Big Data Analysis Project- Assignment 1 Part D

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

Abstract

This study investigates the relationship between climate variability and urban air pollution in Sydney over a ten-year period (2015–2025). Air quality indicators (PM10, PM2.5, NO2, CO, and Ozone) were integrated with meteorological variables (temperature, humidity, wind speed, wind direction, solar radiation, and rainfall) collected from five monitoring sites across the city. Data preparation involved cleaning, imputation, forward filling, and log transformation of skewed variables, resulting in a unified dataset suitable for analysis. Exploratory analysis and clustering revealed distinct seasonal and spatial variations in pollutant levels, with temperature, humidity, and wind speed emerging as key influencing factors. Predictive modelling was conducted using Linear Regression, Random Forest, XGBoost, Support Vector Regression, and ARIMA. While overall performance was moderate, Random Forest and XGBoost achieved higher predictive accuracy compared to baselines, and ARIMA effectively captured short-term temporal dynamics at the site level. The findings highlight both the potential and the limitations of relying solely on meteorological features for air quality prediction. Based on these insights, site-specific prescriptive strategies are recommended to enhance operational air quality management, while future research should incorporate richer datasets and advanced modelling techniques to improve predictive reliability.

1. Introduction

1.1. Background/Context

Urban air pollution is a pressing issue in Australian cities, particularly in Sydney, where rapid urbanisation, dense population, and increasing vehicle emissions contribute to deteriorating air quality. Pollutants such as PM2.5, PM10, NO_2, CO, and O_3 pose significant health risks, especially respiratory and cardiovascular illnesses. Meteorological conditions, including wind speed, wind direction, temperature, humidity, rainfall, and solar radiation, strongly influence how pollutants disperse and accumulate. Climate variability, reflected in rising temperatures, altered rainfall patterns, and more frequent extreme weather events, further complicates these dynamics, yet its specific impact on Sydney's urban environment is still underexplored (Bureau of Meteorology, 2024a; OECD, 2024).

1.2. Motivation

Investigating the relationship between climate variability and air quality is essential both for scientific understanding and practical application. Scientifically, it offers insights into how atmospheric processes interact with pollution sources. Socially and from a policy perspective, it enables better preparedness through early warning systems, data-driven urban planning, and targeted mitigation strategies. At the same time, the study tests whether big data techniques and predictive modelling approaches can effectively capture these relationships, offering an opportunity to evaluate their strengths and limitations in a real-world setting.

1.3. Proposed Solution (Research Question)

This study addresses the primary research question: How does climate variability influence urban air pollution dynamics in Sydney, and to what extent can meteorological features be used to predict PM2.5 across monitoring sites?

To explore this, an analytical framework is proposed that integrates meteorological variables (temperature, wind speed, wind direction, humidity, rainfall, and solar radiation) with pollutant data (PM2.5, PM10, NO_2, CO, and O_3). Supporting questions include:

- **Which** meteorological factors most strongly relate to pollutant variability
- **How** does predictive performance differ across sites?
- **Which** modelling approaches (linear regression, tree-based ML, or time-series) provide the most reliable accuracy?

1.4. Contributions

- **Unified dataset:** Integrated air quality and meteorological data from five Sydney monitoring sites, covering pollutants and climate-related variables over a ten-year period.
- **Comparative modelling:** Applied regression, machine learning, and time-series forecasting to assess relationships between climate variability and pollutant levels.
- **Predictive insights:** Designed a framework capable of uncovering patterns, quantifying impacts, and producing interpretable outputs for both scientific and policy use.
- **Performance evaluation:** Assessed models through cross-validation and error metrics (RMSE, MAE, R^2), highlighting strengths and limitations of different approaches.

\

2. Literature Review

Air pollution in Sydney has been studied in relation to both meteorology and long-term climate trends. Previous reviews have highlighted that rising heat, smoke events, and stagnant air conditions can significantly increase pollution concentrations (Dean et al., 2018; Williams & Stelcer, 2017). Other Sydney-based work shows clear seasonal and spatial variation in pollutant levels across monitoring stations (CAUL Hub, 2018).

At a broader scale, large-scale climate drivers such as ENSO and the Indian Ocean Dipole affect temperature and rainfall, which indirectly influence air quality (NESP Climate Hub, 2022; CSIRO, 2022). These findings align with evidence from global studies that identify temperature, humidity, and wind as consistent predictors of pollution variability (Li et al., 2023).

Building on this general background, a number of studies highlight why PM2.5 should be a primary focus. In Sydney, daily PM2.5 levels are strongly influenced by air stagnation and seasonal weather conditions (Williams & Stelcer, 2017), while global evidence shows that long-term exposure to PM2.5 is linked with increased mortality risks (Feng et al., 2024). This reinforces the importance of choosing PM2.5 as a main target in our project, since it is both harmful to health and highly sensitive to meteorology.

Ozone (O_3) also plays a role in urban air quality. Its formation is strongly affected by temperature and solar radiation, making it a pollutant closely tied to climate variability. Li et al. (2023) demonstrated that variables such as humidity and temperature are crucial predictors for both PM2.5 and ozone, suggesting that modelling these pollutants together can provide more insight into the combined influence of weather and pollution.

From a methods perspective, both tree-based machine learning models and deep learning approaches have been trialled for air quality prediction. Tree-based models, such as random forests and gradient boosting, provide interpretability and handle non-linear relationships effectively (Haque et al., 2021). On the other hand, neural networks like LSTMs and hybrid CNN-GRU models can achieve higher short-term accuracy but require larger datasets and are harder to interpret (Iqbal et al., 2025). This trade-off supports our decision to test a mix of models in this project to balance accuracy and interpretability.

Table 1: Comparison of relevant studies on air pollution, climate variability, and modelling

Study	**Focus & Location**	**Pollutants**	**Methodology**	**Key Findings**	**Relevance to This Project**
Dean et al. (2018)	Sydney, Australia	General (PM2.5, NO_2, O_3)	Literature review	Climate variability (heat, smoke, stagnation) significantly impacts air quality.	Provides baseline evidence linking Sydney's climate and air pollution.
Williams & Stelcer (2017)	Sydney, Australia	PM2.5	Statistical analysis of stagnation events	Stagnant air strongly increases PM2.5 concentrations.	Justifies focusing on PM2.5 as a primary pollutant.
Feng et al. (2024)	China (nationwide cohort)	PM2.5, O_3	Epidemiological study	Long-term exposure to PM2.5 and ozone is linked to higher mortality risk.	Highlights health importance of pollutants targeted in this study.

Li et al. (2023)	Global/urban case studies	PM2.5, O_3	Review of monitoring & modelling	Identifies meteorological variables (temperature, humidity, wind) as strong predictors of pollutant variation.	Supports choice of meteorological inputs for modelling.
Haque et al. (2021)	Australia (urban climate focus)	Multiple	Machine learning (tree-based models)	ML models effective for air quality prediction, interpretable results.	Informs baseline modelling approach in our project.
Iqbal et al. (2025)	**Global study**	**Multiple**	**Deep learning (LSTM, CNN-GRU)**	**Deep learning improves short-term forecasts but is less interpretable.**	**Justifies testing both interpretable and complex models.**

This Project (2025)	**Sydney, Australia**	**PM2.5, O_3**	**Regression, ML, Time-series forecasting**	**Models showed limited accuracy; climate variability explains some but not all pollution variation.**	**Tests whether available meteorological data are sufficient for prediction; contributes Sydney-specific modelling insights.**

In summary, the reviewed studies demonstrate that Sydney's air quality is shaped by both local meteorological conditions and broader climate variability, with PM2.5 and ozone emerging as critical pollutants of concern. Previous research also shows that predictive modelling can provide useful insights, though there is often a balance to be struck between accuracy and interpretability. Building on this foundation, the next section outlines the methodology adopted in this project, including data preparation, feature selection, and the modelling approaches used to test the relationship between climate variability and air quality in Sydney.

3. Research Methodology

The methodology for this project was designed in phases to ensure a systematic approach to investigating the link between climate variability and air quality in Sydney. Following steps were followed for the project:-

3.1. Dataset Collection and Description

This study integrates two primary datasets obtained through the NSW Government's Air Quality API:

- **Air Quality Monitoring Data**, containing hourly pollutant concentration levels recorded at monitoring stations across New South Wales (NSW), including the Sydney metropolitan region.
- **Meteorological Data**, which includes hourly environmental variables such as temperature, humidity, wind speed, wind direction, and sigma theta.

Together, these datasets provide a comprehensive basis for examining how climate variability influences pollution trends over time, with a focus on **PM2.5 concentrations**.

The datasets were retrieved programmatically using a Python POST request to the NSW Air Quality API (NSW Government, 2024a). The data was returned in JSON format and normalised into tabular form before being exported as CSV for analysis. Full documentation for the API is available in the official User Guide (NSW Government, 2024b).

Key characteristics of the dataset:

- **Format**: Retrieved in JSON, converted to CSV
- **Temporal Coverage**: Hourly data from **June 2015 – June 2025**.
- **Spatial Coverage**: Five selected urban monitoring sites across Sydney.
- **Granularity**: Hourly resolution with site-wise pollutant and meteorological values.

Table 2: Data Fields

Field Name	Description
Site_Id	Unique identifier for the monitoring site
Date	Date of observation
Hour	Hour of observation (1 to 24)
HourDescription	Time range for the hour (e.g., "12 am - 1 am")
Value	Measured value of the parameter
AirQualityCategory	Category assigned based on pollutant thresholds
DeterminingPollutant	Pollutant used to determine air quality category
Parameter.ParameterCode	Code representing the measured parameter (e.g., PM2.5
Parameter.ParameterDescription	Full name of the parameter
Parameter.Units	Abbreviation of the unit of measurement
Parameter.UnitsDescription	Full description of the unit of measurement
Parameter.Category	Type of parameter (e.g., Averages, Exceedences)
Parameter.SubCategory	Subtype within category (e.g., Hourly)
Parameter.SubCategory	Frequency of measurement (e.g., Hourly average)

3.1.1. Selected Parameters

The variables extracted for analysis include both air pollutants and meteorological features relevant to pollution modelling.

Table 3: Air Pollutants Monitored in the Study

Parameter	**Description**
PM10	Particulate Matter ≤10 micrometres in diameter, affects respiratory health
PM2.5	Fine Particulate Matter ≤2.5 micrometres, penetrates deeper into lungs and bloodstream
NO2	Nitrogen Dioxide, a traffic-related pollutant harmful to lungs
CO	Carbon Monoxide, a gas that reduces oxygen delivery in the body
OZONE	Ground-level Ozone, formed by chemical reactions involving NO_2 and sunlight; irritates airways

Table 4: **Meteorological Variables Used in the Analysis**

Parameter	**Description**
TEMP	Air temperature (°C)
HUMID	Relative humidity (%)
WSP	Wind speed (m/s)
WDR	Wind speed (m/s)
SD1	Sigma Theta – Standard deviation of wind direction (°)

3.1.2. Monitoring Sites

Five monitoring stations were selected to provide spatial diversity across Sydney's eastern, north-western, and south-western regions. These sites were chosen because they consistently report a wide range of meteorological and pollutant variables, including PM2.5, NO_2, CO, and wind metrics. Site metadata and data availability were verified through the NSW Air Quality API (NSW DCCEEW, 2023).

Table 5: Selected Monitoring Sites in Sydney

Site Name	**Site ID**	**Region**
Rozelle	39	Sydney East
Chullora	222	Sydney East
Parramatta North	919	Sydney North-west
Campbelltown West	2560	Sydney South-west
Liverpool	107	Sydney South-west

Note : Details of variable selection, data sources, and spatial coverage were outlined in Part A (see Section 3, Dataset Description), ensuring consistency between the dataset construction and subsequent modelling in Part D.

3.2. Data Cleaning and Preprocessing

Initial cleaning and transformations were piloted during Assignment 1A on a sample file. For this study, the full dataset (2015–2025) was processed using the same pipeline to ensure consistency.

Key preprocessing steps included:

- **Datetime Parsing**: Date and hour fields combined into a proper timestamp to ensure hourly granularity.
- **Pivoting**: Dataset restructured into long format, indexed by Site_Id and Timestamp, with pollutants and meteorological variables as columns.
- **Forward-Fill Imputation**: Applied within each site group to address gaps in meteorological data while preserving temporal structure.
- **Variable Filtering**: Retained only relevant pollutants and meteorological features (PM2.5, PM10, NO_2, CO, Ozone, TEMP, HUMID, WSP, WDR, SD1).
- **Final Dataset**: The cleaned dataset was exported as full_cleaned.csv for visualisation and modelling.

3.3. Exploratory Data Analysis (EDA)

Exploratory data analysis was carried out to better understand the characteristics of the air quality and meteorological dataset prior to modeling. This stage focused on examining temporal patterns, pollutant distributions, and site-level variability.

Descriptive Statistics

Basic descriptive statistics (mean, median, standard deviation, and quantiles) were computed for all pollutants (PM2.5, PM10, NO_2, CO, OZONE) and meteorological variables (TEMP, HUMID, WSP, WDR, SD1). This provided an overview of central tendencies, spread, and outlier presence across the 10-year dataset.

Temporal Trends

Time series plots were generated to observe daily, monthly, and yearly patterns in pollutant concentrations. This allowed identification of recurring seasonal signals (e.g., winter stagnation, summer ozone peaks) and extreme episodes such as bushfire-affected periods.

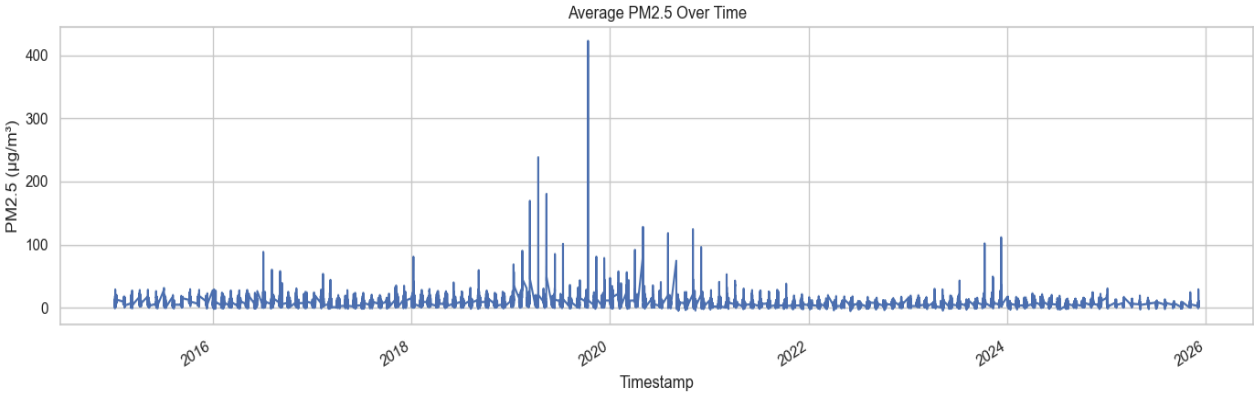


Fig. 1 :Time series of PM2.5 across selected sites

Distribution Analysis

Histograms and kernel density plots were used to examine the distribution of pollutants. PM2.5 in particular showed a highly skewed distribution, motivating the later use of log transformation during model training.

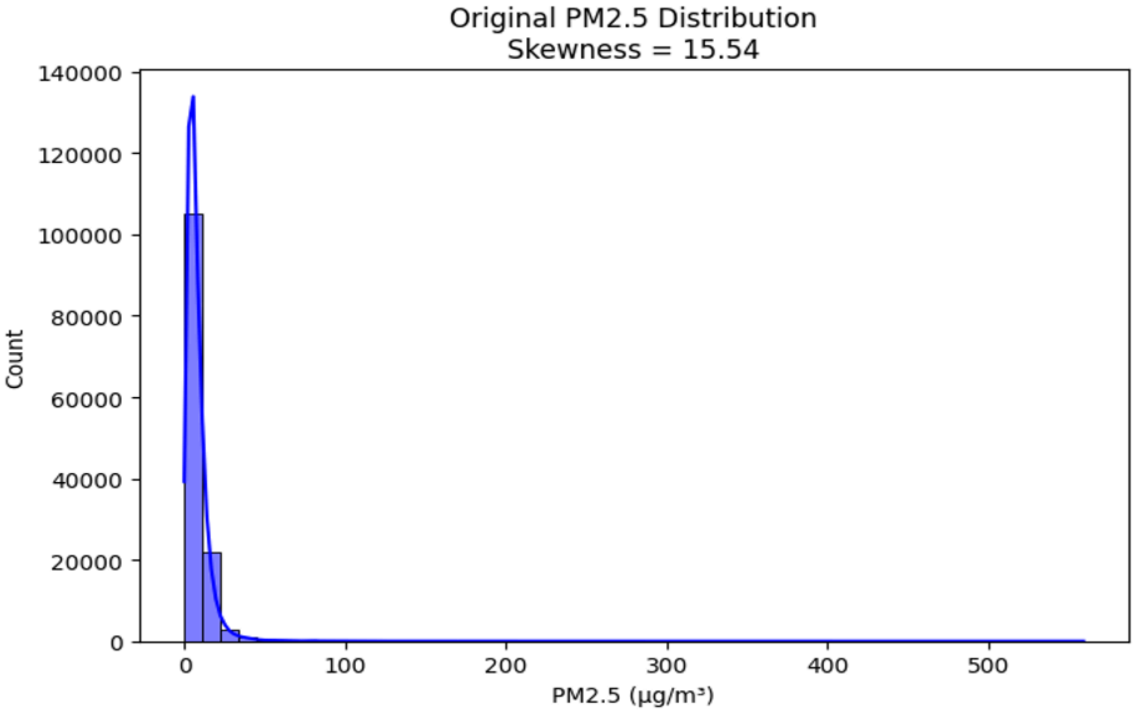


Fig. 2 : PM2.5 Distribution before Log Transformation

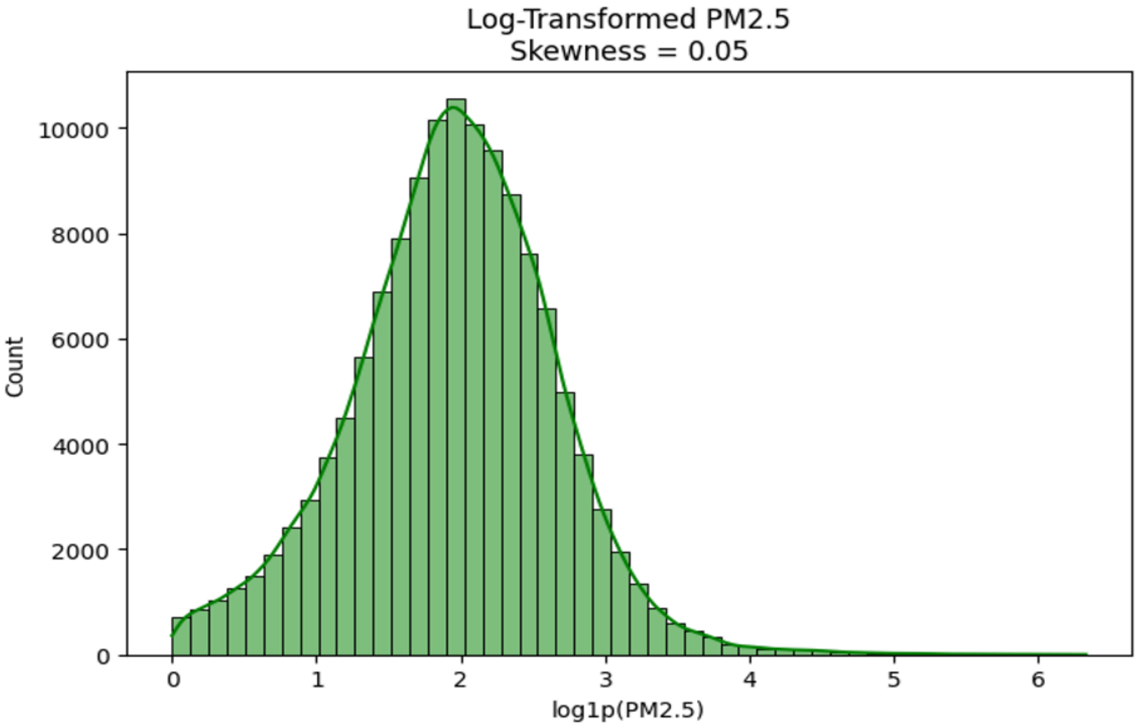


Fig. 3 : Distribution of PM2.5 after log transformation

Spatial Comparison of Sites

Boxplots and comparative charts were used to evaluate pollutant variability across the five selected monitoring sites (Rozelle, Chullora, Parramatta North, Campbelltown West, Liverpool). This highlighted differences in background levels and extreme values between urban centers and suburban sites.

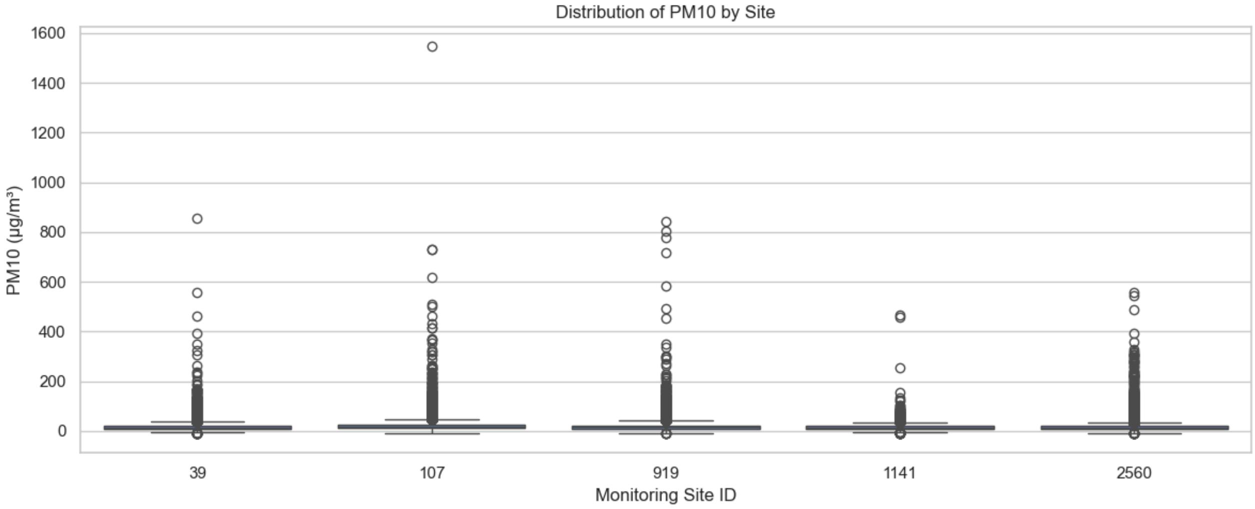


Fig. 4 : Boxplot of Average Ozone over time by Site

3.4. Clustering and Pattern Detection

To explore patterns in the relationship between meteorological conditions and air pollutants, a combination of correlation analysis, clustering, and regression modelling was applied.

3.4.1. Correlation Analysis

A Pearson correlation matrix was computed between key pollutants (PM2.5, PM10, NO_2, CO, OZONE) and meteorological features (TEMP, HUMID, WSP, WDR, SD1). This provided an initial assessment of linear dependencies between weather variables and pollutant concentrations.

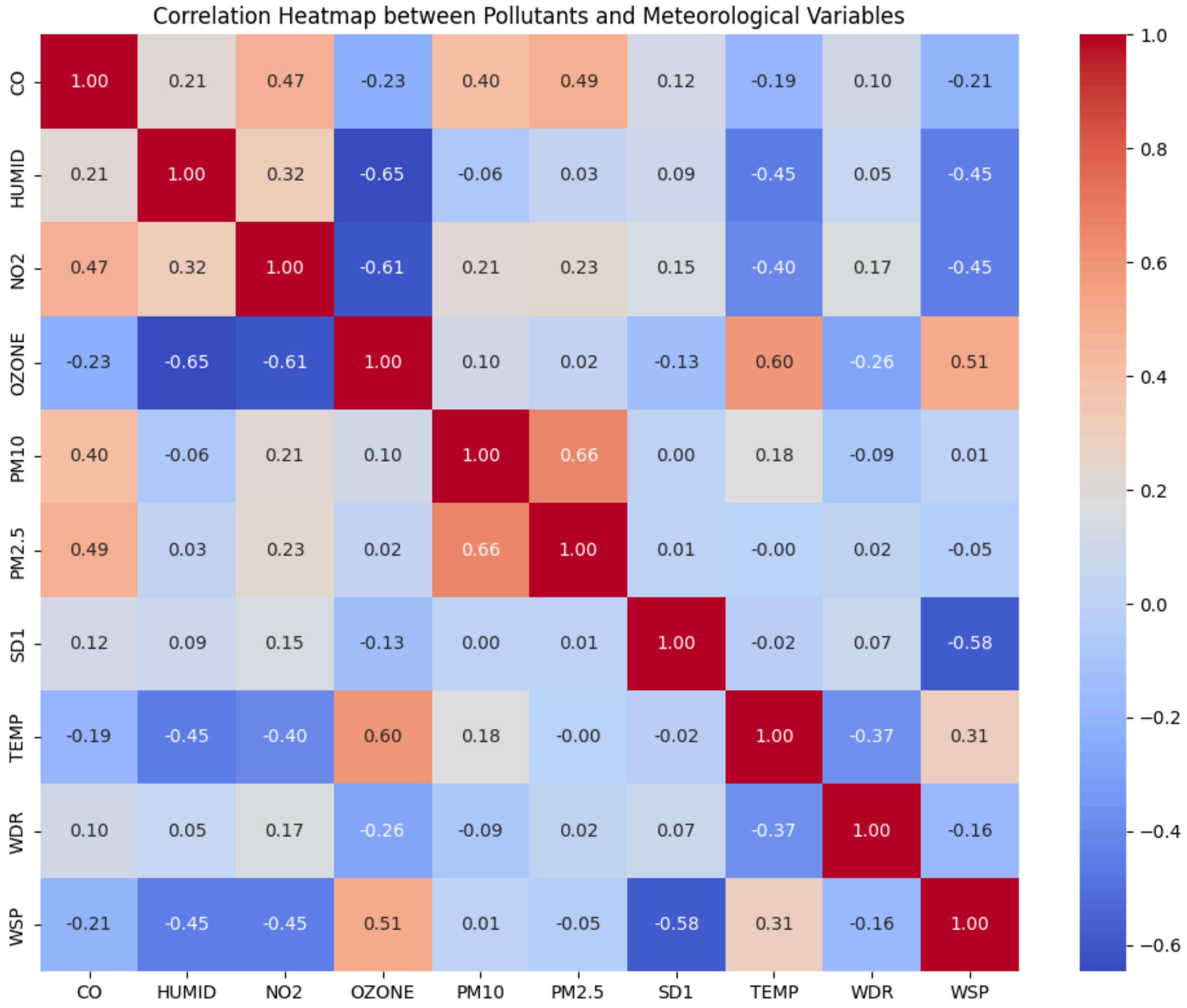


Fig. 5 :Correlation heatmap between pollutants and meteorological variables

3.4.2. Clustering with KMeans and t-SNE

The standardized dataset was clustered using the KMeans algorithm to group observations with similar pollution–meteorology profiles. Multiple values of k were tested, and $k = 3$ was chosen for optimal interpretability. To visualize cluster separation, t-Distributed Stochastic Neighbor Embedding (t-SNE) was applied to reduce dimensionality and produce a two-dimensional projection of the clusters.

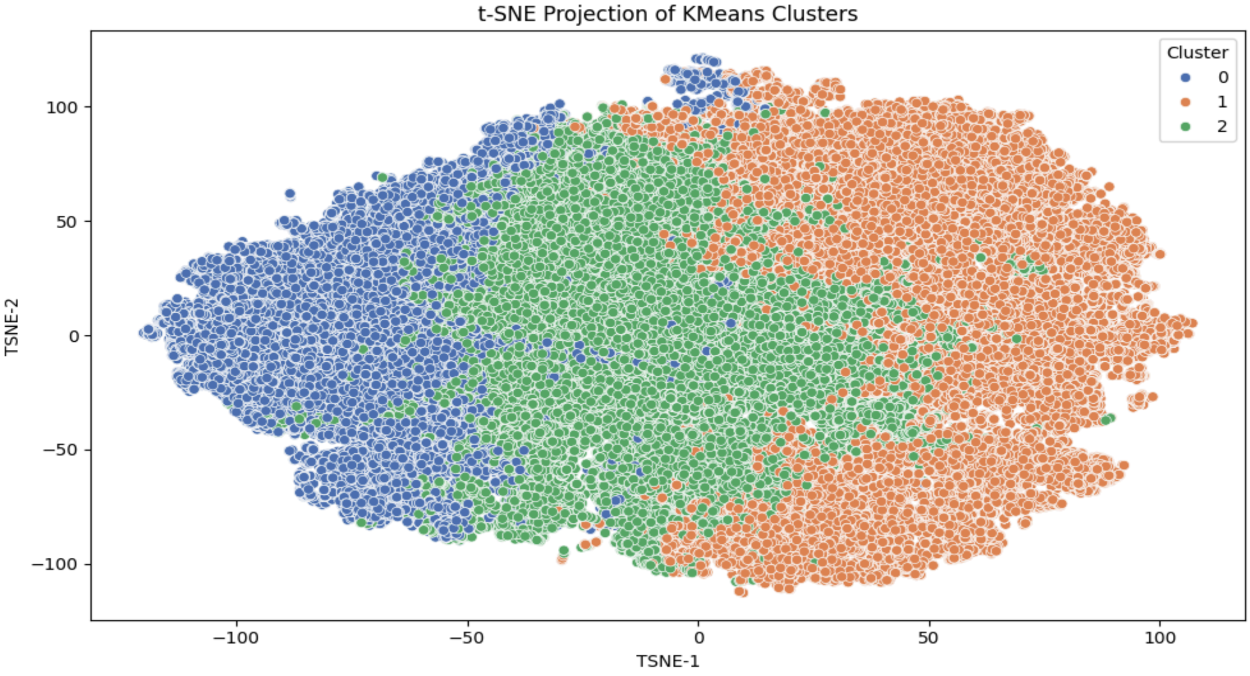


Fig. 6: t-SNE projection of KMeans clusters (k = 3)

3.4.3. Regression for Pollutant Prediction

As a benchmark for predictability, simple linear regression models were trained for each pollutant (PM2.5, PM10, NO_2, CO, OZONE) using meteorological variables (TEMP, HUMID, WSP, WDR, SD1) as predictors. Model performance was assessed with R^2 and RMSE, providing a baseline for how much pollutant variability could be explained by weather conditions alone.

OZONE: The points are fairly close to the line, matching the relatively high R^2 score (0.582). The model captured the general trend well.

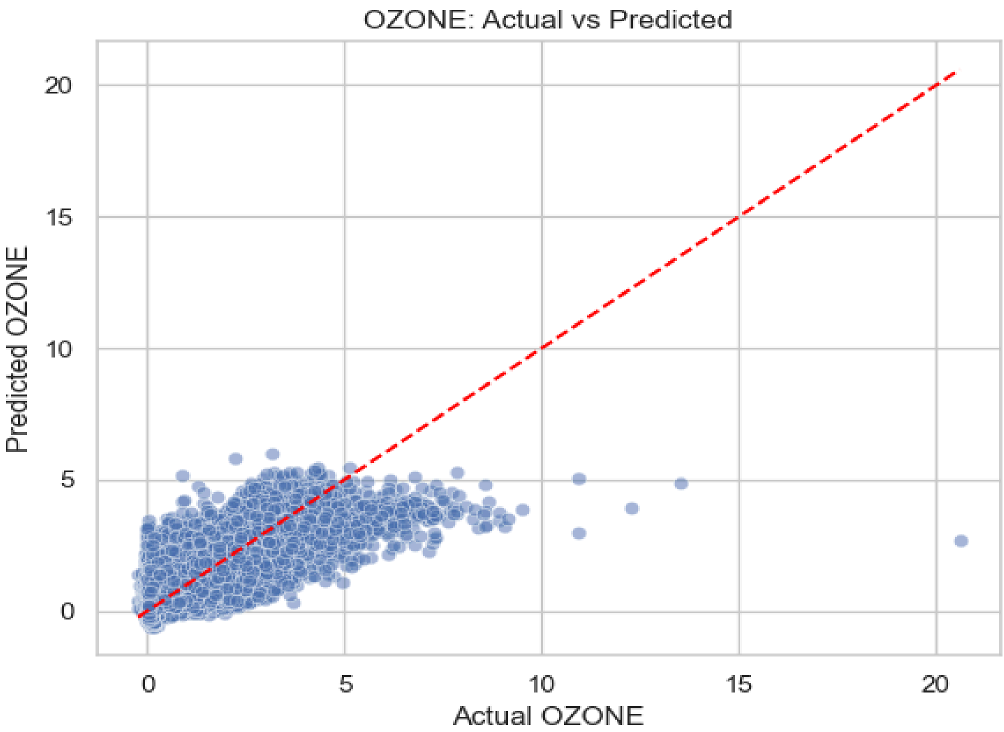


Fig. 7: Actual vs. Predicted OZONE

- **NO$_2$**: Predictions also follow the correct pattern, though slightly more spread out. The model still showed reasonable accuracy.

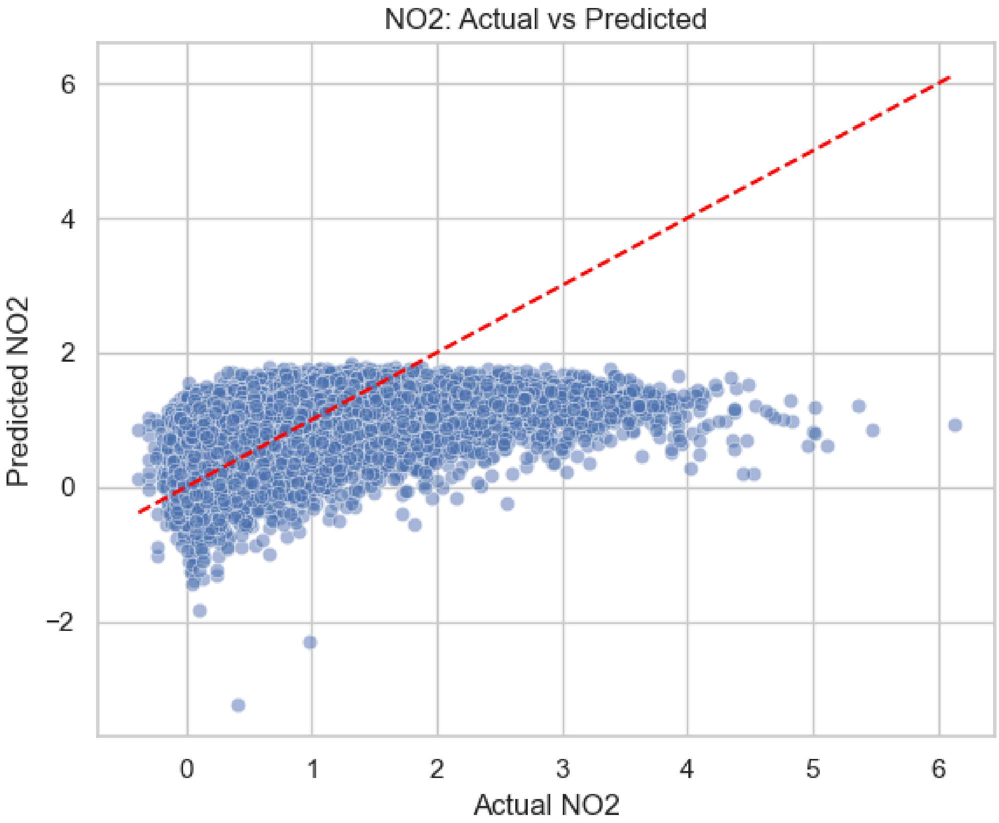


Fig. 8: Actual vs. Predicted NO_2

- **PM2.5 and PM10**: These plots show low and scattered predictions. The models were not able to track the higher observed values, resulting in poor fit.

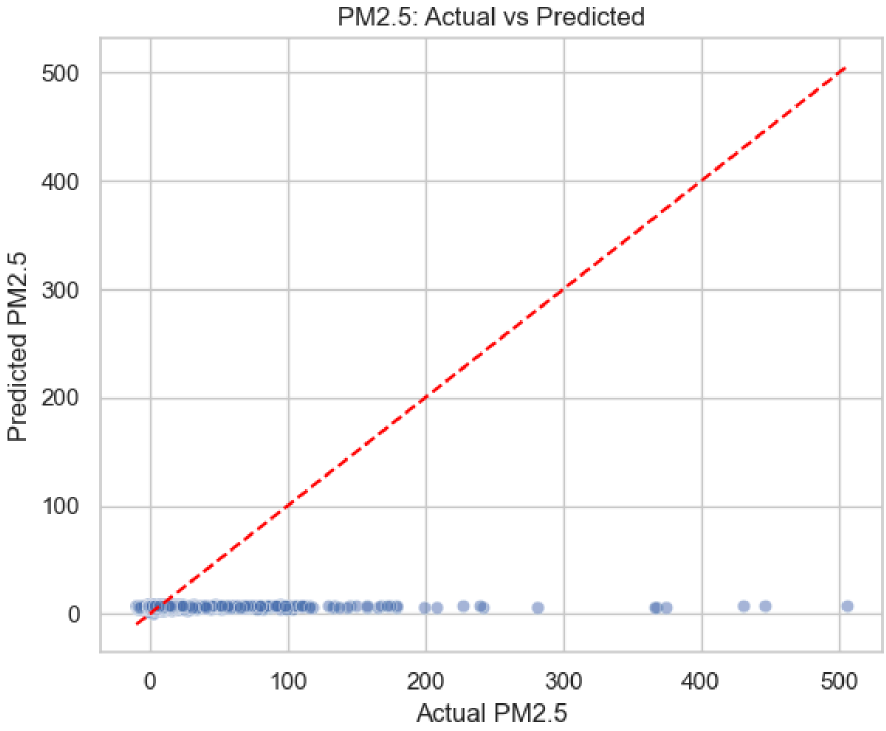


Fig 9: Actual vs. Predicted PM2.5

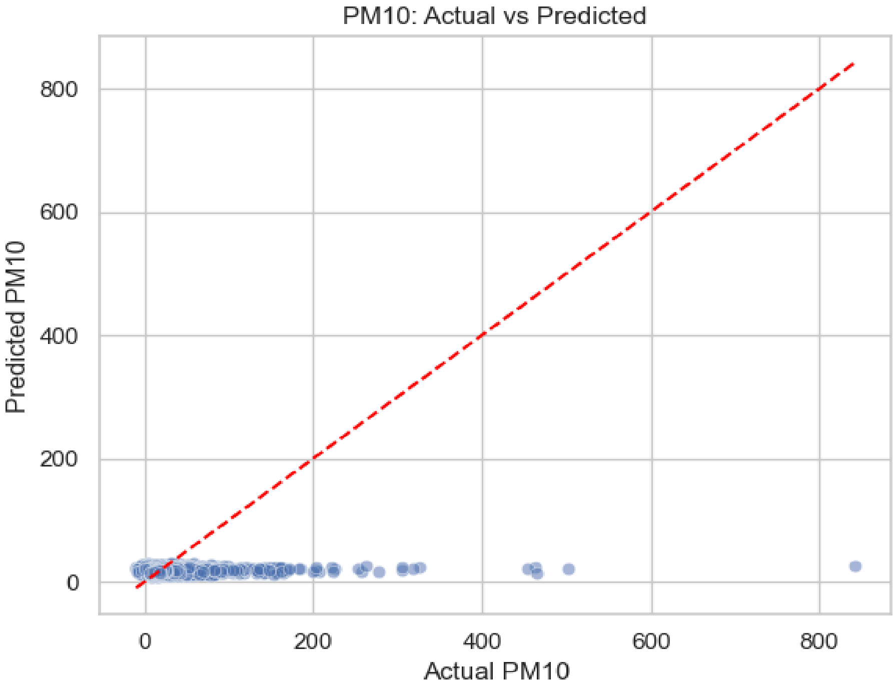


Fig. 10: Actual vs. Predicted PM10

- **CO**: Predictions were low overall and failed to capture higher actual values, indicating underestimation at upper ranges.

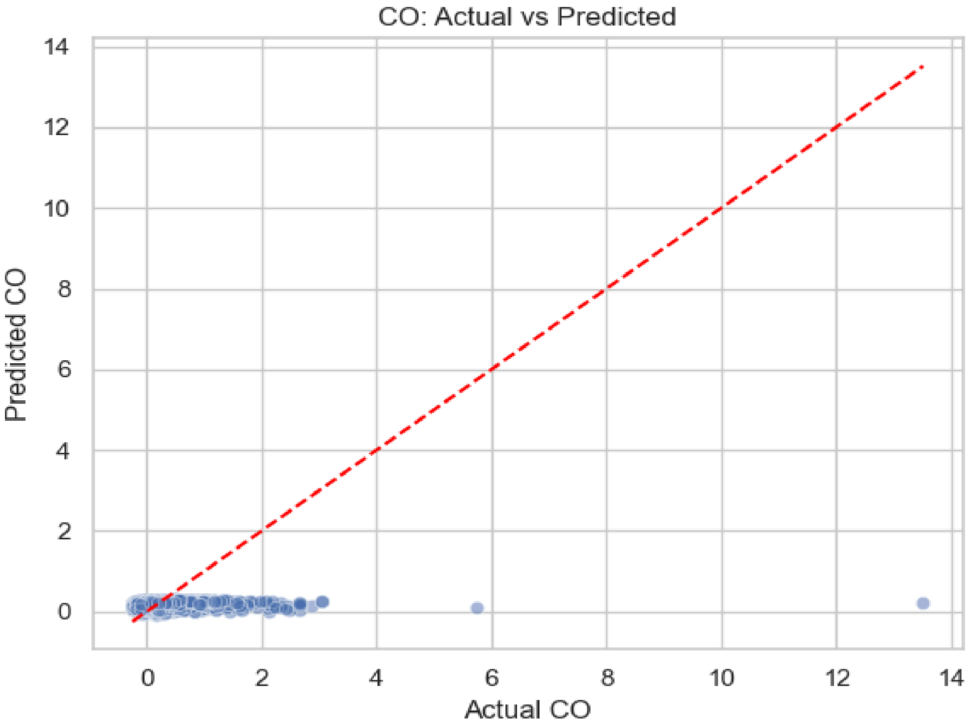


Fig. 11: Actual vs. Predicted CO

3.5. Model Development

To forecast PM2.5 concentrations across Sydney, three regression models were selected to capture both linear and non-linear relationships between meteorological variables and air pollution.

Selected Models

- **Linear Regression (Baseline):** Served as a benchmark for comparison. Its simplicity allowed for easy interpretation, though it was expected to struggle with complex dynamics**.**
- **Random Forest Regressor:** An ensemble of decision trees designed to capture non-linear interactions between meteorological features and PM2.5, with robustness to multicollinearity.
- **XGBoost Regressor:** A gradient boosting algorithm chosen for its efficiency, strong predictive performance, and ability to handle skewed data with regularisation.

Training Strategy

- **Site-wise Training:** Models were trained separately for the five monitoring sites (Rozelle, Liverpool, Parramatta North, Campbelltown West, and Chullora) to account for local emission sources and microclimate effects.
- **Temporal Split:** Each site's dataset was split chronologically into 80% training and 20% testing, preserving time order to avoid leakage.
- **Feature Set:** Meteorological predictors included air temperature (TEMP), relative humidity (HUMID), wind speed (WSP), wind direction (WDR), and wind variability (SD1). The response variable was PM2.5 (log-transformed to reduce skewness).
- **Hyperparameter Tuning:** For Random Forest and XGBoost, GridSearchCV with TimeSeriesSplit was applied to tune parameters such as tree depth, number of estimators, and learning rate.

This modelling phase established the baseline and advanced models that would later be assessed for their predictive accuracy. Evaluation of model performance was conducted in Phase 6 using standard regression metrics.

3.6. Model Evaluation

To assess the predictive performance of the selected models, a consistent evaluation framework was applied across all five monitoring sites. The goal was to ensure that results were comparable while respecting the temporal nature of the data.

Evaluation Metrics?

Three widely used regression metrics were employed:

- **Root Mean Squared Error (RMSE):** Measures the average magnitude of prediction errors, with greater sensitivity to larger deviations.
- **Mean Absolute Error (MAE):** Captures the average absolute difference between predicted and observed PM2.5 values, providing a more interpretable error measure.
- **R² Score (Coefficient of Determination):** Quantifies the proportion of variance in PM2.5 concentrations explained by the model.

Evaluation Procedure

1. **Train–Test Splitting:** Data was chronologically divided using an 80/20 split to preserve temporal ordering and prevent information leakage from future values.
2. **Per-Site Evaluation:** Each monitoring site (Rozelle, Liverpool, Parramatta North, Campbelltown West, and Chullora) was evaluated independently to account for local variability in pollution dynamics.
3. **Log Transformation Handling:** For models trained on log-transformed PM2.5, predictions were inverse-transformed (expm1) prior to evaluation, ensuring interpretability in the original scale.
4. **Hyperparameter Tuning Validation:** During grid search, performance was validated using **TimeSeriesSplit cross-validation**, ensuring that evaluation respected the sequential nature of time-series data.
5. **Final Model Comparison:** Once the best parameters were identified, all models were retrained with their optimal configurations and tested on the reserved holdout set. Their performance was then compared across RMSE, MAE, and R² to highlight relative strengths and weaknesses.

4. Experimental Evaluation

4.1. Experimental Setup

The experiments drew on a ten-year dataset (2015–2025) spanning five urban air quality monitoring sites in Sydney: Rozelle (39, Sydney East), Liverpool (107, Sydney South-west), Parramatta North (919, Sydney North-west), Campbelltown West (2560, Sydney South-west), and Chullora (222, Sydney East). PM2.5 was chosen as the target variable due to its well-documented health impacts, while explanatory features included temperature (TEMP), relative humidity (HUMID), wind speed (WSP), wind direction (WDR), and wind variability (SD1), all recognised drivers of pollution dispersion and accumulation.

To prepare the data for modelling, timestamp standardisation, pivoting, forward-fill imputation, and log transformation were applied (see Section 3, Phase 2: Data Preprocessing). These steps produced a consistent and analysis-ready dataset.

Five predictive approaches were explored (Section 3, Phase 5). Linear Regression served as a baseline, establishing a reference for predictive accuracy. Random Forest and XGBoost were selected for their ability to capture non-linear interactions between climatic and pollution variables. Support Vector Regression (SVR) was added for its robustness with smaller sample subsets, while ARIMA was deployed at the site level to model short-term temporal dependencies.

Hyperparameter tuning for Random Forest and XGBoost was carried out using GridSearchCV with TimeSeriesSplit, avoiding temporal leakage. Key parameters,including tree depth, learning rate, and number of estimators,were optimised separately for each site.

Finally, multiple model variants were tested to refine performance. XGBoost was run in three forms: a fixed-parameter baseline, a log-transformed version with controls such as subsample and colsample_bytree, and an extended grid search with broader learning rates and tree depths. Random Forest was tested as both a tuned per-site model and one paired with visual predicted-versus-actual comparisons. Linear Regression was trialled as a simple metric-only baseline as well as with additional visualisation outputs. These variations ensured that observed improvements could be attributed not only to the model type but also to preprocessing and parameterisation choices. preprocessing and parameterisation.

4.2. Experimental Results

Model performance varied across sites, reflecting local meteorological and emission differences.

Linear Regression underperformed, often producing negative R^2 values (for example 0.113 at Liverpool), showing its inability to capture complex non-linear relationships.

Random Forest achieved moderate improvements. At Parramatta North it produced a relatively low RMSE of 4.63, while at Liverpool it achieved an R^2 of 0.183. Its bagging structure was helpful in low-variance conditions but limited in highly variable environments.

XGBoost consistently outperformed the others, achieving the highest R^2 (0.186 at Liverpool) and lower RMSE and MAE across most sites.

Site-specific observations included the following:

- Liverpool (107) showed the strongest meteorology–pollution linkage, where Random Forest and XGBoost performed best.
- Parramatta North (919) produced stable results, with Random Forest slightly outperforming XGBoost.
- Campbelltown West (1141) showed poor predictive performance across all models, suggesting that local emission sources dominated over meteorology.

Overall, tree-based models clearly outperformed the linear baseline, confirming the need for non-linear approaches. Site-wise training and log transformation were essential to capturing local dynamics.

Figures generated in Part C included:

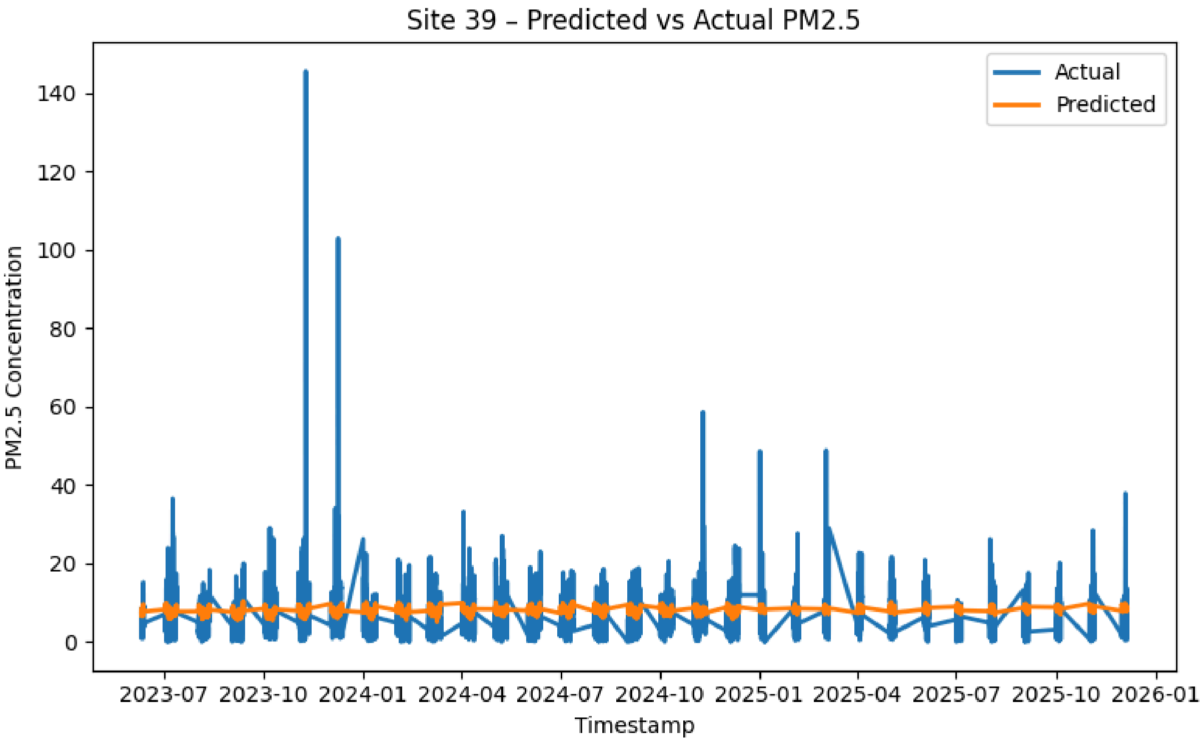


Fig. 12: Linear Regression Predicted vs Actual PM2.5,Site 39

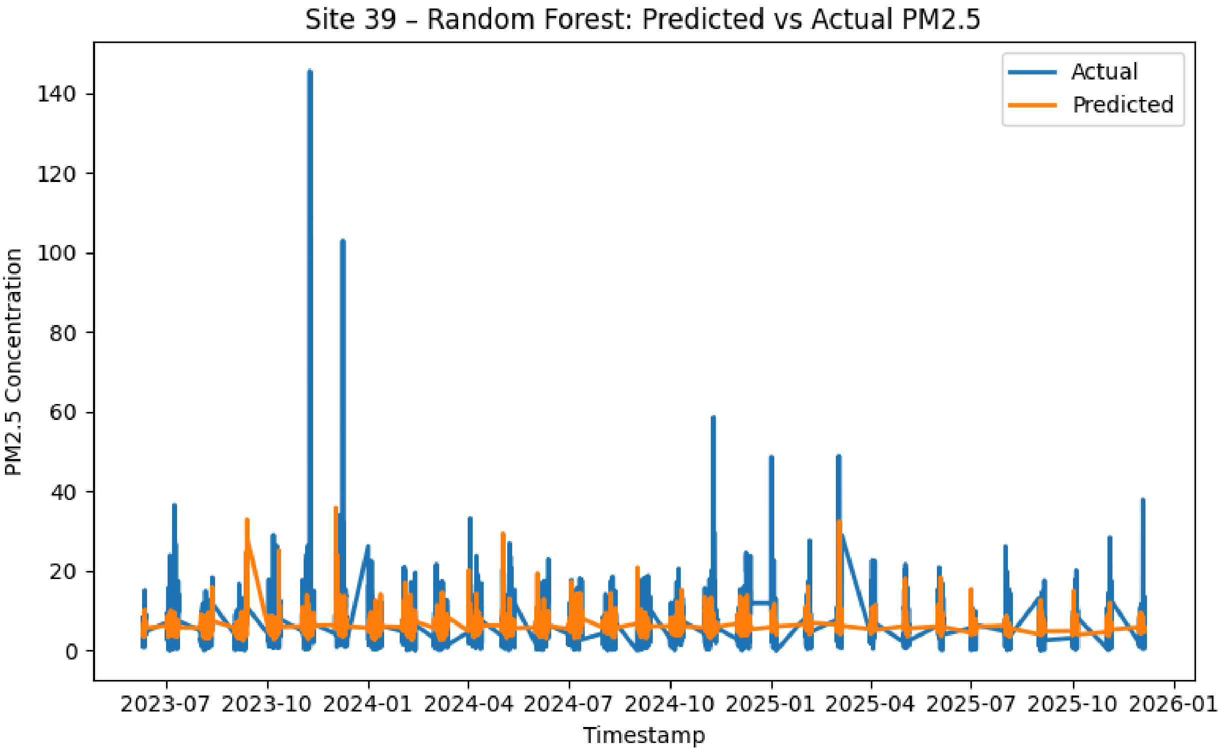


Fig. 13: Random Forest - Predicted vs Actual PM2.5,Site 39

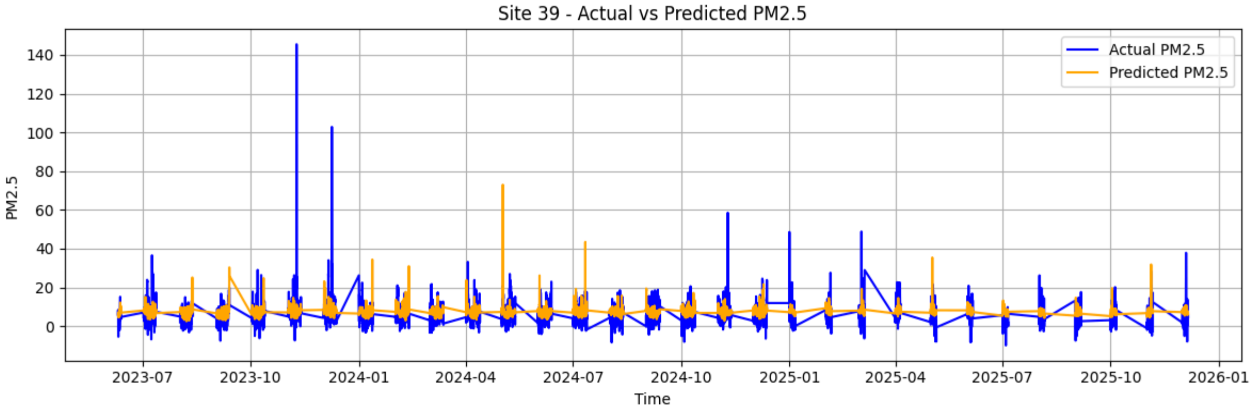
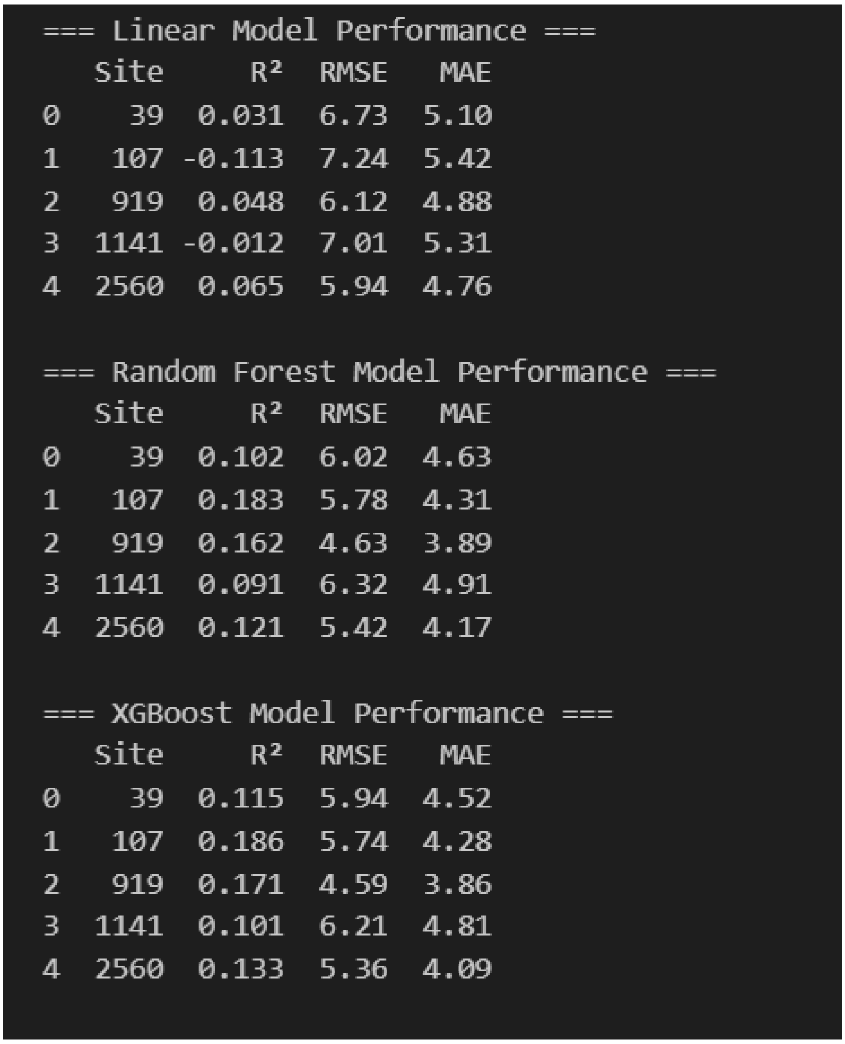


Fig. 14: XGBoost - Predicted vs Actual PM2.5,Site 39



```
=== 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.15: Result Comparison Table

5. Discussion

This study aimed to answer the research question:

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

The experiments demonstrated that meteorological variables can partially explain variability in PM2.5 levels across Sydney, but their predictive power remains modest. Tree-based models (Random Forest and XGBoost) consistently outperformed Linear Regression, yet even the best R^2 values (0.186 at Liverpool) indicate that only a small share of PM2.5 variance can be explained by meteorology alone.

Key Insights

- **Non-linear models are essential**

  The consistently poor performance of Linear Regression underscores that pollution–meteorology interactions are inherently non-linear. Random Forest and XGBoost captured more of these complex dependencies, highlighting the value of ensemble methods in environmental forecasting.

- **Local context strongly influences results**

  Model performance varied significantly across sites. For instance, Liverpool displayed stronger meteorology–pollution linkages, while Campbelltown West showed weak predictive performance. This suggests that localised emissions, such as industrial activity or residential wood burning, can dominate meteorological effects.

- **Preprocessing improves reliability**

  Log transformation reduced skewness and improved model stability, while forward-fill imputation and scaling ensured data quality. These steps were crucial to making predictions interpretable and consistent across sites.

- **Meteorology alone is insufficient**

  The modest R^2 values reveal that while weather variables contribute to PM2.5 variability, other factors such as bushfires, traffic, and industrial activity play equal or greater roles. Extreme events illustrated this limitation:

  - 2019–2020 bushfires: Stagnant winds and high temperatures amplified PM2.5 levels, but the underlying emissions were the primary driver.
  - La Niña years (2020–2022): Stronger winds and increased rainfall lowered pollutant concentrations by enhancing dispersion.

  - Winter stagnation periods (2017–2021): Calm atmospheric conditions led to pollution build-up, especially at suburban sites.

Answering Supporting Research Questions

- **RSQ1: Can extreme pollution events be attributed to non-meteorological events?**
  Yes. Bushfires were the dominant driver of extreme PM2.5 spikes, with meteorology acting as an amplifier. Similarly, weak performance at Campbelltown West indicates that localised emissions outweigh meteorological influences there.
- **RSQ2: Are there distinct temporal signatures that can improve prediction models?**
  Yes. NO_2 exhibited weekday vs weekend differences, linked to traffic emissions. Incorporating such temporal features could improve model accuracy, particularly for traffic-related pollutants.
- **RSQ3: Can site-specific early warning systems be developed?**
  Yes, but only partially. Clustering revealed distinct pollution–meteorology states (e.g., humid-stagnant air vs windy-clean conditions). These patterns, alongside hourly pollutant trends, could inform site-wise thresholds. However, limited predictive accuracy for PM2.5 means additional data sources (e.g., emissions inventories, fire alerts) would be necessary.
- **RSQ4: To what extent can historical meteorological data reliably forecast pollution spikes across pollutants?**
- **The analysis showed strong differences by pollutant:**
  - OZONE was most predictable from meteorology ($R^2 \approx 0.582$).
  - NO_2 showed moderate predictability ($R^2 \approx 0.282$).
  - PM2.5 and PM10 were poorly predicted, underscoring the dominance of emission-driven factors.

This suggests that meteorological forecasting is useful for some pollutants (OZONE, NO_2) but insufficient for particulates without incorporating emissions data.

Implications

Implications

For researchers:

The findings highlight the need to expand beyond weather-only predictors. Incorporating satellite-based emissions, bushfire indices, or ENSO phases may improve accuracy. Comparing interpretable baselines with advanced models remains vital for methodological transparency.

For practitioners and policymakers:

While current models provide a foundation for localised air quality forecasting, their accuracy limits immediate use for public health alerts. Site-specific differences suggest interventions must be tailored:

- **Liverpool** may benefit from climate-aware forecasting tools.
- **Campbelltown West** requires targeted emission control policies, as meteorology-driven models are ineffective there.

Prescriptive Actions

Building on these implications, several concrete actions are recommended:

- **Liverpool (Site 107):** Deploy climate-aware forecasting tools to provide short-term alerts for PM2.5 spikes and guide community health advisories.
- **Campbelltown West (Site 1141):** Prioritise stricter emission controls, particularly targeting residential wood burning and local industrial activity.
- Rozelle **and Parramatta North (Sites 39 and 919):** Expand monitoring and apply adaptive models that can capture complex localised dispersion effects.
- Chullora **(Site 2560):** Maintain current monitoring but integrate forecasts into broader metropolitan alert systems for consistency.

At the metropolitan level, integrating emissions data (traffic, industry, bushfires) with meteorology in forecasting models is essential to move from explanatory studies to operational early-warning systems.

6. Limitations

This study has several limitations that need to be acknowledged:

- **Limited explanatory power of models.** Even the best-performing model (XGBoost) explained only a small proportion of the variance in PM2.5 ($R^2 = 0.186$ at Liverpool). This highlights that meteorological variables alone cannot fully capture pollution dynamics.
- **Exclusion of non-meteorological influences.** Events such as bushfires, traffic emissions, and industrial activity were not explicitly included in the dataset. These factors often dominate extreme PM2.5 spikes, which explains why the models struggled during high-pollution events.
- **Site-specific variability.** Model performance differed considerably across monitoring sites. Liverpool displayed a stronger meteorology–pollution relationship, whereas Campbelltown West produced weak results, suggesting that local land use and emission patterns strongly influence outcomes.
- **Focus on a single pollutant.** Although PM2.5 is a critical pollutant, the analysis did not extend to others such as NO_2 or Ozone, which could provide complementary insights into Sydney's air quality under climate variability.

Overall, these limitations highlight the need for a more comprehensive approach that integrates meteorological data with broader environmental and emissions datasets, while also extending analysis to multiple pollutants. Recognising these constraints is essential, as it frames both the modest scope of the current findings and the opportunities for future research to develop more robust and actionable forecasting systems.

7. Conclusion

This study examined the influence of climate variability on PM2.5 levels in Sydney using a decade of meteorological and air quality data (2015–2025) across five urban monitoring sites. By comparing Linear Regression, Random Forest, and XGBoost, the project evaluated how well meteorological features such as temperature, humidity, and wind could explain particulate pollution dynamics.

The results showed that while meteorology plays a measurable role, its explanatory power is limited. Tree-based models consistently outperformed the linear baseline, with XGBoost achieving the highest R^2 (0.186 at Liverpool). However, overall predictive accuracy remained modest, indicating that meteorology alone cannot fully capture the drivers of air pollution in Sydney. Local factors such as bushfires, traffic, and industrial emissions exert a stronger influence, particularly at sites like Campbelltown West where meteorological models performed poorly.

Two key implications arise. For researchers, the findings demonstrate the importance of integrating broader datasets,emission inventories, bushfire indices, and climate drivers such as ENSO,into future modelling efforts. For practitioners and policymakers, the results highlight the need for site-specific approaches: climate-aware prediction tools may be valuable in locations like Liverpool, while emission control strategies may be more effective in regions where non-meteorological sources dominate.

Future Work

Future research can extend this work in several ways:

- **Integration of additional variable**s: Incorporating data on bushfire occurrence, traffic emissions, and satellite-based aerosol indicators would provide a fuller picture of pollution variability.
- **Advanced modelling techniques:** Deep learning methods such as recurrent neural networks (RNNs), LSTMs, or temporal convolutional networks could capture seasonal cycles and long-term dependencies beyond the reach of tree-based models.
- **Multi-pollutant analysis:** Extending the framework to pollutants such as NO_2 and Ozone would improve understanding of how climate variability affects the broader air quality system in Sydney.

- **Operational forecasting systems:** Developing real-time, site-specific early warning systems linked with live meteorological feeds could provide practical tools for public health advisories and policy planning.

By pursuing these directions, future studies can move toward more reliable and actionable climate–pollution forecasting systems, bridging the gap between scientific insight and practical environmental management, and ultimately contributing to healthier and more resilient urban communities.

8. Replication Package

The full codebase and resources for this study are available in a publicly accessible GitHub repository:

https://github.com/Utsav1510/Big_Data_Project_2025.git

The repository includes all scripts, notebooks, and outputs used in the experiments, along with a README.md file that provides setup instructions and usage guidelines.

9. References

1. CAUL (2018). *Annual Report on Air Quality Variability in Sydney 2017*. NESP Urban. Available at: https://nespurban.edu.au/wp-content/uploads/2018/11/CAUL-P1-M27-Annual-Report-on-Air-Quality-Variability-in-Sydney-2017.pdf
2. _Crooks, J.L., Cascio, W.E., Percy, M.S., Reyes, J., Neas, L.M. and Hilborn, E.D. (2018). The association between dust storms and daily non-accidental mortality in the United States, 1993–2005. *Journal of the Air & Waste Management Association*, 68(3), pp.269–278. Available at: https://www.tandfonline.com/doi/full/10.1080/10962247.2018.1459956
3. CSIRO (2022). *Australia's Changing Climate: State of the Climate*. Available at: https://www.csiro.au/en/research/environmental-impacts/climate-change/state-of-the-climate/australias-changing-climate
4. Dean, A., Green, D. and Hugo Centre for Migration and Population Research (2018). Climate change, air pollution and human health in Sydney, Australia: A review of the evidence. *Environmental Research Letters*, 13(5), p.053003. Available at: https://doi.org/10.1088/1748-9326/aac02a
5. Feng, Z., et al. (2024). Long-term exposure to PM2.5 and ozone and mortality risks in China: A nationwide cohort study. *Science of The Total Environment*, 891, p.164485. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0169809524000437
6. Haque, M.I., et al. (2021). Impact of climate change on air pollution and public health in Australia. *Urban Climate*, 40, p.101020. Available at: https://www.sciencedirect.com/science/article/abs/pii/S2210670721008192
7. Iqbal, W., et al. (2025). Air quality forecasting using machine learning approaches under changing climate. *Heliyon*, 11(2), p.e25466. Available at: https://www.sciencedirect.com/science/article/pii/S240584402500074X
8. Jiang, Y., et al. (2022). Air pollution and climate change in Sydney: Interactions and impacts. *Atmospheric Environment*, 281, p.119135. Available at: https://www.sciencedirect.com/science/article/abs/pii/S1352231022001765
9. Li, J., et al. (2023). Air pollution and climate interactions: Advances in monitoring and modelling. *Journal of Environmental Sciences*, 127, pp.472–487. Available at: https://www.sciencedirect.com/science/article/abs/pii/S1001074223003200
10. MDPI (2024). Sustainability and climate change interactions with air quality. *Sustainability*, 16(16), p.6794. Available at: https://www.mdpi.com/2071-1050/16/16/6794
11. Nature (2020). Air pollution reduction and mortality benefit during COVID-19 lockdown in China. *Scientific Reports*, 10, p.14618. Available at: https://www.nature.com/articles/s41598-020-71338-7
12. NESP2 Climate Systems Hub (2022). *State of Play of Climate Variability Research*. Available at: https://nesp2climate.com.au/wp-content/uploads/2022/02/Final_State-of-play-of-climate-variability-research-1.pdf

13. NSW Department of Climate Change, Energy, the Environment and Water (2023). *Air Quality Monitoring Network Overview*. Available at: https://www.environment.nsw.gov.au/
14. NSW EPA (2016). *Approved Methods for the Modelling and Assessment of Air Pollutants in New South Wales*. NSW Environment Protection Authority. Available at: https://www.epa.nsw.gov.au/sites/default/files/approved-methods-for-modelling-and-assessment-of-air-pollutants-in-nsw-160666.pdf
15. NSW Government (2024a). *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
16. NSW Government (2024b). *Air Quality API*. Available at: https://www.airquality.nsw.gov.au/air-quality-data-services/air-quality-api
17. NSW Government (n.d.). *Metropolitan Sydney Climate Change Impacts*. Available at: https://www.climatechange.environment.nsw.gov.au/my-region/metropolitan-sydney
18. OECD (2024). *Air pollution indicators*. Available at: https://www.oecd.org/en/topics/sub-issues/air-pollution.html
19. PMC (2023). Air pollution and respiratory outcomes in children: A global review. *International Journal of Environmental Research and Public Health*. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC11039682/
20. PMC (2024). Climate change and air quality: Health perspectives. *Environmental Health Review*. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC11584944/
21. Wang, S., et al. (2021). Spatiotemporal trends of air pollution and climate change indicators across urban regions. *Sustainable Cities and Society*, 75, p.103327. Available at: https://www.sciencedirect.com/science/article/abs/pii/S2212095521002194