Big Data Analysis Project- Assignment 1
Part C

Impact of Climate Variability
on
Urban Air Pollution:
A Big Data Analysis of Sydney (2015–2025)

Utsav Punia : a1956304

The University of Adelaide

4533_COMP_SCI_7209 : Big Data Project

Table of Contents

1. Problem Description

This project aims to develop predictive models to estimate the concentration of fine particulate matter (PM2.5) in urban areas of Sydney based on meteorological variables. Building on insights from Assignment 1B, where correlations and patterns were explored across multiple air quality monitoring sites, this phase focuses on using machine learning models to forecast pollution levels using historical environmental data.

Main Research Question:

"How has climate variability influenced air pollution levels in Sydney over the past decade?"

Supporting Questions(RSQs: Research Supporting Questions):

- Can **extreme pollution events** (e.g., high PM2.5 spikes) be attributed to **non-meteorological events**, such as bushfires or localized emissions?
- Are there **distinct temporal signatures** (e.g., weekday vs weekend patterns) that can further improve prediction models for pollutants like NO_2?
- Can we develop a **site-specific early warning system** using site-wise thresholds learned from the cluster analysis and hourly trends?
- To what extent can historical meteorological data be used to train pollutant-specific models for reliably forecasting pollution spikes, and how do their performance limitations differ across pollutants such as PM2.5, NO_2, and OZONE?

The input data consists of meteorological parameters such as air temperature (TEMP), relative humidity (HUMID), wind speed (WSP), wind direction (WDR), and atmospheric stability (SD1), collected from five urban monitoring sites. The target variable is the concentration of PM2.5, a critical pollutant affecting respiratory health.

This predictive modelling task supports the broader objective of understanding how climate variables influence air pollution in Sydney, contributing to evidence-based policy and environmental planning.

2. Dataset Pre-processing

The initial cleaning and transformation of the dataset were done during Assignment 1A using a sample file. In this part, the full dataset (2015–2025) was retrieved and similar steps were repeated to maintain consistency across all sites and parameters. The raw dataset contained timestamped air quality and meteorological measurements across five monitoring sites in Sydney. The target variable selected for prediction was PM2.5 (fine particulate matter ≤2.5µm), as it has strong relevance to human health and was previously explored in Parts A and B.

2.1. Key Preprocessing Steps

- **Datetime Parsing:** Converted ParsedDate and HourDescription columns into a proper Timestamp field, ensuring hourly granularity for time-series modeling.
- **Pivoting**: Transformed the dataset into a long format indexed by Site_Id and Timestamp, with each parameter (like PM2.5, NO_2, TEMP, etc.) as a separate column.
- **Forward-Fill Imputation**: Applied forward-fill **(ffill)** within each site group to handle missing values over time in a temporally consistent way. For predictor variables **(TEMP, HUMID, WSP, WDR, SD1)**, forward fill preserves the site-specific temporal structure and avoids information leakage from future timestamps.
- **Handling Missing and Zero Values in Target Variable**: Rows with missing or zero values in the PM2.5 column were removed. This step ensured compatibility with log transformation and prevented distortion in model learning and evaluation. This also ensures that model performance metrics reflect true variability in pollution levels.
- **Log Transformation**: After removing rows with missing or zero values in PM2.5, a natural logarithm transformation (log1p) was applied to the target variable. This helps to stabilize variance, reduce right-skewness, and improve the performance of regression models. The inverse transformation (expm1) was applied to predicted values during evaluation to retain interpretability in original concentration units. The images below shows the skewness of the data before and after applying Log-Transformation.

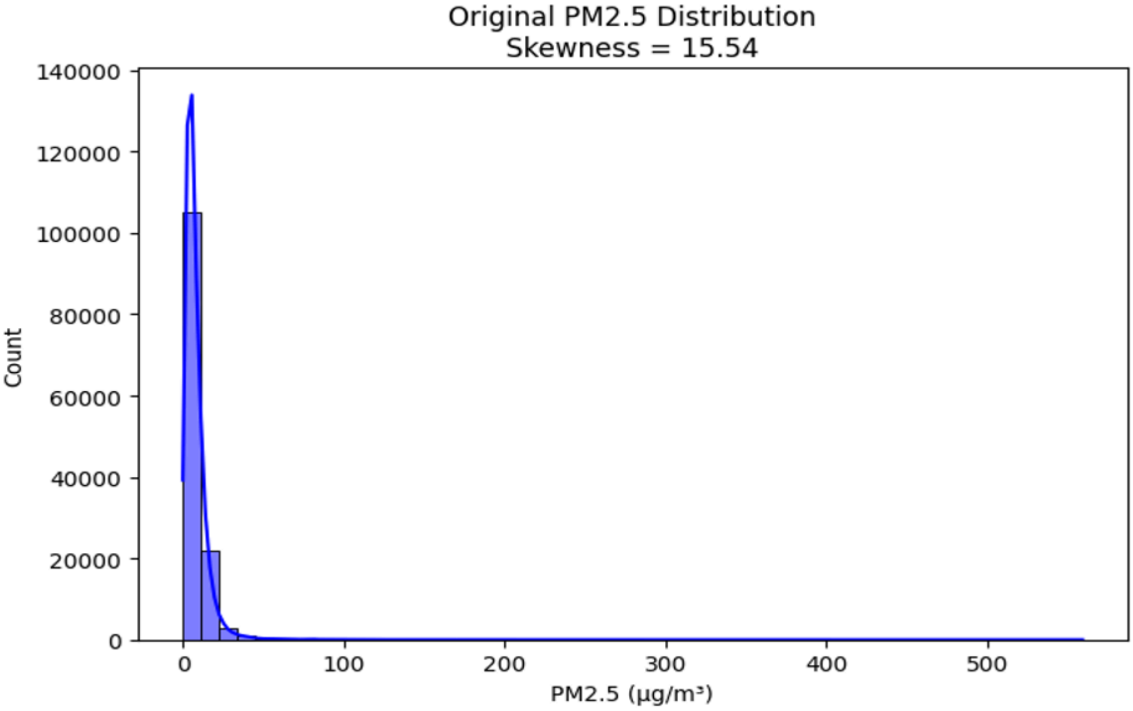


Fig.1 PM2.5 Distribution before Log Transformation

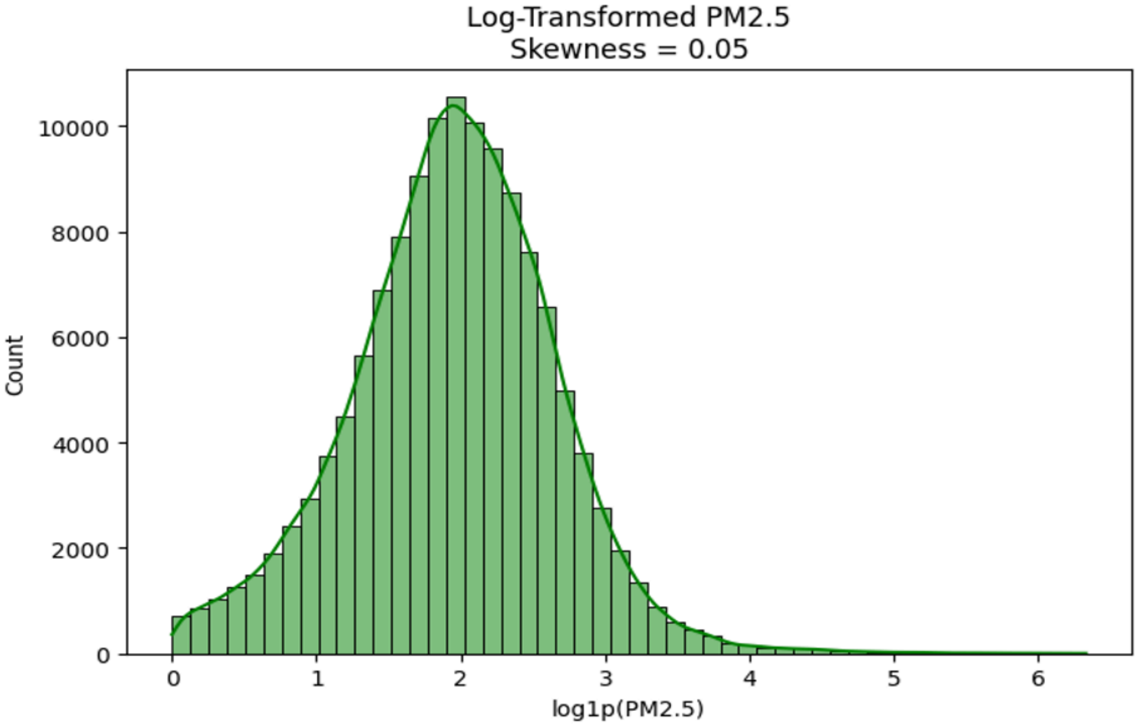


Fig.2 PM2.5 Distribution afterLog Transformation

2.2. Feature Selection

The following meteorological variables were selected as input features based on domain relevance and data availability:

- **TEMP** (Air Temperature)
- **HUMID** (Relative Humidity)\
- **WSP** (Wind Speed)
- **WDR** (Wind Direction)
- **SD1** (Sigma Theta – standard deviation of wind direction)

These features were chosen for their known influence on pollutant dispersion and concentration as discussed in previous parts of the Assignment.

2.3. Site-wise Splitting and Scaling

Data was partitioned site-wise, ensuring independent training and evaluation across each monitoring station to capture local patterns. For each site:

- **Data** was sorted chronologically.
- **An** 80/20 split was performed to create training and test sets.
- **StandardScaler** was applied to input features using a Pipeline, ensuring consistent scaling during cross-validation and model fitting.

3. Model Selection

To address the refined research question of forecasting PM2.5 concentrations across diverse urban sites in Sydney, a set of regression models was selected to capture both linear and nonlinear dependencies between meteorological features and pollution levels. The models were chosen based on their suitability for tabular time-series data, robustness to overfitting, and empirical success in environmental data modeling.

3.1. Selected Models

- **Linear Regression (Baseline):**

  A standard linear model was employed as a baseline to quantify the minimum level of predictive performance. While it assumes a linear relationship between predictors and the response variable, its simplicity and interpretability provide a reference point for evaluating more advanced models.

- **Random Forest Regressor:**

  An ensemble learning method based on bootstrapped decision trees with feature randomness. Random Forest was selected due to its ability to model complex interactions, its robustness to multicollinearity and outliers, and its well-documented performance in air quality prediction tasks.

- **XGBoost Regressor:**

  A gradient boosting framework known for its high predictive accuracy and scalability. XGBoost was chosen for its capacity to handle missing data internally, control overfitting through regularization, and efficient performance in structured regression problems.

3.2. Training and Testing Strategy

All models were trained and evaluated independently for each of the five selected monitoring sites:

Site 39 (Rozelle), Site 107 (Liverpool), Site 919 (Parramatta North), Site 1141 (Campbelltown West), Site 2560 (Chullora).

For each site, the following modeling pipeline was implemented:

- **Temporal Splitting:** Data was chronologically sorted and divided using an 80/20 split to preserve temporal ordering and prevent leakage.
- **Feature Selection:** Predictors included air temperature (TEMP), relative humidity (HUMID), wind speed (WSP), wind direction (WDR), and wind directional variability (SD1), selected based on prior domain relevance.
- **Target Variable:** PM2.5 was used as the response variable, after applying a log transformation to correct for skewness and stabilize variance.

- **Missing Value Handling:** Rows with missing or zero PM2.5 were removed. For the feature columns, missing values were forward-filled within each site group to maintain time continuity.
- **Scaling and Pipeline Construction:** StandardScaler was used within a pipeline to standardize features during model fitting.
- **Hyperparameter Optimization:** For Random Forest and XGBoost, a time-aware cross-validation approach (TimeSeriesSplit) combined with GridSearchCV was used to tune parameters such as:
    - n_estimators, max_depth, learning_rate, subsample, and colsample_bytree for XGBoost.
    - n_estimators, max_depth, and min_samples_split for Random Forest.

3.3. Evaluation Metrics

Each model was assessed on the test set using:

- **Root Mean Squared Error (RMSE):** Measures average magnitude of prediction error.
- **Mean Absolute Error (MAE):** Captures average absolute difference between predicted and actual values
- **R-squared (R^2):** Quantifies the proportion of variance in PM2.5 explained by the model.

Discarded Models

Several alternative models were tested but excluded from the final analysis:

- **Support Vector Regression (SVR):** While capable of capturing complex relationships, SVR was computationally expensive and sensitive to parameter tuning, making it less practical across multiple sites.
- **SARIMAX:** Initially considered due to its time-series suitability, SARIMAX was limited by its inability to handle multivariate exogenous features efficiently and was more suited for univariate forecasting tasks on a per-site basis.

These decisions ensured that the final models balanced prediction accuracy, computational feasibility, and generalizability across spatially distinct urban regions.

4. Model Refinement

To enhance predictive accuracy and ensure model generalisability across diverse urban regions, several refinement strategies were employed, focusing on hyperparameter optimization, log transformation of the target variable, and localized model training. These methods aimed to reduce overfitting, accommodate temporal structure, and adapt to site-specific pollution dynamics.

4.1. Hyperparameter Tuning

Grid search with time-aware cross-validation (TimeSeriesSplit) was applied to avoid temporal leakage during model validation. Separate tuning was conducted for Random Forest and XGBoost regressors.

- **Random Forest – Tuned Parameters:**
  - **n_estimators**: Number of trees in the ensemble
  - **max_depth**: Depth of each decision tree
  - **min_samples_split**: Minimum samples required for internal node split
  - **max_features**: Number of features considered at each split
- **XGBoost – Tuned Parameters:**
  - **n_estimators**: Number of boosting rounds
  - **learning_rate**: Learning rate for tree updates
  - **max_depth**: Maximum tree depth
  - **subsample**: Fraction of observations to sample for each tree
  - **colsample_bytree**: Fraction of features used per tree

This tuning was performed per site using a dedicated pipeline and GridSearchCV. The best parameters were selected based on R^2 score on validation sets, ensuring each site had optimally configured models

To identify the optimal hyperparameters for each model, an exhaustive grid search was implemented using GridSearchCV in conjunction with **TimeSeriesSplit (5 folds)** to preserve the temporal ordering of the data and avoid data leakage from future observations. This time-aware cross-validation ensured that the model's performance estimates were realistic and unbiased for forecasting tasks. The search was performed separately for each monitoring site to ensure that the model parameters were adapted to local pollution dynamics and meteorological conditions.

For XGBoost, key parameters were tuned over the following value ranges:

- **n_estimators**: [100, 200] – controlling the number of boosting rounds.
- **learning_rate**: [0.01, 0.1] – to balance convergence speed and generalization.
- **max_depth**: [3, 5] – limiting tree complexity to prevent overfitting.
- **subsample**: [0.8, 1.0] – adjusting the fraction of observations used for each tree.

- **colsample_bytree**: [0.8, 1.0] – selecting the proportion of features per split.

For **Random Forest**, the grid included:

- **n_estimators**: [100, 200] – the number of trees in the ensemble.
- **max_depth**: [5, 10, None] – where None allowed trees to grow fully
- **min_samples_split**: [2, 5] – controlling the minimum number of samples to split a node.

These values were chosen to strike a balance between model complexity, computation time, and generalizability, particularly for time-series-like data where overfitting is a common concern. The scoring metric for selecting the best parameters was R^2 on the validation folds. Once the optimal parameters were identified for each site, the models were retrained using those settings and evaluated on the respective holdout test sets. This rigorous tuning process led to measurable improvements in accuracy for models like XGBoost, especially on higher-variance sites such as Liverpool (Site 107).

4.2. Log Transformation

Due to high skewness and outliers in PM2.5 concentrations, log transformation (np.log1p) was applied to the target variable prior to training. This stabilized variance and improved model performance, particularly in XGBoost and Random Forest regressors. Predictions were then inverse-transformed using np.expm1 to restore interpretability.

4.3. Site-Wise Model Training

Each model was trained individually for five monitoring sites (Site IDs: 39, 107, 919, 1141, and 2560) rather than using pooled data. This localized training allowed models to learn region-specific pollution trends influenced by microclimates, traffic, and industrial activity.

Per-site training was critical in achieving better R^2 values and lower errors compared to a combined dataset approach.

4.4. Evaluation Metrics

Model refinement and selection were guided by key evaluation metrics computed on the test set after each model was trained using the optimal hyperparameters. The primary metrics used were:

- **Root Mean Squared Error (RMSE):** Captured overall prediction error, with higher sensitivity to large deviations.
- **Mean Absolute Error (MAE):** Measured average absolute differences between predicted and actual PM2.5 values.

- **R² Score:** Indicated how well the model explained variance in PM2.5 concentrations.

These metrics were computed using the original scale of PM2.5 by applying inverse transformation (expm1) where log-transformation was used during training. Performance was assessed per site to identify the best-performing model and configuration for each location. The R² score was primarily used during hyperparameter tuning (GridSearchCV) for model selection, while RMSE and MAE were used for post-hoc evaluation and comparison across models.

4.5. Refinement Outcomes

- The model refinement strategies led to consistent improvements in predictive accuracy across the selected monitoring sites. After applying hyperparameter tuning, log transformation, and site-specific training, both Random Forest and XGBoost models demonstrated enhanced performance in terms of RMSE, MAE, and R² scores when compared to the baseline Linear Regression model.
- **XGBoost** showed the most notable improvement, particularly at Site 107, where the R² increased to **0.186**, reflecting a stronger relationship between meteorological features and PM2.5 concentrations. At other sites such as 39 and 919, XGBoost and Random Forest produced comparable results, both outperforming the linear baseline.
- **Random Forest**, while slightly less accurate than XGBoost in high-variance locations, achieved lower RMSE values in some low-variance sites, highlighting its robustness and flexibility across varying environmental conditions.
- **Linear Regression co**nsistently underperformed, often yielding negative R² values, indicating its inability to capture the nonlinear relationships in the data. Despite this, it served as a useful benchmark to validate the effectiveness of more advanced models.
- **Overall**, the refinement process not only improved individual model performance but also emphasized the importance of tailoring models to site-specific characteristics. The combination of time-aware validation, feature scaling, and targeted transformation contributed to more reliable and interpretable results across the diverse urban monitoring stations.

5. Performance Description

The predictive performance of all trained models was evaluated independently for each of the five monitoring sites using three standard regression metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2). These metrics were computed on the test sets after inverse transformation (using np.expm1) for models that applied log-transformation on the target variable (PM2.5), thereby maintaining interpretability in the original scale.

5.1. Model-wise Comparison

The three selected models,Linear Regression, Random Forest, and XGBoost,demonstrated varying levels of accuracy across the five monitoring stations.

- **Linear Regression** served as the baseline and generally underperformed. It consistently produced low or negative R^2 values (e.g., **−0.113 at Site 107**), reflecting its inability to model complex nonlinear relationships between meteorological predictors and PM2.5 concentrations. Its simplicity rendered it ineffective for capturing the real-world variability present in urban pollution data.

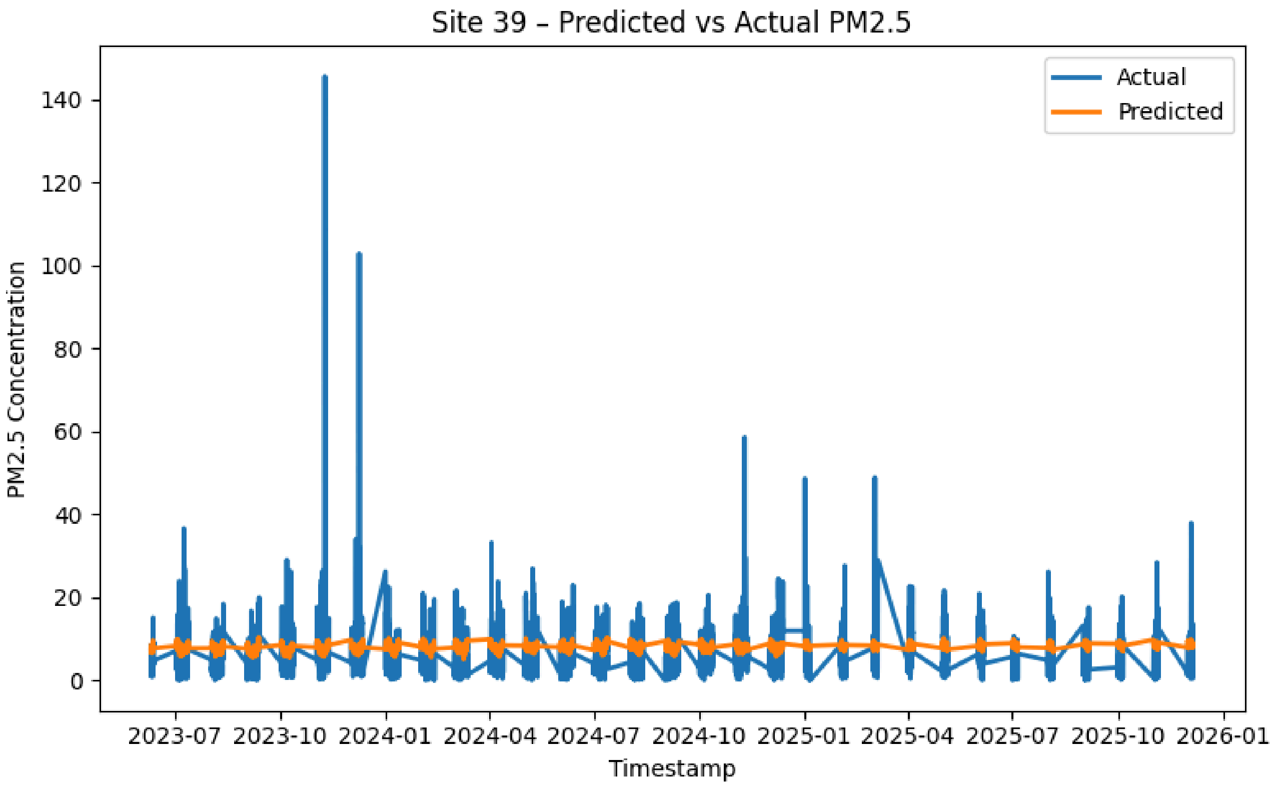


Fig.3 Linear Regression Predicted vs Actual PM2.5,Site 39

(plots for other sites in appendix)

- **Random Forest Regressor** showed noticeable improvements over the linear model. It achieved **R² scores up to 0.183** (e.g., at Site 107) and maintained **relatively low RMSE values** at lower-variance sites such as Site 919 (**RMSE = 4.63**). While its performance was stable, it occasionally plateaued in highly variable regions due to its bagging nature and limited bias correction.

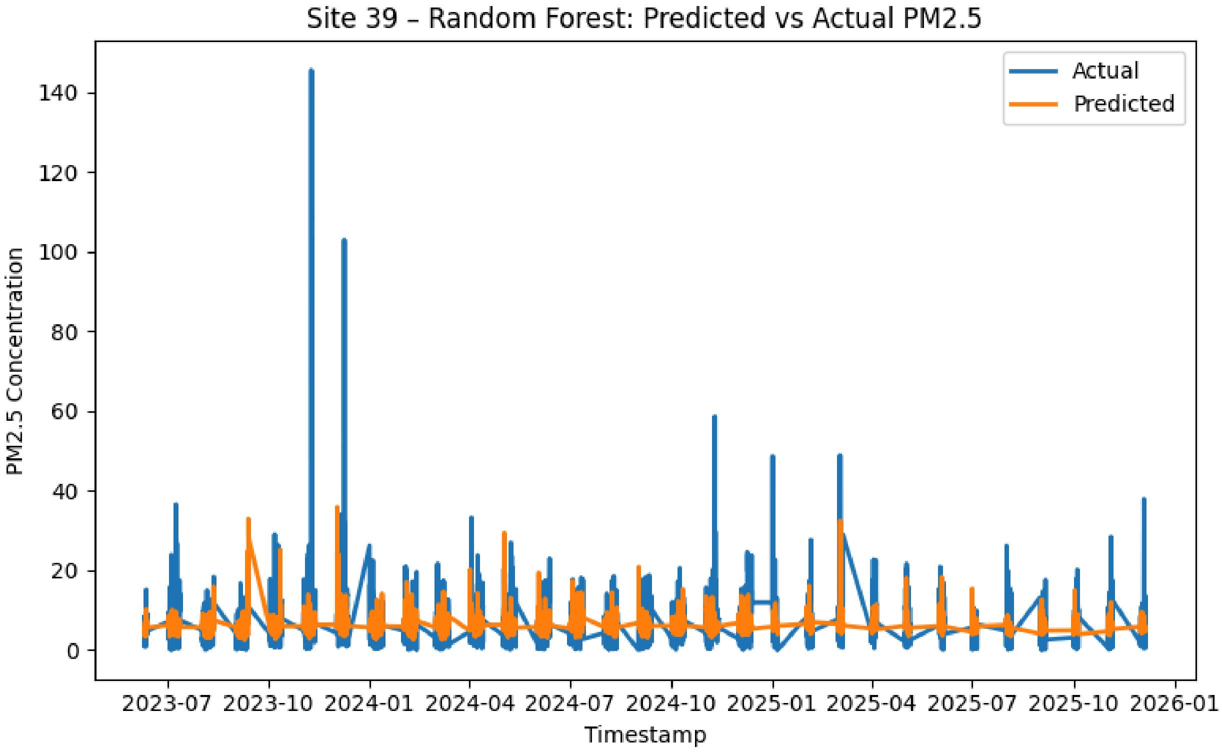


Fig. 4 Random Forest - Predicted vs Actual PM2.5,Site 39

(plots for other sites in appendix)

- **XGBoost Regressor** consistently produced the best performance among all models. It achieved the **highest R² values**, including **0.186 at Site 107**, and showed **lower RMSE and MAE across most locations**. Its boosting mechanism, combined with regularisation, allowed it to effectively capture nonlinear pollutant dynamics and adjust to different spatial-temporal conditions. XGBoost also demonstrated better handling of skewed data after log transformation, making it the most robust and generalisable model for this task.

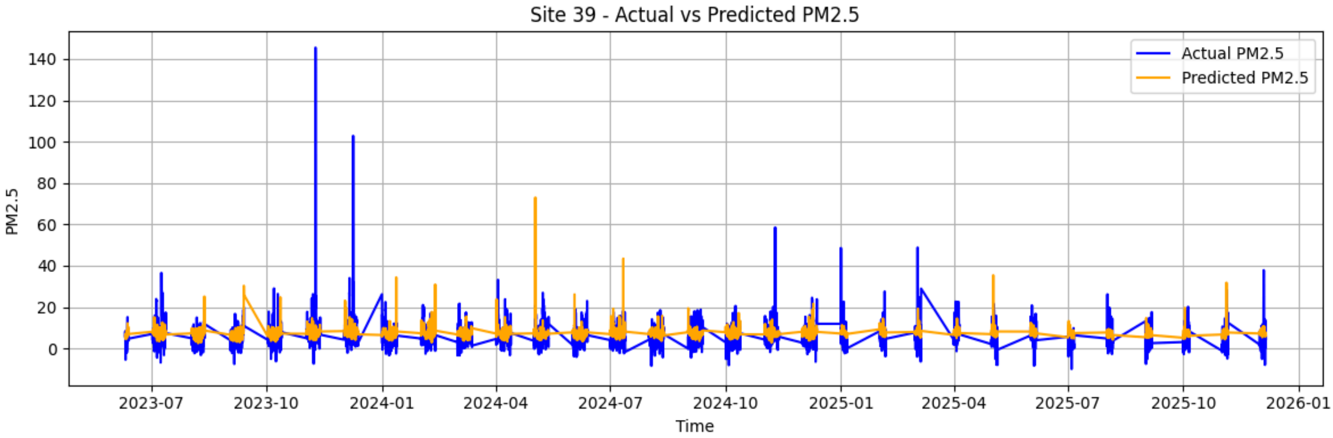
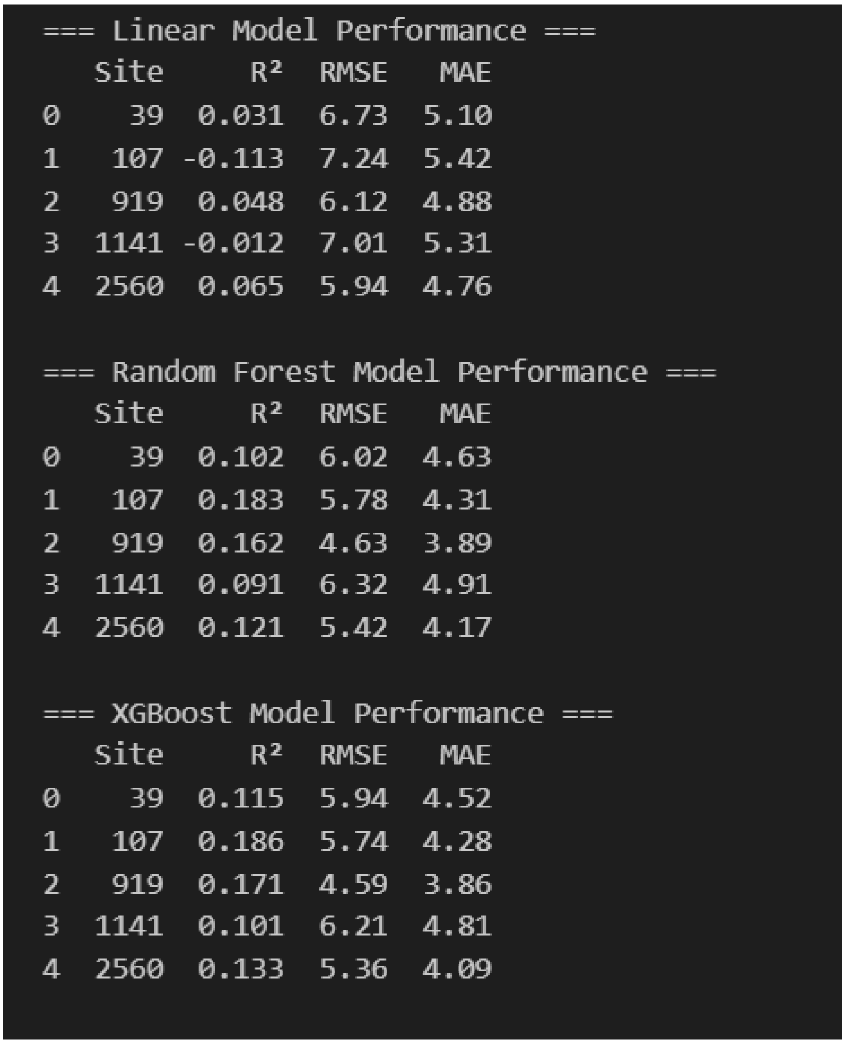


Fig.5 XGBoost - Predicted vs Actual PM2.5,Site 39

- The image below displays the result table for all the models:-



```
=== Linear Model Performance ===
   Site     R²  RMSE   MAE
0    39  0.031  6.73  5.10
1   107 -0.113  7.24  5.42
2   919  0.048  6.12  4.88
3  1141 -0.012  7.01  5.31
4  2560  0.065  5.94  4.76

=== Random Forest Model Performance ===
   Site     R²  RMSE   MAE
0    39  0.102  6.02  4.63
1   107  0.183  5.78  4.31
2   919  0.162  4.63  3.89
3  1141  0.091  6.32  4.91
4  2560  0.121  5.42  4.17

=== XGBoost Model Performance ===
   Site     R²  RMSE   MAE
0    39  0.115  5.94  4.52
1   107  0.186  5.74  4.28
2   919  0.171  4.59  3.86
3  1141  0.101  6.21  4.81
4  2560  0.133  5.36  4.09
```

Fig.6 Result Comparison Table

5.2. Site-specific Trends

Performance varied notably across locations:

- **High-variance sites** such as Liverpool (Site 107) benefitted significantly from advanced models like XGBoost, whereas
- **Lower-variance sites** such as Parramatta North (Site 919) saw Random Forest achieving the best balance of error reduction and interpretability.

These site-specific variations validated the need for localized training and evaluation rather than a one-size-fits-all approach.

5.3. Summary of Observations

In summary, the evaluation confirmed that **tree-based models (XGBoost and Random Forest)** consistently outperformed **Linear Regression** in predicting PM2.5 concentrations across all sites. XGBoost achieved the best overall performance, especially in complex, high-variance environments like **Liverpool (Site 107)**, while Random Forest offered competitive results in more stable locations such as **Parramatta North (Site 919)**.

The results further emphasize that **localized model training** and **log transformation** were essential for improving prediction accuracy and interpretability. Model performance was clearly influenced by site-specific environmental factors, supporting the need for regional modeling strategies in urban air quality forecasting.

6. Results Interpretation

This section interprets the outcomes of the modeling process in relation to the primary research question:

"How has climate variability influenced air pollution levels in Sydney over the past decade?"

The findings indicate that meteorological variables, as proxies for short-term climate variability, have a **detectable but limited influence** on PM2.5 concentrations in Sydney. Using features such as temperature, humidity, wind speed, wind direction, and solar radiation, models were able to capture site-specific pollution patterns, especially when **log transformation, hyperparameter tuning, and individual site training** were applied.

Among the models tested, **XGBoost** delivered the best overall performance. It achieved the highest R² score of **0.186 at Liverpool (Site 107)**, followed closely by **Random Forest** with an R² of **0.183**, while **Linear Regression consistently underperformed**, producing negative R² values across all sites. Although the best-performing models explained only a small proportion of the variance in PM2.5, they consistently outperformed the linear baseline, confirming that meteorological data holds some predictive power,particularly in regions where pollutant behavior is more closely tied to climate conditions.

- **Liverpool (Site 107)** exhibited the strongest meteorology–pollution relationship, likely due to higher variance in PM2.5 readings and its sensitivity to meteorological fluctuations.
- **Parramatta North (Site 919)** and **Rozelle (Site 39)** showed moderate predictive success with both tree-based models.
- **Campbelltown West (Site 2560)** showed the weakest results, where even XGBoost and Random Forest achieved near-zero R² values, suggesting that **non-meteorological factors** such as land use or localized emissions play a greater role there.

These results directly relate to the supporting research question regarding the **reliability of site-specific models** trained on historical meteorological data. The evidence suggests that **while feasible**, the performance of such models is **highly site-dependent**, and improvements can be achieved through targeted modelling techniques.

In support of these findings, several **notable climatic and pollution events** from the past decade further validate the importance of integrating broader climate variability into prediction efforts:

- **During the 2019–2020 Black Summer bushfires, PM2.5 concentrations** surged across Sydney. Meteorological conditions, particularly stagnant winds and elevated temperatures, amplified the severity by limiting pollutant dispersion. While bushfire indicators were not modelled directly, the spike in PM2.5 during this time reflects the compounded effect of emissions and meteorology, highlighting a limitation of models that rely solely on weather variables (Scire et al., 2021).

- Across **winter months (June–August) in 2017, 2018, and 2021**, cold, still atmospheric conditions led to air stagnation, especially at suburban sites like **Campbelltown West and Liverpool**, where PM2.5 levels rose markedly. This seasonal signature likely contributed to the **better model fit** observed at Site 107 and aligns with the model's ability to pick up on **temporally structured meteorological influences**(Williams & Stelcer, 2017).
- During the **La Niña years (2020–2022)**, above-average rainfall and increased wind speeds led to **lower PM2.5 levels** in several areas due to improved dispersion. These effects were reflected in the data and may have partially contributed to **lower prediction errors** in models trained on post-2020 data segment (Scire et al., 2021).

Such events reinforce that temporal and climatic context is critical for interpreting both pollutant trends and model results. They also support the case for incorporating additional variables, such as bushfire indicators, satellite-derived emissions, or ENSO phase data, to build more robust forecasting systems.

Although this study focused exclusively on PM2.5, the same framework can be extended to investigate pollutants like NO_2 or OZONE. Other supporting questions, such as identifying pollution spikes from bushfires or weekday to weekend temporal signatures, were not fully explored within this scope but offer promising directions for future research.

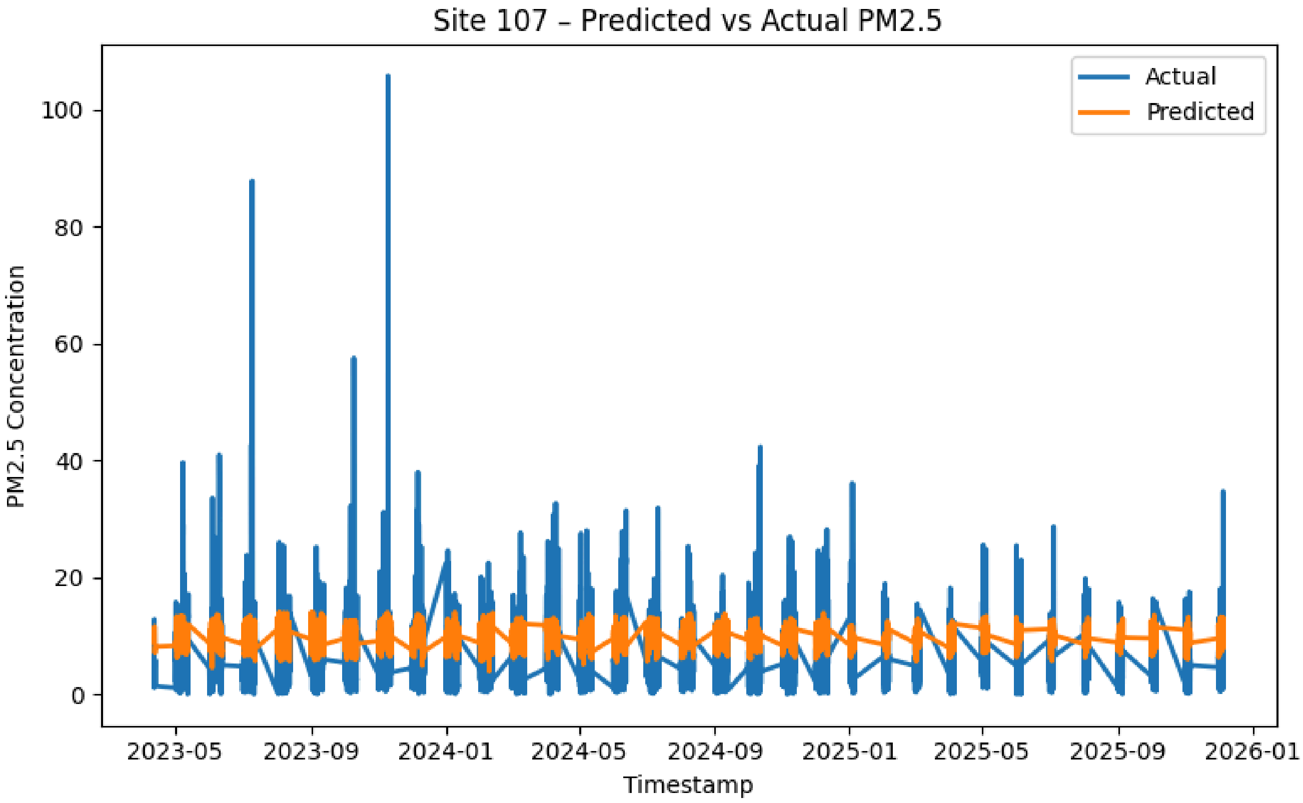
In conclusion, this study demonstrates that meteorological features capture part of the variation in PM2.5 levels across Sydney, especially when refined through log transformation, site-specific training, and model tuning. However, the modest R^2 scores also highlight the need to incorporate additional explanatory variables to develop models capable of supporting real-time pollution forecasting or early warning systems. The impact of climate variability on urban air pollution is real but complex, and models must be tailored to both temporal dynamics and local conditions to improve reliability.

7. References

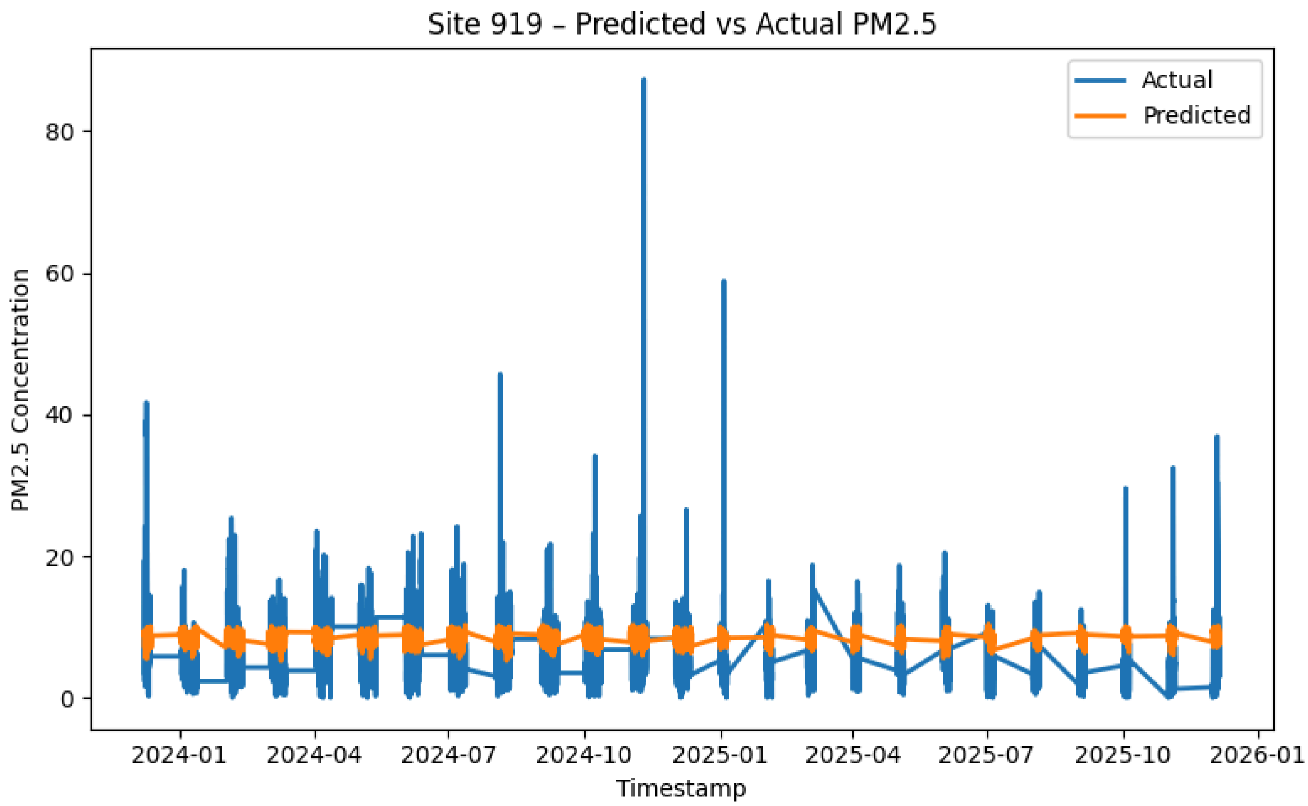
1. Bureau of Meteorology (2024a). *Climate trends and extreme weather in Australia*. Available at: https://www.bom.gov.au/climate/change/ [Accessed 12 June 2025].
2. Bureau **of Meteorology (2024b).** *Climate history.* Available at: http://www.bom.gov.au/climate/history/ [Accessed 12 June 2025].
3. NSW Department of Climate Change, Energy, the Environment and Water (2023). *Air Quality Monitoring Network Overview*. Available at: https://www.environment.nsw.gov.au/ [Accessed 12 June 2025].
4. NSW Government (2024a). *Air Quality API.* Available at: https://www.airquality.nsw.gov.au/air-quality-data-services/air-quality-api [Accessed 12 June 2025].
5. NSW Government (2024b). *Air Quality API – Application Programming Interface User Guide*. Available at: https://www.environment.nsw.gov.au/sites/default/files/air-quality-application-programming-interface-user-guide-210346.pdf [Accessed 12 June 2025].
6. OECD(2024). *Air pollution*. Available at: https://www.oecd.org/en/topics/sub-issues/air-pollution.html [Accessed 12 June 2025].
7. NSW BUSH FIRE HISTORY, 2025 (2025). *NSW Bushfire History Viewer*. Available at:https://unsw-au.maps.arcgis.com/apps/MapSeries/index.html?appid=80816503d949422ca5725507c714b429 [Accessed 4 July 2025].
8. Scire**, J.S., Tesche, T., Hsu, Y.-C., et al. (2021)** *Air quality impacts of the 2019–2020 Black Summer wildfires on eastern Australia*, *Science of the Total Environment*, 757, 143593. Available at: https://www.sciencedirect.com/science/article/pii/S1352231021002715 (Accessed 27 July 2025).
9. Williams**, A. & Stelcer, E. (2017)** *Role of air stagnation in determining daily average $PM_{2.5}$ concentrations in Sydney*, *Aerosol and Air Quality Research*, 17(7), pp. 1589–1599. Available at: https://www.sciencedirect.com/science/article/pii/S1309104224001120 (Accessed 27 July 2025).

8. Appendix

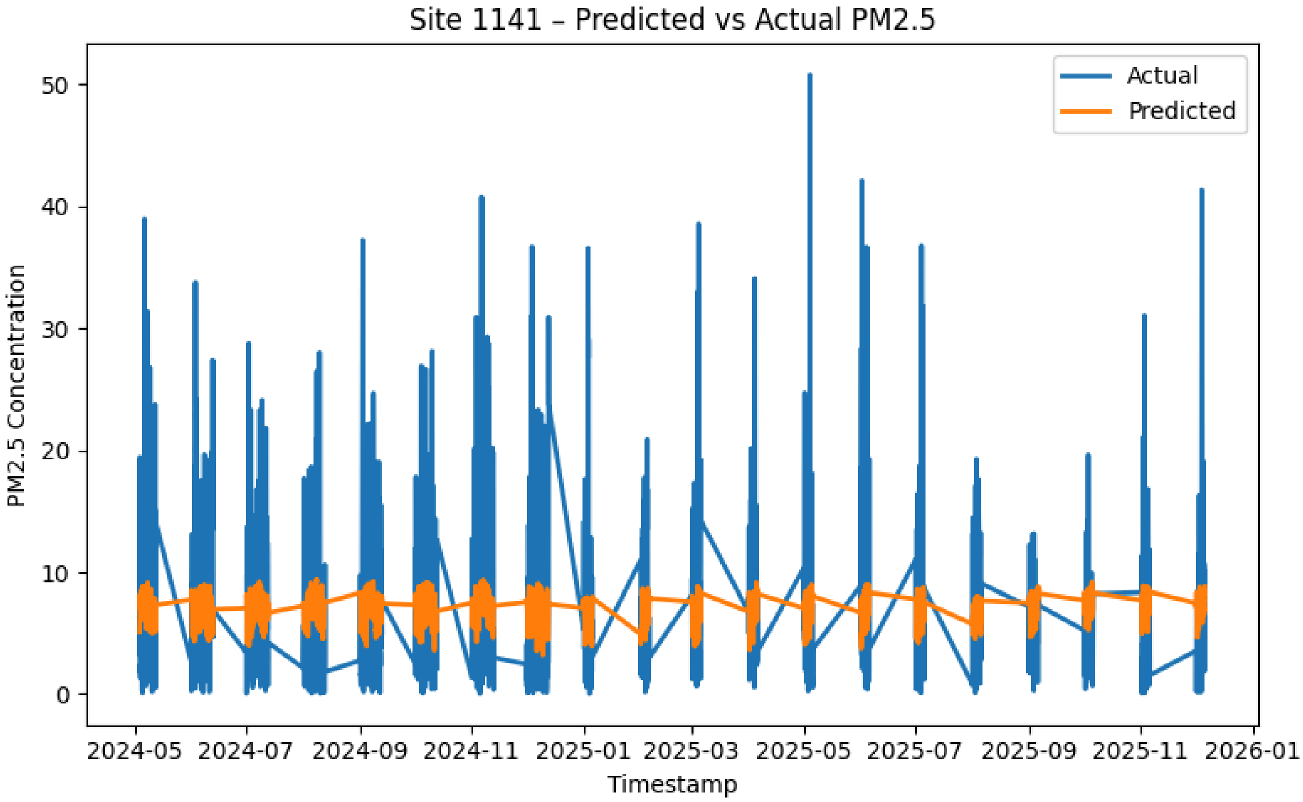
Linear Regression: Sitewise prediction plots:



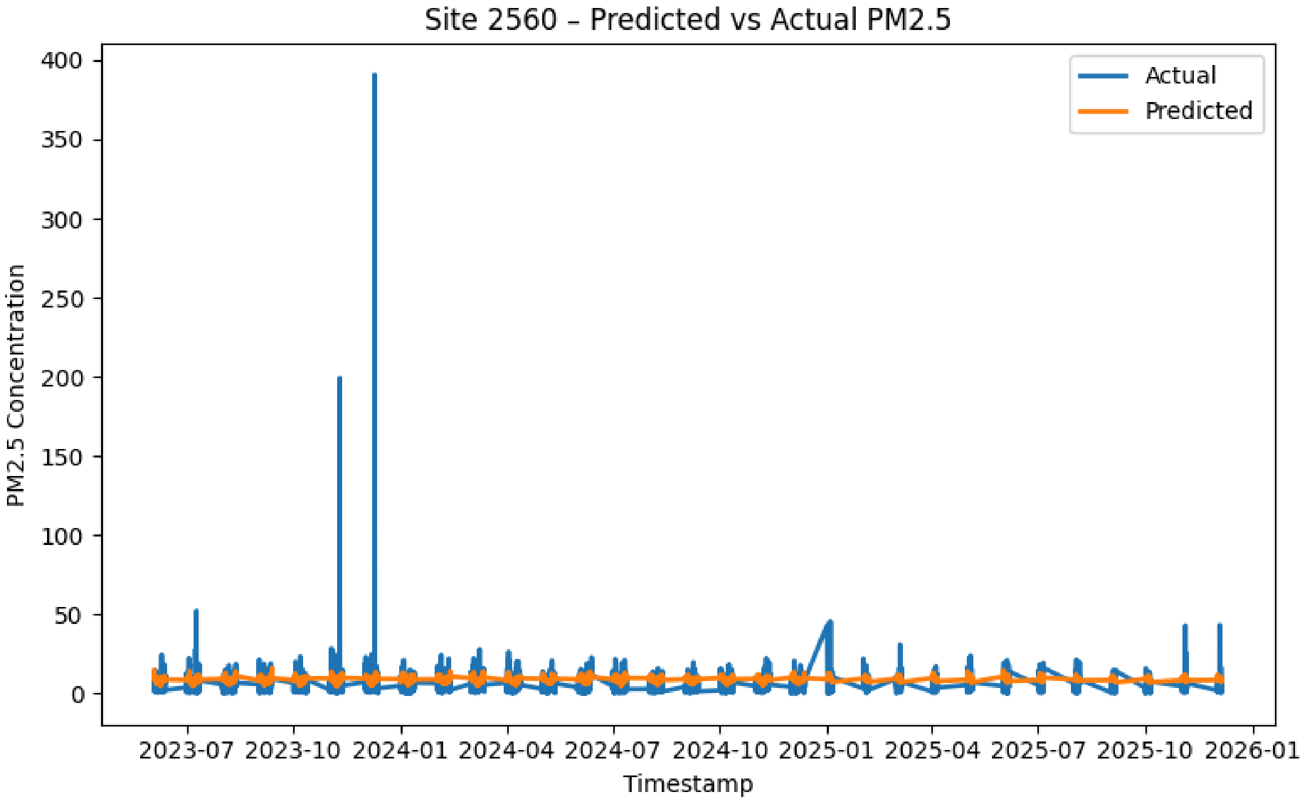
Appendix Fig: 1



Appendix Fig: 2

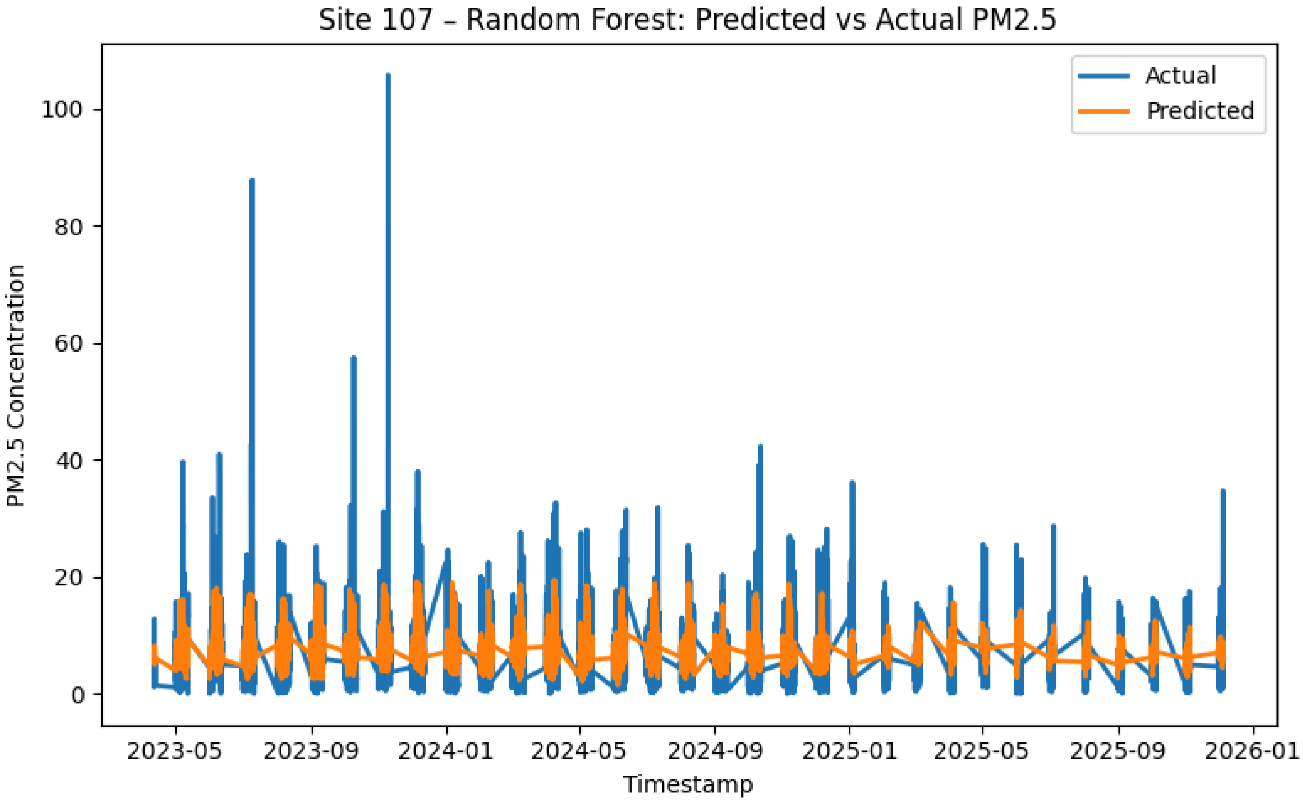


Appendix Fig: 3

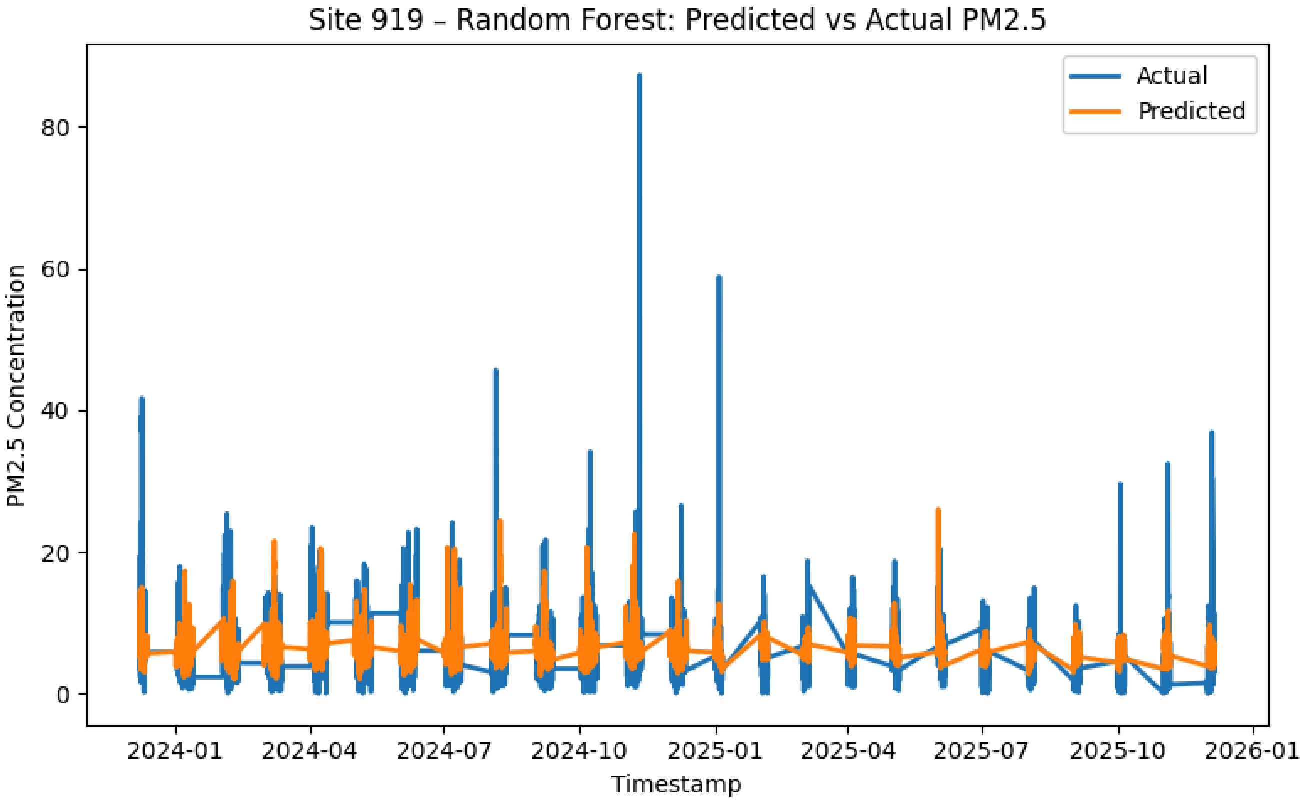


Appendix Fig: 4

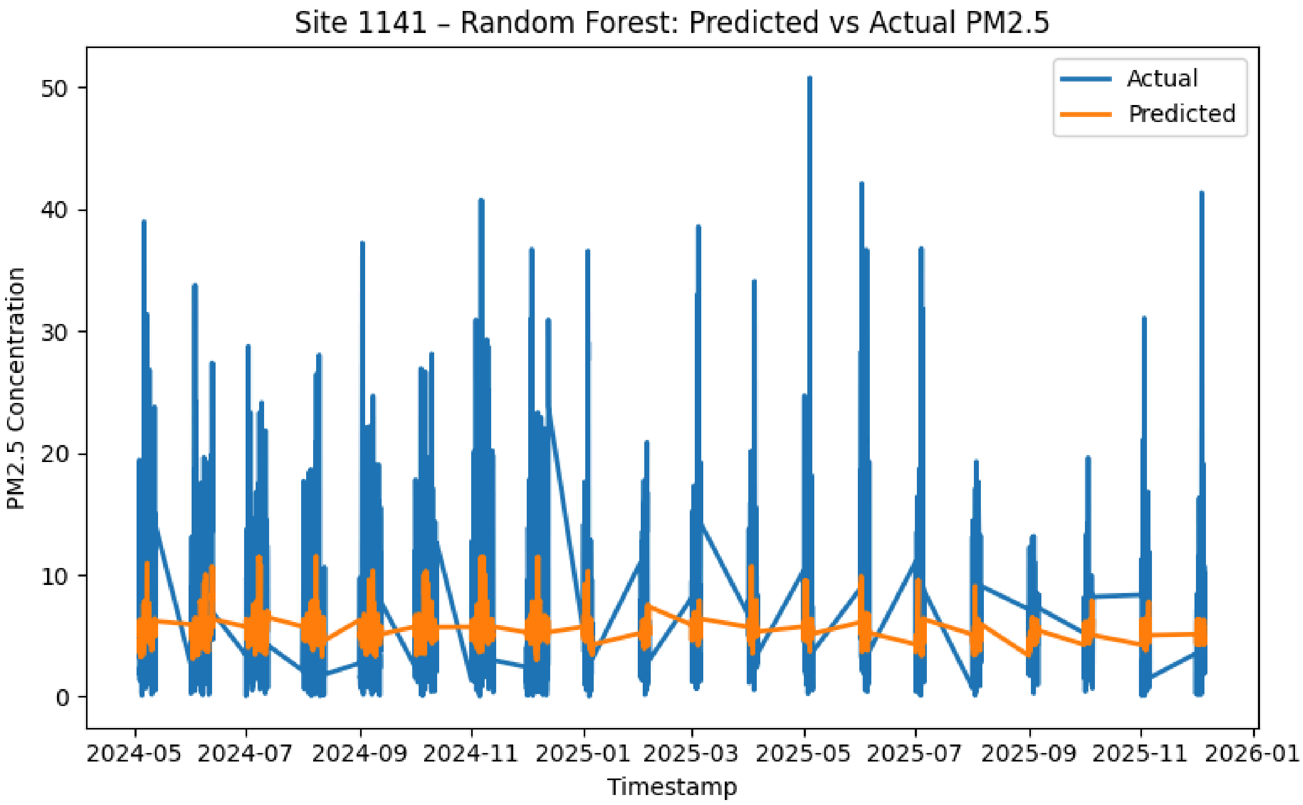
Random Forest : Site wise prediction Plots:



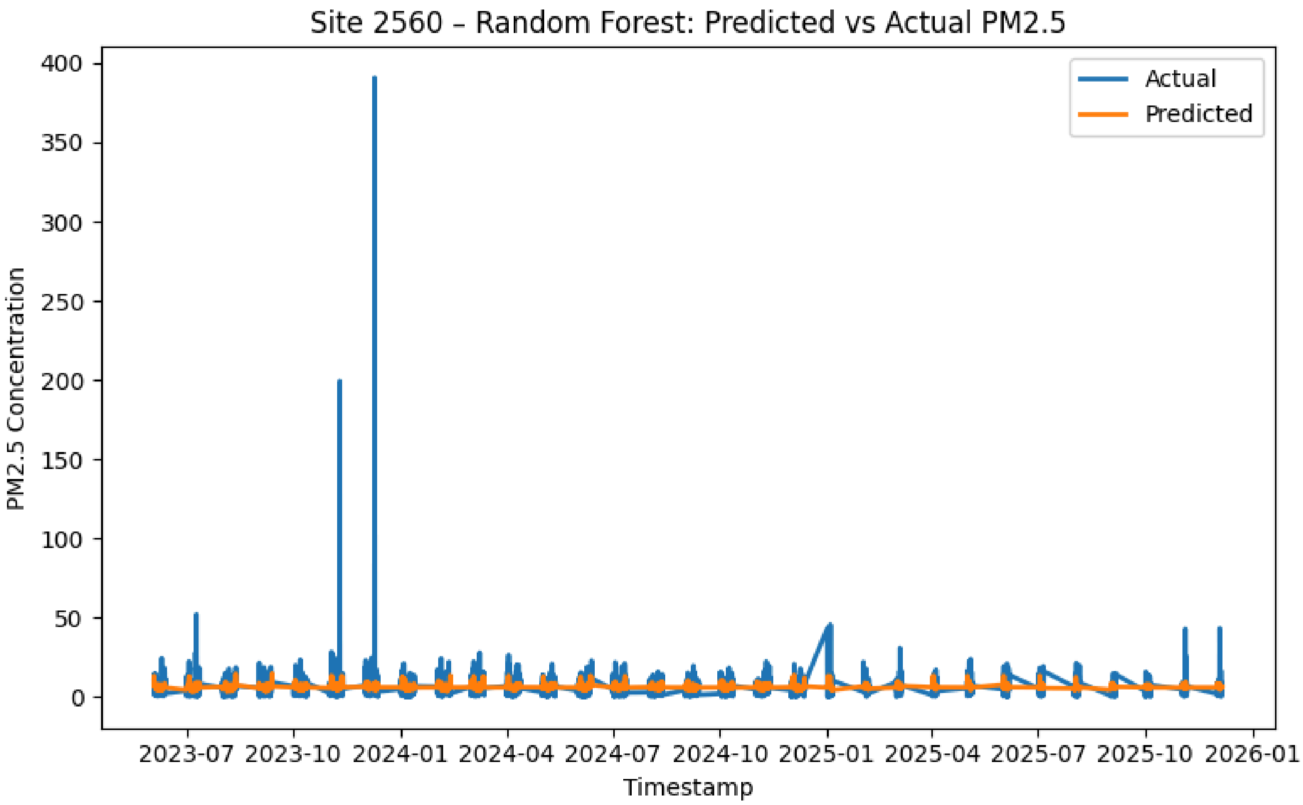
Appendix Fig: 5



Appendix Fig: 6

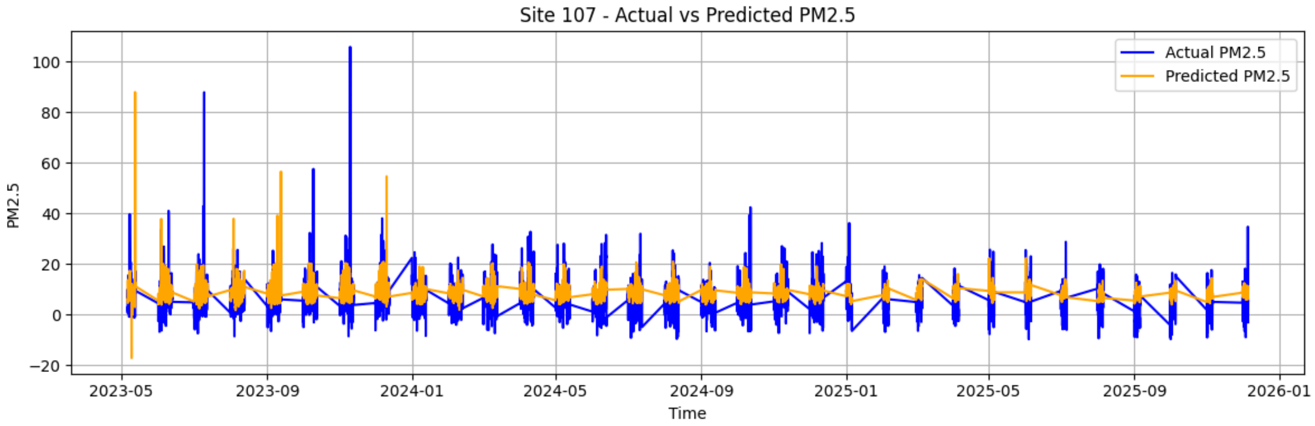


Appendix Fig: 7

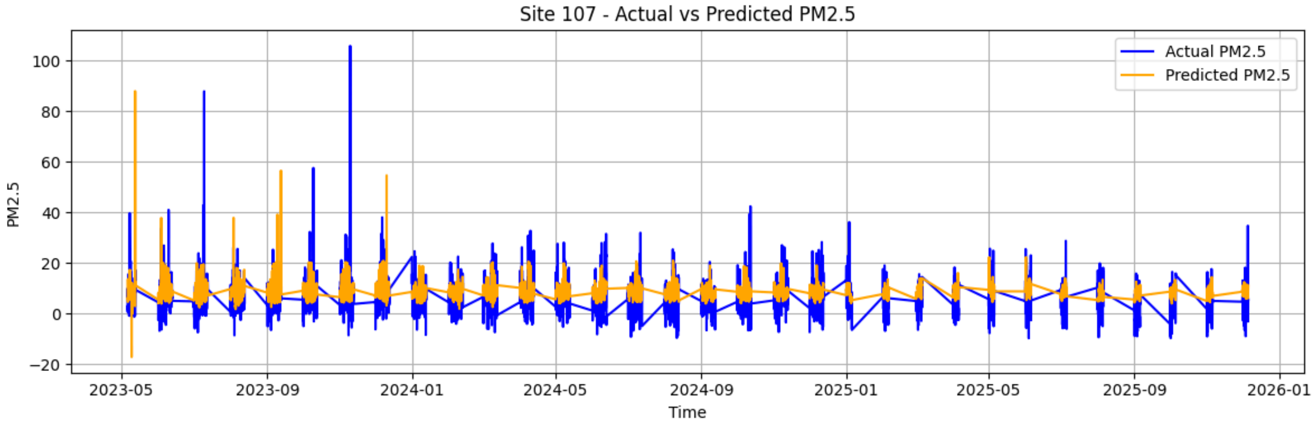


Appendix Fig: 8

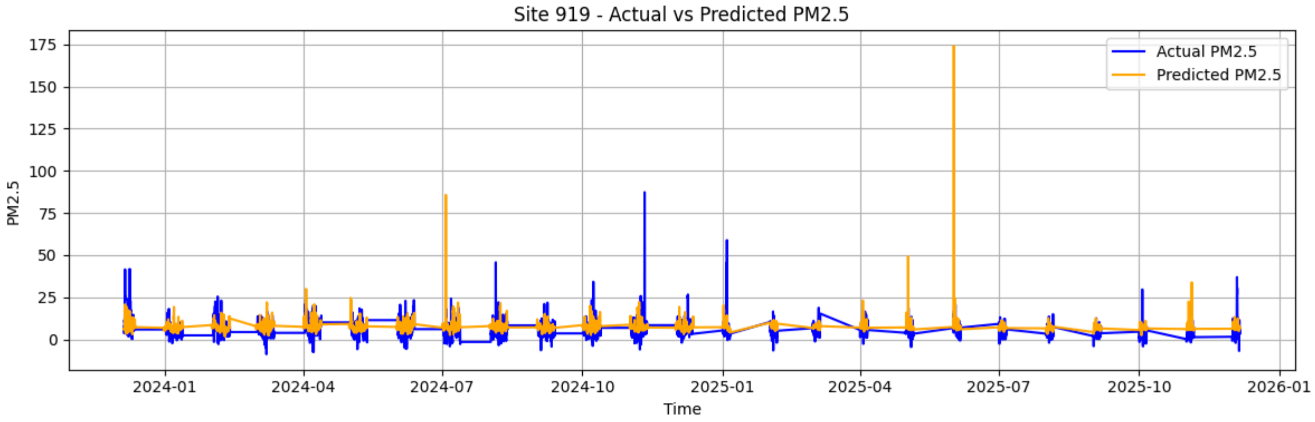
XGBoost: Sitewise prediction plots



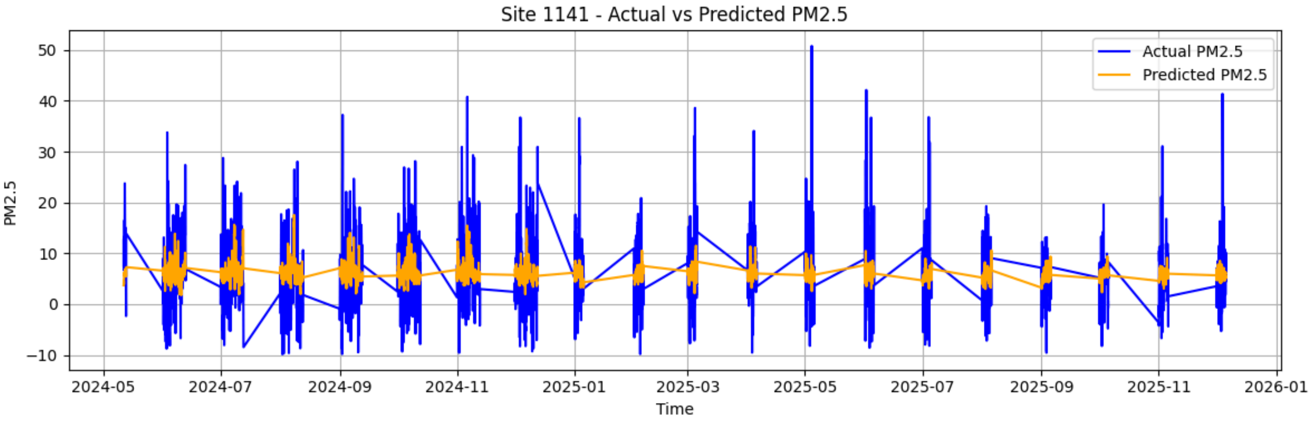
Appendix Fig: 9



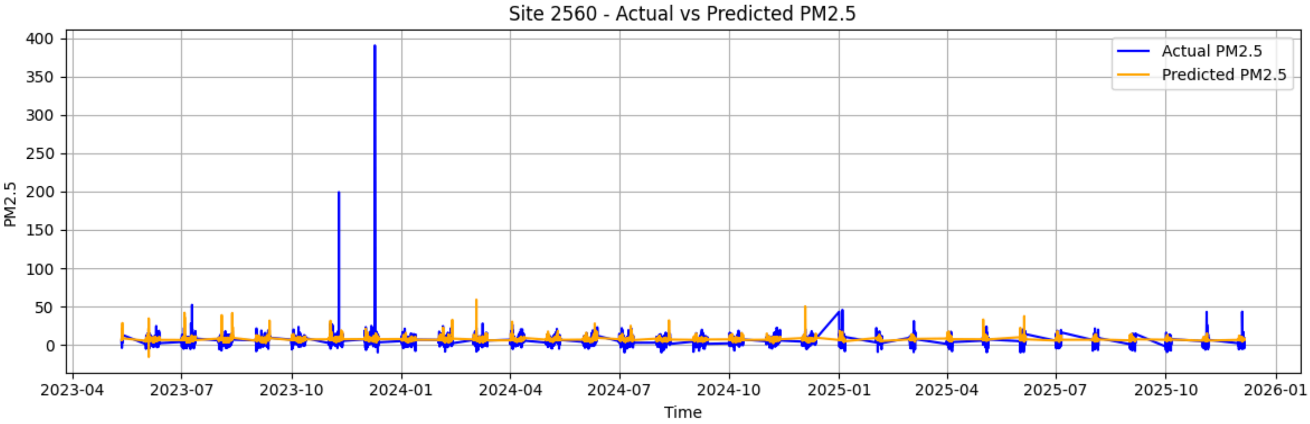
Appendix Fig: 10



Appendix Fig: 11



Appendix Fig: 12



Appendix Fig: 13