COMP9417 – Group Project

Aayush Veturi, z5428469

Arnav Raina, z5476671

Neha Gajendra, z5477706

Vansh Kalra, z5583207

2025 T3

# Introduction

Air pollution forecasting plays a critical role in environmental monitoring, urban planning and public health protection. Accurate short-term predictions allow governments to issue warnings, mitigate exposure risks, and better understand how human activity and meteorological conditions influence pollutant behaviour. The significance of pollutant forecasting is reinforced.

By established public health evidence. As Pope and Dockery (2006) note, exposure to fine airborne particles is "persuasively linked to adverse cardiopulmonary outcomes," with effects observed even at low-to-moderate pollution levels and across both short and long-term exposure windows [1]. This underscores the relevance of developing models capable of anticipating pollution fluctuations before they accumulate into harmful exposure conditions.

The data used in this analysis is sourced from the UCI Machine Learning Repository and is an Air Quality dataset containing 9,358 hourly observations between March 2004 and 2005 [2]. The air quality is measured from a road-level monitoring station based in Italy [2]. Each hourly entry provides data pertaining to 4 key pollutants (CO, C6H6, NOx, NO2), meteorological attributes of the environment (temperature, relative humidity, and absolute humidity). Additionally, sensor responses for the pollutants are also provided.

Exploratory data analysis is used to examine the complex relationships connecting pollutants, and meteorological conditions, supported by anomaly and event detection to verify data reliability. This context is used to better understand how these relationships drive pollution concentrations. Feature engineering is utilised to better organise the data for efficient learning by the employed ML models. This analysis focuses on evaluating Linear regression, gradient boosting regression models, and logistic regression and random forest classification models. This report aims to analyse, compare and interpret the performance of these different models, and assess limitations to each model as well as the potential ways to improve prediction accuracy.

# Exploratory Data Analysis

The dataset contains hourly measurements of gaseous pollutants (CO, C6H6, NO2 and NOx) alongside meteorological attributes recorded over 2004-2005. This EDA focused on understanding the quality of the data, and identifying relationships among pollutants and environmental variables, as these insights are what guide the preprocessing and modelling strategies applied later on.

Given in the raw data, missing values, sensor faults, or invalid data was represented by sentinel vales (-200), a heatmap was generated (Appendix A: figure 1) to assess data reliability (by matching sensor data with pollutant concentrations), how much of the data was missing and the pre-processing otherwise required. An initial assessment revealed that significant data was missing for some of the pollutants. NMHC(GT) exhibited prolonged periods of absent readings. NOx, NO2 as well as CO, also demonstrated relatively large sections of missing data, however not as high of an extent as seen NMHC. Meteorological variables as well as the pollutant benzene, C6H6, possessed little to no missing data in contrast. The other pollutant sensors (e.g., PT08.S2(NHMC), PT08.S3(NOx), etc) possess some white streaks within the data, which indicate instances of potential sensor faults; these gaps do not necessarily correspond to white gaps present in pollutant data, which could be indicative of faulty or inaccurate measurements in the pollutant concentration data, at the core of this analysis. While the DateTime property also showed missing values in the heatmap, it is worth noting this is due to the variable not being heatmap compatible; therefore, no DateTime data was missing when manually examined.

Analysis of average pollutant concentration over different time frames levels reveals strong diurnal patterns. Observing hourly trends (Appendix A: Figure 5), NOx and NO2 display dominant concentration peaks around 7-10 am, as well as 4-9 pm. CO and C6H6 show similar dynamics in terms of peak behaviour, however the average hourly concentrations are significantly lower as compared to NOx and NO2. Observing weekly trends (Appendix A: Figure 6) in average pollutant concentration, concentration remains relatively consistent all days of the week for all pollutants, excluding Wednesday where the NOx concentration is elevated.

A monthly scatter plot of pollutant concentrations (Appendix A: Figure 4) highlights the absence of a smooth seasonal trend and instead reveals a repeating pattern of sharp short-term spikes, particularly for NOx. Some of these spikes exceed typical concentrations by large margins, indicating either transient pollution surges or potential sensor-related anomalies. Because these extreme values can disproportionately influence error metrics in regression models, identifying and treating these events through anomaly detection is essential for robust longer-horizon forecasting.

Finally, to investigate relationships between pollutant concentrations, and meteorological variables, a correlation matrix was computed (Appendix A: Figure 3). Strong positive correlations were observed among the pollutants themselves, with correlations ranging from 0.61-0.93. The strongest positive correlations were between CO and C6H6, as well as NOx and CO. In contrast, correlations between pollutants and meteorological variables were comparatively very weak. Temperature showed mild negative associations with NOx and NO2, while humidity variables exhibited only modest relationships with pollutant concentrations. These correlation insights imply substantial multicollinearity among pollutant features, reinforcing the need for models capable of handling correlated predictors or incorporating regularisation, and confirming the importance of temporal features over meteorological drivers.

# Methodology

## Preprocessing

For data preprocessing, pollutant concentration and meteorological measurements containing missing or invalid sensor readings (encoded as –200 in the raw dataset) were first converted to NaN. These missing values were then handled using time-based interpolation, which was appropriate given that the gaps occurred in continuous blocks rather than as isolated random points. This preserves the temporal structure of the series and avoids introducing artificial discontinuities. All columns were standardised by cleaning inconsistent naming conventions and removing duplicate records. The separate date and time fields were merged into a single DateTime attribute, which served as the primary index for sorting and organising the dataset.

To support downstream modelling, several temporal structural features including hour of day, weekday, month, and year, were engineered from the DateTime field. These features were essential for capturing the strong diurnal and weekly patterns identified during the exploratory data analysis, enabling the models to learn recurring behavioural cycles in pollutant concentrations.

## Feature Engineering

The feature engineering for this assignment was directly guided by the temporal patterns, pollutant relationships and behaviours interpreted and identified in the EDA. The primary features programmed into our codebase were cyclical time features, lag features, and rolling average features.

Given the EDA demonstrated strong diurnal peaks in pollutant concentration as well as repeating weekly and monthly patterns, the cyclical time representation enables us to maintain a periodic structure, and capture interdependencies between past and future data. To create this feature, hour and month time frames were encoded through sinusoidal transformations.

The purpose of the lag features is to capture short-term and long-term temporal dependence seen with pollutant concentration. Through the EDA, specifically the hourly concentration plots, the average concentration levels were relatively continuous in appearance, demonstrating how previous hours concentration of pollutants can affect future concentrations. For this reason, to capture autocorrelation between current and past concentrations, 1 and 2-hour lag features were generated and added to the data for all pollutants and meteorological variables. Additionally, a 24-hour lag was also added to reflect reoccurrence of hourly trends.

## Temporal Data Splitting

We used a chronological train–test split, training the model on all 2004 data (73.56%) and testing it on all 2005 data (26.44%). This avoids leakage by ensuring the model only learns from past information and is evaluated on truly future observations.

## Model Selection Rationale

The EDA highlighted several structural characteristics of the dataset that determine which models are most suitable. The strong diurnal structure of pollutant concentrations suggests that models must accommodate sequential dependencies, lag relationships, and non-linear temporal patterns. The correlation analysis revealed both high pollutant-pollutant correlations and much weaker pollutant meteorological associations, indicating the need for models that can regularise or tolerate correlated predictors. Finally, given the extent of missing and interpolated data, models must operate robustly following temporal interpolation and lag-window feature construction.

This project evaluates both regression and classification models. Regression models are used to forecast pollutant concentrations 1, 6, 12, and 24 hours ahead, while classification models categorise CO concentrations into low, mid, and high.

For regression, Linear Regression and Gradient Boosting were chosen. Linear Regression provides an essential baseline: the EDA showed strong linear relationships among pollutants, meaning concentrations tend to rise and fall proportionally. As a fast, regularizable, and interpretable model, it allows us to measure the added value of more complex models. Gradient Boosting complements this by addressing the dataset's non-linearity. The EDA revealed sharp hourly peaks and irregular short-term spikes patterns that linear models cannot capture. Gradient Boosting can model non-linear behaviour, exploit interactions introduced through lag and rolling features, and perform well for multi-horizon forecasting.

For classification, Logistic Regression and Random Forest were selected. Logistic Regression is appropriate because the temporal features (hour, weekday, cyclical encodings) and pollutant interactions create largely monotonic and separable relationships after preprocessing. Its regularisation helps manage multicollinearity, and its interpretability provides a clear baseline for CO-level classification. Random Forest, in contrast, captures the non-linear structures identified in the EDA including short-term spikes, threshold effects, and interactions between lagged features. It is robust to noisy or interpolated predictors and naturally handles correlated inputs without requiring feature scaling.

Together, these four models reflect a balanced progression from interpretable baselines to more expressive non-linear learners, aligned with the behavioural structure uncovered in the EDA and the engineered feature set used in this analysis.

## Evaluation of Performance

Model performance for both regression and classification is assessed by comparing each models prediction for horizons of 1hr, 6hr, 12hr and 24hrs against a naïve baseline that uses discretized pollutant concentration for time t, as the prediction for t +1, t+6, t+12 and t+24. This provides a meaningful metric by which to quantify home much improvement, if any, is achieved through the application of temporal features, and trend modelling to predicting pollutant concentration.

For regression, 1hr, 6hr, 12hr and 24hr horizon forecasting is achieved and evaluated using root mean squared error (RMSE). The naïve baseline mentioned prior assumes future pollutant concentration is equal to current concentration, and hence a regression model is preferred when the RMSE is lower than the naïve RMSE at the same horizon. A model performing worse than the naïve baseline would indicate it has not learned useful temporal structure and could potentially be overfitting or underfitting the data.

For classification, given it pertains to the pollutant CO in particular, concentration is classified by classes Low, Mid and High. Performance is measured by accuracy, weighted f1, and macro f1 score. Here, a good model is when accuracy exceeds the naïve baseline without disproportionately favour the majority classes, and when the weighted f1 improves (better than corresponding naïve baseline f1 score). Once again, poor performance is when naïve baseline performance is not beaten, indicating that the model isn't learning from temporal features and non-linear model structures.

# Results

*Table 1: Regression Model Performance (RMSE) Across Pollutants and Forecast Horizons*

| Pollutant | Horizon | Gradient Boosting RMSE | Linear Regression RMSE | Naive RMSE |
|---|---|---|---|---|
| CO(GT) | 1h | 0.789 | 0.699 | 0.796 |
| CO(GT) | 6h | 1.343 | 1.300 | 1.840 |
| CO(GT) | 12h | 1.265 | 1.232 | 1.934 |
| CO(GT) | 24h | 1.257 | 1.443 | 1.581 |
| NMHC(GT) | 1h | 308.167 | 153.518 | 0.000 |
| NMHC(GT) | 6h | 227.192 | 168.017 | 0.000 |
| NMHC(GT) | 12h | 276.657 | 183.299 | 0.000 |
| NMHC(GT) | 24h | 242.164 | 243.959 | 0.000 |
| C6H6(GT) | 1h | 3.541 | 3.369 | 3.346 |
| C6H6(GT) | 6h | 7.064 | 6.387 | 7.688 |
| C6H6(GT) | 12h | 6.571 | 6.602 | 8.272 |
| C6H6(GT) | 24h | 5.687 | 7.264 | 7.009 |
| NOx(GT) | 1h | 118.129 | 124.180 | 118.855 |
| NOx(GT) | 6h | 190.195 | 223.284 | 302.740 |
| NOx(GT) | 12h | 214.163 | 286.193 | 304.650 |
| NOx(GT) | 24h | 230.790 | 283.278 | 240.026 |
| NO2(GT) | 1h | 37.116 | 31.351 | 25.657 |
| NO2(GT) | 6h | 53.575 | 59.531 | 71.725 |
| NO2(GT) | 12h | 55.482 | 70.143 | 73.470 |
| NO2(GT) | 24h | 65.490 | 71.267 | 57.155 |

*Table 2: Regression Model Performance (RMSE) Across Pollutants and Forecast Horizons*

| **Pollutant** | **Horizon** | **Logistic Regression Accuracy** | **Logistic Regression F1 Score** | **Random Forest Accuracy** | **Random Forest F1 Score** | **Naive Baseline Accuracy** | **Naïve Baseline F1 score** |
|---|---|---|---|---|---|---|---|
| CO(GT) | 1h | 0.707 | 0.682 | 0.724 | 0.730 | 0.748 | 0.748 |
| CO(GT) | 6h | 0.498 | 0.392 | 0.577 | 0.584 | 0.402 | 0.402 |
| CO(GT) | 12h | 0.520 | 0.425 | 0.535 | 0.535 | 0.346 | 0.346 |
| CO(GT) | 24h | 0.518 | 0.446 | 0.486 | 0.492 | 0.496 | 0.497 |

Across the regression models (linear regression and gradient regression), performance was evaluated using RMSE for each pollutant and forecast horizon. Linear regression produced RMSE values of 0.473 (CO), 2.217 (C6H6), 85.642 (NOx) and 19.989 (NO2). Gradient Boosting achieved PMSEs of 0.789-1.257 (CO), 3.541-5.687 (C6H6), 118.129-230.790 (NOx) and 37.116-65.490 (NO2) across the 1h, 6h, 12h, and 24h horizons. Corresponding naïve baseline RMSEs ranged from 0.796–1.934 (CO), 3.346–8.272 (C6H6), 118.855–304.650 (NOx), and 25.657–57.155 ($NO_2$), with full results summarised in Table 1 above.

For the CO classification task, prediction accuracy was the desired performance metric. Logistic regression achieved accuracies of 0.707 (t+1), 0.498 (t+6), 0.520 (t+12), and 0.518 (t+24). Random Forest reported accuracies of 0.724 (t+1), 0.577 (t+6), 0.535 (t+12), and 0.486 (t+24). The naïve baseline accuracies for the same horizons were 0.748, 0.402, 0.346, and 0.496, respectively. These results are shown in Table 2 above.

# Discussion

## Model Performance Overview (Regression)

Both Linear Regression (LR) and Gradient Boosting Regressor (GBR) follow consistent trends across pollutants and forecast horizons. At short-term horizons (1 h), pollutant concentrations change slowly, so naïve forecasts perform unexpectedly well. In several cases, the naïve model even beats both learned models, for example, **$NO_2$ at 1 h** (Naïve RMSE = 25.66 vs LR = 31.35 and GBR = 65.5). This pattern is visible in the **Actual vs Predicted plots in Figure 21-22** where all models closely follow the observed 1-hour behaviour.

As the forecast window increases to 6 h, 12 h, and 24 h, naïve performance deteriorates sharply, while LR and especially GBR remain more stable. For instance, **CO(GT) at 6 h** shows naïve RMSE = 1.84 compared with GBR at 1.34, visible in **Figure 20**, where GBR tracks the smoothed trend more closely than LR. These figures highlight that learned models handle longer-horizon, less autocorrelated dynamics more effectively.

Excluding NMHC due to invalid RMSE values, GBR achieves the lowest RMSE in **9 out of 16** pollutant–horizon combinations. LR wins **4/16**, mostly at shorter horizons. The naïve model performs well only at 1 h (winning **3/16**), confirming that machine-learning and econometric models generalise far better once the prediction task becomes non-trivial.

## LR vs GBR Performance

LR performs competitively at short horizons because pollutant behaviour is strongly autocorrelated and mostly linear. This is visible in the 1-hour plots (e.g., **Figures 17–26**), where LR and naïve nearly overlap. However, as horizons lengthen, GBR consistently outperforms LR. The **12 h and 24 h figures (Figures 17–26)** show clearer deviations where LR underfits turning points and nonlinear changes, while GBR adapts more flexibly. Tree-based models are also more robust to multicollinearity, which LR suffers from e.g., **CO(GT) and C6H6(GT)** correlate at **0.93**, inflating LR's variance and reducing stability.

## Pollutant-Specific Insights (with References)

- **CO(GT):** The clearest and most predictable pollutant. In **Figures 19 and 20**, both LR and GBR track actual values well, but GBR handles the longer-term drift more accurately.
- **$NO_2$(GT):** More erratic and noisy. In **Figures 21 and 22**, the naïve model performs surprisingly well at 1 h and 24 h because of strong short-run autocorrelation, but GBR improves performance at 6 h and 12 h.
- **C6H6(GT):** Moderate predictability. In **Figure 17 and 18**, GBR shows noticeably smoother long-horizon predictions compared with LR, which tends to overshoot.
- **NOx(GT):** Highly volatile. As shown in **Figure 25**, LR struggles at longer horizons, while GBR stabilises the worst fluctuations.

## Limitations & Future Directions

- **Data Quality:** NMHC was excluded due to severe missingness and invalid RMSE values.
- **Anomalies:** The anomaly plots (**Figures 7-9,11,13,15**) clearly show clusters of sensor spikes and irregular values in $NO_2$, NOx, and C6H6. These outliers distort model training and contribute to the volatility seen in LR and naïve predictions, especially at 24 h. Incorporating anomaly-filtering (e.g., Isolation Forest pre-cleaning) before training would likely reduce RMSE substantially.
- **Feature Gaps:** Important environmental drivers (wind, temperature, traffic, seasonality) were not included. Their absence reduces long-horizon accuracy.
- **Model Scope:** While GBR handles nonlinearity well, deeper temporal structure could benefit from LSTM/sequence models or boosted hybrid approaches.
- Multicollinearity is clearly present in the dataset, particularly among pollutants such as CO(GT), C6H6(GT), and NOx(GT). As shown in **Figure 3**, CO and C6H6 are correlated at **0.93**, and several other pollutant pairs exceed **0.70**, which is high enough to destabilise Linear Regression coefficients. This helps explain why LR performs inconsistently at longer horizons, whereas tree-based models like GBR remain stable.

For CO-level classification, performance follows the temporal structure observed in the EDA. Logistic Regression achieved accuracies of 0.707, 0.498, 0.520 and 0.518 across the 1 h, 6 h, 12 h and 24 h horizons, while Random Forest achieved 0.724, 0.577, 0.535 and 0.486. The naive baseline performed highest at 1 h (0.748), consistent with the strong short-term class persistence seen in the diurnal concentration curves, before dropping to 0.402 and 0.346 at 6 h and 12 h and recovering slightly to 0.496 at 24 h.

At 1 h, both models classify Low and High levels effectively. Logistic Regression predicts Low correctly 378 times but misclassifies Medium as Low 140 times, showing its tendency to collapse borderline cases into the majority short-term class. Random Forest distributes errors more evenly, predicting Low correctly 312 times and Medium correctly 150 times, which reflects its ability to model the nonlinear pollutant–meteorology relationships identified in the EDA correlation heatmaps.

At 6 h and 12 h, temporal dependence weakens and classification becomes more difficult. For Logistic Regression at 6 h, Medium is again strongly misclassified as Low (219 Medium to Low errors), and at 12 h the model predicts zero Medium across the entire test set, instead assigning 195 Medium cases to Low. This pattern aligns with the diminishing influence of lag features and cyclical encodings at medium-range horizons. Random Forest retains better balance, correctly identifying 86 Medium cases at t+6 and 89 at t+12, though confusion between Medium and High increases, consistent with the EDA's observation of volatility and overlapping concentration ranges around transitional hours.

By 24 h, the predictive value of engineered features has largely decayed. Logistic Regression overwhelmingly predicts Low, assigning 254 Low cases correctly but also misclassifying 86 Medium cases and 62 High cases as Low. Random Forest exhibits broader misclassification patterns, including 130 High to Medium errors, indicating that class boundaries have become less separable. The partial improvement of the naive baseline at this horizon suggests a re-emergence of daily periodicity, although the models do not fully capture it.

The anomaly plots in the EDA show irregular sensor spikes and noisy clusters, particularly in NO2, NOx and C6H6. These anomalies were not filtered prior to training and likely blurred class boundaries, contributing to Logistic Regression's collapse toward Low at longer horizons and Random Forest's growing confusion between Medium and High. While the magnitude of this effect cannot be isolated, the misclassification trends match the locations of instability highlighted in the EDA.

Overall, Logistic Regression with L2 regularisation provides a stable and interpretable baseline at short horizons where pollutant behaviour is smooth. Random Forest performs best at short and medium horizons by capturing nonlinear interactions and short-term spikes. Performance at 24 h is limited for both models, suggesting future improvements such as anomaly filtering, incorporating additional environmental drivers and the use of sequence-based models capable of modelling long-range temporal structure.

# Conclusion

This project set out to forecast air pollutant behaviour using exploratory analysis, preprocessing, feature engineering, and machine learning models. The data showed clear diurnal and weekly cycles, along with sharp short-term spikes linked to transient events or sensor irregularities. These patterns guided the use of lagged features, rolling averages, and cyclical time encodings to properly represent temporal structure.

The results show that both Linear Regression and Gradient Boosting learned meaningful pollutant dynamics. Linear Regression exceeded the naïve baseline for stable pollutants, achieving RMSE values of roughly 0.7–1.4 (CO) and 2–3 (C6H6) with $R^2$ scores near 0.85–0.88 at shorter horizons. Gradient Boosting performed better for more volatile pollutants such as NOx, reducing RMSE by 60+ units compared to Linear Regression at longer horizons. In the CO classification task, Logistic Regression and Random Forest improved on the naïve baseline across 6–24 hour forecasts, with Random Forest reaching accuracies up to 0.72 at 1 hour.

The findings show that pollutant forecasting benefits strongly from temporal feature engineering and from models capable of capturing non-linear behaviour. Linear models offer reliable baselines for smoother pollutants, while Gradient Boosting and Random Forest are better suited to pollutants with high variability or measurement noise. With further refinement of features or the integration of advanced sequence models, prediction accuracy could be improved even more. This project demonstrates the effectiveness of combining domain-aware preprocessing with targeted machine learning techniques to better understand and predict air quality patterns.

# Bibliography

[1] C. A. Pope and D. W. Dockery, “Health Effects of Fine Particulate Air Pollution: Lines that Connect,” February 2006. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/10473289.2006.10464485. [Accessed 23 November 2025].

[2] S. Vito, “UCI Machine Learning Repository,” 2016. [Online]. Available: https://archive.ics.uci.edu/dataset/360/air+quality.. [Accessed 22 November 2025].

# Appendices

## Appendix A: EDA

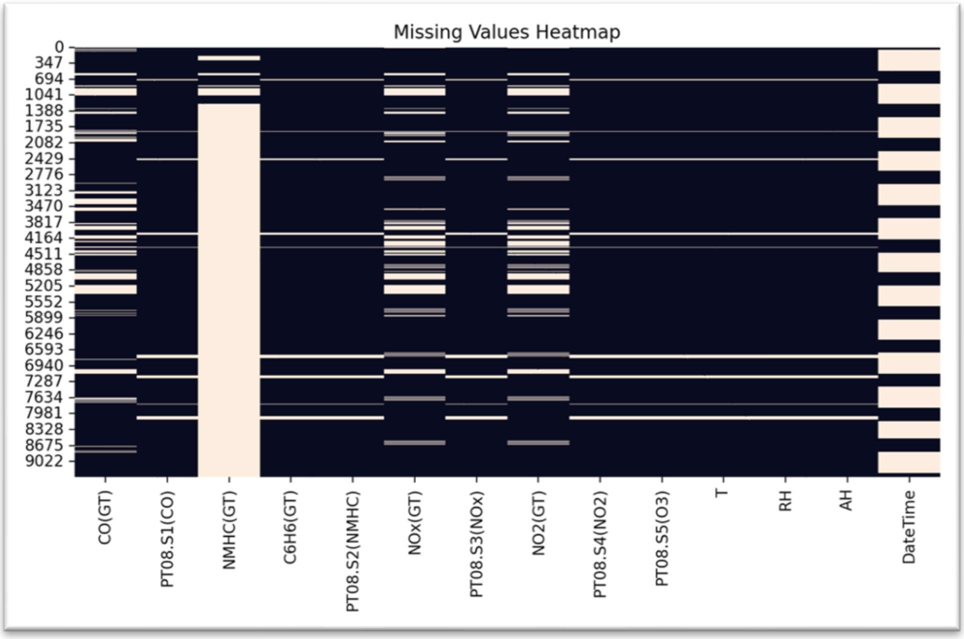

*Figure 1: Heat map of missing values from UC Irvine Air Quality Data*

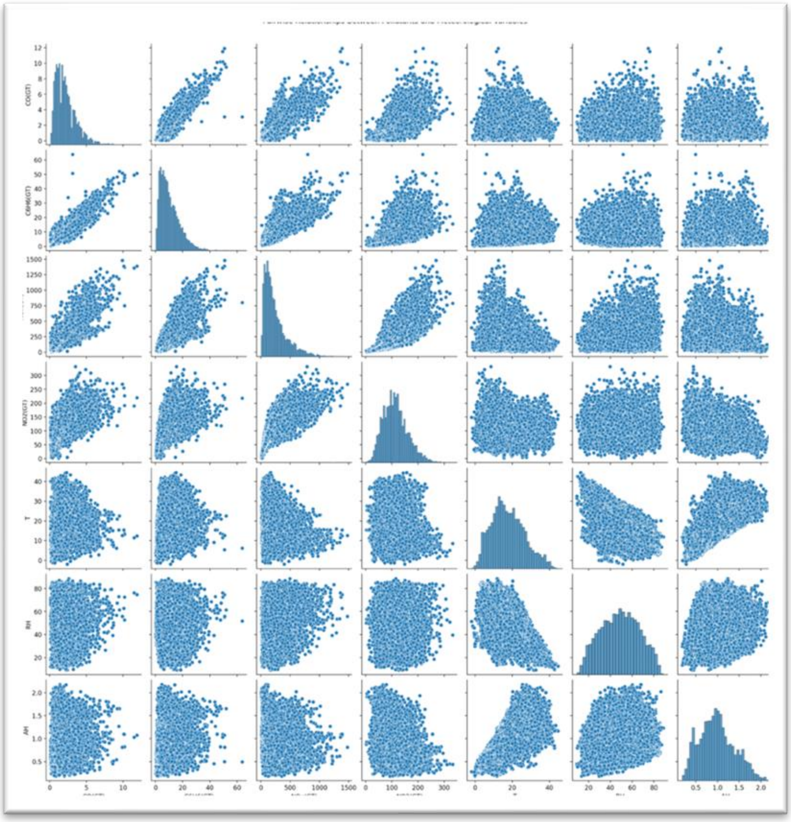

*Figure 2: Pairwise relationships between pollutants and meteorological variables*

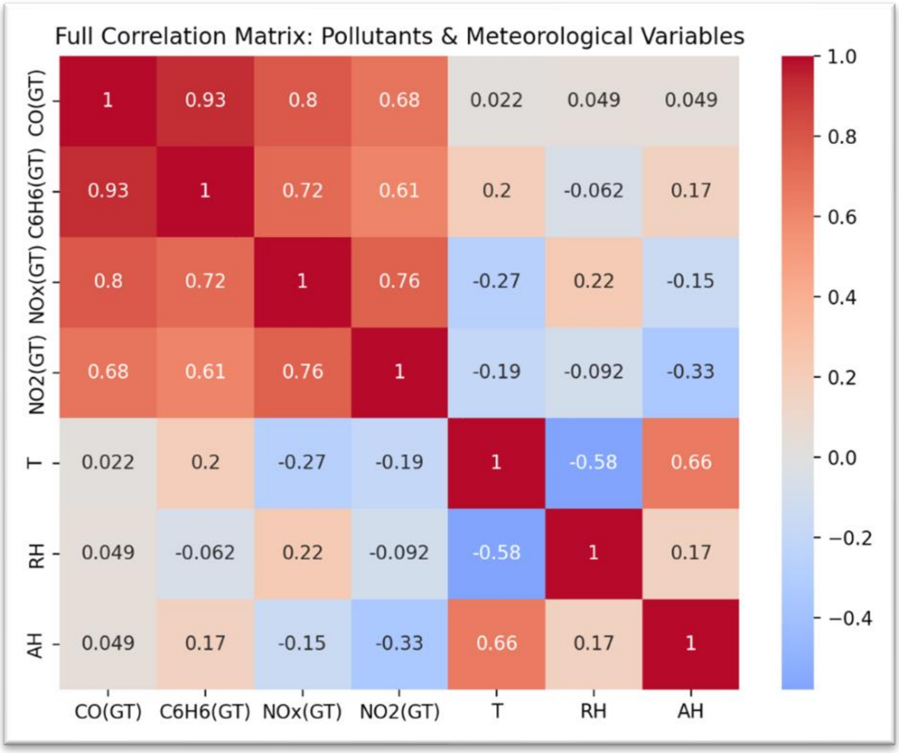

*Figure 3: Correlation matrix between pollutants and meteorological variables*

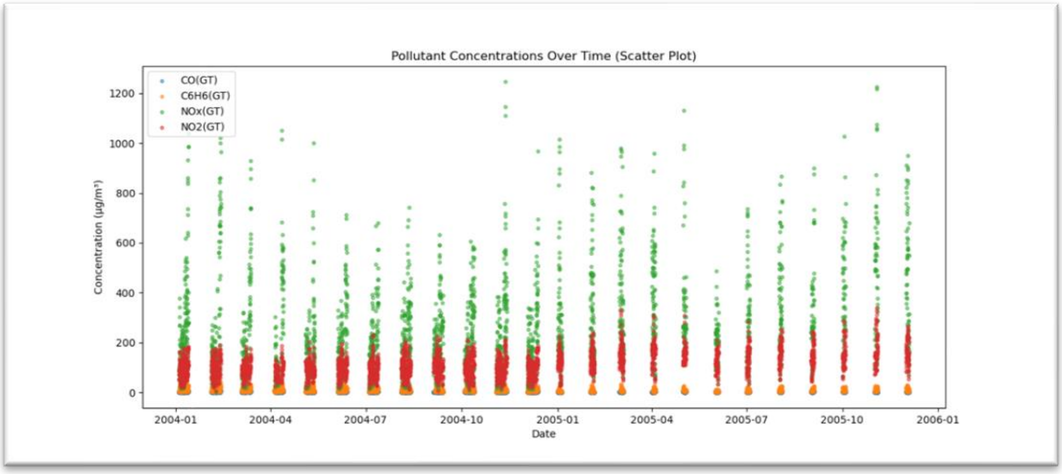

*Figure 4: Pollutant concentrations over time grouped by specific pollutant*

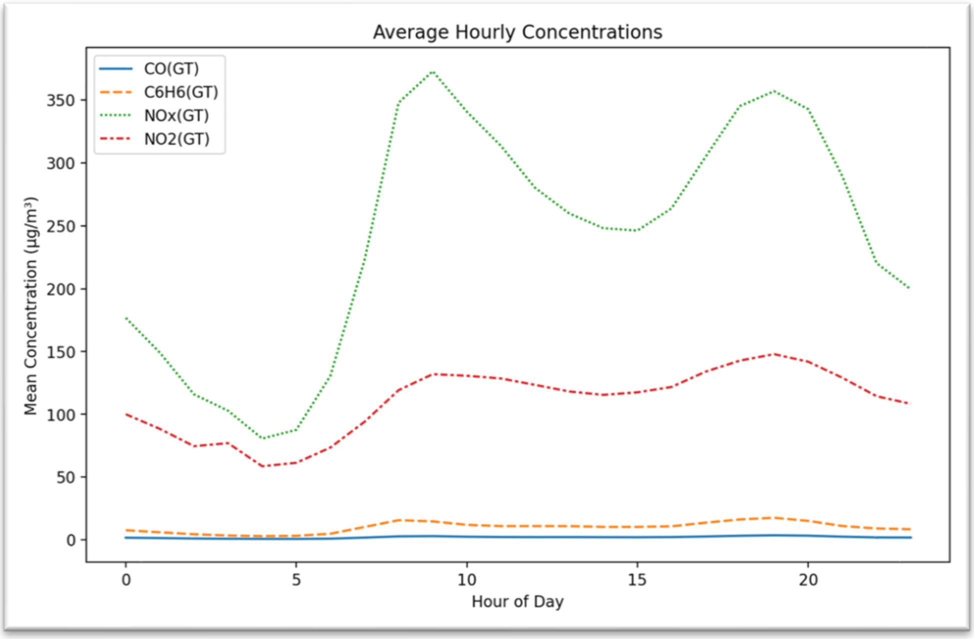

*Figure 5: Pollutant concentrations (average) vs. hour of day*

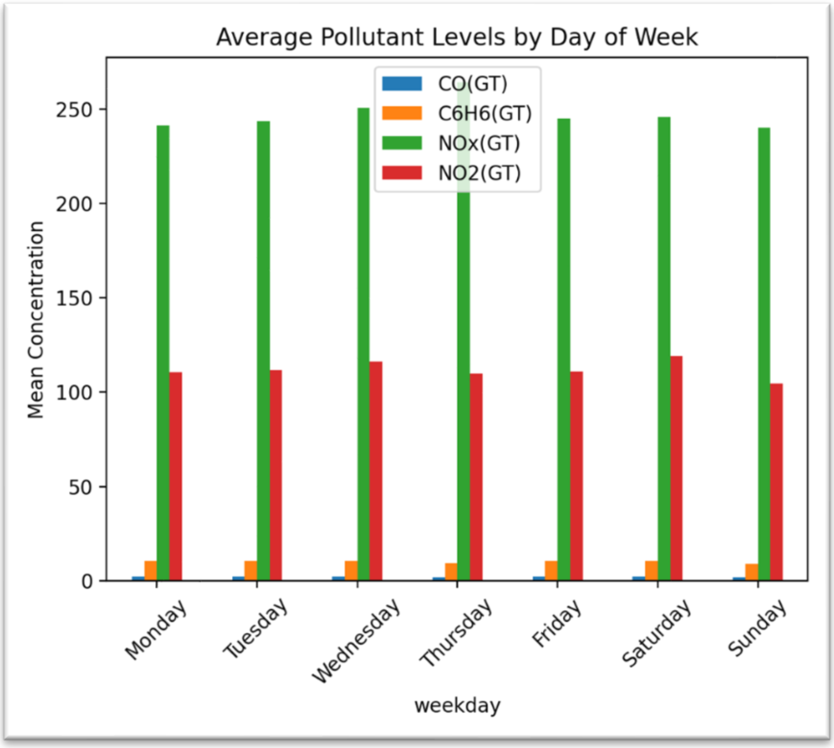

*Figure 6: Pollutant concentration (mean) vs. weekday*

## Appendix B: Anomaly Detection Analysis Plots

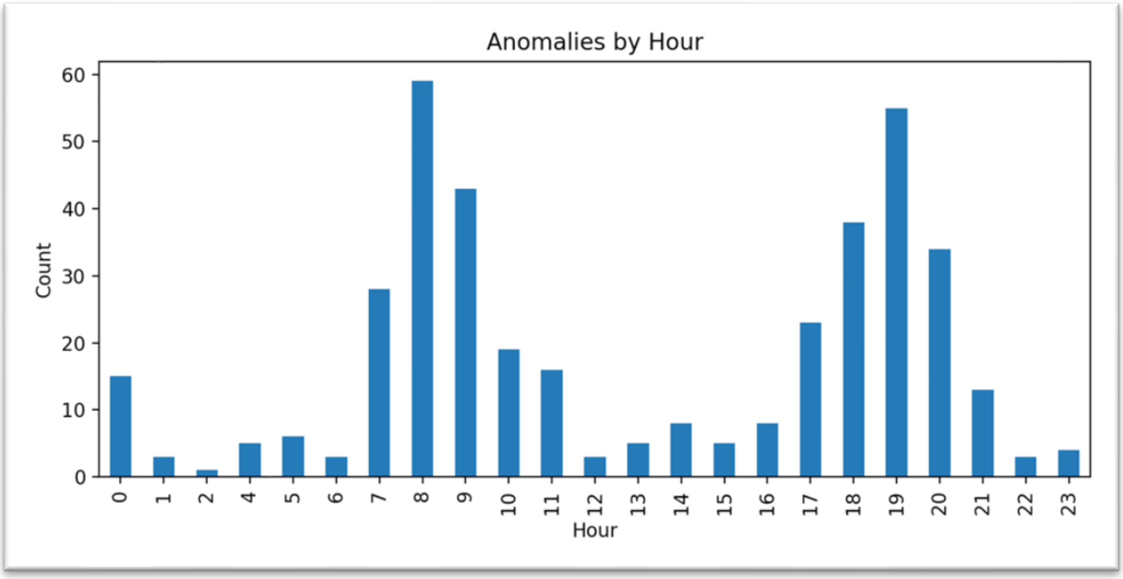

*Figure 7: Anomalies count by hour between all pollutants*

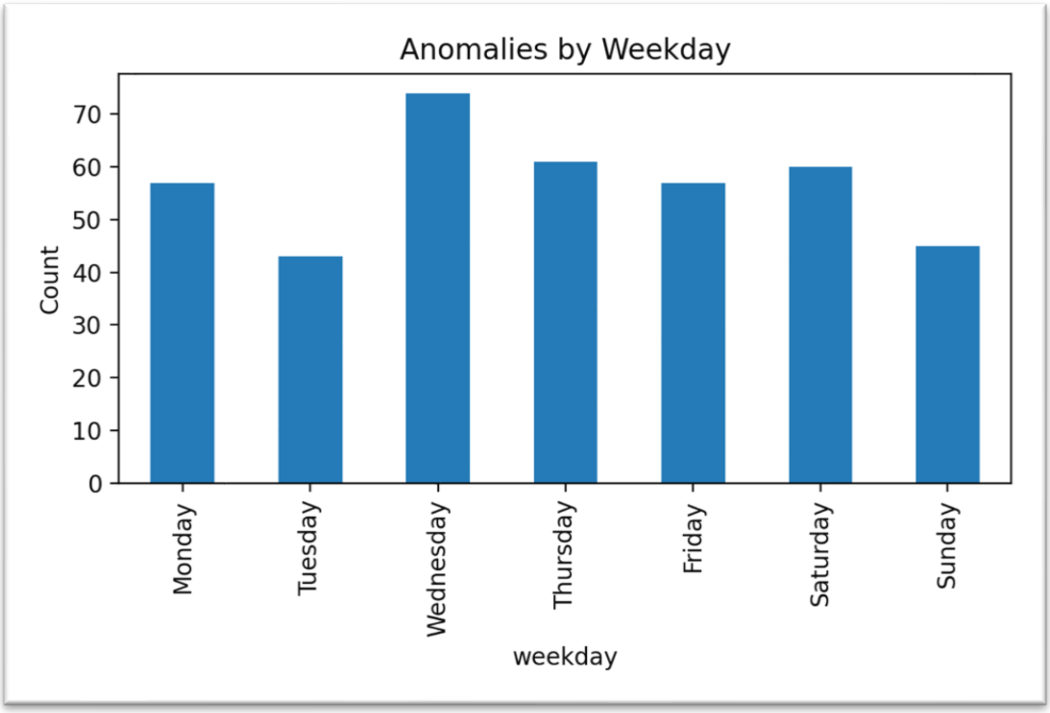

*Figure 8: Anomalies count per week between all pollutants*

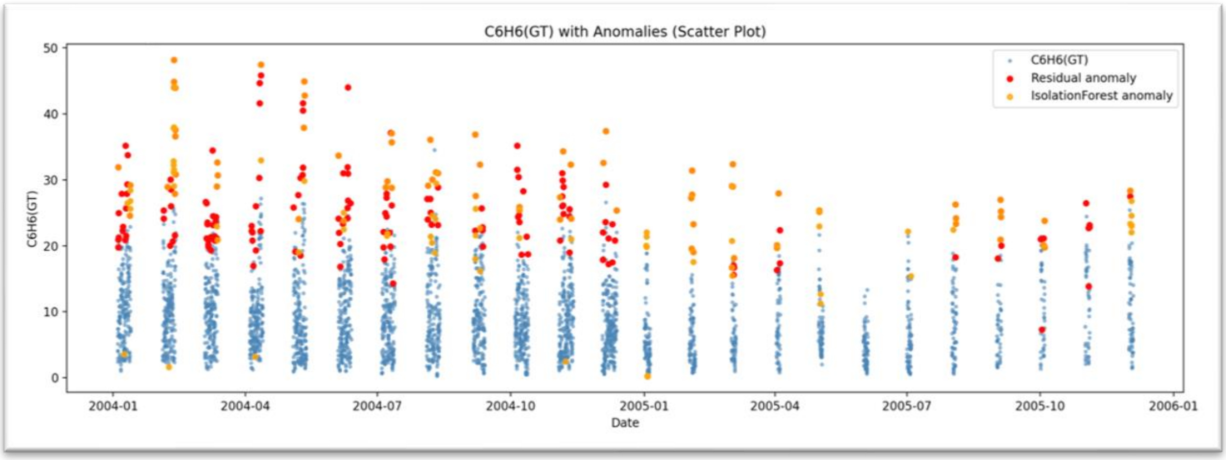

*Figure 9: Anomalies for C6H6*

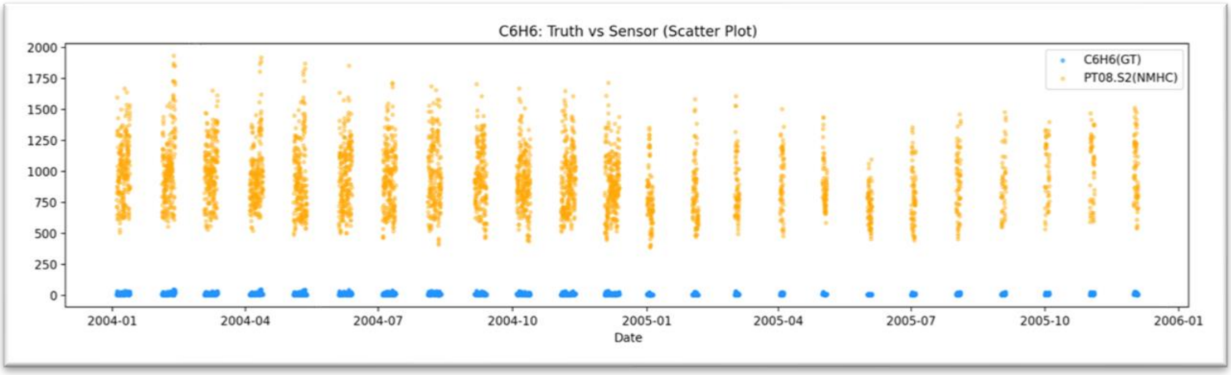

*Figure 10: Truth vs. Sensor Data for C6H6*

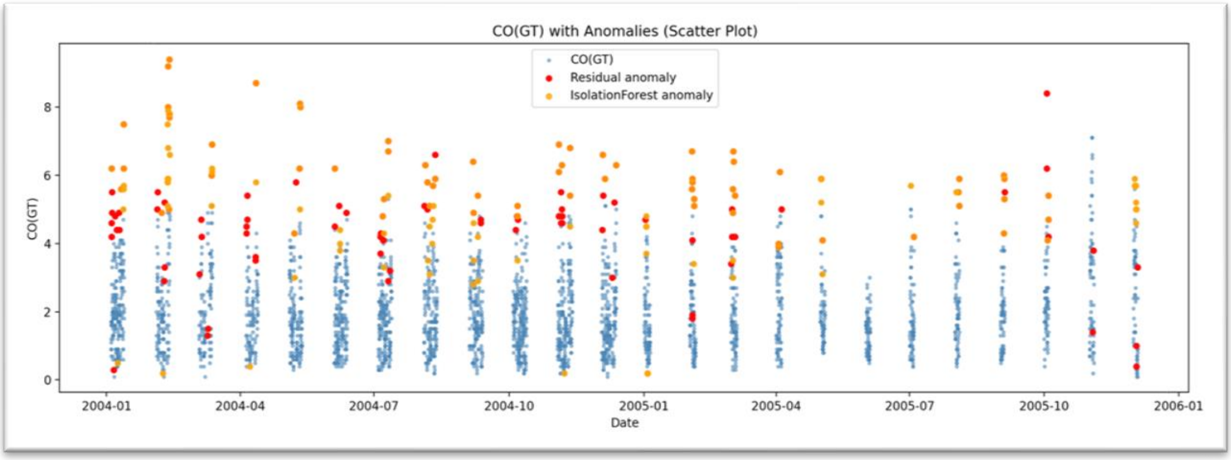

*Figure 11: Anomalies for CO*

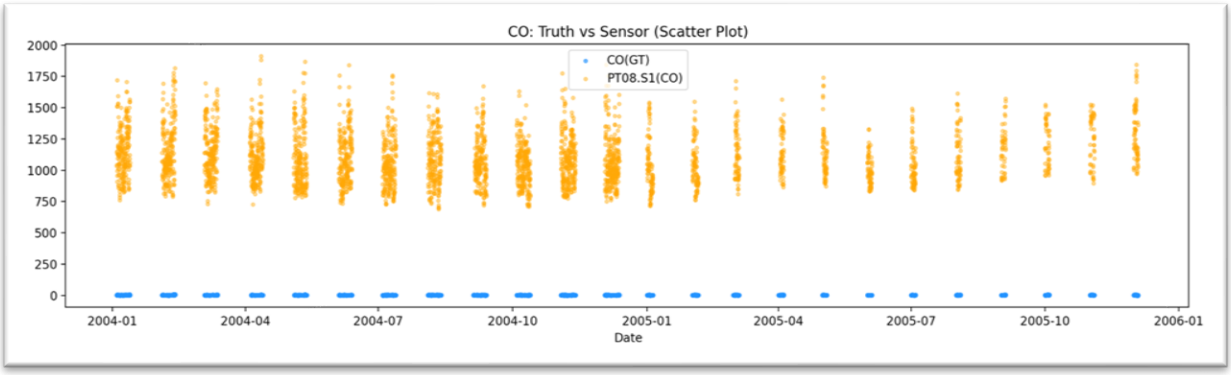

*Figure 12: Truth vs. Sensor for CO*

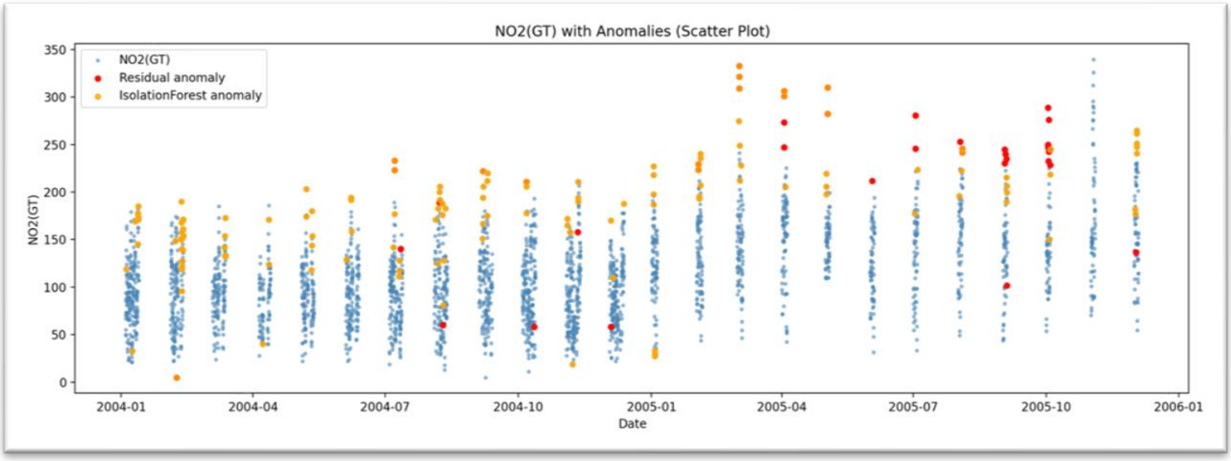

*Figure 13: Anomalies for NO2*

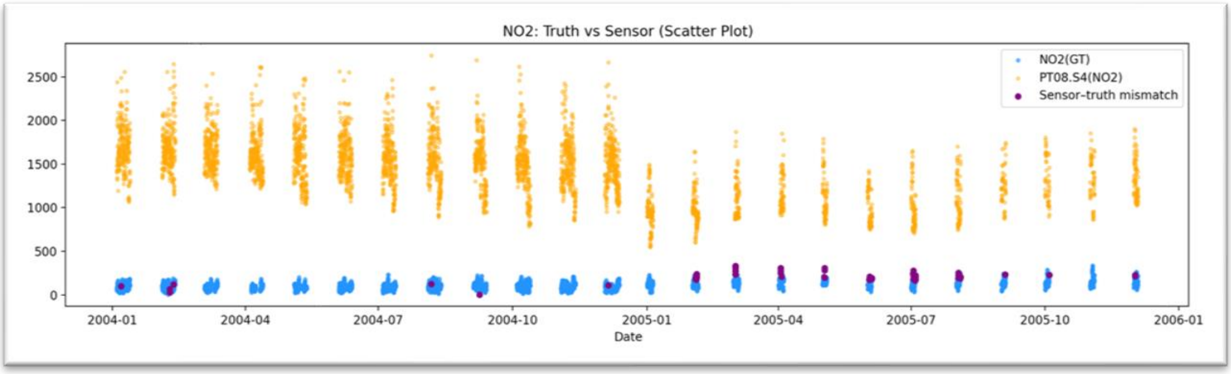

*Figure 14: Truth vs. Sensor for NO2*

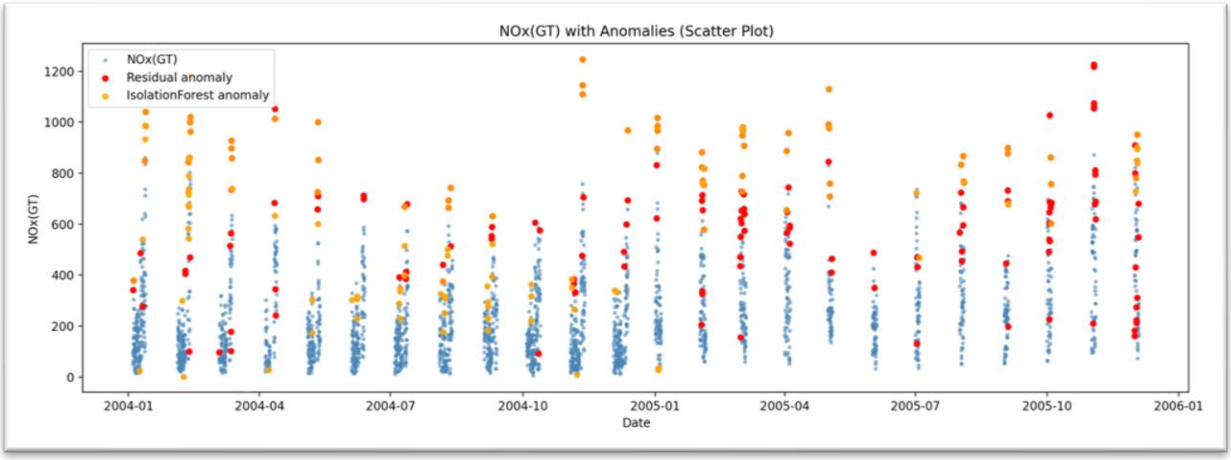

*Figure 15: Anomalies for NOx*

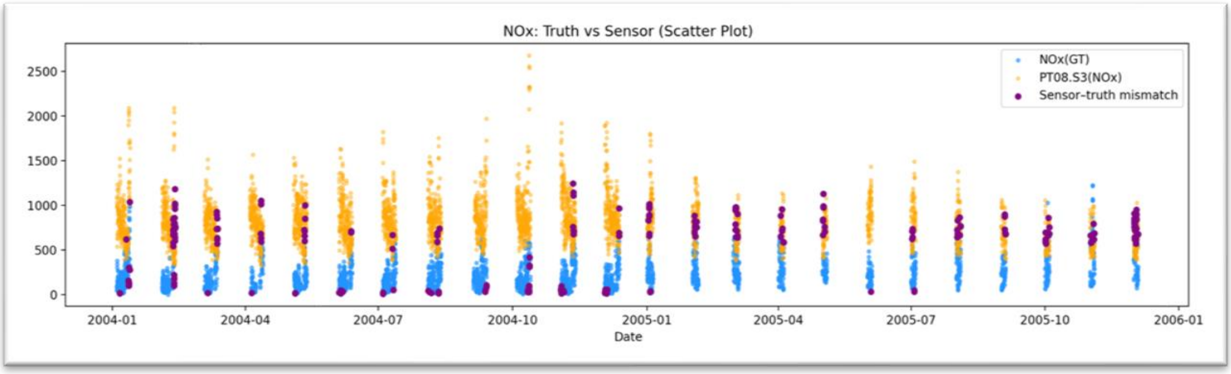

*Figure 16: Truth vs. Sensor for NOx*

# Appendix C: All Modelling Results and Plots

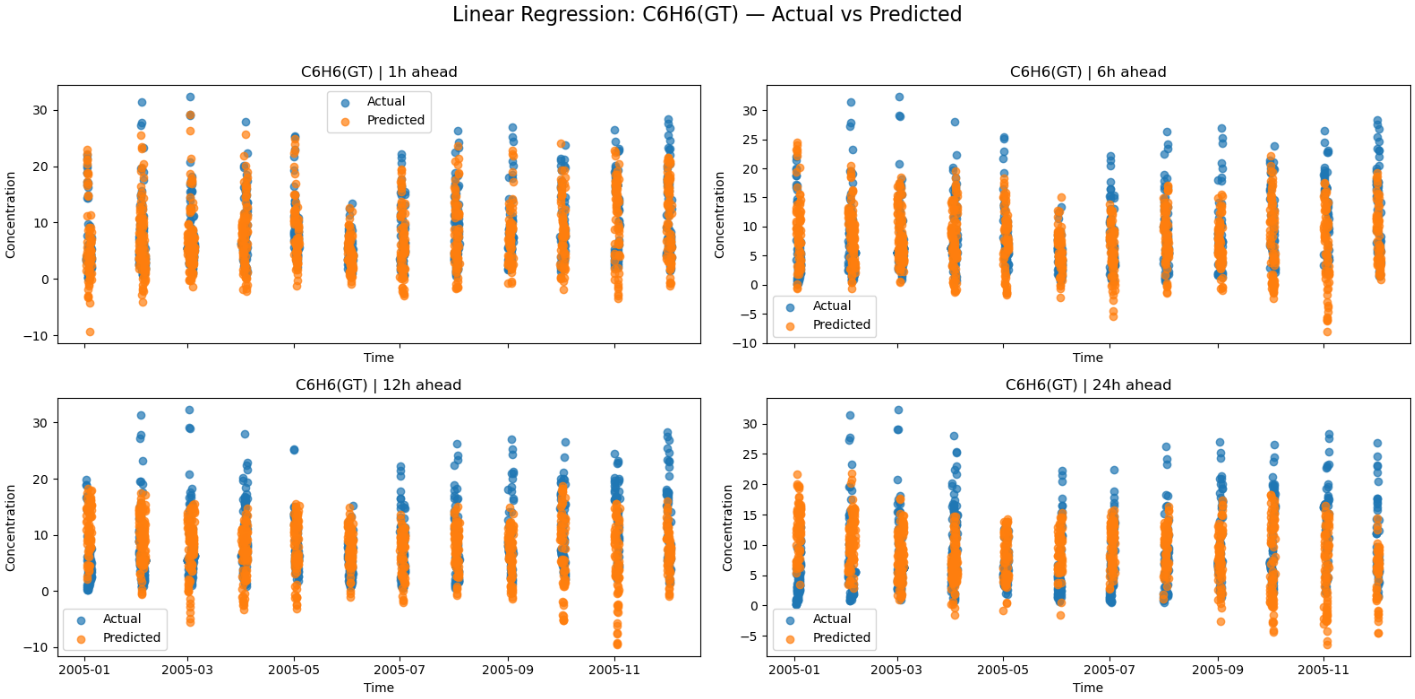

*Figure 17: Actual vs. Predicted for C6H6 (linear regression)*

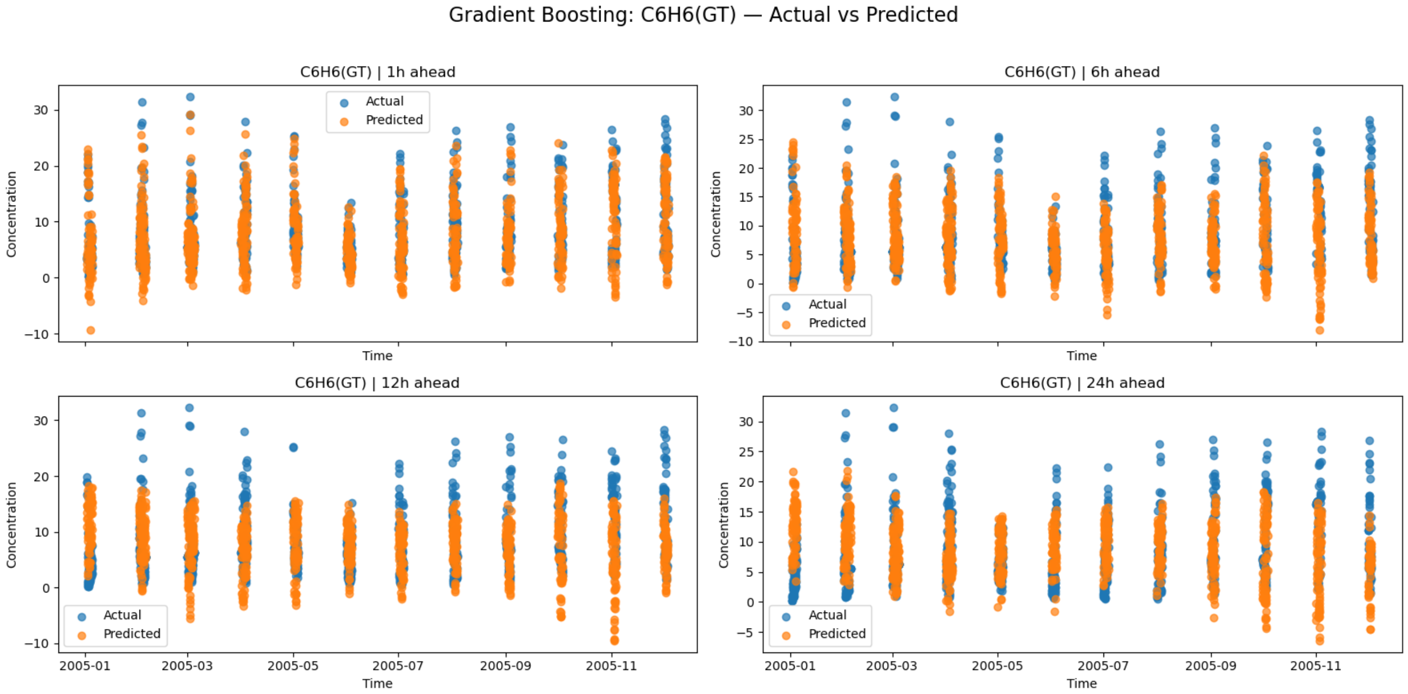

*Figure 18: Gradient Boosting C6H6 Actual Concentration v. Prediction*

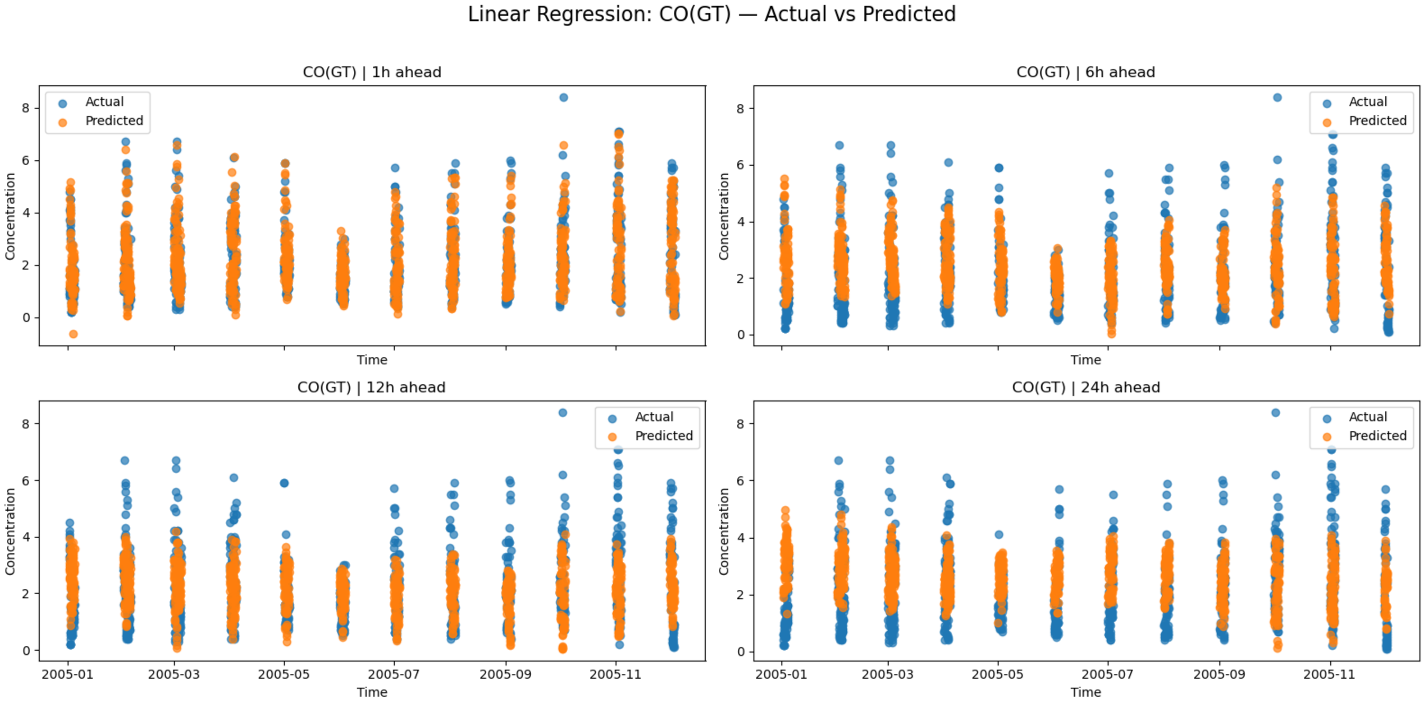

*Figure 189: Linear Regression CO(GT) Actual Concentration v. Predicted*

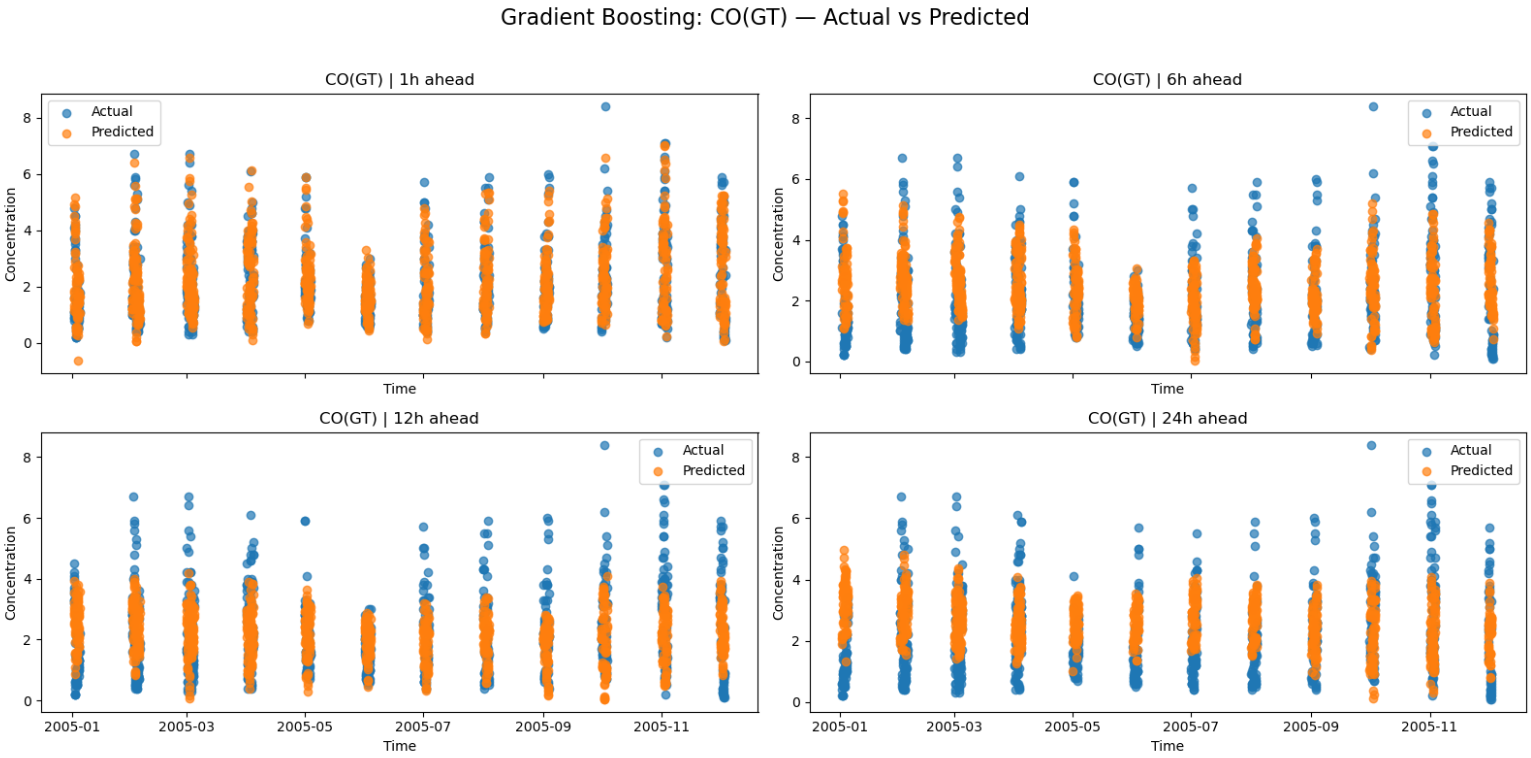

*Figure 2019: Gradient Boosting Actual vs. Predicted for CO(GT)*

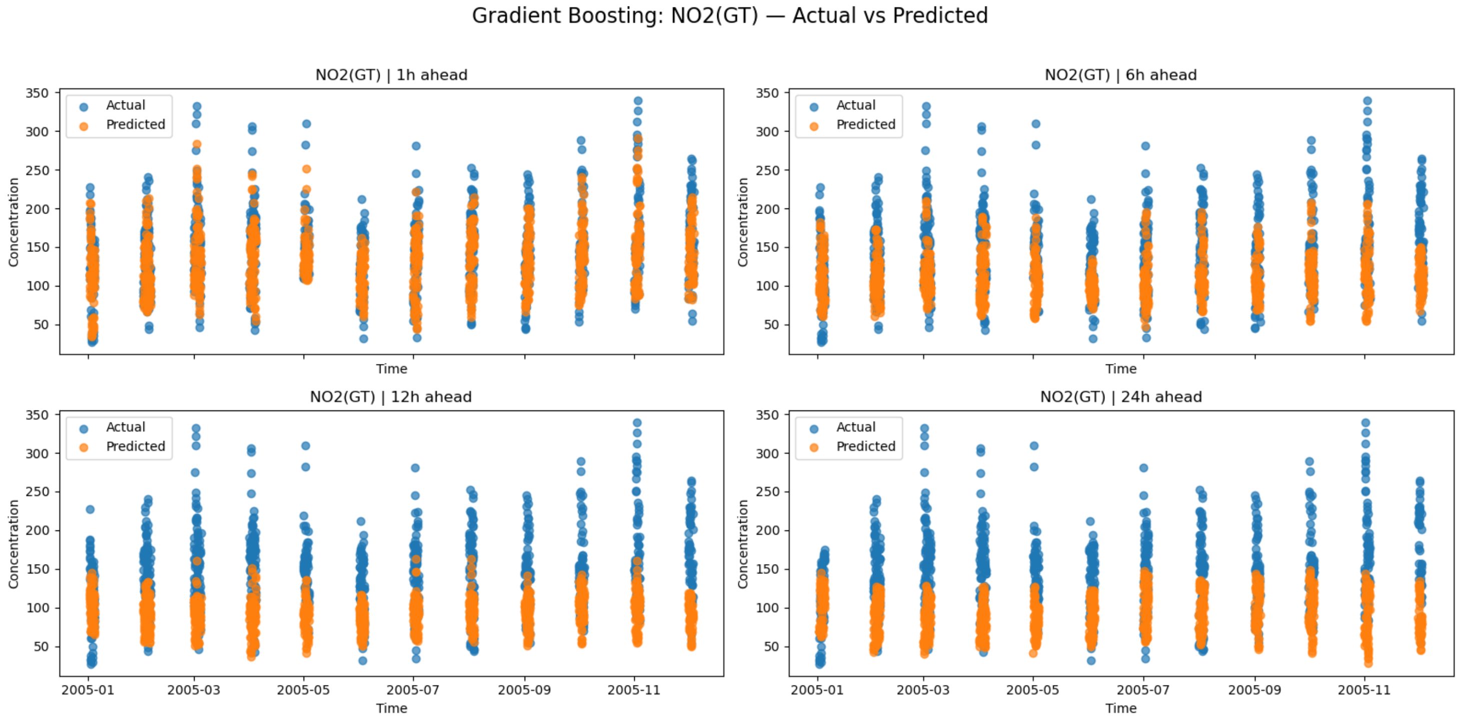

*Figure 2120: Actual vs. Predicted for N02 (Gradient Boosting)*

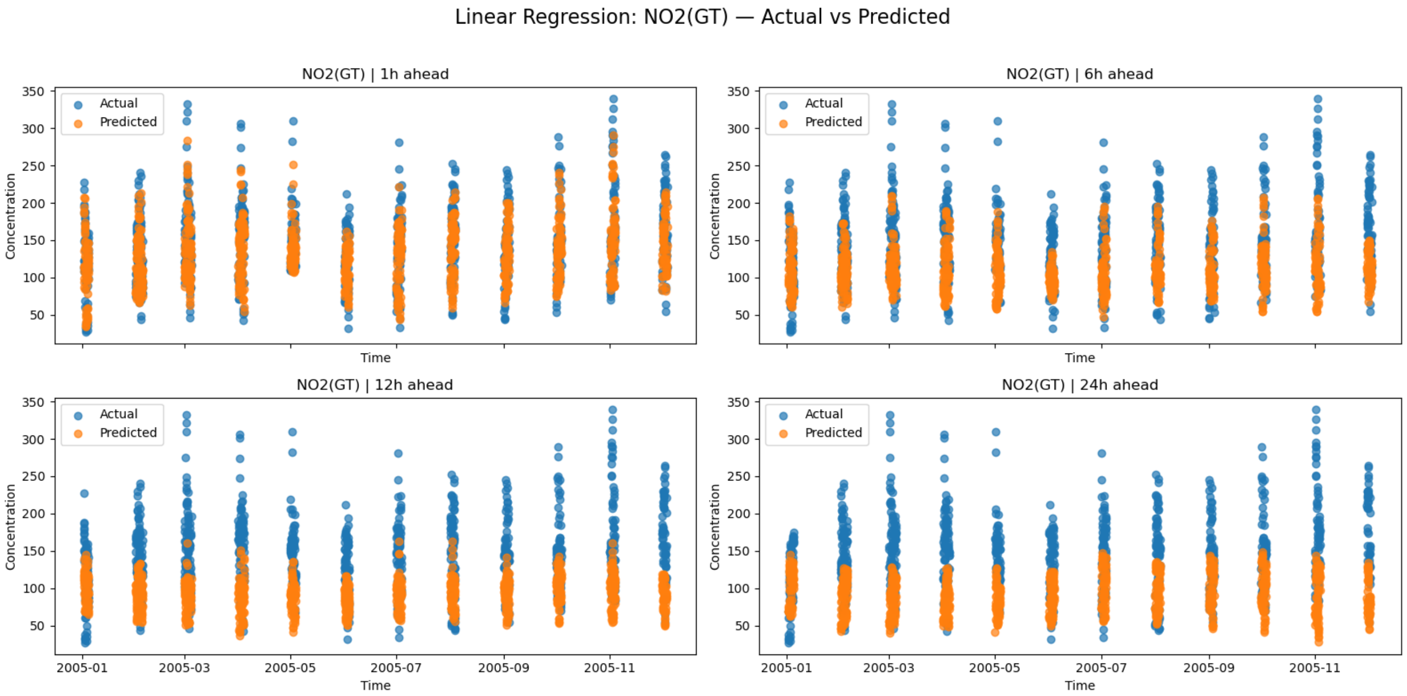

*Figure 2221: Actual vs. Predicted for NO2 (linear regression)*

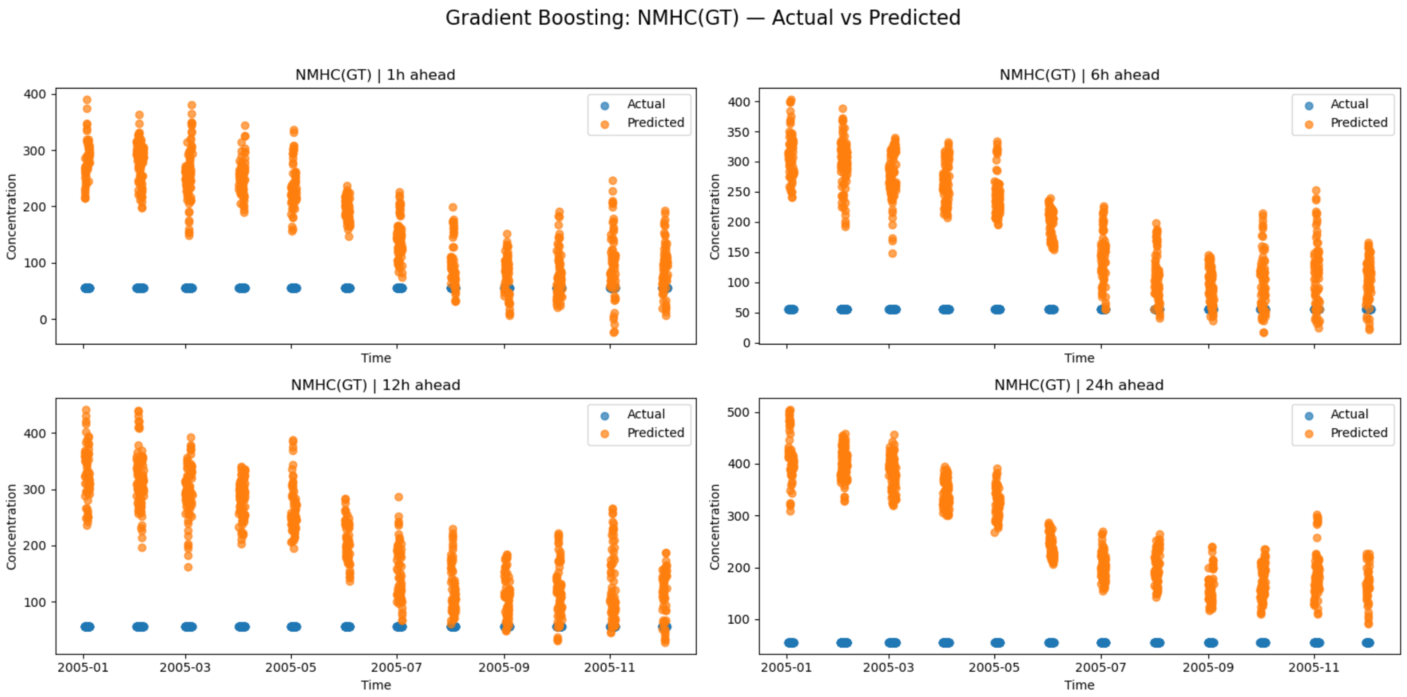

*Figure 22: Gradient Boosting NMHC(GT) Actual Concentration v. Prediction*

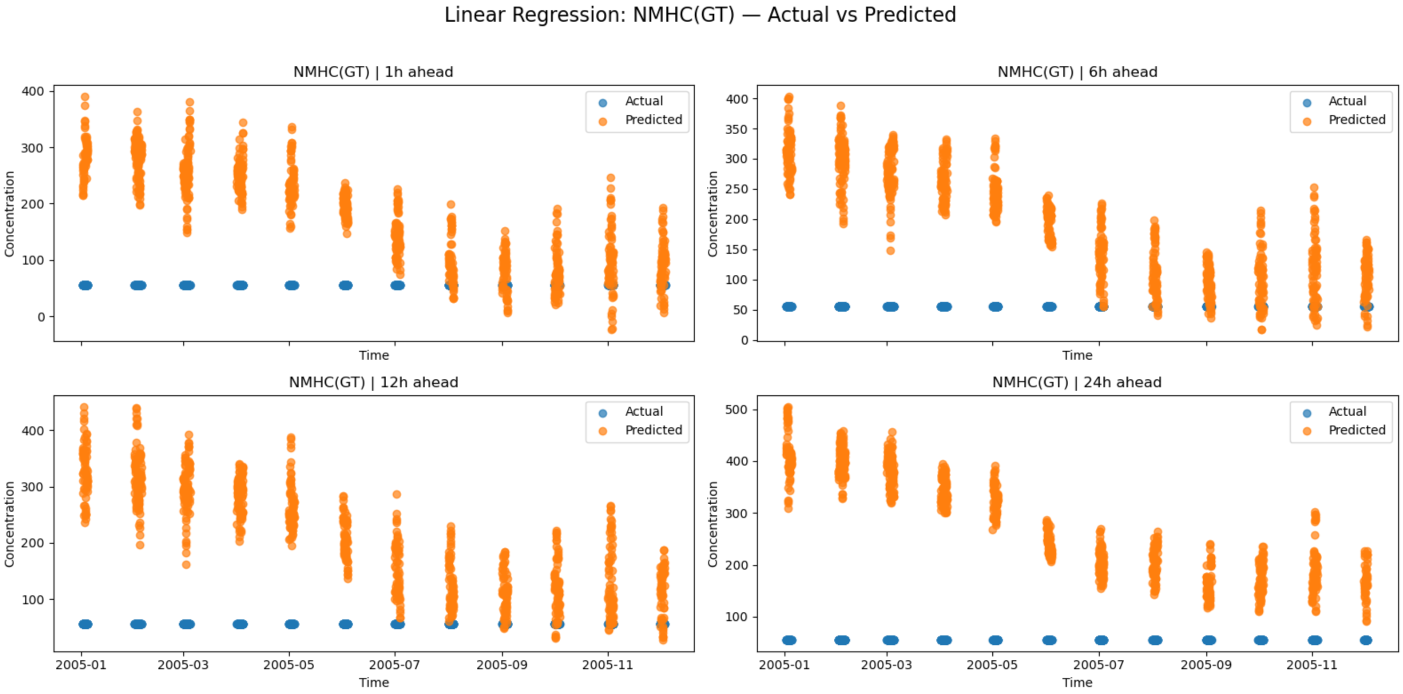

*Figure 23: Linear Regression NMHC(GT) Actual Concentration v. Prediction*

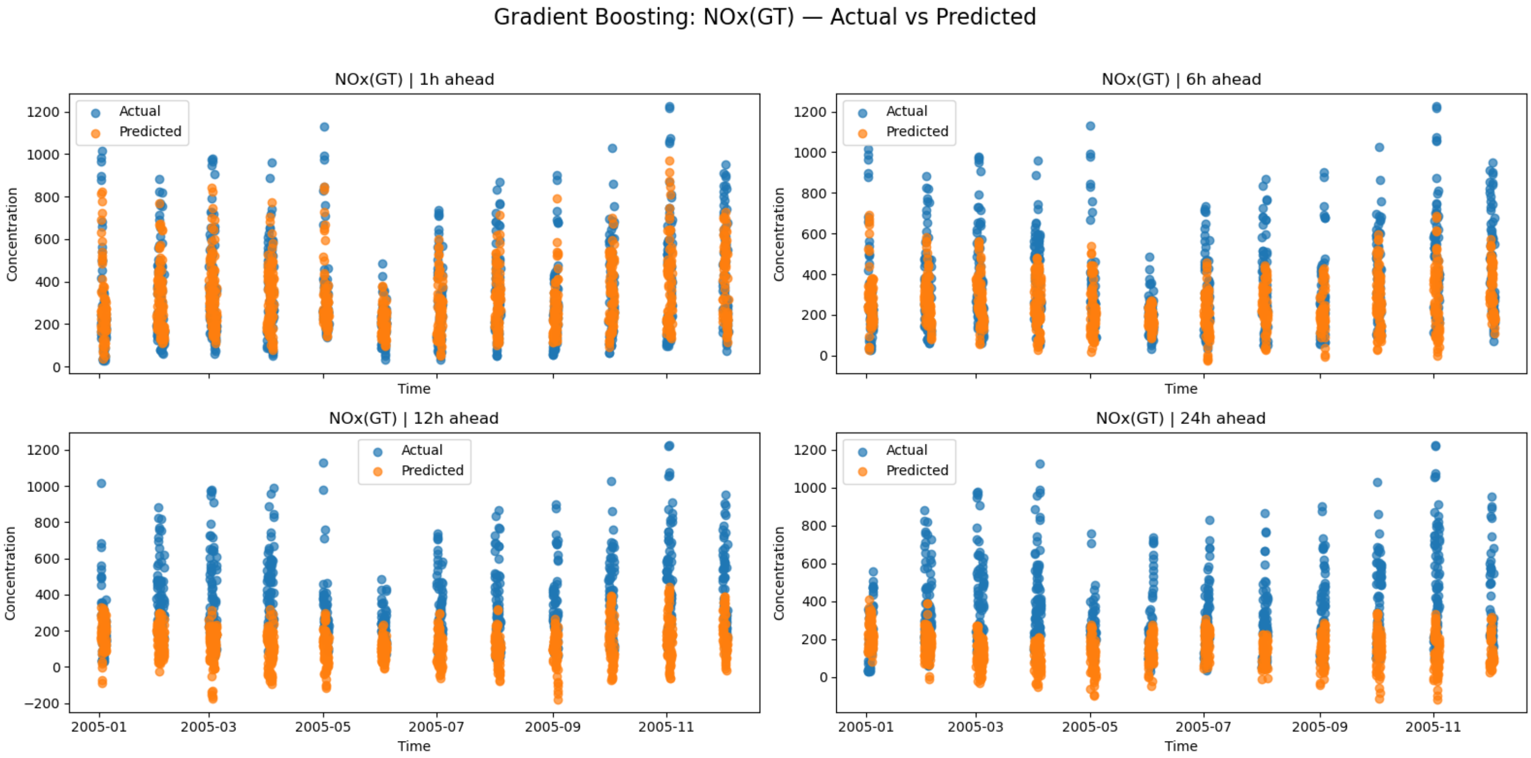

*Figure 25: Gradient Boosting NOx(GT) Actual Concentration v. Prediction*

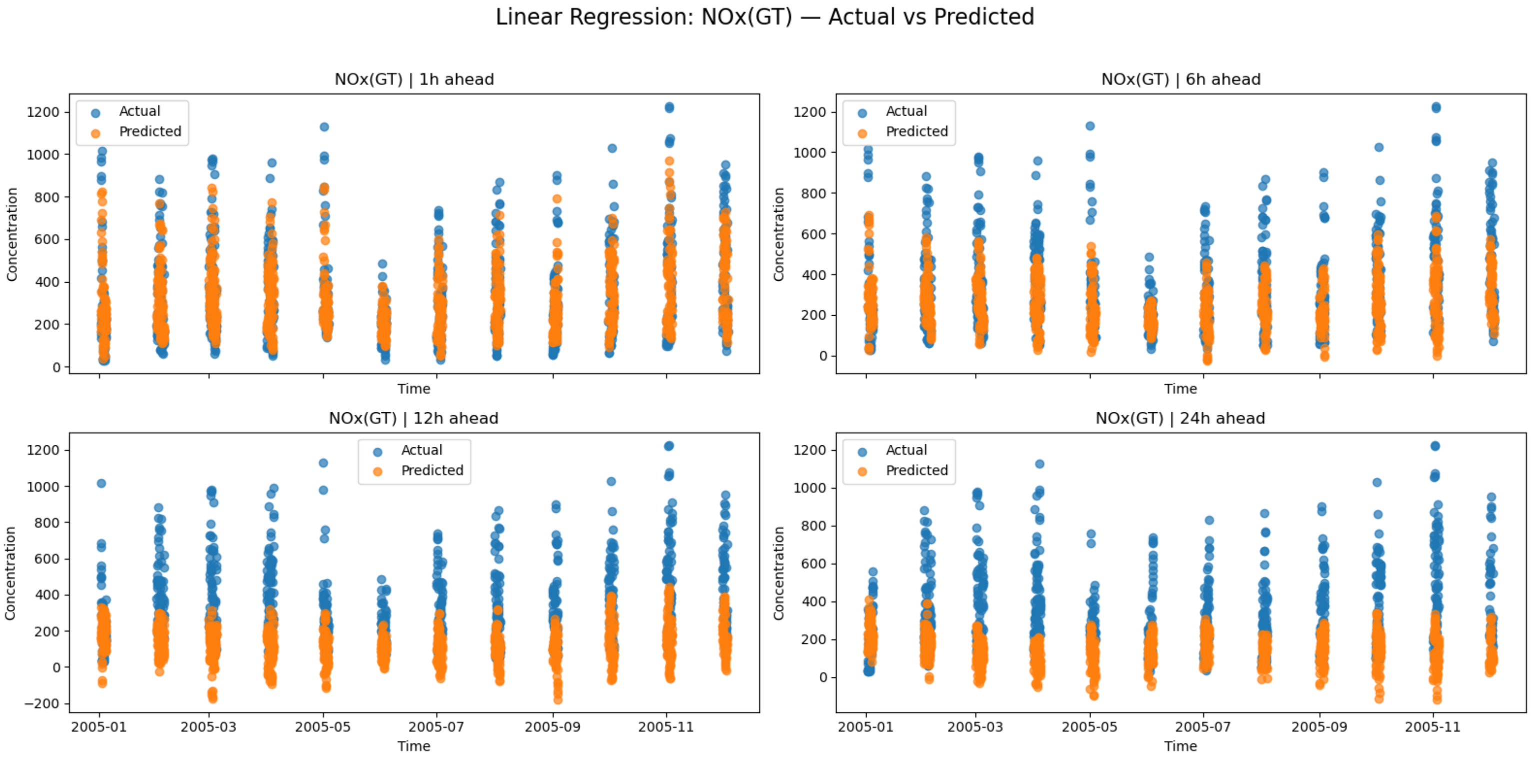

*Figure 26: Linear Regression NOx(GT) Actual Concentration v. Prediction*

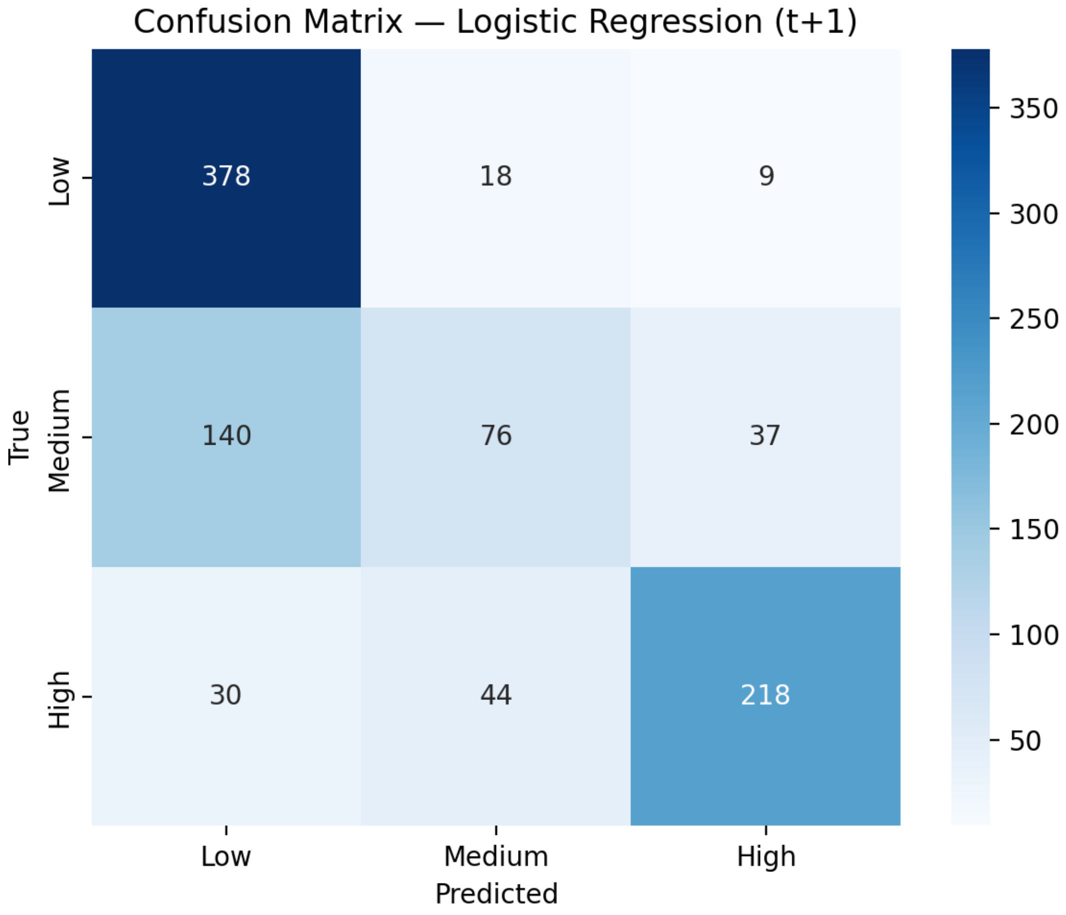

*Figure 27: Actual vs Predicted Logistic Regression t + 1*

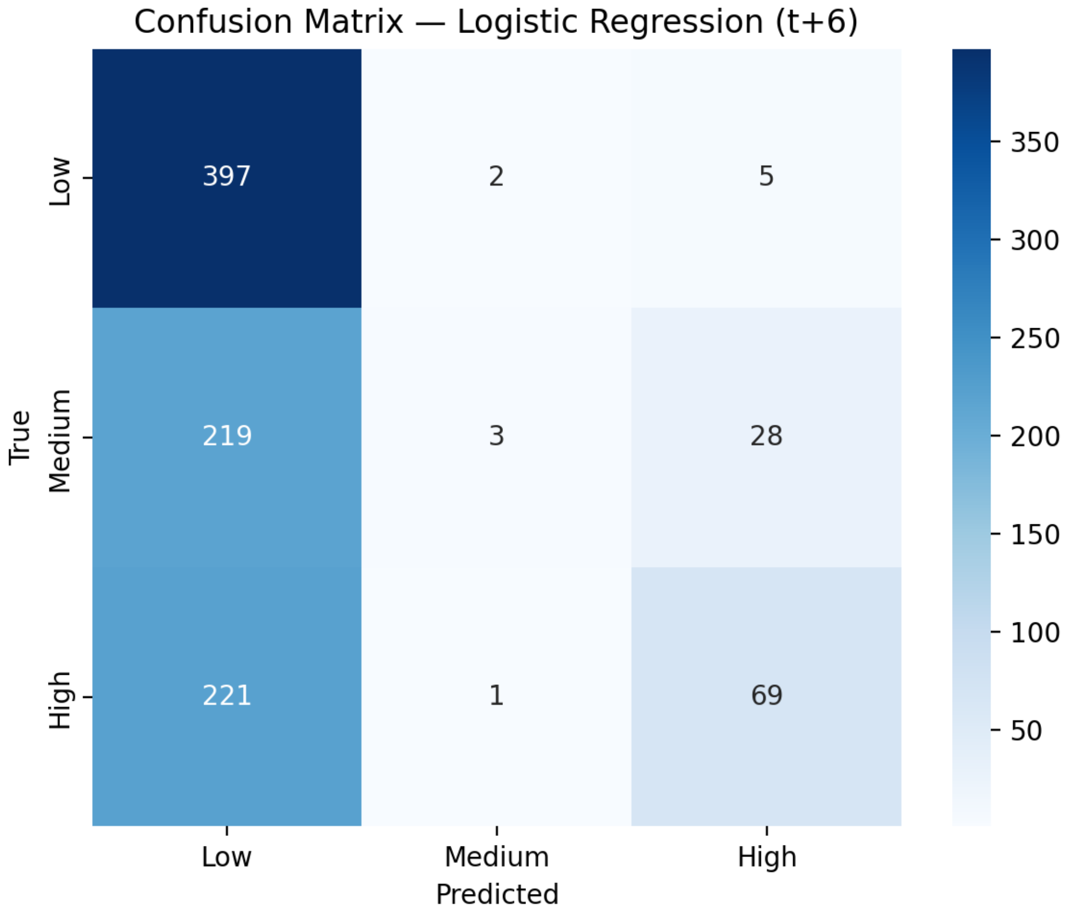

*Figure 28: Actual vs Predicted Logistic Regression t + 6*

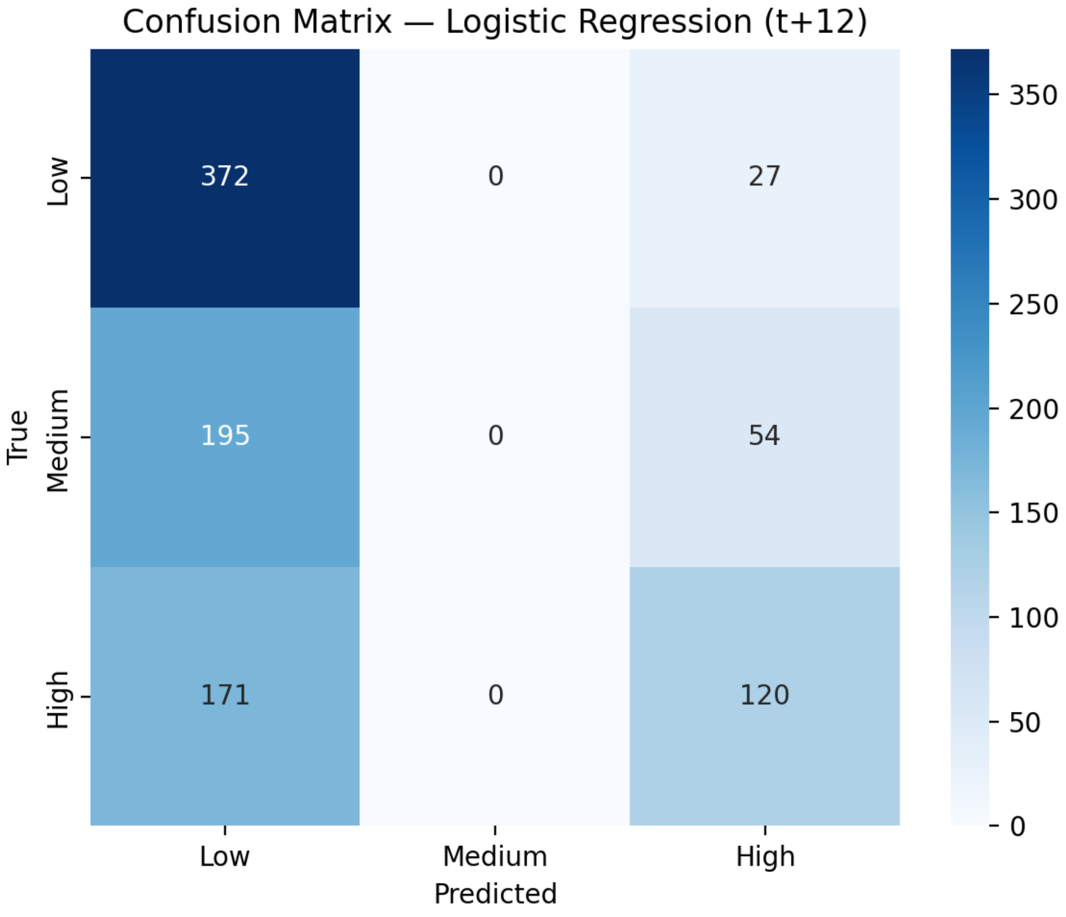

*Figure 29: Actual vs Predicted Logistic Regression t + 12*

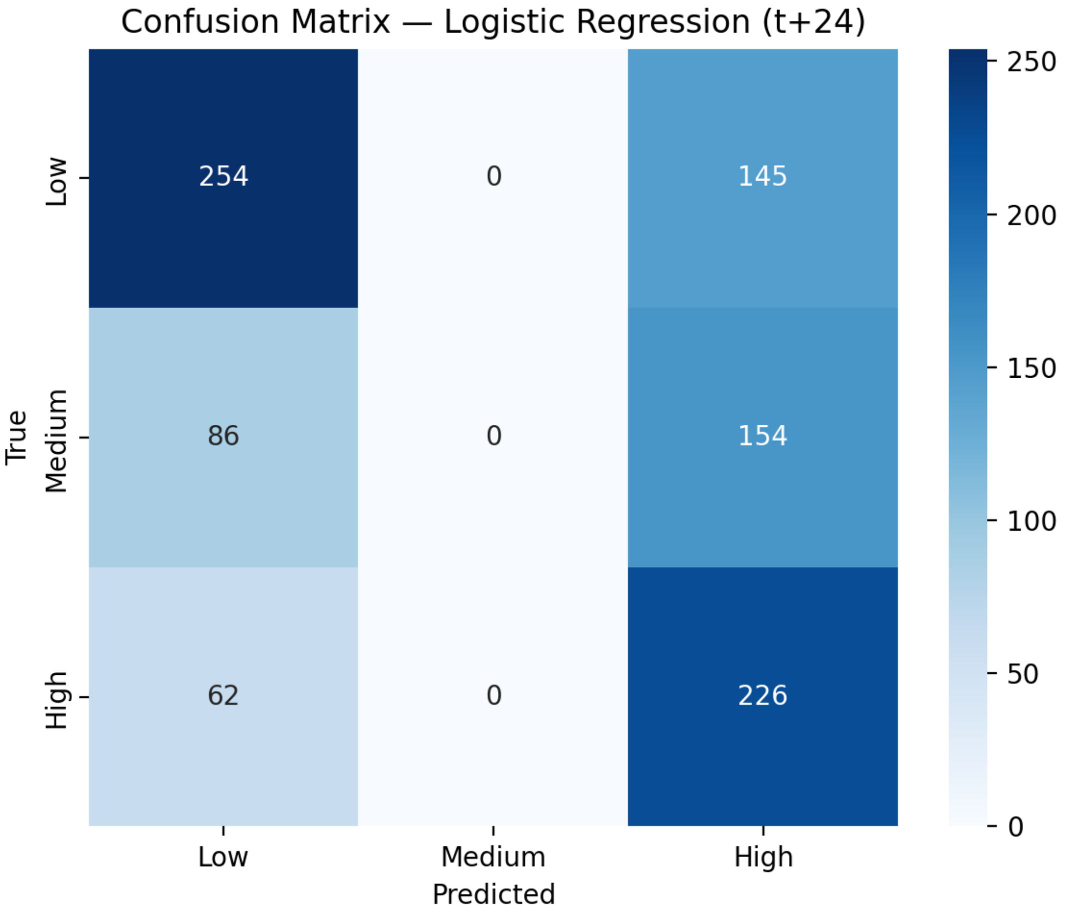

*Figure 30: Actual vs Predicted Logistic Regression t + 24*

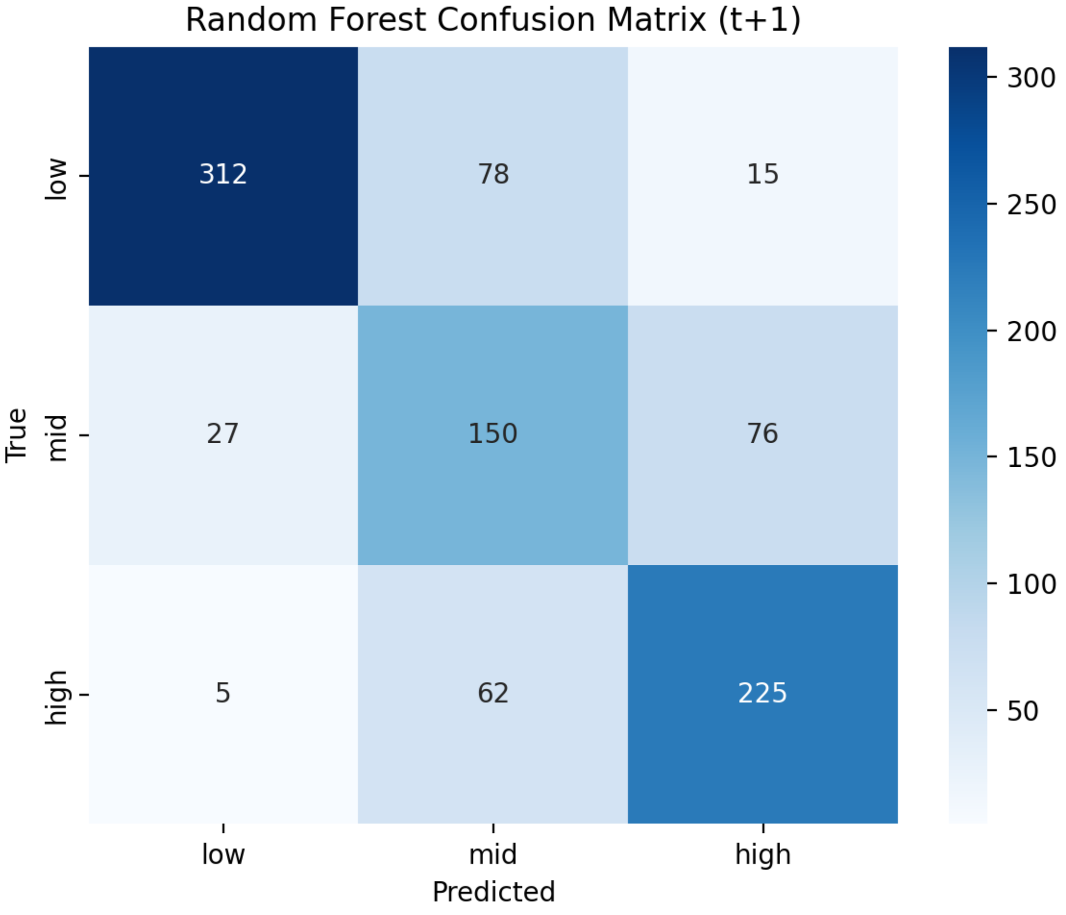

*Figure 31: Actual vs Predicted Random Forest t + 1*

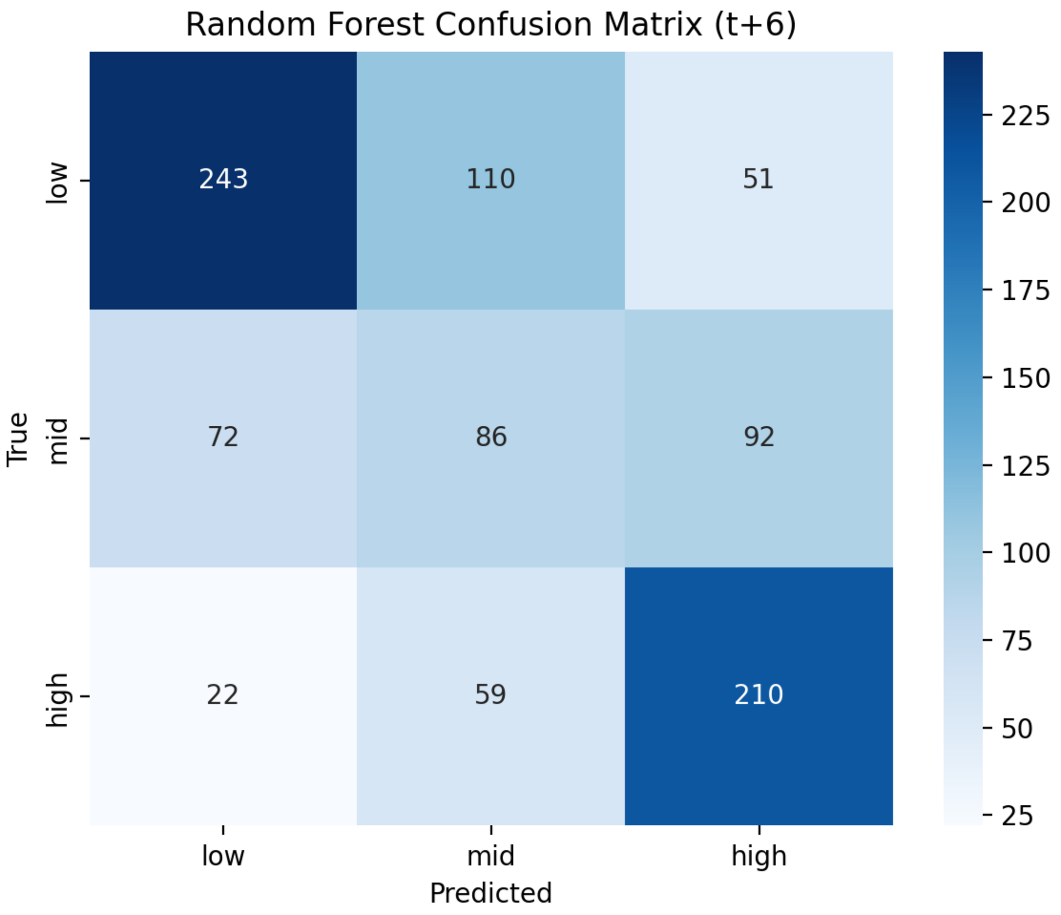

*Figure 32: Actual vs Predicted Random Forest t + 6*

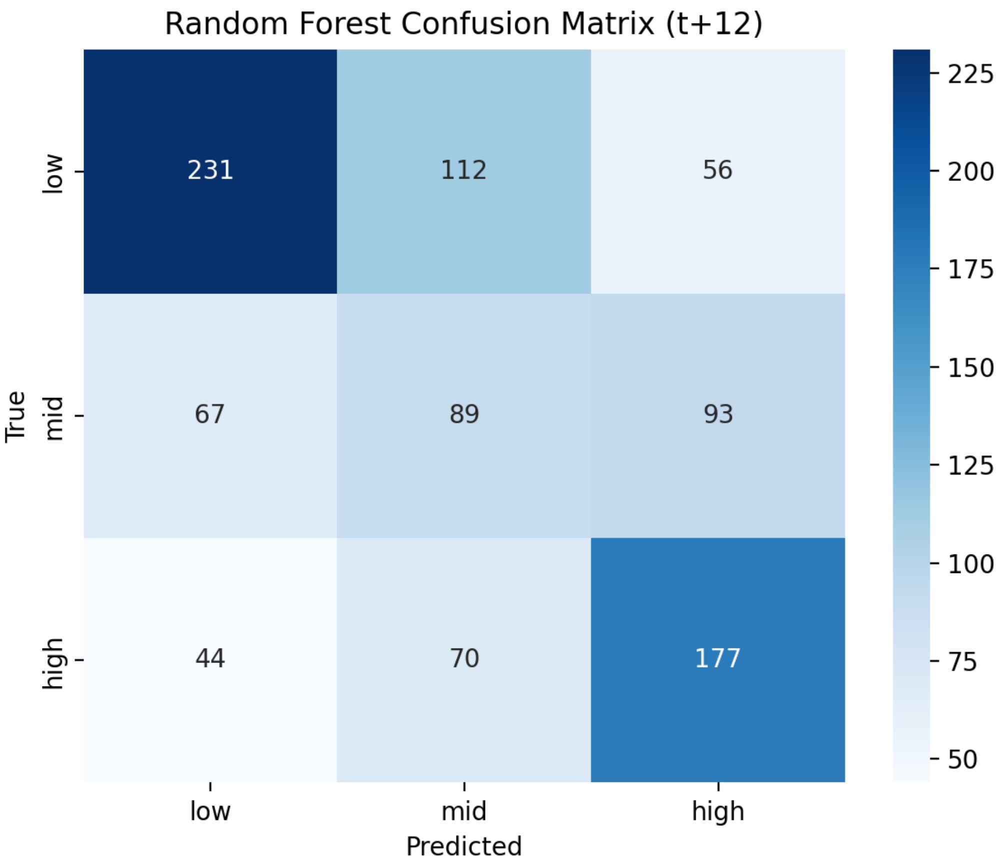

*Figure 33: Actual vs Predicted Random Forest t + 12*

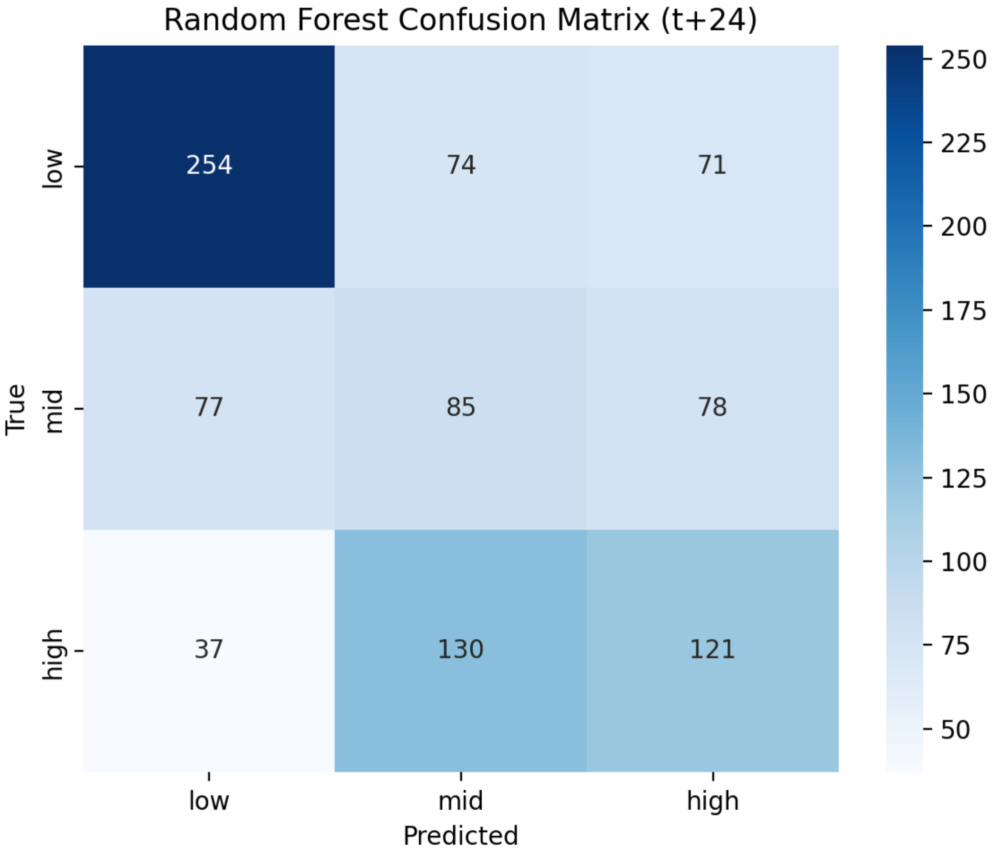

*Figure 34: Actual vs Predicted Random Forest t + 24*