4.33	0.17
	BBAU	AMS	0.95	1.23	0.90	6.31	10.18	0.16
		Xact	0.11	0.16	0.96	1.68	2.91	0.20
	Talkatora	AMS	1.50	2.05	0.96	18.80	22.40	0.07
		Xact	0.61	1.31	0.92	2.30	3.71	0.33
LightGBM	CSIR-CIMAP	AMS	0.62	0.42	0.97	6.45	8.91	0.21
		Xact	0.18	0.32	0.95	2.55	4.19	0.20
	BBAU	AMS	1.44	2.70	0.98	7.16	9.38	0.19
		Xact	0.13	0.19	0.92	3.10	4.77	0.26
	Talkatora	AMS	1.91	3.59	0.93	14.88	17.37	0.31
		Xact	0.21	0.60	0.97	3.49	5.56	0.11

5.2.2 Comparison of Prediction Models

The prediction of PM2.5 is performed using three machine learning models, Random Forrest, XGBoost, and LightGBM. Table 5 shows the performance metrics of all three models with and without STL decomposition across sites and datasets. The empirical results reveal interpretable patterns in model performance across sensors, sites, and source categories. A few salient observations follow. First, performance varies strongly with sensor type and scale. Errors for AMS-derived targets (MAE in the range $4-12$, RMSE up to 15) are substantially larger in absolute terms than those reported for Xact targets (MAE typically ≤ 1.5, RMSE ≤ 3). This difference reflects the disparate units and magnitudes of quantities measured by AMS (species mass

https://doi.org/10.5194/egusphere-2026-1103
Preprint. Discussion started: 27 April 2026

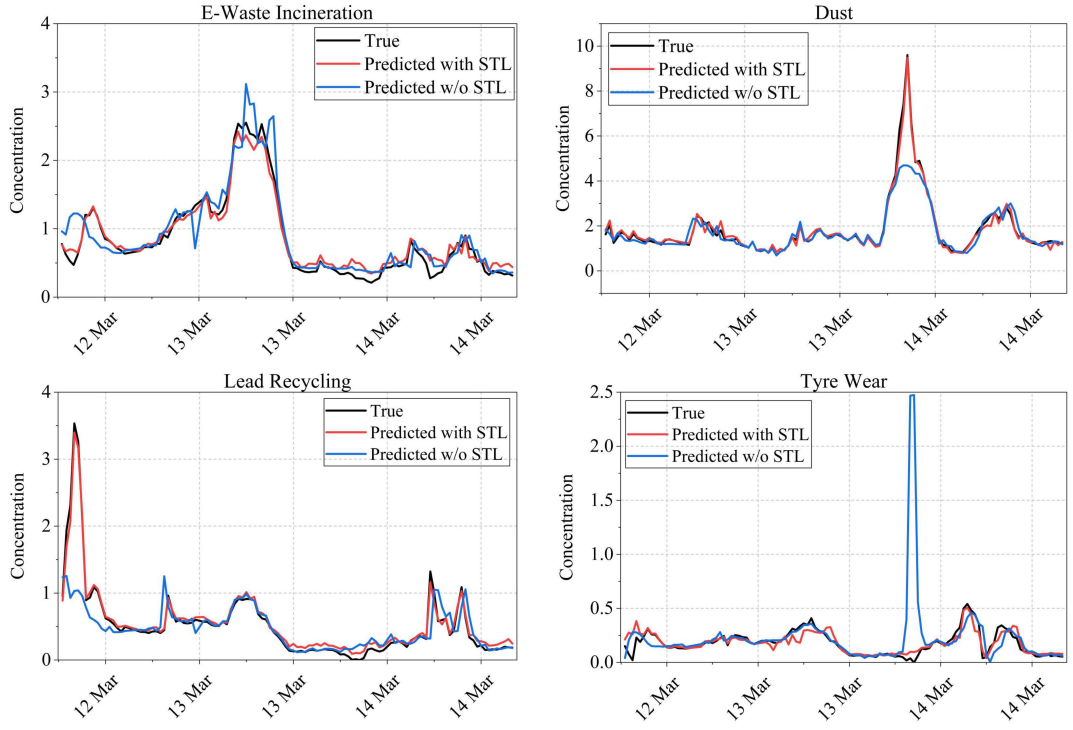

Figure 4. Predicted and actual sources in the Xact dataset in Cycle 2 after recombination.

concentrations) versus Xact (trace-element signals), and implies that direct comparison of absolute error magnitudes across sensors is not meaningful without normalization. Relative metrics (e.g., R^2) are therefore more informative for cross-sensor comparison.

Random Forest achieves high R^2 values at the Talkatora AMS site ($R^2 \approx 0.97$) and the Lalbagh Xact site ($R^2 \approx 0.96$), indicating that tree ensembles can capture local temporal structure at both sites. Conversely, XGBoost shows superior performance at CSIR-CIMAP for AMS-derived targets ($R^2 \approx 0.97$ versus Random Forest's $R^2 \approx 0.92$), suggesting that gradient-boosting variants may better exploit subtle nonlinearities at some sites. LightGBM exhibits intermediate behavior (e.g., $R^2 \approx 0.90$ for AMS-Talkatora), reinforcing the idea that no single learner dominates across all site-sensor combinations. These per-site or model differences argue for a model-selection or averaging strategy applied separately to each site and source component. Decomposition into trend, seasonal, and residual components materially aids interpretability and error attribution. Thus, for each component, the predictions with the highest individual R^2 score were selected. These selections were used to obtain the final predictions.

https://doi.org/10.5194/egusphere-2026-1103
Preprint. Discussion started: 27 April 2026

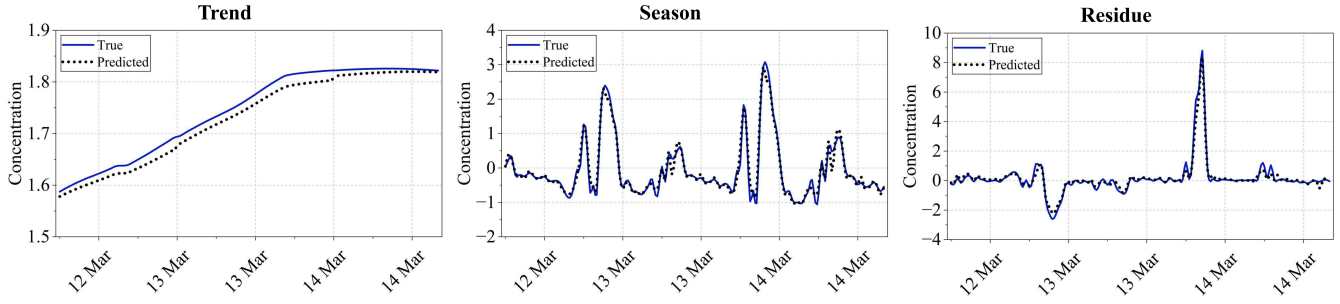

Figure 5. Predicted and actual components of Dust in the Xact C-II dataset.

5.2.3 Comparison of Temporal Prediction

The temporal variation of true and predicted PM2.5 emission for various sectors is shown in Fig. 4. Inspection of the component plots shows that trend and seasonal components are generally well reproduced: predicted trend and seasonal envelopes follow the observed low-frequency behaviour and recurring cycles. Residual components (high-frequency fluctuations and peaks), however, remain the principal source of forecast error; their reconstructed contribution propagates into the recombined series in Fig. 4 as occasional extreme shoots during sharp events. This aligns with the quantitative finding that component-level prediction yields higher accuracy for trend/seasonal parts than for residuals. Again, recombination performance demonstrates that, for well-modelled sources, the sum of predicted components produces close agreement with observations over test intervals. For sources dominated by anthropogenic cycles (e.g., traffic, certain oxidized components), cyclic features and STL-derived seasonality improve fidelity. For episodic sources (e.g., sudden biomass-burning plumes or short-lived industrial releases), residual variance increases and reconstructed peaks are underpredicted, indicating the need for additional high-frequency predictors (nowcasting inputs) or probabilistic models that explicitly model heavy-tailed residuals.

The proposed STL-ensemble framework offers advantages by providing a structured understanding of temporal behavior, thereby improving interpretability. Additionally, ensemble methods capture nonlinear and interactive dynamics across multiple predictors. Further, multi-output models leverage cross-source relationships for improved accuracy.

6 Conclusions

This study presents a hybrid modeling framework that integrates Seasonal-Trend decomposition using Loess (STL) to extract sectoral emission components from speciated datasets. The proposed architecture is trained and validated on three independently source-apportioned datasets. Prediction performance is systematically evaluated using multiple quantitative metrics and complementary visual analyses. The results demonstrate that incorporating STL decomposition within the hybrid modeling approach leads to a substantial improvement in predictive accuracy and robustness. However, the present work is constrained by the use of emission inventories derived solely from speciated data and region-specific variables, and therefore does not ac-

https://doi.org/10.5194/egusphere-2026-1103
Preprint. Discussion started: 27 April 2026

count for meteorological influences. Future research will focus on integrating region-specific meteorological and environmental parameters to further enhance the generalization capability and applicability of the proposed framework.

Code availability. The codes may be made available upon request.

Author contributions. Jintu Borah and Deepali Kushwaha developed the theoretical & experimental framework. Sachchida Nand Tripathi and Rajesh M. Hegde analyzed the results. All authors prepared and approved the final version.

Competing interests. The authors declare that they have no conflict of interest.

https://doi.org/10.5194/egusphere-2026-1103
Preprint. Discussion started: 27 April 2026

References

Alfoldy, B., Gregorič, A., Ivančič, M., Ježek, I., and Rigler, M.: Source apportionment of black carbon and combustion-related CO 2 for the determination of source-specific emission factors, Atmospheric Measurement Techniques, 16, 135–152, https://doi.org/10.5194/amt-16-135-2023, 2023.

Antad, K. Y., Chinmaye, N., and Pankaja, R.: AirQ with TGNN: Air Quality Prediction and Carbon Credit Management Using AI, in: 2025 International Conference on Computing Technologies (ICOCT), pp. 1–7, IEEE, https://doi.org/10.1109/ICOCT64433.2025.11118648, 2025.

Bhandari, S., Gani, S., Patel, K., Wang, D. S., Soni, P., Arub, Z., Habib, G., Apte, J. S., and Hildebrandt Ruiz, L.: Sources and atmospheric dynamics of organic aerosol in New Delhi, India: insights from receptor modeling, Atmospheric Chemistry and Physics, 20, 735–752, https://doi.org/10.5194/acp-20-735-2020, 2020.

Bousiotis, D., Singh, A., Haugen, M., Beddows, D. C., Diez, S., Murphy, K. L., Edwards, P. M., Boies, A., Harrison, R. M., and Pope, F. D.: Assessing the sources of particles at an urban background site using both regulatory instruments and low-cost sensors–a comparative study, Atmospheric Measurement Techniques, 14, 4139–4155, https://doi.org/10.5194/amt-14-4139-2021, 2021.

Bousiotis, D., Beddows, D., Singh, A., Haugen, M., Diez, S., Edwards, P., Boies, A., Harrison, R., and Pope, F.: A study on the performance of low-cost sensors for source apportionment at an urban background site, Atmospheric Measurement Techniques Discussions, 2022, 1–40, https://doi.org/10.5194/amt-15-4047-2022, 2022.

Bukhari, A. H., Raja, M. A. Z., Shoaib, M., and Kiani, A. K.: Fractional order Lorenz based physics informed SARFIMA-NARX model to monitor and mitigate megacities air pollution, Chaos, Solitons & Fractals, 161, 112 375, https://doi.org/10.1016/j.chaos.2022.112375, 2022.

Chakraborty, S., Tripathi, S. N., Sethi, D., Lakra, A., Kumar, A., Srivastava, P. K., Rao, N. T., Tripathi, A., and Kar, P.: A Framework for Dynamic Hyper-local Source Apportionment using Low-cost Sensors for Real-time Policy Action, EGUsphere, 2025, 1–51, https://doi.org/10.5194/egusphere-2025-5677, 2025.

Chauhan, A., Hsu, C.-Y., and Lee, C.-H.: New methodology to quantify source contributions to nitrate, sulfate and ammonium using PMF-AI, Environmental Pollution, p. 127018, https://doi.org/10.1016/j.envpol.2025.127018, 2025.

Cheng, S., Shi, J., Cheng, Q., Zhou, X., Zhou, J., and Liu, Q.: Improved Short-Term Load Forecasting Using a Composite Cascaded Neural Network with STL Decomposition, in: 2025 IEEE Kiel PowerTech, pp. 1–6, IEEE, https://doi.org/10.1109/PowerTech59965.2025.11180591, 2025.

Cohen, D. D., Crawford, J., Stelcer, E., and Bac, V. T.: Characterisation and source apportionment of fine particulate sources at Hanoi from 2001 to 2008, Atmospheric Environment, 44, 320–328, https://doi.org/10.1016/j.atmosenv.2009.10.037, 2010.

Das, A.: Pollution source apportionment and application of machine learning approaches in surface water suitability for irrigation based on hydro chemical analysis, Green Technology, Resilience, and Sustainability, 5, 1–24, https://doi.org/10.1007/s44173-025-00021-9, 2025.

Diapouli, E., Manousakas, M., Vratolis, S., Vasilatou, V., Maggos, T., Saraga, D., Grigoratos, T., Argyropoulos, G., Voutsa, D., Samara, C., et al.: Evolution of air pollution source contributions over one decade, derived by PM10 and PM2. 5 source apportionment in two metropolitan urban areas in Greece, Atmospheric environment, 164, 416–430, https://doi.org/10.1016/j.atmosenv.2017.06.016, 2017.

Dinh, V. N. T., Uzu, G., Dominutti, P., Sauvage, S., Elazzouzi, R., Darfeuil, S., Voiron, C., Samaké, A., Zhang, S., Socquet, S., et al.: Toolbox for accurate estimation and validation of Positive Matrix Factorization solutions in Particulate Matter source apportionment, Atmospheric Measurement Techniques, 18, 6817–6833, https://doi.org/10.5194/amt-18-6817-2025, 2025.

https://doi.org/10.5194/egusphere-2026-1103
Preprint. Discussion started: 27 April 2026

Dragomiretskiy, K. and Zosso, D.: Variational Mode Decomposition, IEEE Transactions on Signal Processing, 62, 531–544, https://doi.org/10.1109/TSP.2013.2288675, 2014.

Edwards, T. D.: Novel Methods for Air Emissions Measurement and Source Apportionment Applied at a Major Urban Highway, Ph.D. thesis, University of Toronto (Canada), 2025.

Gao, S., Zhang, Q., Tian, R., Ma, Z., Liu, Y., and Hao, Z.: Collaborative apportionment noise-based soft sensor framework, IEEE Transactions on Instrumentation and Measurement, 71, 1–12, https://doi.org/10.1109/TIM.2022.3200088, 2022.

Jin, S., Nong, K., Ma, Y., Lei, C., Yang, Q., Wei, J., Liu, B., Gong, W., and Wang, L.: Synergistic Estimation of Surface Particulate Matter and Ozone Pollutants to enhance accuracy and interpretability by a Deep Learning Approach, IEEE Transactions on Geoscience and Remote Sensing, https://doi.org/10.1109/TGRS.2026.3657522, 2026.

Lahmiri, S.: Comparing Variational and Empirical Mode Decomposition in Forecasting Day-Ahead Energy Prices, IEEE Systems Journal, 11, 1907–1910, https://doi.org/10.1109/JSYST.2015.2487339, 2017.

Lawrence, S. and Bhathmanabhan, S.: Harnessing deep learning for air pollution forecasting: trends, techniques, and future prospects, Artificial Intelligence Review, 2026.

Li, Y., Zhang, W., Tan, M. H. Y., and Chien, P.: Sequential Decomposition of Multiple Seasonal Components Using Spectrum-Regularized Periodic Gaussian Process, IEEE Transactions on Signal Processing, https://doi.org/10.1109/TSP.2025.3540720, 2025.

Liu, X., Li, Y., Wang, F., Qin, Y., and Lyu, Z.: Decomposition-reconstruction-optimization framework for hog price forecasting: Integrating STL, PCA, and BWO-optimized BiLSTM, PloS one, 20, e0324 646, https://doi.org/10.1371/journal.pone.0324646, 2025.

Madukpe, V. N., Ugoala, B. C., and Zulkepli, N. F. S.: Topological Approach and Kernel Principal Component Analysis for Air Pollution Source Apportionment, International Journal of Environmental Research, 19, 260, https://doi.org/10.1007/s41742-025-00912-6, 2025.

Manousakas, M. I., Zografou, O., Canonaco, F., Diapouli, E., Papagiannis, S., Gini, M., Vasilatou, V., Tobler, A., Vratolis, S., Slowik, J. G., et al.: Implementation of real-time source apportionment approaches using the ACSM–Xact–Aethalometer (AXA) setup with SoFi RT: the Athens case study, Atmospheric Measurement Techniques, 18, 3983–4002, https://doi.org/10.5194/amt-18-3983-2025, 2025.

Mei, H., Gali, N. K., Ghadikolaei, M. A., Wei, P., Qin, X., Pan, J., Fu, Q., and Ning, Z.: Local-Scale Pollution Source Apportionment in Complex Port Environments Using High-Density Low-Cost Sensor Networks and Statistics Analysis, ACS ES&T Air, 2, 2504–2516, https://doi.org/10.1021/acsestair.5c00192, 2025.

Mohapatra, A., Tripathi, A. N., Khan, U., Kumar, R., and Singh, I.: An IoT–Machine Learning Proof-of-Concept Framework for Real-Time Urban Particulate Matter Prediction Using Low-Cost Sensors, IEEE Access, 14, 6683–6697, https://doi.org/10.1109/ACCESS.2026.3651996, 2026.

Morain, A., Nedd, R., Poole, K., Hawkins, L., Jones, M., Washington, B., and Anandhi, A.: Artificial Intelligence Application in Nonpoint Source Pollution Management: A Status Update, Sustainability, 17, 5810, https://doi.org/10.3390/su17135810, 2025.

Muhammad, S., Ahmad, Z., Tokatli, C., Khan, A., Iqbal, Z., and Ahmad, A.: Spatial distribution, source characterization, and risk indices of heavy metals pollution in the sediments of the Hindu Kush Lacustrine ecosystem, Environmental Geochemistry and Health, 48, 10, https://doi.org/10.1007/s10653-025-02898-8, 2026.

Qiao, W., Wang, Y., Zhang, J., Tian, W., Tian, Y., and Yang, Q.: An innovative coupled model in view of wavelet transform for predicting short-term PM10 concentration, Journal of Environmental Management, 289, 112 438, https://doi.org/10.1016/j.jenvman.2021.112438, 2021.

Rabhi, L., Lemou, A., Ladji, R., Bonnaire, N., Sciare, J., and Yassaa, N.: Source apportionment of PM2. 5 in a coastal City of Algeria using principal component analysis model, Journal of Atmospheric Chemistry, 82, 13, https://doi.org/10.1007/s10874-025-09477-2, 2025.

https://doi.org/10.5194/egusphere-2026-1103
Preprint. Discussion started: 27 April 2026

Ryoo, I., Kim, T., Ryu, J., Cheong, Y., Moon, K.-j., Jeon, K.-h., Hopke, P. K., Yi, S.-M., and Park, J.: Source apportionment of PM2. 5 using dispersion normalized positive matrix factorization (DN-PMF) in Beijing and Baoding, China, Journal of Environmental Sciences, 155, 395–408, https://doi.org/10.1016/j.jes.2024.10.029, 2025.

Sharma, D., Thapar, S., and Sachdeva, K.: Enhancing particulate matter prediction in Delhi: insights from statistical and machine learning models, Environmental Monitoring and Assessment, 197, 723, https://doi.org/10.1007/s10661-025-14121-3s, 2025.

Shukla, B. P. and Pal, P. K.: A source apportionment approach to study the evolution of convective cells: An application to the nowcasting of convective weather systems, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 5, 242–247, https://doi.org/10.1109/JSTARS.2011.2170661, 2011.

Song, Y., Wang, G., Li, H., Sun, Y., Jin, H., Su, T., Wu, X., and Zhang, Q.: New energy and CCUS thermal power synergistic peaking cost model and apportionment optimization strategy, IEEE Access, https://doi.org/10.1109/ACCESS.2025.3585989, 2025.

Tongkhow, P. and Kantanantha, N.: Bayesian model for time series with trend, autoregression and outliers, in: 2012 Tenth International Conference on ICT and Knowledge Engineering, pp. 90–94, https://doi.org/10.1109/ICTKE.2012.6408577, 2012.

Yuan, L., Yang, Q., Ji, H., Li, M., and Gu, Q.: A novel hybrid approach for quantitative source apportionment and source-oriented health risk of heavy metal (loid) s pollution in an old industrial city, Journal of Hazardous Materials, p. 139521, https://doi.org/10.1016/j.jhazmat.2025.139521, 2025.

Zhao, P., Zhao, P., Zhan, Z., Dai, Q., Casuccio, G. S., Gao, J., Li, J., He, Y., Qian, H., Bi, X., et al.: Advancing Source Apportionment of Atmospheric Particles: Integrating Morphology, Size, and Chemistry Using Electron Microscopy Technology and Machine Learning, Environmental Science & Technology, 59, 3645–3655, https://doi.org/10.1021/acs.est.4c10964, 2025a.

Zhao, P., Zhao, P., Zhang, W., Tong, M., Yang, Y., Casuccio, G. S., Li, L., Gao, J., Li, J., and Feng, Y.: Source Identification of Atmospheric Particles via Low-Voltage Electron Microscopy Image Recognition: A Case Study of Submicrometer Particles, Environmental Science & Technology Letters, https://doi.org/10.1021/acs.estlett.5c00700, 2025b.

Zhou, W., Yu, R., Guo, F., Shen, C., Liu, Y., and Huang, Y.: Source apportionment and risk assessment of soil heavy metals in the Huangshui River Basin using a hybrid model, Ecological Indicators, 160, 111 906, https://doi.org/10.1016/j.ecolind.2024.111906, 2024.

Zhou, Y., Jiang, Y., Wei, Y., Bai, X., and Du, C.: Short-Term Load Forecasting Using STL Decomposition and Stacking Ensemble Learning, in: 2025 IEEE 12th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), vol. 12, pp. 446–450, IEEE, https://doi.org/10.1109/ITAIC64559.2025.11163320, 2025.

Zhu, Q., Huang, X.-F., Cao, L.-M., Wei, L.-T., Zhang, B., He, L.-Y., Elser, M., Canonaco, F., Slowik, J. G., Bozzetti, C., et al.: Improved source apportionment of organic aerosols in complex urban air pollution using the multilinear engine (ME-2), Atmospheric Measurement Techniques, 11, 1049–1060, https://doi.org/10.5194/amt-11-1049-2018, 2018.

Zografou, O., Gini, M., Manousakas, M. I., Chen, G., Kalogridis, A. C., Diapouli, E., Pappa, A., and Eleftheriadis, K.: Combined organic and inorganic source apportionment on yearlong ToF-ACSM dataset at a suburban station in Athens, Atmospheric Measurement Techniques, 15, 4675–4692, https://doi.org/10.5194/amt-15-4675-2022, 2022.