AI-Driven PM2.5 Forecasting with Weather Signals

Xuanming Zhang

Abstract

This project presents a complete pipeline for hourly $PM_{2.5}$ nowcasting, encompassing data ingestion, cleaning, exploratory data analysis (EDA), feature engineering, scaling, and supervised learning with time-aware evaluation. The system integrates pollution and weather signals to generate accurate, near-real-time predictions of air quality, providing actionable insights for public health monitoring and environmental management.

1 Introduction

Air pollution remains one of the most pressing environmental health challenges worldwide, with fine particulate matter ($PM_{2.5}$) posing significant risks to respiratory and cardiovascular health. As urban populations grow and industrial activity intensifies, the ability to monitor and forecast air quality in real time becomes increasingly critical—not only for informing public advisories but also for enabling data-driven environmental governance.

This report focuses on developing a complete AI-powered pipeline for $PM_{2.5}$ nowcasting—that is, predicting current-hour pollution levels using real-time meteorological and air-quality data. The motivation behind this work lies in addressing the temporal gap between pollutant measurement and actionable response. Traditional air quality forecasts often operate on daily scales and are limited in resolution. By contrast, nowcasting provides high-frequency, localized predictions that can support more agile and responsive public health decisions.

The pipeline includes all stages of data processing: ingestion, cleaning, exploratory data analysis (EDA), time- and cycle-based feature engineering, scaling, and supervised learning with time-aware model validation. Through this structured approach, we aim to demonstrate how machine learning techniques can be applied not just for forecasting future air quality, but for real-time environmental awareness and early intervention.

2 Data and Preprocessing

I used an hourly dataset that includes both air quality and meteorological data. The original columns were labeled `pm2.5`, `DEWP`, `TEMP`, `PRES`, `cbwd`, `Iws`, `Is`, and `Ir`. To improve clarity and consistency throughout the project, I renamed these columns. I also dropped the first day of records to eliminate partial and incomplete hourly entries.

To prepare the data for modeling, I coerced all relevant columns to numeric types and clipped negative values in the `pollution` column at zero. I addressed missing values by applying both forward and backward filling methods. For categorical consistency, I standardized wind direction entries into a fixed set of compass directions such as `N`, `NNE`, and so on.

3 Exploratory Data Analysis

3.1 Time-Series Structure

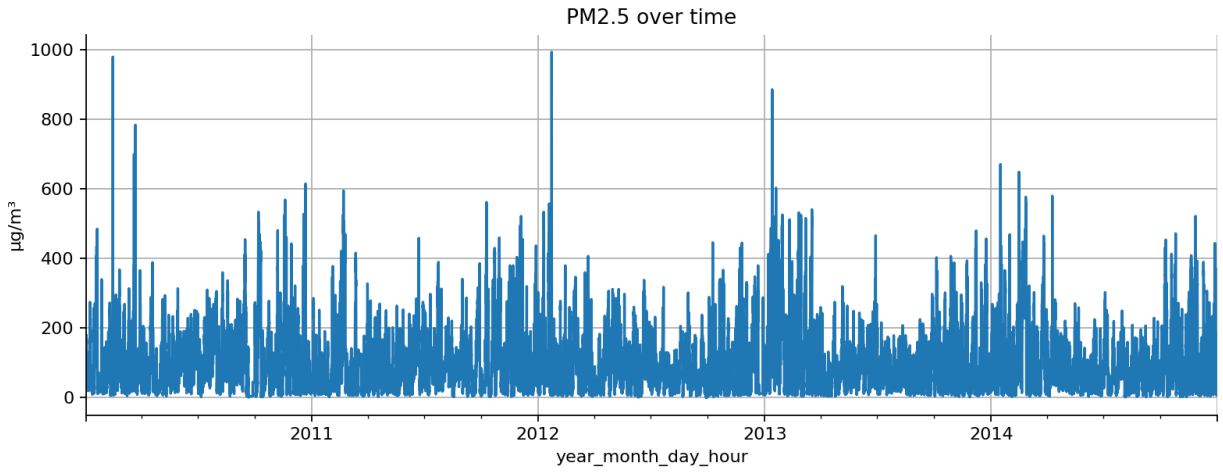

Figure 1: Hourly $PM_{2.5}$ timeline. I observe clear episodes of spikes interleaved with low baselines, suggesting a mix of persistent background and transient events (e.g., traffic rushes, meteorological shifts).

3.2 Distributions and Outliers

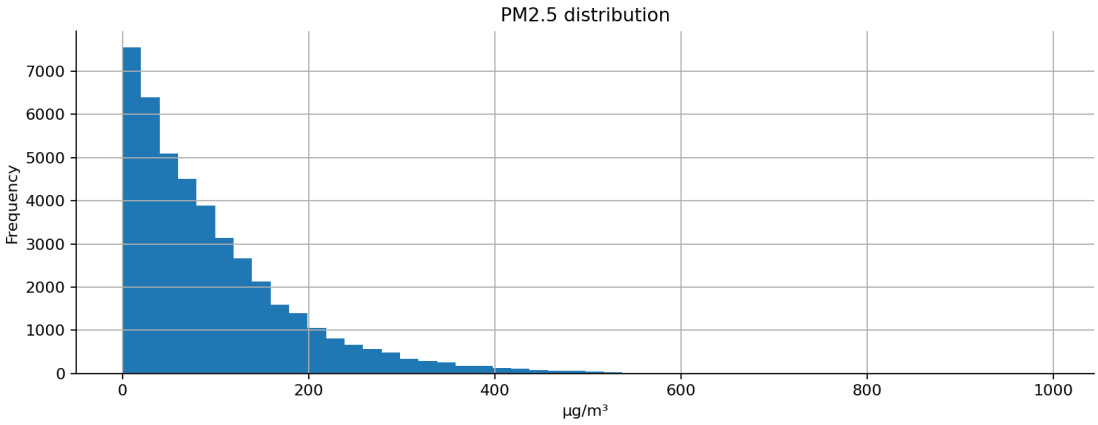

Figure 2: $PM_{2.5}$ histogram. The right tail is heavy, indicating occasional high-pollution episodes that will dominate RMSE more than MAE.

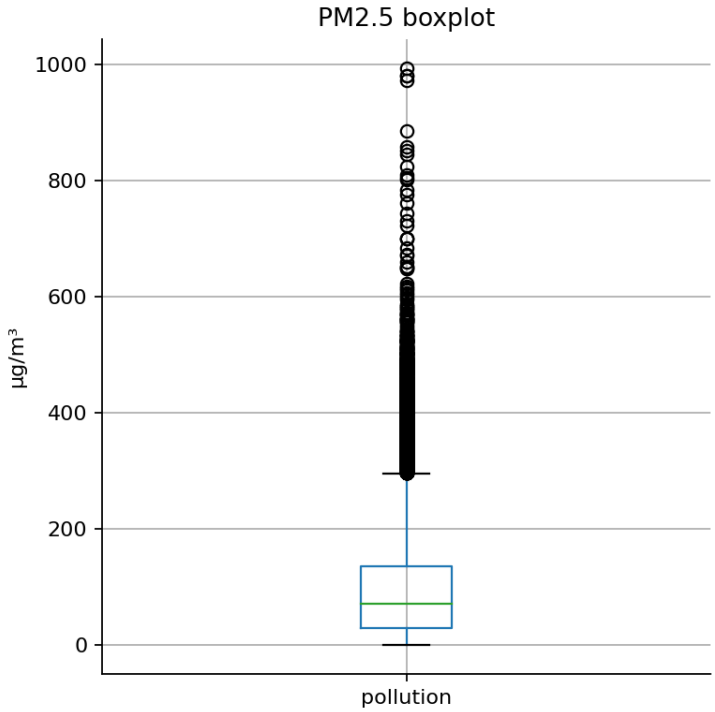

Figure 3: $PM_{2.5}$ boxplot. Outliers are frequent during episodes, reinforcing the value of robust models and error metrics reported together.

3.3 Seasonality and Periodicity

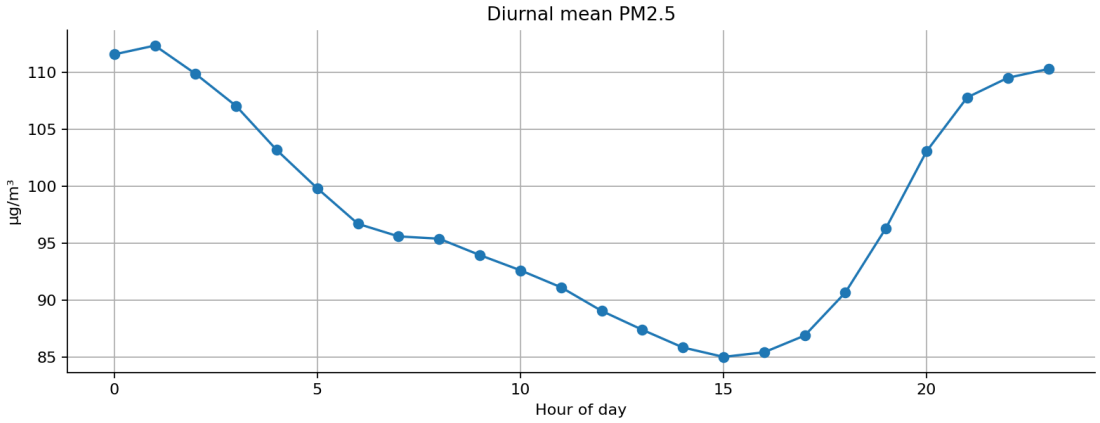

Figure 4: Diurnal mean $PM_{2.5}$. I see higher means around late evening/early night and lower values mid-day, consistent with boundary-layer dynamics and traffic patterns.

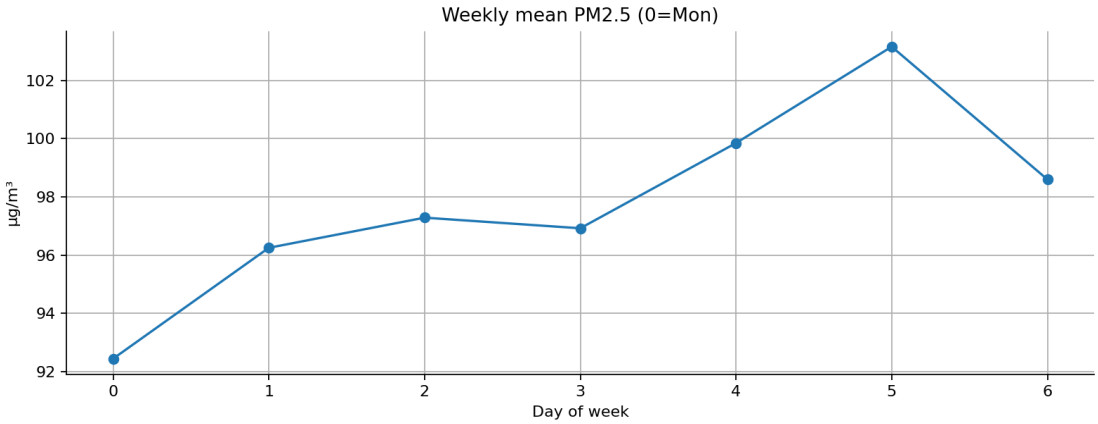

Figure 5: Weekly mean $PM_{2.5}$. Differences across weekdays are modest; weekends show slightly altered profiles, which I capture with cyclical features.

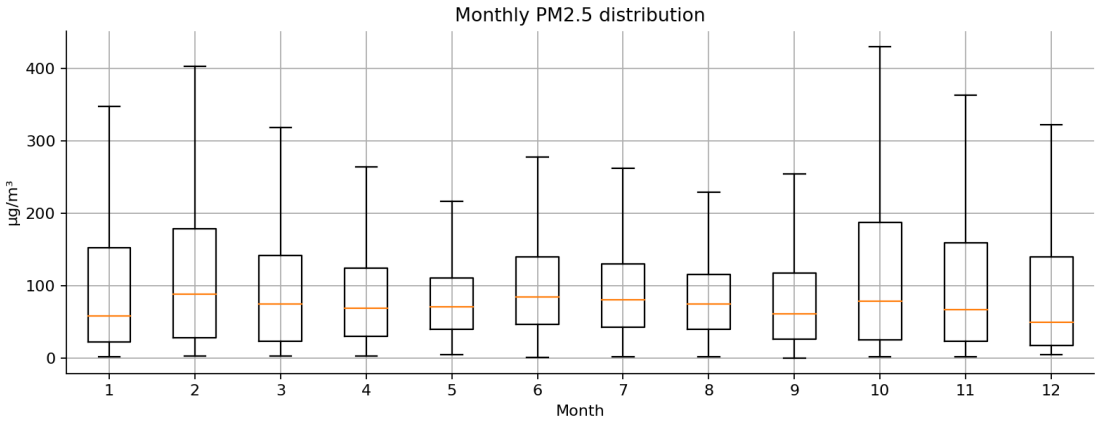

Figure 6: Monthly distributions. Broader boxes and higher medians in specific months indicate seasonal regimes that justify including month cycles.

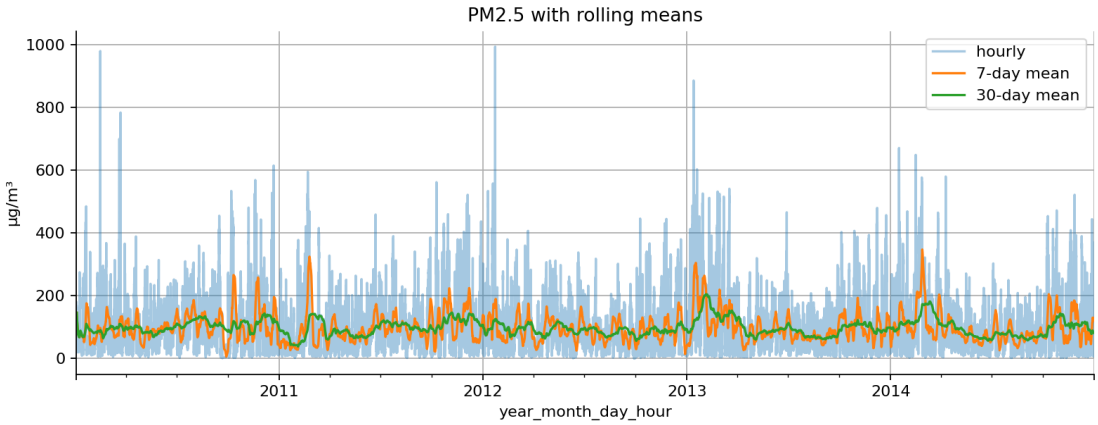

Figure 7: 7/30-day rolling means. The smoothed curves highlight persistent seasonal envelopes overlaid on high-frequency variability.

3.4 Autocorrelation Proxies and Covariates

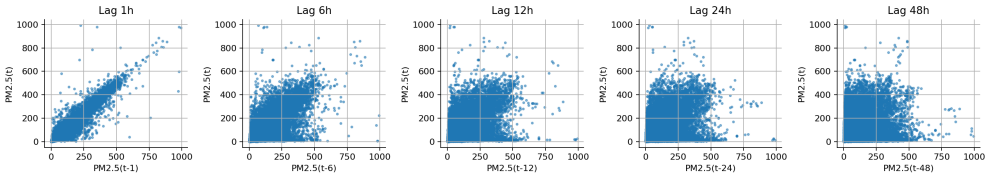

Figure 8: Lag scatter panels. Strong alignment near the diagonal for short lags indicates high persistence, underpinning the value of a lag-1 feature and the persistence baseline.

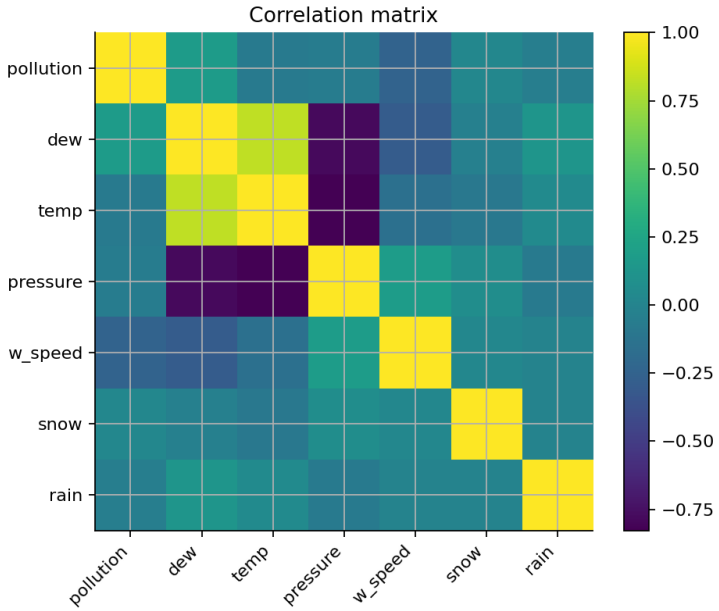

Figure 9: Correlation matrix. Temperature, pressure, and wind speed show meaningful associations with $PM_{2.5}$, while precipitation proxies (`snow`/`rain`) relate to scavenging episodes.

3.5 Wind Effects

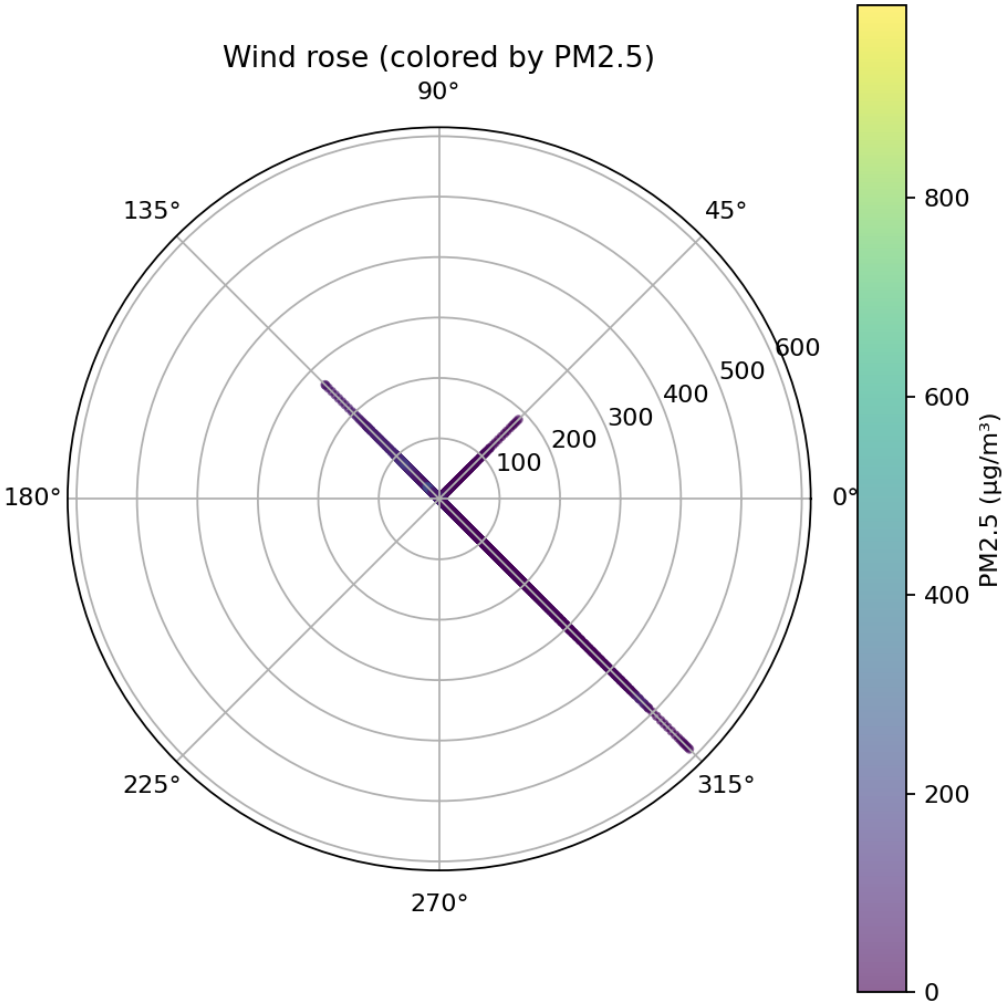

Figure 10: Polar wind rose (colored by $PM_{2.5}$). Specific sectors co-occur with higher concentrations, motivating directional encoding via `w_sin`/`w_cos`.

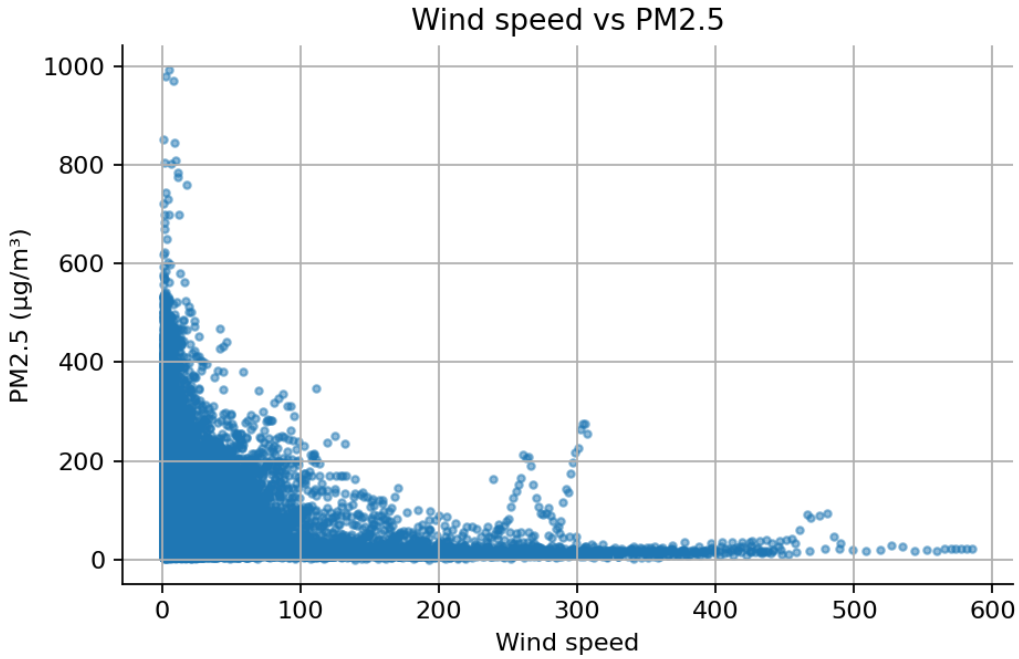

Figure 11: Wind speed vs $PM_{2.5}$. Higher speeds generally reduce concentrations (ventilation), but strong events can coincide with transported plumes.

4 Feature Engineering, Scaling, and Selection

I use cyclic encodings for time (`hour_sin/cos`, `dow_sin/cos`, `mon_sin/cos`) and wind (`w_sin`/`w_cos`). I optionally add `pollution_lag1`. This project applies median imputation and min–max scaling fit on the training split only. I select up to 20 features using mutual information to retain the most informative signals for nowcasting.

5 Modeling and Evaluation

I frame the task as predicting next-hour $PM_{2.5}$ (y_{t+1}) from current-hour features—pollution, weather variables, cyclical time encodings, and wind direction encoded numerically—with an optional lag-1 term, using a chronological split of 70% training, 15% validation, and 15% testing. I compare a baseline **Linear Regression** model with a **Random Forest Regressor** that captures nonlinear effects while remaining robust after normalization. For each model, I report MAE, RMSE, MAPE, and R^2, and I provide diagnostic visualizations including timeline overlays, true-versus-predicted scatter plots, residual traces and histograms, and hour-of-day error profiles.

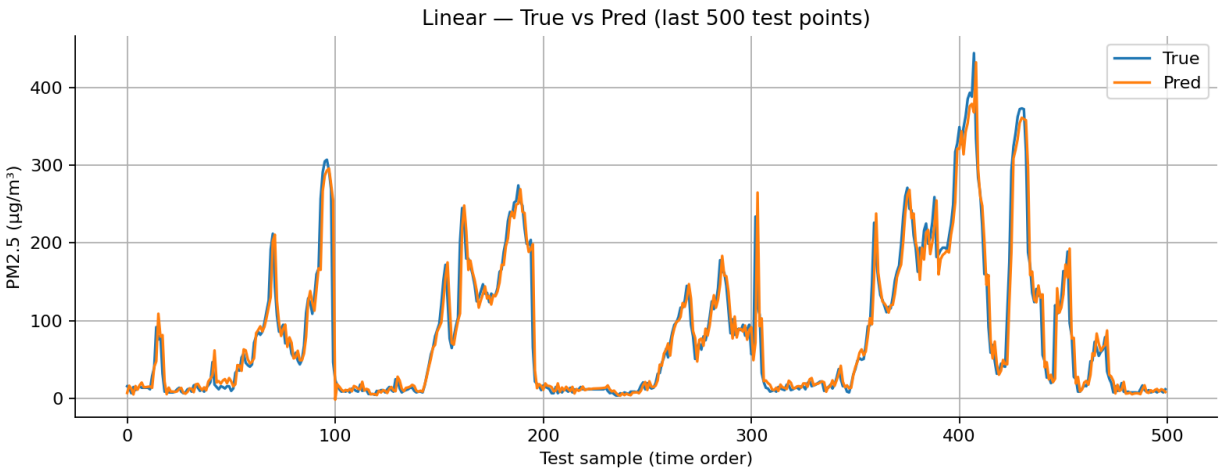

Figure 12: Linear model: timeline overlay (test). (file: `models_out/linear_ts.png`)

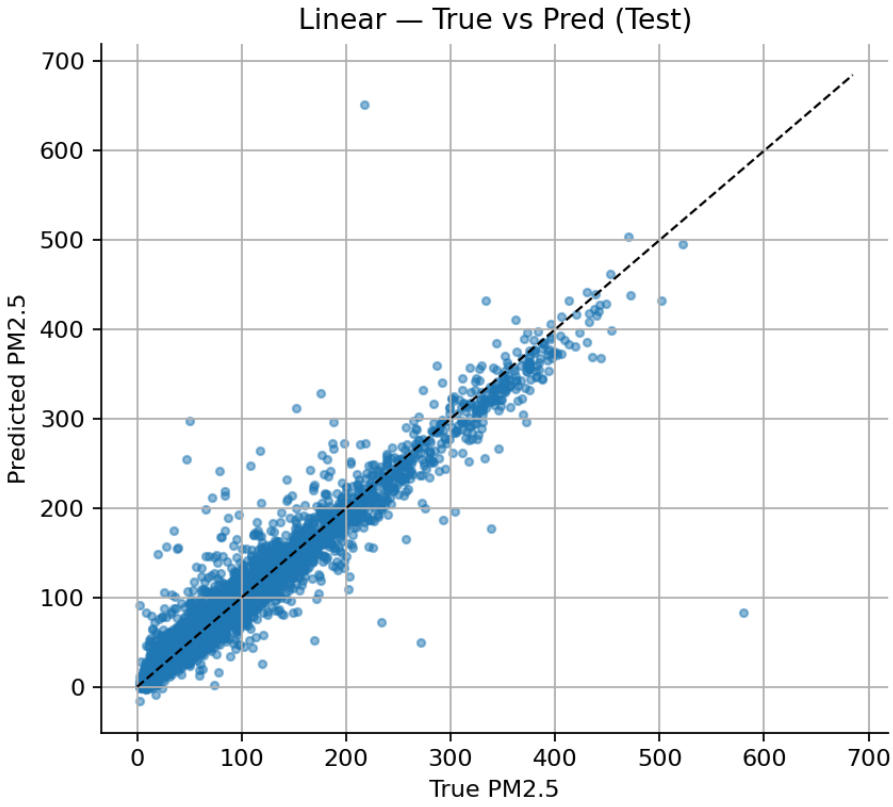

Figure 13: Linear model: true vs predicted (test). (file: `models_out/linear_scatter.png`)

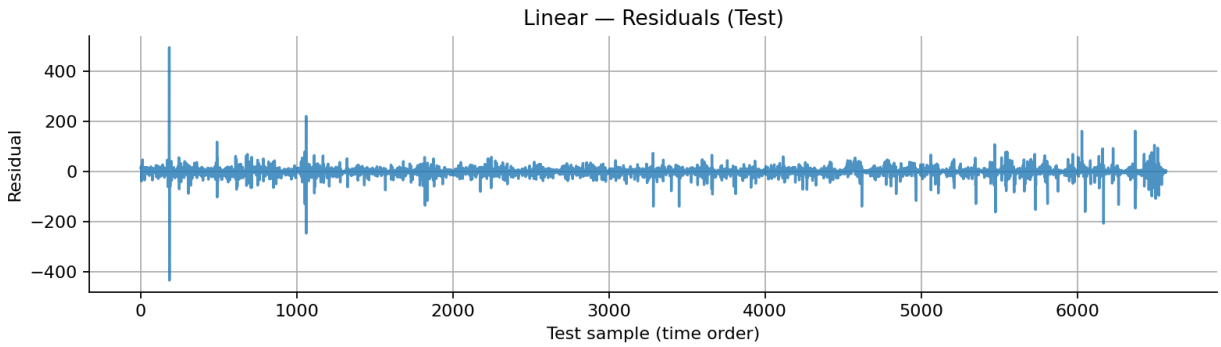

Figure 14: Linear model: residual trace (test). (file: `models_out/linear_residuals.png`)

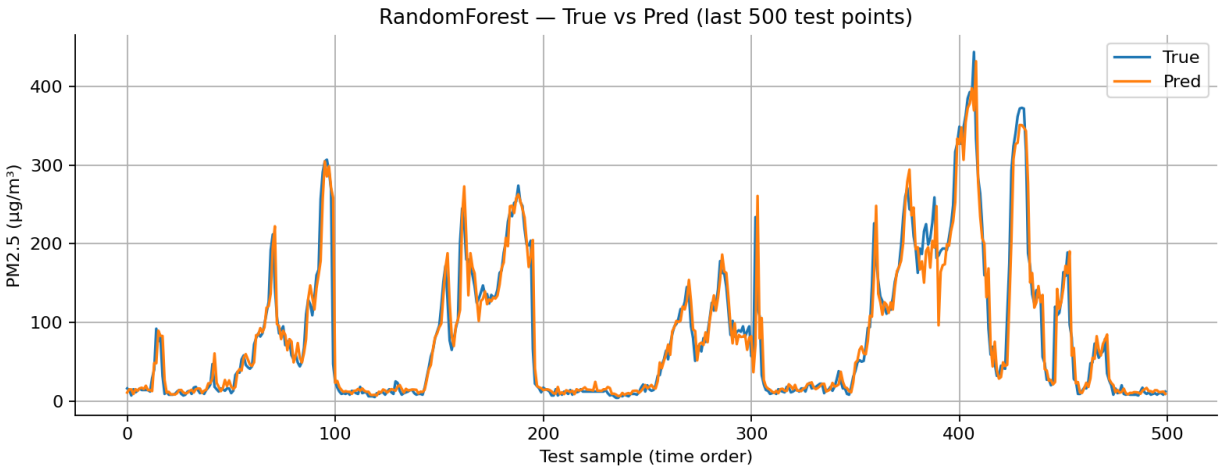

Figure 15: Random Forest: timeline overlay (test).

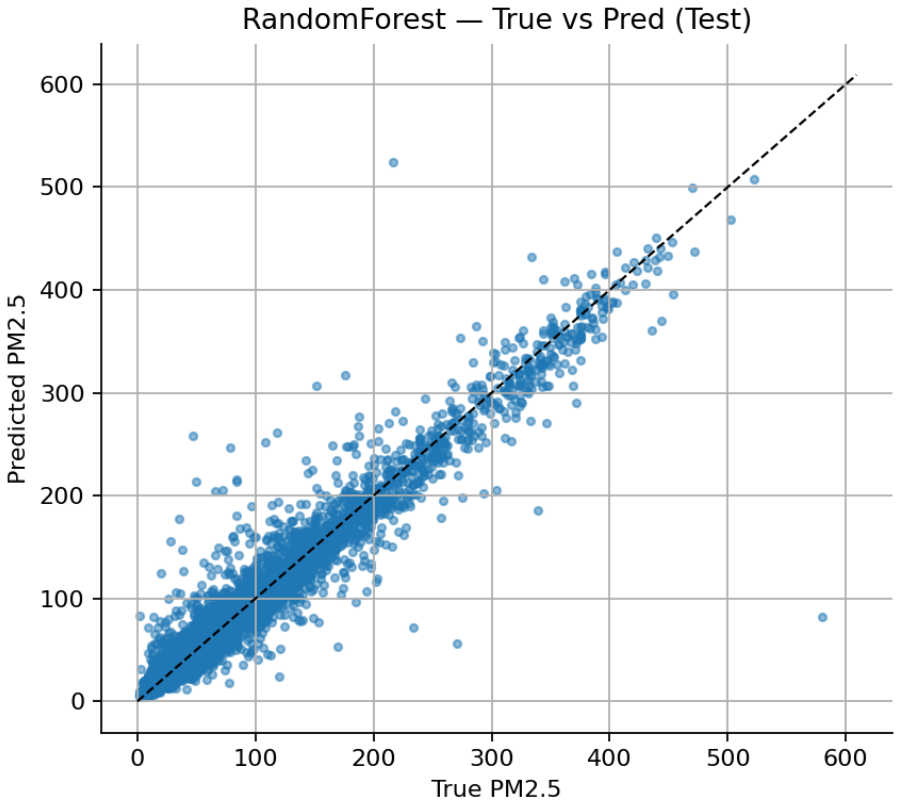

Figure 16: Random Forest: true vs predicted (test).

6 Findings and Discussion

I observe strong diurnal and seasonal structure, non-Gaussian pollution distributions with heavy right tails, and material associations with meteorology and wind. Lag scatter panels confirm short-term persistence.

Model Performance The Random Forest generally improves upon the linear baseline and naive persistence, with lower MAE/RMSE and better hour-of-day error profiles. Feature importance indicates substantial weight on recent pollution level, wind-derived features, and temperature/pressure covariates.