Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Meteorological and HYSPLIT-Derived Transport Influences on $PM_{2.5}$ Exposure in a Khulna City, Bangladesh

Sharfan Upaul

s.upaul@urp.kuet.ac.bd

Khulna University of Engineering and Technology

Anika Tabassum

Khulna University of Engineering and Technology

Mahmudul Hasan

Khulna University of Engineering and Technology

Article

Keywords: PM2.5, Meteorological drivers, HYSPLIT backward trajectories, Seasonal-Trend decomposition, K-means clustering, Khulna, Bangladesh

Posted Date: October 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11078552/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

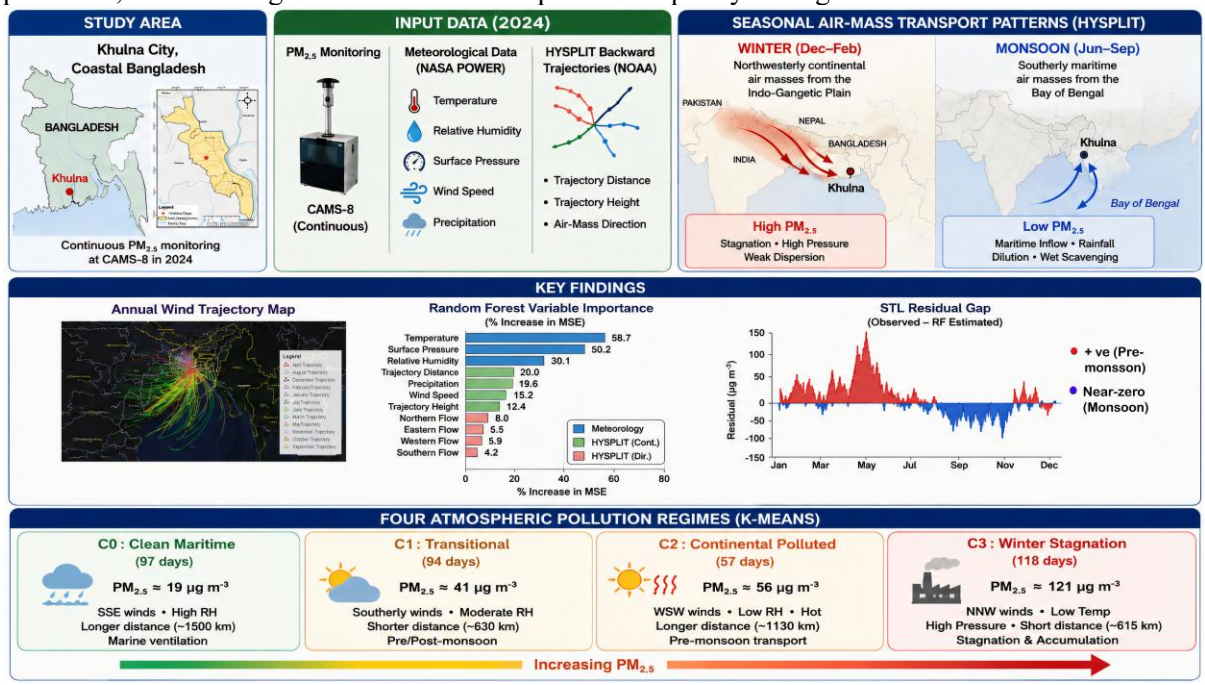
Additional Declarations: No competing interests reported.

Meteorological and HYSPLIT-Derived Transport Influences on $PM_{2.5}$ Exposure in a Khulna City, Bangladesh

Sharfan Upaul[1*], Anika Tabassum[2], Mahmudul Hasan[3]

[123]Department of Urban and Regional Planning, Khulna University of Engineering and Technology, Khulna-9203, Bangladesh.
*Corresponding Author: s.upaul@urp.kuet.ac.bd Tel: +8801798331038

Abstract:
Fine particulate matter ($PM_{2.5}$) is a major public health concern in South Asian cities, yet its meteorological drivers and transport dynamics remain poorly characterized in coastal Bangladesh. This study examined the temporal variability of $PM_{2.5}$ in Khulna City using continuous hourly measurements from the CAMS-8 station (2024), combined with NASA POWER meteorological data and NOAA-HYSPLIT backward trajectories, through multiple linear regression, Random Forest, Seasonal-Trend Decomposition (STL), and K-means clustering. $PM_{2.5}$ showed strong seasonality, peaking in winter (~121 µg m^{-3}, nearly eight times the WHO 24-hour guideline) and falling to ~15 µg m^{-3} during the monsoon. Temperature and surface pressure were the dominant Random Forest predictors (58.7% and 50.2% of variance; OOB R^2 = 0.64), while temperature and relative humidity consistently suppressed $PM_{2.5}$ in all MLR models. HYSPLIT analysis revealed a bimodal transport pattern: short-range northwesterly continental trajectories in winter versus long-range southerly maritime inflows during the monsoon. K-means clustering identified four pollution-meteorology regimes (~19 to ~121 µg m^{-3}), and STL residuals detected pre-monsoon pollution spikes (~138 µg m^{-3}) linked to long-range continental transport. These findings show that $PM_{2.5}$ in Khulna is jointly governed by thermodynamic and transport processes, underscoring the need for season-specific air quality management.



Keywords: PM2.5, Meteorological drivers, HYSPLIT backward trajectories, Seasonal-Trend decomposition, K-means clustering, Khulna, Bangladesh

Acknowledgement
The authors like to express their gratitude toward the Department of Urban and Regional Planning, Khulna University of Engineering & Technology, for its extreme help during this study.

List of Abbreviation

$PM_{2.5}$	:	Particulate Matter with aerodynamic diameter ≤ 2.5 μm
NASA	:	National Aeronautics and Space Administration
NCEI	:	National Centers for Environmental Information
CAMS	:	Continuous Air Monitoring Station
MLR	:	Multiple Linear Regression
RF	:	Random Forest
STL	:	Seasonal-Trend Decomposition using Loess
HYSPLIT	:	Hybrid Single-Particle Lagrangian Integrated Trajectory
GFS	:	Global Forecast System
AGL	:	Above Ground Level
WS	:	Wind Speed
WD	:	Wind Direction
TMP	:	Temperature
RH	:	Relative Humidity
CWT	:	Concentration Weighted Trajectory
PSCF	:	Potential Source Contribution Function
R^2	:	Coefficient of Determination
RMSE	:	Root Mean Square Error
MAE	:	Mean Absolute Error
DBI	:	Davies-Bouldin Index
HAC	:	Heteroskedasticity and Autocorrelation Consistent
ENSO	:	El Niño–Southern Oscillation
WCSS	:	Within-Cluster Sum of Squares
BNAAQS	:	Bangladesh National Ambient Air Quality Standards
MERRA-2	:	Modern-Era Retrospective Analysis for Research and Applications, Version 2

1. Introduction

Air pollution has emerged as one of the most critical environmental challenges globally, posing severe risks to human health, ecosystems, and climate systems (Borchers-Arriagada et al. 2024). Among various air pollutants, fine particulate matter ($PM_{2.5}$), defined as airborne particles with aerodynamic diameters ≤2.5 μm, is of particular concern due to its strong association with adverse health outcomes (Wang and Ogawa 2015a; Choi et al. 2024). Owing to its extremely small size and long atmospheric residence time, $PM_{2.5}$ can penetrate deep into the respiratory system and enter the bloodstream, thereby increasing the risk of cardiovascular and respiratory diseases (Ali et al. 2025).

The concentration and composition of $PM_{2.5}$ are influenced by both natural and anthropogenic sources, including vehicular emissions, industrial activities, biomass burning, and construction processes (Espinoza-Guillen et al. 2025). Industrial activities contribute to both primary $PM_{2.5}$ emissions and secondary aerosol formation via precursor gases such as NO_x and SO_2. In addition to emissions, meteorological conditions such as temperature, wind speed, humidity, and precipitation play a crucial role in governing the formation, dispersion, and accumulation of particulate matter (Wang and Ogawa 2015a). Furthermore, regional atmospheric transport significantly modifies local $PM_{2.5}$ levels, indicating that $PM_{2.5}$ variability is shaped not only by local emission sources but also by cross-regional atmospheric transport processes (Gao et al. 2024).

In Bangladesh, rapid urbanization, industrial expansion, and increasing energy demand have intensified $PM_{2.5}$ pollution, particularly in major urban centers. Observational studies have consistently reported that $PM_{2.5}$ concentrations in several Bangladeshi cities exceed both national and World Health Organization (WHO) air quality guidelines (Al-Muyeed 2016; Gupta et al. 2020; Fahmida Khatun et al. 2021; Ali et al. 2025). The U.S. Environmental Protection Agency (EPA) has established $PM_{2.5}$ standards of 12 μg m^{-3} annually (3-year average) and 35 μg m^{-3} for 24-hour exposure (EPA 2012), while

the WHO 2021 guidelines have further tightened thresholds to 5 μg m^{-3} (annual) and 15 μg m^{-3} (24-hour) (WHO 2021). According to Bangladesh's national ambient air quality standard (BNAAQS), the daily and annual average standard values of $PM_{2.5}$ are 65 μg m^{-3} and 15 μg m^{-3} (Begum et al. 2010; Md. Masud Rana et al. 2018). In contrast, measured concentrations in Bangladesh significantly exceed these limits, highlighting a severe public health concern.

Khulna City, the third-largest urban center in Bangladesh and a key industrial and transportation hub, has experienced particularly elevated $PM_{2.5}$ levels. Studies report concentrations reaching approximately 302 μg m^{-3}, far exceeding both national and international standards (Saju et al. 2023; Ali et al. 2025). Spatial and temporal analyses indicate strong variability in $PM_{2.5}$ levels across residential, industrial, and traffic-dominated zones, with winter concentrations substantially higher than monsoon levels due to reduced dispersion and increased emissions (Saju et al. 2023). Anthropogenic activities such as vehicular emissions, construction, waste burning, and industrial operations are major local contributors (Mazumder et al. 2022; Ali et al. 2025). However, accumulating evidence indicates that emission sources alone cannot fully explain observed variability, as meteorological conditions significantly regulate pollutant accumulation and removal processes (Wang and Ogawa 2015a; Islam et al. 2023b).

Meteorological influences on $PM_{2.5}$ exhibit strong spatial and seasonal heterogeneity, particularly in South Asian urban environments. In Bangladesh, variables such as relative humidity, rainfall, wind speed, temperature, and solar radiation have been identified as key drivers of particulate matter variability (Islam et al. 2023a; Rahman and Meng 2024). Studies in Dhaka have demonstrated significant negative correlations between $PM_{2.5}$ concentrations and precipitation, humidity, and wind speed, indicating the importance of wet deposition and atmospheric dispersion processes (Afrin et al. 2021; Borchers-Arriagada et al. 2024; Rahman and Meng 2024). Seasonally, winter conditions favor pollutant accumulation due to stagnant air and low temperatures, while monsoon rainfall promotes pollutant removal, although secondary aerosol formation may still occur under high humidity and radiation conditions (Islam et al. 2023b; Ali et al. 2025). Similar meteorological dependencies have also been observed in other regions, such as Japan and China (Wang and Ogawa 2015a; Zhan et al. 2021). Seasonal-trend decomposition techniques have been used in recent studies to clearly separate the trend, seasonal, and residual parts of $PM_{2.5}$ time series. This approach allows for more accurate attribution of pollution variability (Borchers-Arriagada et al. 2024).

In addition to local meteorological controls, regional air mass transport plays a critical role in shaping urban air quality. Backward trajectory-based approaches, including the HYSPLIT model, are widely used atmospheric transport and dispersion models developed by NOAA (Roland R. Draxler and G.D.Hess 1998; Stein et al. 2015; Rolph et al. 2017). It computes the trajectories of air parcels forward or backward in time using Lagrangian methods. This allows identification of pollutant source regions and transport pathways, along with statistical techniques such as cluster analysis, the Potential Source Contribution Function (PSCF), and the Concentration Weighted Trajectory (CWT), etc (Hopke 2016; Dihan et al. 2021; Gao et al. 2024). These methods demonstrate that long-range transport from surrounding regions, including the Indo-Gangetic Plain, contributes significantly to $PM_{2.5}$ loading in Bangladesh (Hopke 2016; Afrin and Islam 2020). Moreover, wind trajectories are widely incorporated into air quality forecasting models to simulate pollutant transport and predict exposure risk under varying meteorological conditions (Rheingrover and Gordon 1988; Roland R. Draxler and G.D.Hess 1998; Stein et al. 2015; Bucchignani 2024).

While significant progress has been made in understanding $PM_{2.5}$ dynamics in Dhaka and across the country, city-level studies in Khulna are still limited (Bushra et al. 2019; Mazumder et al. 2022; Saju et al. 2023). Existing studies in Khulna primarily focused on mapping $PM_{2.5}$ distributions and identifying local emission sources (Mazumder et al. 2022; Hassan and Bhuiyan 2023; Islam et al. 2023b; Ali et al. 2025). However, these studies have not systematically examined the temporal variability of $PM_{2.5}$ concentrations or the combined influence of meteorological conditions and regional air-mass transport on air quality (Mazumder et al. 2022; Saju et al. 2023). To address these gaps, this study adopts a broder analytical framework to examine the temporal dynamics of $PM_{2.5}$ in Khulna City. It assesses the effects of daily and seasonal meteorological conditions, regional air-mass transport pathways, and their combined influence on surface $PM_{2.5}$ concentrations. The study further integrates machine learning with Seasonal-Trend Decomposition using LOESS (STL) to characterize temporal variability. STL decomposes a time series into trend, seasonal, and residual components using locally weighted

regression (LOESS) and is robust to outliers and complex seasonality (Robert B. Cleveland et al. 1990). This combined approach helps isolate pollution spikes driven by regional transport that may not be captured by conventional regression methods.

This approach is especially important for coastal cities, where local and external factors interact. Given Khulna's unique coastal location, industrial background, and closeness to pollution corridors from the Indo-Gangetic Plain, such an integrated analysis is critically needed to inform targeted air quality management strategies. This study has two main goals. First, it aims to describe the seasonal and diurnal variability of $PM_{2.5}$ concentrations in Khulna City using continuous monitoring data from 2024. Second, it aims to assess the impact of local weather conditions and regional air-mass movement on $PM_{2.5}$ behavior using a combined approach that includes regression modeling, machine learning, backward trajectory analysis, and time-series decomposition. Together, these goals aim to lay the foundation for season-specific air quality management in a coastal South Asian city, where both local emissions and cross-border transport contribute to pollution exposure.

2. Materials and methods

2.1 Study area and stations

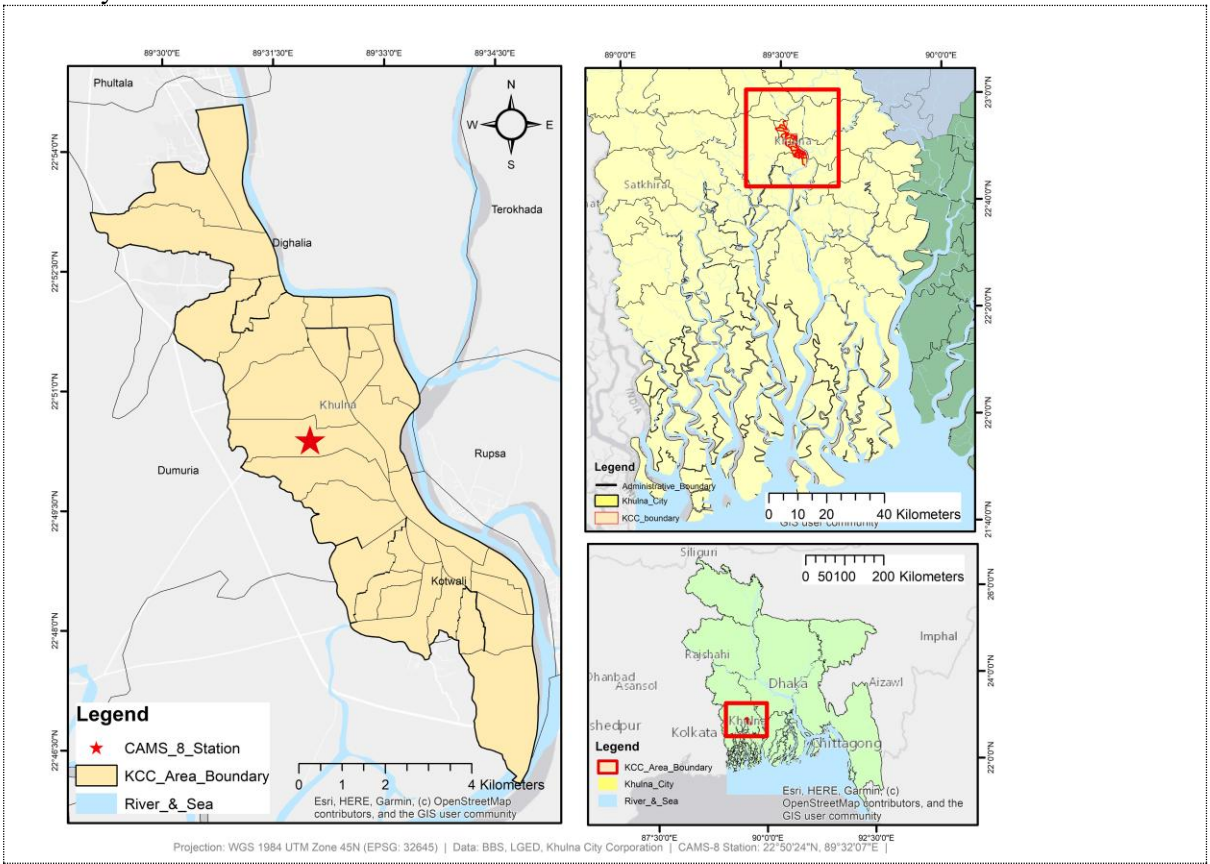


Fig. 1 Location of the CAMS-8 monitoring station and study area, Khulna City, Bangladesh. The main panel shows the KCC administrative boundary and ward-level subdivisions. The upper inset displays the KCC boundary within Khulna District, and the lower inset shows the location of Khulna within Bangladesh. The red star indicates the CAMS-8 station (22.83N, 89.53E).

Fig. 1 shows the Khulna City Corporation (KCC) boundaries (22°48′–22°58′ N, 89°28′–89°38′ E) with the CAMS-8 monitoring station (22.83°N, 89.53°E) marked (Saju et al. 2023; Ali et al. 2025). Khulna, Bangladesh's third-largest city, is an industrial and transportation hub situated along the Rupsha and Bhairab rivers near the Bay of Bengal (Mazumder et al. 2022). Rapid urbanization, vehicular traffic, and industrial activities have deteriorated local air quality (Saju et al. 2023; Ali et al. 2025). The tropical monsoon climate produces distinct seasonal $PM_{2.5}$ variability, further modulated by transboundary transport from the Indo-Gangetic Plain (Hopke 2016; Afrin and Islam 2020; Islam et al. 2023b). The

2024 study period coincided with an ENSO-neutral to weak La Niña transition, with monsoon timing closely following climatological averages (Hu et al. 2024a). Table 1 presents summary statistics for all variables.

Meteorological and transport-related variables for Khulna in 2024 were obtained from multiple authoritative sources and incorporated into the modelling framework. Daily air temperature ranged from 13.90 to 37.23 °C (mean: 25.96 °C), relative humidity from 30.81% to 95.34% (mean: 76.21%), wind speed from 0.85 to 11.53 m s^{-1} (mean: 2.79 m s^{-1}), precipitation from 0 to 126.50 mm h^{-1} (mean: 5.39 mm h^{-1}), and surface pressure from 986.54 to 1019.22 hPa (mean: 1007.64 hPa). Meteorological data were collected from the NASA POWER Project (MERRA-2) and precipitation data from the National Centers for Environmental Information (NCEI; Khulna Station SYN41947). Air-mass transport characteristics were derived using the NOAA HYSPLIT model. The distance of wind trajectories varied between 152.22 and 2679.86 km day^{-1} (mean: 933.36 km day^{-1}), while trajectory height ranged from 59.95 to 3242.29 m above ground level (mean: 675.08 m). Continuous variables were standardized prior to modelling. Wind directions were represented as dichotomous variables (0/1), with the highest frequencies observed for southerly (92 days) and northerly (77 days) trajectories, followed by south-westerly (46 days), north-westerly (44 days), south-easterly (40 days), westerly (25 days), easterly (21 days), and north-easterly (21 days) flows.

2.2 Data and methods for analysis

The air quality monitoring data used in this study were obtained from the Department of Environment (DoE) Continuous Air Monitoring Station (CAMS-8) in Bangladesh. The dataset consists of hourly $PM_{2.5}$ concentrations (μg m^{-3}) from January 1 to December 31, 2024. The measurements followed standard air quality monitoring procedures set by the national authority. To understand temporal variability, hourly $PM_{2.5}$ concentrations were combined to create seasonal averages based on established climatological divisions. We screened the data quality before analysis to ensure consistency over time. Missing observations represented 18.32% of the total dataset (1,609 hours). The distribution of missing data demonstrated substantial temporal heterogeneity, with a pronounced peak in May (75.67%) and increased gaps during the monsoon transition period (June to July). These patterns suggest non-random missingness likely resulting from instrument downtime and adverse environmental conditions. To fill in data gaps, three techniques were applied and compared: (i) Local mean smoothing with a 5-point window, (ii) Linear interpolation, and (iii) Multivariate Imputation by Chained Equations (MICE). Linear interpolation estimates missing values using nearby observations. It is commonly used in atmospheric time-series reconstruction because it is simple and effective for short gaps (Junninen et al., 2004) . A 10% random holdout validation was conducted, and model performance was evaluated with root mean square error (RMSE). Although the 10% random holdout validation does not fully replicate the observed temporal structure of missingness, it was implemented as a controlled baseline to enable consistent comparison across imputation methods. Due to the highly variable gap structure, especially in May, interpolation performance may differ under conditions of extreme missingness. Linear interpolation consistently outperformed both local mean smoothing and MICE in reconstructing masked observations. It achieved the lowest RMSE (25.19 μg m^{-3}) and was therefore selected for the final dataset. Extreme $PM_{2.5}$ values were identified using a percentile-based threshold method often used in air pollution studies (Rautela et al., 2025). Observations exceeding the 99.5th percentile were classified as outliers. However, we kept these values in the analysis to capture high-pollution episodes linked to poor weather conditions and regional transport processes.

Meteorological conditions have a considerable impact on the transportation and diffusion of air pollutants (H. L. Wang et al., 2016; Wu et al., 2013; Zhan et al., 2021). Fifteen meteorological and transport predictors were selected for hourly analysis (n = 8,784 hourly observations) between 12:00 AM on 01/01/2024 and 12:00 PM on 12/31/2024 (Yang et al.). 15 predictor variables were used daily in the analysis, including the wind trajectory component (n=365). The data for the meteorological variables were collected from the 'NASA POWER' website (https://power.larc.nasa.gov/data-access-viewer/) using the MERRA-2 reanalysis dataset. Data were extracted at a spatial resolution of 0.5° × 0.625° for the entire study period. Precipitation data were collected from the 'Synoptic' website

(https://download.synopticdata.com), which processes and provides the 'National Centers for Environmental Information (NCEI) data. The nearest available station, SYN41947 (Khulna), was selected for this study (Latitude: 22.78333; Longitude: 89.53333; Elevation: 9.84m). Satellite/reanalysis-based precipitation estimates (e.g., MERRA-2) may introduce uncertainties due to indirect measurement approaches and were therefore excluded, as they can misrepresent actual rainfall variability.

All statistical analyses were conducted using Python 3.10. Key packages included stats models for OLS regression, VIF computation, and Durbin-Watson testing; statsmodels.tsa.seasonal.STL for seasonal-trend decomposition; scikit-learn for Ridge regression (RidgeCV with 5-fold cross-validation), Random Forest modeling (RandomForestRegressor with out-of-bag validation), and K-Means clustering (with WCSS, DBI, and silhouette-based k-selection); and seaborn for Pearson correlation heatmaps. HYSPLIT version 5 was accessed through the NOAA READY web platform for back-trajectory analysis. Trajectory analysis was performed using the Global Forecast System (GFS) meteorological dataset (0.25° resolution), operated by NOAA (Saha et al. 2010), which provided the atmospheric input for HYSPLIT trajectory calculations.

The use of gridded reanalysis data with resolutions coarser than the point receptor, along with an 18.3% imputation rate in the $PM_{2.5}$ record, is an inherent limitation on the accuracy of the findings.

2.3 Evaluating Meteorological and Transport Effects

While machine learning algorithms were found to be used for showing the causal inferences and relations among the meteorological factors and particulate matter, (Islam et al. 2023b; Gulati et al. 2023)Multiple linear regression offers a comprehensible linear framework for measuring the distinct contributions of meteorological and transport variables to $PM_{2.5}$ variability, making it especially effective for identifying primary drivers within seasonal and diurnal patterns (Afrin et al. 2021; Kliengchuay et al. 2021). Multiple linear regression (MLR) is a statistical method that models the linear relationship between a dependent variable, such as $PM_{2.5}$, and multiple independent predictor variables simultaneously. This approach yields standardized coefficients that quantify each predictor's relative contribution while controlling for the others (Afrin et al. 2021; Kliengchuay et al. 2021). MLR was used to assess linear relationships between $PM_{2.5}$ concentrations and meteorological and transport predictor variables across seasons and times of day. This approach follows established methods in South Asian urban air quality studies (Afrin and Islam 2020; Afrin et al. 2021; Kliengchuay et al. 2021; Islam et al. 2023). The general model is as follows:

$$PM_{2.5} = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \in$$

where β_0 is the intercept, β_i are regression coefficients, x_i represents the i-th predictor, and ε is the error term. The coefficient of determination (R^2) was used to quantify the proportion of variation in PM concentration explained by the meteorological variables.

Seasonal models were developed separately for Winter (December to February), Pre-monsoon (March to May), Monsoon (June to September), and Post-monsoon (October to November) using daily data, following established Bangladesh climatological season classifications (Rahman et al. 2020). Diurnal models were created for Morning (06:00 AM to 11:00 AM), Afternoon (12:00 PM to 5:00 PM), Evening (6:00 PM to 11:00 PM), and Late Night (12:00 AM to 05:00 AM) using hourly data. The predictors included surface temperature, precipitation, wind speed, relative humidity, surface pressure, HYSPLIT trajectory displacement distance, trajectory height, and eight wind direction indicators (N, E, W, NE, NW, SE, SW). The southerly direction was excluded as the reference category to avoid perfect multicollinearity(Damodar N. Gujarati and Dawn C. Porter 2009). Models were estimated using OLS in Python (statsmodels). Ridge regression with 5-fold cross-validated alpha selection was also applied to the diurnal models as a robustness check.

The Variance Inflation Factor (VIF) was calculated using mean-centered continuous predictors to evaluate multicollinearity among explanatory variables. A VIF value of 1 indicates the absence of collinearity, whereas higher values signify greater interdependence among predictors. Although VIF

values exceeding 10 are typically regarded as problematic (O'Brien 2007). In atmospheric studies, meteorological variables such as temperature, relative humidity, and pressure are inherently linked due to physical connections in the atmosphere (Graham 2003; Zhang and Cao 2015). Residual diagnostics, Durbin-Watson autocorrelation tests, and ANOVA F-tests were done for all models. The analyses were performed in Python.

Pearson correlation coefficients were also calculated across the four seasons. The results were visualized as seasonal heatmaps using a diverging colormap scaled from -1 to +1(Zhang and Cao 2015; Yu et al. 2024; Rahman and Meng 2024). Wind direction from NASA POWER (in degrees) was reclassified into directional sectors (N, S, E, W, NE, NW, SE, and SW) using predefined angular ranges and encoded as binary dummy variables (0/1) for regression analysis.

2.4 NOAA-HYSPLIT Trajectory Model:

Backward trajectory analysis was done using the NOAA Hybrid Single-Particle Lagrangian Integrated Trajectory (HYSPLIT) model. This helped identify possible source areas and describe regional air mass transport pathways that affect $PM_{2.5}$ variability in Khulna City (Roland R. Draxler and G.D.Hess 1998; Yerramilli et al. 2012; Stein et al. 2015; Rolph et al. 2017).

The NOAA READY online platform (https://www.ready.noaa.gov) was used to compute the trajectories. Meteorological input data were obtained from the Global Forecast System (GFS) reanalysis at 0.25° spatial resolution, a widely used global atmospheric dataset for trajectory-based transport analysis (Saha et al. 2010).

For each day during the study period (1 January-31 December 2024), 24-hour backward trajectories were calculated using hourly inputs from 00:00 to 23:00 UTC, at a height of 500 m above ground level (AGL) coordinates (22.83N, 89.53E). The 500 m AGL height was chosen to represent the lower boundary layer, where pollutant-laden air masses most affect surface $PM_{2.5}$ concentrations (Ommi et al., 2017).

The daily trajectory displacement distance (km) was computed from the exported trajectory endpoint coordinates using the Haversine formula. The Haversine formula calculates the great-circle distance between two points on a sphere using their latitudes and longitudes, accounting for Earth's curvature. This gave the great-circle distance between the 24-hour upwind air mass origin and the receptor site:

$$d = 2r\arcsin\left(\sqrt{\sin^2\left(\frac{\Delta\phi}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\Delta\lambda}{2}\right)}\right)$$

Where, ϕ and λ represent latitude and longitude (radians), and $r = 6{,}371$km is the Earth's radius. This metric reflects transport distance, with larger values indicating long-range (maritime or continental) sources and smaller values suggesting local or regional influence. The daily mean trajectory height (m AGL) was derived from trajectory metadata. Trajectories were further classified into Bangladesh's four climatological seasons (winter, pre-monsoon, monsoon, and post-monsoon) to assess seasonal transport patterns, and the results were displayed in Google Earth Pro.

2.5 K-Means Cluster Analysis for Pollution Regime Classification

K-means clustering was used to classify different pollution-weather interaction regimes based on $PM_{2.5}$ concentrations and related environmental factors. This unsupervised machine learning algorithm partitions observations into k clusters by minimizing intra-cluster variance over multiple iterations. It does this without requiring pre-labeled target data, thereby helping identify significant environmental patterns (Choi et al., 2024). This method is common in atmospheric science for defining pollution regimes with varying meteorological and transport conditions, from high-pollution stagnant episodes to clear-air ventilated states (Cui et al. 2021; Choi et al. 2024). Prior to clustering, all features were standardized using z-score normalization to ensure equal contribution of variables with differing units and magnitudes.

The optimal number of clusters was determined through joint evaluation of three complementary indices across k=2 to k=10, with 10 random initialization configurations per k value to ensure solution

stability (Rendana et al. 2022). The Elbow method examines the within-cluster sum of squares (WSS) for a range of k values. The point where the rate of decrease slows down indicates the optimal number of clusters. However, high-dimensional environmental datasets can make it difficult to identify a clear elbow point, so the DBI was also used for additional insight. The DBI measures clustering quality by calculating the ratio of intra-cluster scatter to inter-cluster separation. Lower values show clusters that are more compact and clearly separated (Davies and Bouldin 1979).

The Silhouette score assessed cluster cohesion and separation on a scale of -1 to +1, with higher values indicating greater intra-cluster compactness relative to neighboring clusters (Rousseeuw 1987; Choi et al. 2024). The automated consensus procedure identified k=6 as the best choice. However, the resulting solution showed two small clusters with only 3 and 10 observations, respectively. These tiny clusters were outliers rather than real atmospheric patterns and were therefore excluded. Consequently, k=4 was selected as the final solution. This provided four significant pollution-meteorology groups that matched common atmospheric clustering methods (Cui et al. 2021; Choi et al. 2024). Clusters were relabeled in order of their average $PM_{2.5}$ concentrations. Dominant wind direction for each cluster was calculated using circular mean statistics to avoid the angular averaging error inherent in arithmetic means of degree values. Wind rose diagrams were generated for each cluster to characterize directional wind frequency and speed distributions across pollution regimes.

2.6 Random Forest Model for $PM_{2.5}$ Variability

2.6.1 Random Forest Model for $PM_{2.5}$ Variability

A Random Forest (RF) regression model was developed to capture nonlinear relationships. RF is an ensemble learning method that builds multiple decision trees using bootstrap sampling and random feature selection. Final predictions come from averaging the results across these trees (Breiman 2001). This method is popular in air pollution studies because it can handle complex interactions and non-linear dependencies among predictors (Gulati et al. 2023; Sreenivasulu and Rayalu 2024). Compared to single-model linear methods, like multiple linear regression (MLR), RF is well-suited for environmental datasets that have complex predictor interactions, varying variances, and non-Gaussian distributions(Gulati et al. 2023; Borchers-Arriagada et al. 2024).

The model used 11 predictor variables, which included surface meteorological parameters and HYSPLIT-derived trajectory metrics. The RF model was created with 800 trees, and $\sqrt{p}$ variables were randomly selected at each split, where p is the number of predictors, following standard regression settings. Out-of-bag (OOB) validation was used to evaluate the model's performance, giving an unbiased estimate without needing a separate test dataset (Breiman 2001). To assess model performance, the OOB coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE) were used. Variable importance was evaluated by measuring the percentage increase in mean squared error (%IncMSE) through permutation of predictor variables (Islam et al. 2023a; Sreenivasulu and Rayalu 2024).

2.6.2 Seasonal-Trend Decomposition of $PM_{2.5}$ Time Series Using STL

Seasonal-Trend decomposition using Loess (STL) was applied to both observed and RF-estimated daily $PM_{2.5}$ time series to separate long-term trend, seasonal variation, and residual components (Robert B. Cleveland et al. 1990). The decomposition follows an additive structure:

$$Y_t = T_t + S_t + R_t$$

where T_t represents the trend component, S_t the seasonal component, and R_t the residual component.

A 30-day seasonal period was selected to capture sub-monthly variability within the one-year dataset. The use of an annual seasonal cycle (365 days) was avoided due to the limited temporal coverage, as STL requires multiple seasonal cycles for stable and reliable decomposition (Robert B. Cleveland et al. 1990). Similar adaptations have been applied in recent $PM_{2.5}$ studies, where seasonal window selection was adjusted based on dataset length to ensure robust decomposition (Borchers-Arriagada et al. 2024).

Robust fitting was enabled to reduce the influence of extreme values. The seasonal smoother was defined as periodic, and the trend smoother window was automatically determined (31 days) following standard STL implementation.

STL decomposition was applied separately to the observed and RF-estimated $PM_{2.5}$ series. A residual gap metric was defined as:

$$\Delta R_t = R_t^{obs} - R_t^{RF}$$

where ΔR_t represents the difference between observed and model-estimated residual components.

3. Results and Discussion:

3.1 Multiple Linear Regression Analysis of Meteorological Influence on $PM_{2.5}$

3.1.1 Diurnal MLR Performance

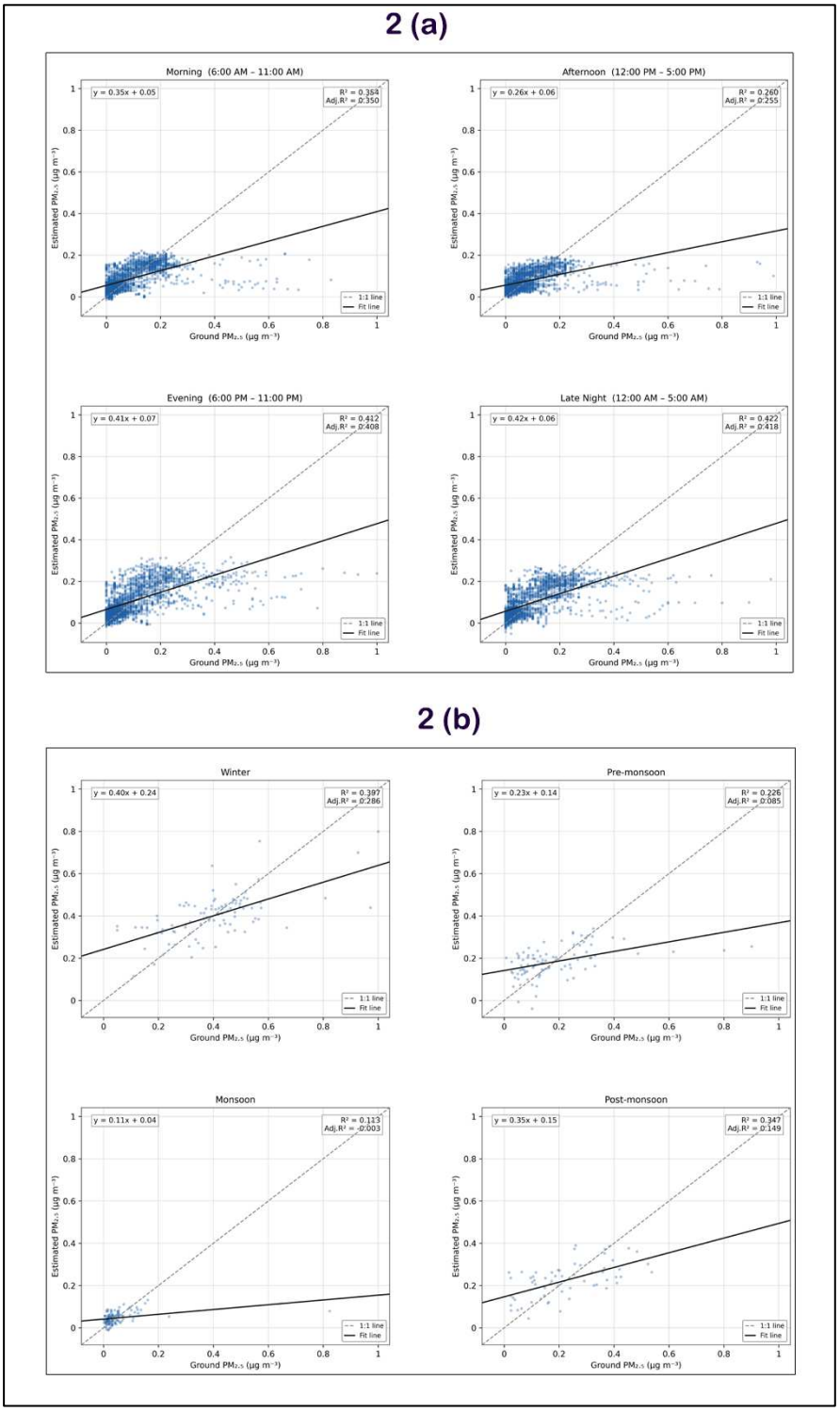
To analyze intra-daily variability in the meteorological control of $PM_{2.5}$, ordinary least squares (OLS) regression models were employed for four distinct diurnal periods: morning (06:00 AM-11:00 AM), afternoon (12:00 PM-05:00 PM), evening (06:00 PM-11:00 PM), and late night (12:00 AM-05:00 AM). Strong positive autocorrelation (Durbin-Watson = 0.44-0.60) required Newey-West HAC standard errors (maxlags = 1) for reliable inference (Newey and West 1987; Junninen et al. 2004). Ridge regression was also tested as a robustness check for possible multicollinearity. However, since the centered VIF values for all diurnal predictors stayed below 7.0 (Table 1), Ridge regularization didn't improve OLS much ($\Delta R^2 \leq 0.003$; VIF < 7). OLS was kept as the main analytical framework because it was easier to understand (Kliengchuay et al. 2021)

Table 1: Diurnal OLS regression results for daily $PM_{2.5}$ models at CAMS-8, Khulna, 2024.

	Morning	**Afternoon**	**Evening**	**Late Night**
	(6:00 AM–11:00 AM)	*(12:00 PM–5:00 PM)*	*(6:00 PM–11:00 PM)*	*(12:00 AM–5:00 AM)*
R^2	0.354	0.260	0.412	0.422
VIF (Range)	1.1–6.1	1.0–5.8	1.0–6.0	1.0–5.6
	B	**B**	**B**	**B**
(Constant)	0.255*	0.321*	0.566*	0.383*
Air Temperature	-0.240*	-0.301*	-0.527*	-0.355*
Relative Humidity	-0.136*	-0.143*	-0.195*	-0.178*
Wind Speed	-0.008	0.070*	0.009	-0.058*
Precipitation	0.036	-0.032	-0.077*	-0.008
Surface Pressure	0.053	0.001	-0.073	0.029
HYSPLIT Distance	-0.077*	-0.056*	-0.091*	-0.222*
HYSPLIT Height	0.024	0.089*	0.069	0.024
Wind Direction N	0.028*	0.003	-0.006	0.013
Wind Direction E	0.011	0.005	0.014	0.008
Wind Direction W	0.005	-0.007	-0.019	-0.020*
NE Direction	0.021*	-0.008	0.008	0.020
NW Direction	0.024*	-0.008	-0.015	-0.008
SE Direction	-0.005	-0.010*	-0.017*	-0.013*
SW Direction	0.010	0.003	0.002	-0.005

* $p < 0.05$ (Newey-West HAC standard errors). Coefficients are standardized (β). R^2 and centered VIF range reported at the top of each column.

The daily performance outperformed seasonal models because of larger sample sizes (N = 2,196 per period). The best explanatory power was at night (R^2 = 0.417), followed by evening (0.403), morning



(0.351), and afternoon (0.252) (Figure 2 (a)).

Fig. 2 - 2(a): Diurnal Performance of the $PM_{2.5}$ Estimation Model Across Different Time Periods. The solid black line represents the ordinary least squares (OLS) regression fit between observed and model-estimated $PM_{2.5}$ concentrations with Newey-West HAC standard errors applied, and the grey dashed line indicates the 1:1 reference line. R^2 and adjusted R^2 are shown in each panel. Fig. 2(b) Scatter plots of estimated versus ground-measured daily $PM^{2.5}$ (µg m^{-3}) for four seasonal OLS regression models at

CAMS-8, Khulna, 2024. Dashed line = 1:1 reference; solid line = regression fit. R^2 and adjusted R^2 are shown in each panel.

The mean bias was small. The higher performance at night indicates stable boundary-layer conditions and less mixing. Lower afternoon performance is due to increased convective turbulence and variability in the boundary layer (Islam et al. 2023b; Rahman and Meng 2024; Ali et al. 2025). Temperature was the main negative predictor across all periods (β = -0.240 to -0.527, $p < 0.001$), consistent with increased atmospheric mixing and boundary-layer development. Relative humidity also had consistent negative impacts (β = -0.136 to -0.195, $p < 0.001$) likely due to wet scavenging processes (Wang and Ogawa 2015b; Islam et al. 2023b). Wind speed showed mixed effects, being positive in the afternoon (β = 0.070, $p = 0.026$) and negative at night (β = -0.058, $p = 0.034$). This suggests advection during the day and dispersion in stable nighttime conditions (Rheingrover and Gordon 1988). Precipitation had a negative relationship with $PM_{2.5}$ in the evening (β = -0.077, $p = 0.011$) in line with wet deposition.

HYSPLIT trajectory distance displayed consistent negative associations across all periods (β = -0.056 to -0.222), indicating lower $PM_{2.5}$ levels due to long-range transport, dilution, and deposition (Cui et al. 2021; Dihan et al. 2021). Trajectory height was significant only in the afternoon (β = 0.089, $p = 0.005$), suggesting the sinking of elevated continental air.

Directional transport analysis reveals clear regional effects. Northerly and northeasterly winds increased morning $PM_{2.5}$, in line with transport from the Indo-Gangetic Plain (Ommi et al. 2017; Afrin and Islam 2020). In contrast, southeasterly winds lowered concentrations in later periods, indicating maritime dilution. Westerly winds reduced nighttime $PM_{2.5}$, while northwesterly winds raised morning levels.

3.1.2 Seasonal MLR Performance

Table 2 shows the OLS multiple linear regression model using 14 meteorological and HYSPLIT-derived predictors were estimated for four seasons: winter (n = 91), pre-monsoon (n = 92), monsoon (n = 122), and post-monsoon (n = 61). Standard errors were kept because Durbin-Watson statistics showed less autocorrelation than in diurnal models, which ranged from 1.503 to 1.950. Winter had moderate values at 1.194 while post-monsoon had a value of 0.983. All centered VIFs were below 7, indicating no multicollinearity issues estimates (Gulati et al. 2023). The mean bias was zero across all seasons (Fig. 2(b)). Model performance varied widely.

Winter had the highest explanatory power (R^2 = 0.397, Adj. R^2 = 0.286) and was the only statistically significant model ($F = 3.574$, $p < 0.001$). Pre-monsoon (R^2 = 0.226, $p = 0.097$) and post-monsoon (R^2 = 0.347, $p = 0.078$) were marginally significant, while the monsoon model was not significant (F = 0.974, $p = 0.484$; Adj. R^2 = -0.003). The poor performance during the monsoon reflects near-zero $PM_{2.5}$ concentrations and strong wet scavenging, which removes most explainable variance (Wang and Ogawa 2015b; Islam et al. 2023a).The gaps between R^2 and Adj. R^2 for the transitional season reflect smaller sample sizes and weaker predictor influence studies (Rahman and Meng 2024). The monsoon coefficients are not interpreted, and the post-monsoon results are treated as exploratory due to the small sample size.

Temperature was the main negative predictor in winter (β = -0.475, $p = 0.023$) and pre-monsoon (β = -0.657, $p = 0.002$). This indicates improved mixing and boundary-layer growth at higher temperatures (Yang et al.; Ali et al. 2025). Wind speed was negatively associated in winter (β = -0.842, $p = 0.053$), indicating mechanical dispersion under stronger flow. Surface pressure had a negative effect in pre-monsoon (β = -0.608, $p = 0.020$), consistent with enhanced vertical mixing under low-pressure conditions. Transport processes were also important. HYSPLIT trajectory distance was negatively related in pre-monsoon (β = -0.262, $p = 0.011$), indicating cleaner long-range air masses (Ommi et al. 2017; Cui et al. 2021). Easterly trajectories were positively associated in winter (β = +0.273, $p = 0.046$), in line with regional pollution transport from eastern industrial areas. No predictors were significant in post-monsoon, consistent with the limited sample size (n = 61) and weak explanatory structure.

Table 2 Seasonal OLS regression results for daily $PM_{2.5}$ models at, Khulna City, 2024.

	Winter	**Pre-monsoon**	**Monsoon**	**Post-monsoon**
	(Dec–Feb)	*(Mar–May)*	*(Jun–Sep)*	*(Oct–Nov)*
N	91	92	122	61
R^2	0.397	0.226	0.113	0.347
Adj. R^2	0.286	0.085	-0.003	0.149
VIF (Range)	1.14–2.46	1.07–6.69	1.04–5.28	1.09–6.85
Durbin–Watson	1.194	1.503	1.950	0.983
	B	**B**	**B**	**B**
(Constant)	0.787*	1.192*	-0.285	0.749
Air Temperature	-0.475*	-0.657*	0.366	-0.080
Relative Humidity	-0.167	-0.204	-0.053	-0.597
Wind Speed	-0.842*	0.025	0.022	-0.094
Precipitation	-0.480	-0.300	0.053	0.505
Surface Pressure	-0.101	-0.608*	0.265*	-0.063
HYSPLIT Distance	-0.029	-0.262*	0.021	-0.239
HYSPLIT Height	-0.135	0.175	0.065	-0.046
Wind Direction N	0.096	-0.042	0.041	0.118
Wind Direction E	0.273*	-0.080	-0.007	0.068
Wind Direction W	-0.050	-0.086	0.038	0.049
NE Direction	0.043	,	-0.009	0.098
NW Direction	0.004	-0.039	0.065	0.139
SE Direction	,	-0.074	-0.021	0.126
SW Direction	-0.028	-0.008	-0.004	0.029

* $p < 0.10$. B = standardized regression coefficient. VIF reported as the range of centered variance inflation factors across all predictors per season. SE direction is the excluded reference category in winter; NE direction in pre-monsoon. , = variable dropped due to perfect collinearity. Post-monsoon results should be regarded as exploratory given the ~4:1 observations-to-predictors ratio (n = 61).

3.2 Backward Air Mass Trajectory Assessment Using the HYSPLIT Model:

Backward trajectory analysis revealed pronounced seasonal variability in the origin, travel distance, and transport height of air masses arriving in Khulna during 2024, indicating a strong seasonal modulation of regional pollutant transport.

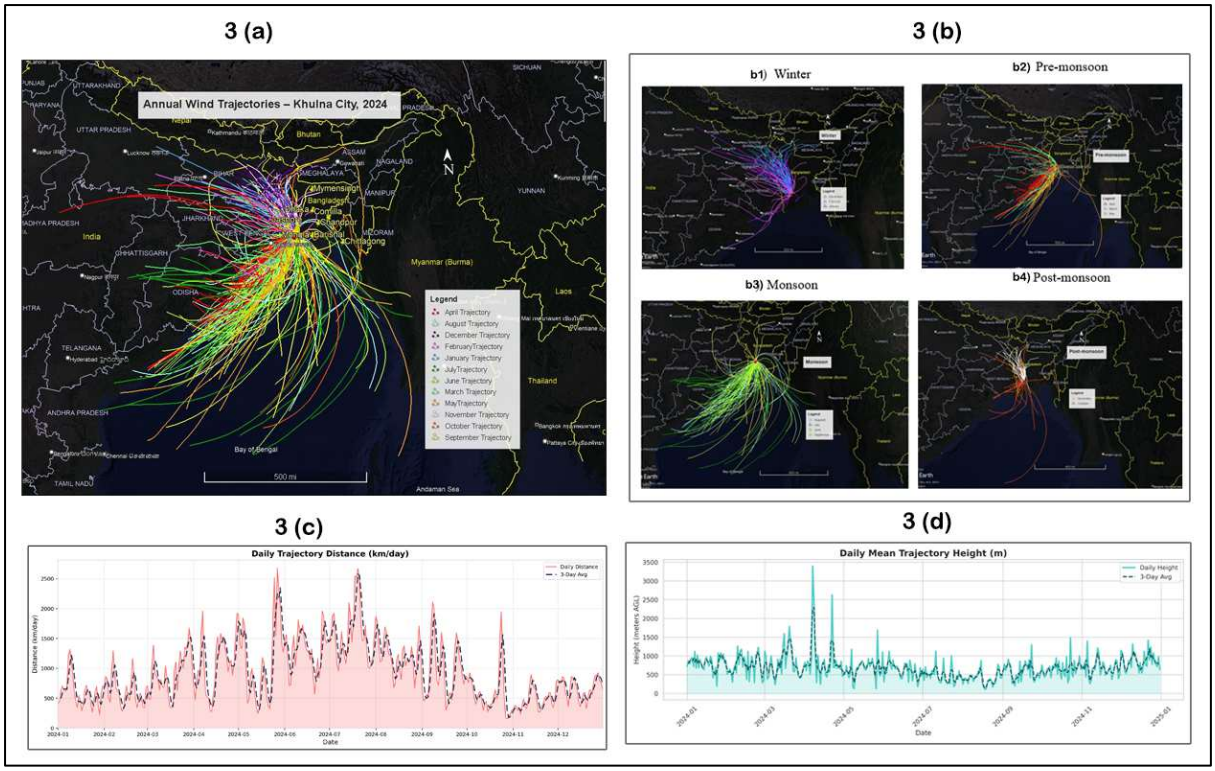


Fig.3 - 3(a) Annual HYSPLIT 24-hour backward trajectory map, Khulna City, 2024. Backward trajectories computed at 500 m above ground level (AGL) at Khulna City, for the full study period (1 January-31 December 2024), using the NOAA-HYSPLIT model driven by Global Forecast System (GFS) reanalysis at 0.25° spatial resolution. 3(b) Seasonal HYSPLIT backward trajectory maps showing dominant transport pathways, Khulna City, 2024, classified by climatological season: (b1) Winter, (b2) Pre-monsoon, (b3) Monsoon, and (b4) Post-monsoon. 3(c) Daily HYSPLIT trajectory displacement distances at Khulna City, for the full study period (1 January-31 December 2024). 3(d) Daily mean HYSPLIT trajectory heights at Khulna City, for the full study period (1 January-31 December 2024)

The annual trajectory map indicated that Khulna receives air masses from a broad geographic region, including the Bay of Bengal, the Indo-Gangetic Plain, peninsular India, and the northeastern Indian Ocean. This pattern reflects Khulna's position at the intersection of continental and maritime circulation systems (Fig. 3(a)). The origins of these trajectories displayed a clear bimodal geographic structure: northerly and northwesterly continental trajectories were predominant during the dry season, whereas southerly and southeasterly maritime trajectories dominated during the monsoon period. This bimodal distribution aligns with the seasonal reversal of the South Asian monsoon circulation system and has been previously documented for Bangladesh as a whole (Hopke 2016; Afrin and Islam 2020).

Seasonal trajectory maps show distinct pathways influencing $PM_{2.5}$ variability (Fig. 3(b)). In winter (December to February), air masses mainly originated from the northwest and traveled over the Indo-Gangetic Plain, passing through Bihar, Jharkhand, and West Bengal before reaching Khulna (Fig. 3(b1)). These short continental pathways pass through heavily populated and industrial areas, which explains the large buildup of pollution in winter. During the pre-monsoon season (March to May), trajectories became more scattered, extending over western and central India, the Arabian Sea, and northwestern inland areas (Fig. 3(b2)). Longer transport distances and occasional higher trajectory heights indicate increased long-range continental flow due to unstable atmospheric conditions. A clear change took place during the monsoon (June to September). Air masses shifted to mostly southerly and south-southeasterly maritime pathways coming from the Bay of Bengal and the northern Indian Ocean (Fig. 3(b3)). These long, steady trajectories, along with frequent rain, led to significant dilution and wet scavenging, resulting in the lowest annual $PM_{2.5}$ levels. In the post-monsoon period (October to November), trajectories quickly returned to shorter northerly continental routes (Fig. 3(b4)).

The daily trajectory distance showed a clear seasonal pattern, along with some high-frequency fluctuations (Fig. 3(c)). Distances were lowest in winter (January to February: about 300 to 600 km per day) and then rose sharply during the transition from pre-monsoon to monsoon, peaking above 2500 km per day (Hu et al. 2024b). The inverse relationship with $PM_{2.5}$ suggests that longer trajectories are associated with cleaner air, whereas shorter continental trajectories are associated with higher pollution levels. This indicates more pollutant buildup over land routes than over sea routes, where dilution and fewer emission sources are present (Hopke 2016; Afrin and Islam 2020). Trajectory height also showed clear seasonal changes, with occasional increases (Fig. 3(d)). Mean heights stayed fairly stable (700 to 1000 m above ground level) during winter and monsoon but increased sharply during pre-monsoon, with two major peaks reaching around 3400 m and 2600 m above ground level. These events point to intermittent elevated transport followed by settling or vertical mixing into the boundary layer (Pandey et al. 2017).

3.3 K-Means Cluster Analysis for Atmospheric Regime Classification:

K-means clustering was used to identify dominant atmospheric patterns affecting $PM_{2.5}$ in Khulna City, based on meteorological, $PM_{2.5}$, and HYSPLIT variables. Model selection based on WCSS, Davies-Bouldin Index, and silhouette score (k = 2-10; Fig. 4(a)) indicated that k = 6 was the best choice statistically. However, the k = 6 solution showed too many clusters and limited meteorological coherence. It had two very small clusters (3 and 10 days) that overlapped in $PM_{2.5}$ and synoptic features across several groups. This made the distinct atmospheric states harder to interpret. A robustness check was conducted with k = 4, yielding more stable and consistent atmospheric patterns while retaining all 366 days(Cui et al. 2021; Choi et al. 2024). The k = 4 solution provided clearer separation of meteorological conditions and a more interpretable $PM_{2.5}$ gradient, making it the simplest representation Table 3. Cluster $PM_{2.5}$ increased consistently from C0 (~19 µg m^{-3}) to C3 (~120 µg m^{-3}) (Fig. 4(b)), confirming that the clusters represent real atmospheric patterns rather than just statistical artifacts.

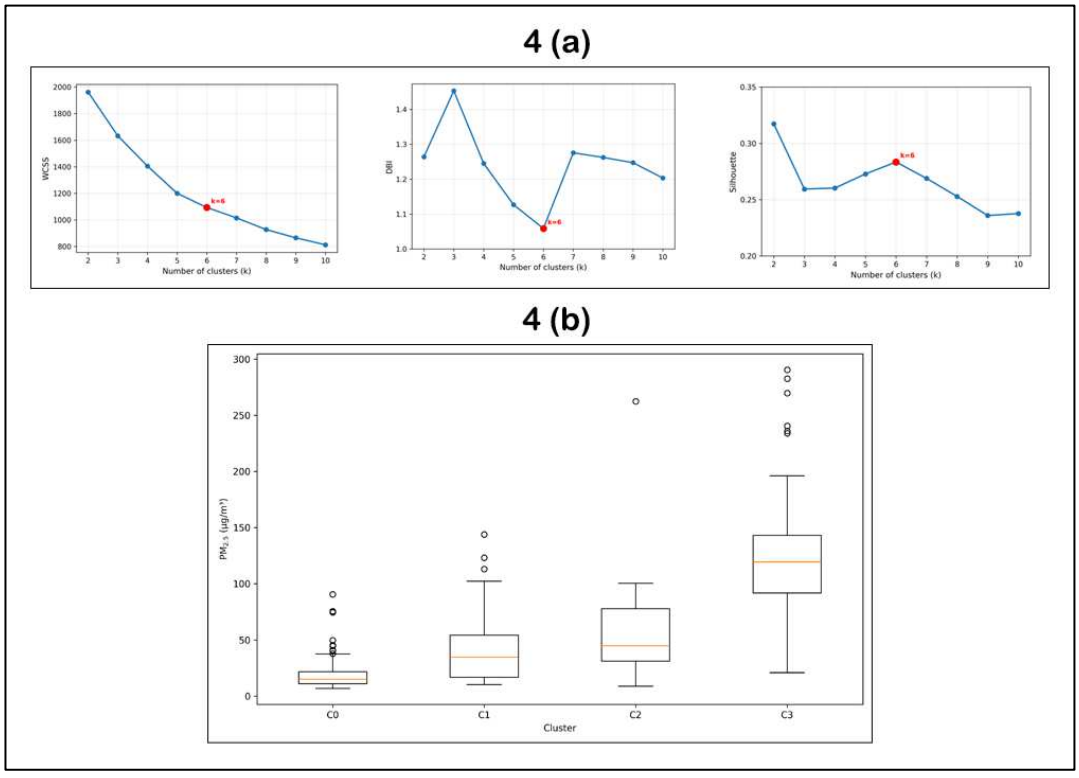


Fig.4 – 4(a) K-means cluster optimization diagnostics for atmospheric pollution regime classification. 4(b) $PM_{2.5}$ concentration distributions across four K-means pollution regimes, Khulna City, 2024. Boxplots of daily $PM_{2.5}$ concentrations (µg m^{-3}) for the four K-means pollution regimes (C0-C3), where n = 366 days. Clusters are ordered by increasing mean $PM_{2.5}$ concentration: C0 (clean maritime), C1 (moderate transitional), C2 (high-pollution continental), and C3 (Most polluted regime). Boxes show

the interquartile range (IQR), horizontal lines indicate medians, and whiskers extend to 1.5 × IQR, where points beyond whiskers represent extreme observations.

Table 3 Meteorological and atmospheric transport characteristics of the four K-means pollution regimes at Khulna city, 2024.

Cluster	Days	$PM_{2.5}$ Mean ($\mu g\ m^{-3}$)	Min	Max	Wind Speed ($m\ s^{-1}$)	Temp (°C)	RH (%)	Pressure (hPa)	Traj. Dist. (km)	Traj. Height (m)	Wind Dir. (deg)
C0	97	19.37	6.79	90.63	4.15	28.81	89.00	1000.55	1498.73	470.42	SSE (162.06°)
C1	94	40.87	10.27	143.97	2.07	27.85	83.89	1006.98	630.09	675.09	S (185.89°)
C2	57	55.51	8.85	262.47	2.83	31.10	48.92	1006.78	1130.30	760.92	WSW (247.67°)
C3	118	120.69	20.95	290.48	2.22	19.63	72.77	1014.42	615.05	801.84	NNW (346.04°)

From Table 3, Cluster C0 represents the cleanest regime. It features the highest wind speed (4.15 $m\ s^{-1}$), the lowest pressure (1000.6 hPa), the highest relative humidity (89.0%), and the longest transport distance (approximately 1499 km day^{-1}). These conditions show strong marine ventilation during active monsoon periods. This ventilation leads to effective dilution and low $PM_{2.5}$ levels. Cluster C1 represents a transitional regime with weak southerly winds and shorter transport paths (around 630 km day^{-1}). As a result, it has moderate $PM_{2.5}$ levels during the pre- and post-monsoon periods. Cluster C2 is a high-pollution regime. It has low humidity (48.9%), high temperature (31.1°C), and long transport from the west-southwest (about 1130 km day^{-1}). This reflects hot, dry conditions and continental inflow during the pre-monsoon season. Cluster C3 represents the most polluted winter regime, lasting 118 days. It shows the highest pressure (1014.4 hPa), the lowest temperature (19.6°C), and the shortest transport distance (around 615 km day^{-1}). This indicates stagnant synoptic conditions and weak dispersion. This regime also had the highest $PM_{2.5}$ levels, with a mean of 120.7 $\mu g\ m^{-3}$ and a maximum of 290.5 $\mu g\ m^{-3}$. These levels far exceed WHO and Bangladesh standards (WHO 2021), consistent with regional winter stagnation patterns (Zhou et al. 2024).

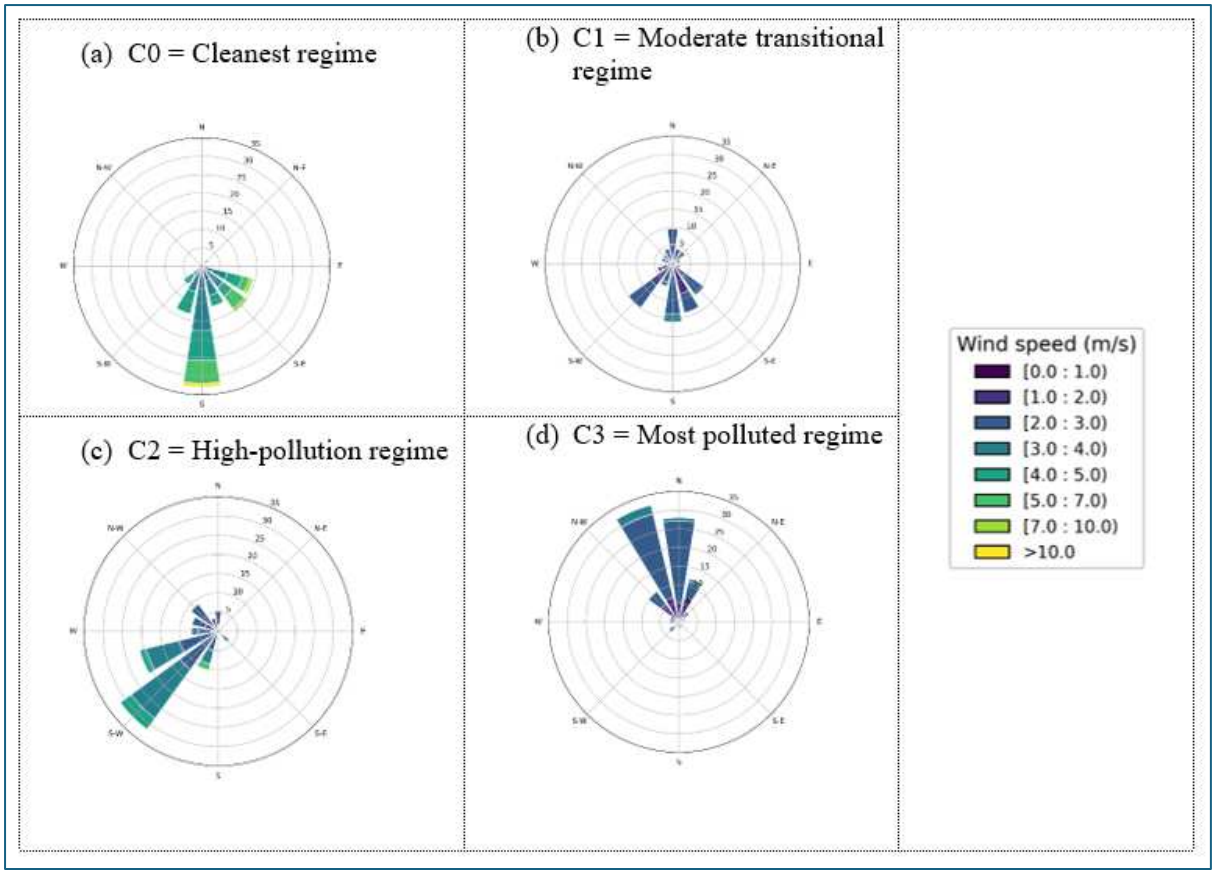


Fig. 5 Wind rose diagrams showing the frequency and speed distribution of wind directions for each of the four K-means atmospheric pollution regimes, Khulna City, 2024: (a) C0, clean maritime regime, (b) C1, moderate transitional regime, (c) C2, high-pollution continental regime, (d) C3, most polluted regime. Radial arms represent wind frequency by directional sector; color gradients represent wind speed bins (m s^{-1}). Dominant directions are calculated using circular mean statistics to avoid angular averaging errors inherent in arithmetic means of degree values. r

Wind rose analysis confirmed the clear wind patterns for each of the four regimes (Fig. 5), SSE high-speed maritime flows in C0, diffuse southerly light winds in C1, WSW moderate transport in C2, and concentrated NNW slow flows in C3, indicating that both wind direction and speed jointly determine the pollution regime identity in Khulna. Directional transport pathways influence the contribution from source regions, while wind speed governs local dispersive capacity. These cluster-identified regimes establish the physical context for interpreting the seasonal patterns observed in the Pearson correlation analysis and the STL residual gap decomposition discussed in subsequent sections.

3.4 Random Forest and STL Decomposition Analysis:

The Random Forest (RF) model accounted for a substantial proportion of $PM_{2.5}$ variability, achieving an out-of-bag (OOB) R^2 of 0.64 (RMSE = 32.39 μg m^{-3}; MAE = 19.87 μg m^{-3}). This result demonstrates that combined meteorological and transport predictors explain approximately 64% of daily $PM_{2.5}$ variance in Khulna, aligning with findings from previous RF-based studies in South and East Asia (Gulati et al. 2023; Borchers-Arriagada et al. 2024). The unexplained variance is attributable to non-meteorological factors, such as emission variability and episodic transport.

Variable importance analysis identified temperature as the main factor affecting $PM_{2.5}$ variability (%IncMSE = 58.7%), followed by surface pressure (50.2%), relative humidity (30.1%), HYSPLIT trajectory distance (20.0%), and precipitation (19.6%). These results confirm that thermodynamic and synoptic-scale processes primarily control conditions at this coastal site. Among transport variables, trajectory distance ranked fourth overall, while northerly flow showed the greatest directional

importance (%IncMSE = 8.0%), aligning with the influence of the Indo-Gangetic Plain (Fig. 6a). STL decomposition showed a strong annual cycle. $PM_{2.5}$ dropped from about 168 µg m^{-3} in winter to around 15 µg m^{-3} during the monsoon, before rising again to about 128 µg m^{-3} by year-end (Fig. 6b). The 30-day seasonal component captured sub-monthly variability. The close match between observed and RF-estimated components suggests that meteorological predictors accurately represent intra-monthly fluctuations (Fig. 6c). The trend emphasizes the strong meteorological controls during the monsoon, driven by maritime inflow, precipitation, and improved mixing. Residual analysis provided further insight into seasonal driver partitioning (Fig. 6d).

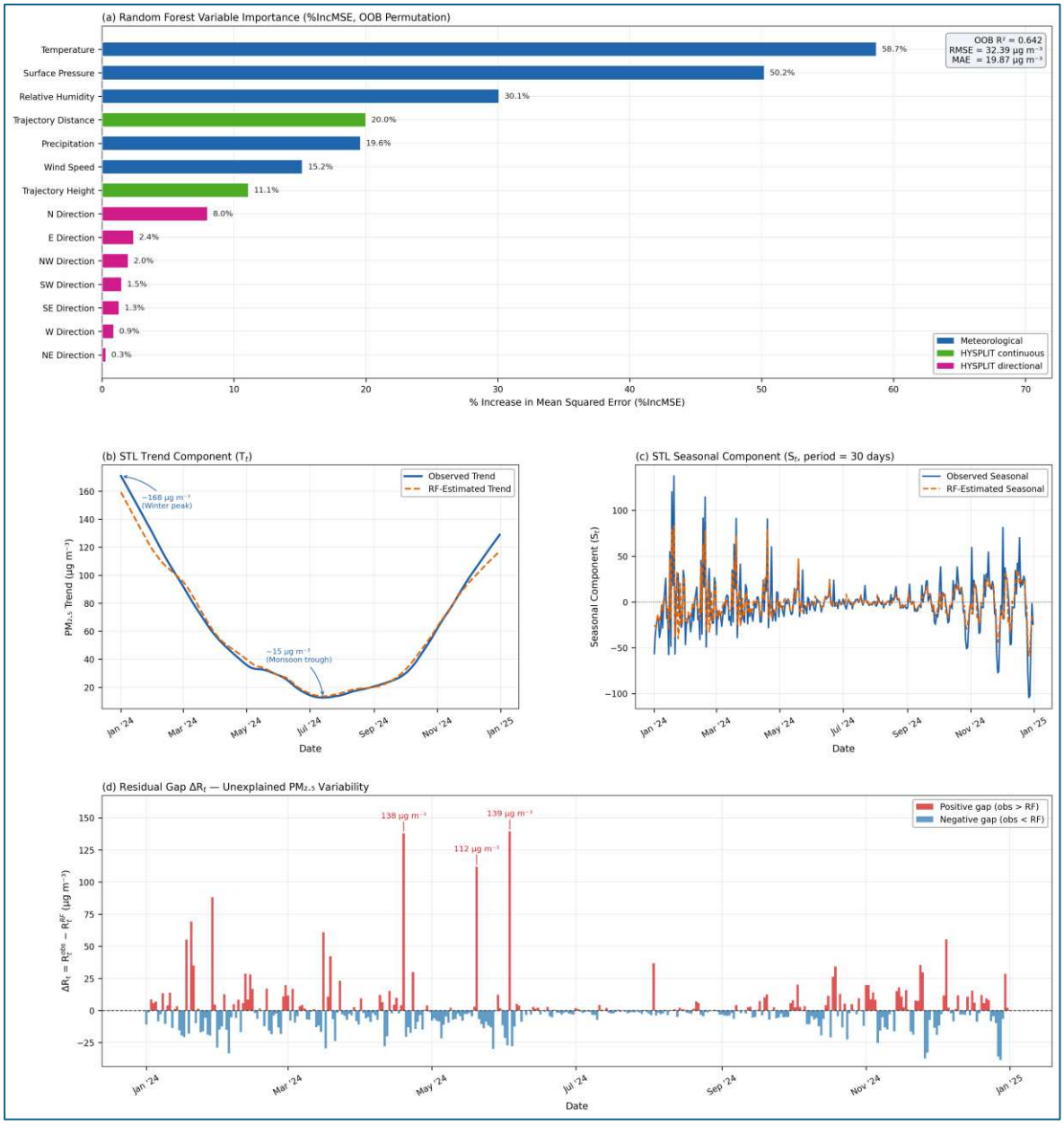


Fig. 6 Random Forest variable importance and STL decomposition of $PM_{2.5}$, Khulna City, 2024. (a) Variable importance scores, expressed as the percentage increase in mean squared error (%IncMSE). Blue bars represent meteorological predictors, green bars represent HYSPLIT continuous transport metrics, and pink bars represent HYSPLIT directional indicators. Error bars indicate ±1 standard deviation across 30 permutation repeats. (b) STL trend component (Tt) of observed (solid blue) and Random Forest-estimated (dashed orange) daily $PM_{2.5}$ time series (c) STL seasonal component (St, period = 30 days) for observed and Random Forest-estimated series, illustrating sub-monthly variability

throughout 2024. (d) Residual gap, where positive values (red) indicate observed $PM_{2.5}$ exceeds Random Forest estimates and negative values (blue) indicate Random Forest overestimation.

Residuals were near zero during the monsoon, confirming strong meteorological control and the model's adequacy. In winter, observed $PM_{2.5}$ was about 8-12 μg m^{-3} higher than RF estimates, indicating additional non-meteorological factors like emissions and episodic transport. The largest deviations occurred in the pre-monsoon season, with extreme positive residuals (about 138 μg m^{-3}). This suggests episodic pollution is likely tied to long-range transport and possible local sources such as biomass burning and industry. These events coincided with higher trajectory heights and westerly flow, which is consistent with descending air masses from the Indo-Gangetic Plain (Zhan et al. 2021; Gao et al. 2024; Liu et al. 2025). Winter and post-monsoon periods showed intermediate residuals, reflecting a mix of meteorological and emission influences.

Conclusion

This study examined the temporal dynamics of $PM_{2.5}$ concentrations in Khulna City, southwestern Bangladesh, using 2024 continuous monitoring data. The central aim was to quantify the combined influence of local meteorological conditions and regional air-mass transport on $PM_{2.5}$ variability through an integrated framework incorporating MLR, Random Forest, HYSPLIT trajectory analysis, STL decomposition, and K-means clustering. The findings demonstrate that $PM_{2.5}$ pollution in Khulna is governed by thermodynamic boundary-layer controls, synoptic-scale meteorology, and cross-regional atmospheric transport, with driver dominance shifting markedly across seasons.

$PM_{2.5}$ exhibited a strong seasonal cycle, with winter means (~121 μg m^{-3}) exceeding the WHO 2021 24-hour guideline nearly eightfold, while monsoon concentrations declined to near-compliant levels (~15 μg m^{-3}). Temperature and relative humidity were the dominant negative meteorological predictors across all diurnal and seasonal models. The Random Forest model explained 64% of daily variance, with temperature, surface pressure, and HYSPLIT trajectory distance as the three leading predictors. STL residual analysis revealed systematic winter underestimation and extreme pre-monsoon pollution episodes linked to long-range continental transport and possible biomass burning from the Indo-Gangetic Plain. HYSPLIT analysis confirmed a bimodal transport regime: short-range northwesterly continental pathways during dry seasons and extended southerly maritime inflows during the monsoon. K-means clustering identified four coherent pollution-meteorology regimes, reinforcing that wind direction and transport distance jointly determine pollution exposure.

Key limitations include reliance on a single monitoring station, one year of data, an 18.3% data gap, and coarse-resolution reanalysis inputs. Future research should prioritize multi-station networks, multi-year datasets to resolve ENSO-driven variability (Islam et al., 2023a; Hu et al., 2024b), and integration of chemical source apportionment with trajectory analysis. Climate-projection-informed air quality modeling would further clarify how shifting monsoon dynamics may alter future $PM_{2.5}$ exposure. Season-specific mitigation strategies targeting winter stagnation and pre-monsoon transport episodes are urgently needed to reduce the health burden across Khulna's non-monsoon period.

Statements and Declarations

Funding Information

This research received no external funding; the study was self-funded by the authors.

Authors' Contributions

Conceptualization: Sharfan Upaul; Methodology: Sharfan Upaul; Formal analysis and investigation: Sharfan Upaul, Anika Tabassum; Data curation: Anika Tabassum; Writing –

original draft preparation: Sharfan Upaul, Anika Tabassum; Writing-review and editing: Sharfan Upaul, Anika Tabassum, Mahmudul Hasan. Manuscript formatting was carried out by Sharfan Upaul.

Ethical Approval

This is not applicable

Consent to Participate

This is not applicable

Consent to Publish

This is not applicable

Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

Data will be made available on reasonable request

Declaration of generative AI use

Only abstract and conclusion portion of this article has been polished with AI tool "Claude" for increasing readability. Premium version of 'Grammarly' has been used to improve the writing throughout the manuscript.

References

1. Afrin S, Islam MM (2020) LINKING AIR QUALITY TO METEOROLOGY: A MULTILINEAR REGRESSION APPROACH. Khulna University of Engineering & Technology, Khulna, Bangladesh
2. Afrin S, Islam MM, Ahmed T (2021) A meteorology based particulate matter prediction model for megacity dhaka. Aerosol Air Qual Res 21:. https://doi.org/10.4209/aaqr.2020.07.0371
3. Ali MA, Assiri ME, Bilal M, Tariq S, de Leeuw G, Islam MN, Wang Y, Alamri L, Ghulam AS, Shahid S (2025) Long-term PM2.5 exposure in Bangladesh: identification of pollution hotspots, trends, sources and health risk assessment. Air Qual Atmos Health 18:2229–2246. https://doi.org/10.1007/s11869-025-01768-7
4. Al-Muyeed A (2016) Urban air pollution: a Bangladesh perspective. https://doi.org/10.2495/AIR050611
5. Begum BA, Biswas SK, Nasiruddin M (2010) TREND AND SPATIAL DISTRIBUTION OF AIR PARTICULATE MATTER POLLUTION IN DHAKA CITY
6. Borchers-Arriagada N, Morgan GG, Van Buskirk J, Gopi K, Yuen C, Johnston FH, Guo Y, Cope M, Hanigan IC (2024) Daily PM2.5 and Seasonal-Trend Decomposition to Identify Extreme Air Pollution Events from 2001 to 2020 for Continental Australia Using a Random Forest Model. Atmosphere (Basel) 15:. https://doi.org/10.3390/atmos15111341
7. Breiman L (2001) Random Forests Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

8. Bucchignani E (2024) A Wind Field Reconstruction from Numerical Weather Prediction Data Based on a Meteo Particle Model. Meteorology 3:70–82. https://doi.org/10.3390/meteorology3010003
9. Bushra F, Rahman SM Roy S, Roy TK (2019) Assessment of Relative Risk Exposure to PM 2.5 in Khulna City Assessment of Relative Risk Exposure to PM 2.5 in Khulna City ASSESMENT OF RELATIVE RISK EXPOSURE TO PM 2.5 IN KHULNA CITY
10. Choi W, Ho CH, Lee Y (2024) Temporal pattern classification of PM2.5 chemical compositions in Seoul, Korea using K-means clustering analysis. Science of the Total Environment 927:. https://doi.org/10.1016/j.scitotenv.2024.172157
11. Cleveland RB, Cleveland WS, McRae JE, Terpenning I (1990) STL: A Seasonal-Trend Decomposition Procedure Based on Loess. 6:3–73
12. Cui L, Song X, Zhong G (2021) Comparative analysis of three methods for hysplit atmospheric trajectories clustering. Atmosphere (Basel) 12:. https://doi.org/10.3390/atmos12060698
13. Davies DL, Bouldin DW (1979) A Cluster Separation Measure. IEEE Trans Pattern Anal Mach Intell PAMI-1:224–227. https://doi.org/10.1109/TPAMI.1979.4766909
14. Dihan MR, Nayeem SA, Choudhury MS (2021) Impact of Trans-Boundary Pollution (PM-2.5) on the Air Quality of Dhaka City in Bangladesh. Chemical Engineering Research Bulletin 114–120. https://doi.org/10.3329/cerb.v22i1.54309
15. Draxler RR, Hess GD (1998) An overview of the HYSPLIT_4 modelling system for trajectories, dispersion, and deposition. Aust Meteorol Mag 47:295–308. https://doi.org/10.1071/ES98032
16. EPA (2012) The National Ambient Air Quality Standards for Particle Pollution REVISED AIR QUALITY STANDARDS FOR PARTICLE POLLUTION AND UPDATES TO THE AIR QUALITY INDEX (AQI)
17. Espinoza-Guillen JA, Alderete-Malpartida MB, Roncal-Romero FD, Vilcanqui-Sarmiento JC (2025) Identification of particulate matter (PM10 and PM2.5) sources using bivariate polar plots and k-means clustering in a South American megacity: Metropolitan Area of Lima-Callao, Peru. Environ Monit Assess 197:. https://doi.org/10.1007/s10661-025-13696-1
18. Gao A, Gao B, Li S, Yao W, Sun W, Cao X, Liu Y, Zhang H, Zhang B (2024) Pollution characteristics, potential source areas, and transport pathways of PM2.5 and O3 in an inland city of Shijiazhuang, China. Air Qual Atmos Health 17:1307–1323. https://doi.org/10.1007/s11869-024-01508-3
19. Graham MH (2003) CONFRONTING MULTICOLLINEARITY IN ECOLOGICAL MULTIPLE REGRESSION. Ecology 84:2809–2815. https://doi.org/10.1890/02-3114
20. Gujarati DN, Porter DC (2009) Basic Econometrics, 5th edn. McGraw-Hill, New York
21. Gulati S, Bansal A, Pal A, Mittal N, Sharma A, Gared F (2023) Estimating PM2.5 utilizing multiple linear regression and ANN techniques. Sci Rep 13:. https://doi.org/10.1038/s41598-023-49717-7
22. Gupta A, Moniruzzaman M, Hande A, Rousta I, Olafsson H, Mondal KK (2020) Estimation of particulate matter (PM2.5, PM10) concentration and its variation over urban sites in Bangladesh. SN Appl Sci 2:. https://doi.org/10.1007/s42452-020-03829-1
23. Hassan MS, Bhuiyan M (2023) Spatiotemporal Mapping and Modeling Hotspot of PM2.5 in the Central part of Bangladesh. https://doi.org/10.29150/2237-2202.2023.257250
24. Hopke PK (2016) Review of receptor modeling methods for source apportionment. J Air Waste Manage Assoc 66:237–259. https://doi.org/10.1080/10962247.2016.1140693
25. Hu R, Lian T, Liu T, Wang J, Song X, Chen H (2024a) Predicting the 2023/24 El Niño from a multi-scale and global perspective. Commun Earth Environ 5:. https://doi.org/10.1038/s43247-024-01867-w
26. Hu Y, Yu H, Kang S, Yang J, Chen X, Yin X, Chen P (2024b) Modeling the transport of PM10, PM2.5, and O3 from South Asia to the Tibetan Plateau. Atmos Res 303:107323. https://doi.org/10.1016/j.atmosres.2024.107323

27. Islam MA, Mallick J, Pal SC, Chakraborty R, Fattah MA, Ghose B, Kakoli MKA, Islam MA, Naqvi HR, Bilal M, Elbeltagi A (2023a) Estimating ground-level PM2.5 using subset regression model and machine learning algorithms in Asian megacity, Dhaka, Bangladesh. Air Qual Atmos Health 16:1117–1139. https://doi.org/10.1007/s11869-023-01329-w
28. Islam MS, Roy S, Tusher TR, Sultana S, Kafy AA, Ashrafi TJ, Rahman MM (2023b) Assessment of Spatio-Temporal Variations in PM2.5 and Associated Long-Range Air Mass Transport and Mortality in South Asia. Remote Sens (Basel) 15:. https://doi.org/10.3390/rs15204975
29. Islam N, Toha TR, Islam MM, Ahmed T (2023c) Spatio-temporal Variation of Meteorological Influence on PM2.5 and PM10 over Major Urban Cities of Bangladesh. Aerosol Air Qual Res 23:. https://doi.org/10.4209/aaqr.220082
30. Junninen H, Niska H, Tuppurainen K, Ruuskanen J, Kolehmainen M (2004) Methods for imputation of missing values in air quality data sets. Atmos Environ 38:2895–2907. https://doi.org/10.1016/j.atmosenv.2004.02.026
31. Khatun F, Saadat SY, Ashraf K (2021) BREATHING UNEASY: AN ASSESSMENT OF AIR POLLUTION IN BANGLADESH
32. Kliengchuay W, Srimanus R, Srimanus W, Niampradit S, Preecha N, Mingkhwan R, Worakhunpiset S, Limpanont Y, Moonsri K, Tantrakarnapa K (2021) Particulate matter (PM10) prediction based on multiple linear regression: a case study in Chiang Rai Province, Thailand. BMC Public Health 21:. https://doi.org/10.1186/s12889-021-12217-2
33. Liu X, Wen W, Ma X, Qian D, Zhang W, Wang S (2025) Air Pollution in a Northwest Chinese Valley City (2020–2024): Integrated WRF-HYSPLIT Modeling of Pollution Characteristics, Meteorological Drivers, and Transport Pathways in Yining. Toxics 13:. https://doi.org/10.3390/toxics13100868
34. Mazumder P, Kabir SB, Saju JA, Rafizul IM, Emon AI (2022) Measuring and Modeling PM2.5 Zonal Distributions, Assembling Geospatial and Meteorological Variables in the Khulna Metropolitan Area. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4210956
35. Newey WK, West KD (1987) A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. Econometrica 55:703. https://doi.org/10.2307/1913610
36. O'Brien RM (2007) A caution regarding rules of thumb for variance inflation factors. Qual Quant 41:673–690. https://doi.org/10.1007/s11135-006-9018-6
37. Ommi A, Emami F, Zíková N, Hopke PK, Begum BA (2017) Trajectory-based models and remote sensing for biomass burning assessment in Bangladesh. Aerosol Air Qual Res 17:465–475. https://doi.org/10.4209/aaqr.2016.07.0304
38. Pandey SK, Vinoj V, Landu K, Babu SS (2017) Declining pre-monsoon dust loading over South Asia: Signature of a changing regional climate. Sci Rep 7:. https://doi.org/10.1038/s41598-017-16338-w
39. Rahman M, Meng L (2024) Examining the Spatial and Temporal Variation of PM2.5 and Its Linkage with Meteorological Conditions in Dhaka, Bangladesh. Atmosphere (Basel) 15:. https://doi.org/10.3390/atmos15121426
40. Rahman SM, Ali MS, Al Mamun MA, Uddin MN (2020) Estimation of seasonal boundaries using temperature data: a case of northwest part of Bangladesh. Mathematics of Climate and Weather Forecasting 6:50–62. https://doi.org/10.1515/mcwf-2020-0102
41. Rana MM, Biswas SK, Azad MAK, Khan SMMH (2018) CLEAN AIR AND SUSTAINABLE ENVIRONMENT PROJECT Department of Environment Ministry of Environment, Forest and Climate Change Government of the People's Republic of Bangladesh AMBIENT AIR QUALITY IN BANGLADESH Draxler
42. Rautela KS, Goyal MK, Nagpure AS (2025) Unequal spatio-temporal distribution of population-weighted pollution extremes through deep learning. NPJ Clim Atmos Sci 8:. https://doi.org/10.1038/s41612-025-01183-w

43. Rendana M, Razi Idris WM, Rahim SA (2022) Clustering analysis of PM2.5 concentrations in the South Sumatra Province, Indonesia, using the Merra-2 Satellite Application and Hierarchical Cluster Method. AIMS Environ Sci 9:754–770. https://doi.org/10.3934/environsci.2022043
44. Rheingrover SW, Gordon GE (1988) Wind-trajectory method for determining compositions of particles from major air pollution sources. Aerosol Science and Technology 8:29–61. https://doi.org/10.1080/02786828808959170
45. Rolph G, Stein A, Stunder B (2017) Real-time Environmental Applications and Display sYstem: READY. Environmental Modelling and Software 95:210–228. https://doi.org/10.1016/j.envsoft.2017.06.025
46. Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65. https://doi.org/10.1016/0377-0427(87)90125-7Yang et all
47. Saha S, Moorthi S, Pan HL, Wu X, Wang J, Nadiga S, Tripp P, Kistler R, Woollen J, Behringer D, Liu H, Stokes D, Grumbine R, Gayno G, Wang J, Hou YT, Chuang HY, Juang HMH, Sela J, Iredell M, Treadon R, Kleist D, Van Delst P, Keyser D, Derber J, Ek M, Meng J, Wei H, Yang R, Lord S, Van Den Dool H, Kumar A, Wang W, Long C, Chelliah M, Xue Y, Huang B, Schemm JK, Ebisuzaki W, Lin R, Xie P, Chen M, Zhou S, Higgins W, Zou CZ, Liu Q, Chen Y, Han Y, Cucurull L, Reynolds RW, Rutledge G, Goldberg M (2010) The NCEP climate forecast system reanalysis. Bull Am Meteorol Soc 91:1015–1057. https://doi.org/10.1175/2010BAMS3001.1
48. Saju JA, Bari QH, Mohiuddin KABM, Strezov V (2023) Measurement of ambient particulate matter (PM1.0, PM2.5 and PM10) in Khulna City of Bangladesh and their implications for human health. Environmental Systems Research 12:. https://doi.org/10.1186/s40068-023-00327-2
49. Sreenivasulu T, Rayalu GM (2024) Enhanced PM2.5 prediction in Delhi using a novel optimized STL-CNN-BILSTM-AM hybrid model. Asian Journal of Atmospheric Environment 18:. https://doi.org/10.1007/s44273-024-00048-7
50. Stein AF, Draxler RR, Rolph GD, Stunder BJB, Cohen MD, Ngan F (2015) Noaa's hysplit atmospheric transport and dispersion modeling system. Bull. Am. Meteorol. Soc. 96:2059–2077
51. Wang HL, Qiao LP, Lou SR, Zhou M, Ding AJ, Huang HY, Chen JM, Wang Q, Tao SK, Chen CH, Li L, Huang C (2016) Chemical composition of PM2.5 and meteorological impact among three years in urban Shanghai, China. J Clean Prod 112:1302–1311. https://doi.org/10.1016/j.jclepro.2015.04.099
52. Wang J, Ogawa S (2015) Effects of meteorological conditions on PM2.5 concentrations in Nagasaki, Japan. Int J Environ Res Public Health 12:9089–9101. https://doi.org/10.3390/ijerph120809089
53. WHO (2021) WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. World Health Organization, Geneva. https://www.who.int/publications/i/item/9789240034228Bushra
54. Wu DL, Lin M, Chan CY, Li WZ, Tao J, Li YP, Sang XF, Bu CW (2013) Influences of commuting mode, air conditioning mode and meteorological parameters on fine particle (PM2.5) exposure levels in traffic microenvironments. Aerosol Air Qual Res 13:709–720. https://doi.org/10.4209/aaqr.2012.08.0212
55. Yang Q, Yuan Q, Li T, Shen H, Zhang L (2017) The relationships between PM2.5 and meteorological factors in China: Seasonal and regional variations
56. Yerramilli A, Dodla VBR, Challa VS, Myles LT, Pendergrass WR, Vogel CA, Dasari HP, Tuluri F, Baham JM, Hughes RL, Patrick C, Young JH, Swanier SJ, Hardy MG (2012) An integrated WRF/HYSPLIT modeling approach for the assessment of PM2.5 source regions over the Mississippi Gulf Coast region. Air Qual Atmos Health 5:401–412. https://doi.org/10.1007/s11869-010-0132-1

57. Yu Y, Zhou H, Zhao Z, Chang Y, Wu D, Li Z, Wang F, Fang M, Zhou X (2024) Spatiotemporal Distribution, Meteorological Influence, and Potential Sources of Air Pollution over Hainan Island, China. Atmosphere (Basel) 15:. https://doi.org/10.3390/atmos15111336
58. Zhan Y, Xie M, Gao D, Wang T, Zhang M, An F (2021) Characterization and source analysis of water-soluble inorganic ionic species in PM2.5 during a wintertime particle pollution episode in Nanjing, China. Atmos Res 262:. https://doi.org/10.1016/j.atmosres.2021.105769
59. Zhang YL, Cao F (2015) Fine particulate matter (PM 2.5) in China at a city level. Sci Rep 5:. https://doi.org/10.1038/srep14884
60. Zhou M, Xie Y, Wang C, Shen L, Mauzerall DL (2024) Impacts of current and climate induced changes in atmospheric stagnation on Indian surface PM2.5 pollution. Nature Communications 15:. https://doi.org/10.1038/s41467-024-51462-y