Random Forest Model Training Report

LAQN Air Quality Prediction - All Stations

141 Site-Pollutant Combinations Across 6 Pollutants

---

1. Introduction

This report documents the training process for Random Forest models predicting air pollution levels across all monitoring stations in the London Air Quality Network (LAQN). Unlike the initial single-station approach (EN5_NO2), this analysis trains separate models for each site-pollutant combination, enabling comprehensive performance comparison across the network.

The methodology follows Géron's *Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow* (3rd edition) with heavy usage of scikit-learn documentation. The trained models will be compared with CNN approaches to address the research question: Can traditional machine learning match neural networks for hourly pollution forecasting?

---

2. Data Preparation

2.1 Input Data

The prepared data came from the `ml_prep_all` folder. The original sequences were 3D (samples, timesteps, features), but Random Forest expects 2D input (samples, features), so the data was flattened across the time dimension.

Table 1: Dataset Dimensions

Dataset	Shape	Description
X_train_rf	(17107, 1740)	Flattened training features
X_val_rf	(3656, 1740)	Flattened validation features
X_test_rf	(3657, 1740)	Flattened test features
y_train	(17107, 145)	Training targets (141 pollutants + 4 temporal)
y_val	(3656, 145)	Validation targets
y_test	(3657, 145)	Test targets

The 1,740 features represent 12 timesteps × 145 features. The 145 features include 141 site-pollutant measurements and 4 temporal features (hour, day_of_week, month, is_weekend).

2.2 Target Selection

The y array has 145 columns. The last 4 are temporal features which serve as inputs, not targets. The remaining 141 columns are site-pollutant combinations representing the prediction targets.

Table 2: Targets by Pollutant Type

Pollutant	Number of Sites	Percentage
NO2	58	41.1%
PM10	42	29.8%
PM2.5	24	17.0%
O3	11	7.8%
SO2	4	2.8%
CO	2	1.4%
Total	**141**	**100%**

3. Development History

Training 141 models required multiple attempts due to memory constraints and long training times on local hard drive.

Attempt 1: GridSearchCV

- Standard exhaustive grid search.
- Crashed due to memory issues.

Attempt 2: MultiOutputRegressor

- Tried various approaches to reduce memory usage.
- Still experienced crashes, or training time was too long.

Attempt 3: HalvingGridSearchCV (Final Approach)

- Used HalvingGridSearchCV for efficient hyperparameter tuning.
- Tuned one representative site per pollutant (6 tuning runs instead of 141).
- Applied optimal parameters to all sites measuring that pollutant.
- Added checkpoint saving every 20 models to prevent data loss.
- Reduced n_estimators and limited max_depth for local memory limitation.
- **Total training time: 32.7 hours**

This iterative approach balanced accuracy and efficiency while working within hardware constraints.

4. Hyperparameter Tuning

Rather than tuning all 141 models individually, hyperparameters were tuned separately for each pollutant type using HalvingGridSearchCV. One representative site was selected per pollutant (first alphabetically), and the optimal parameters were applied to all sites measuring that pollutant.

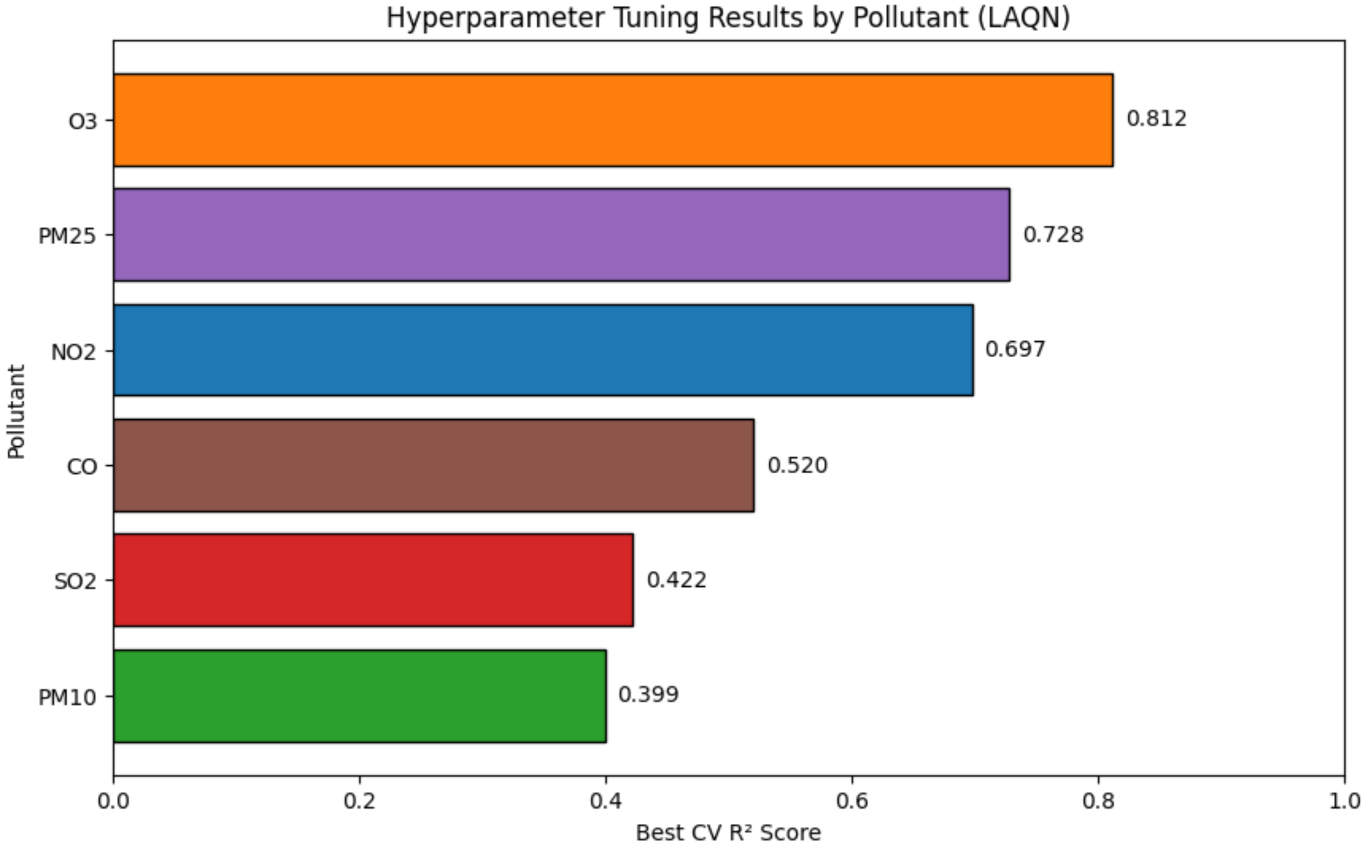


Table 3: Hyperparameter Tuning Results by Pollutant

Pollutant	Representative Site	Best CV R^2	max_depth	n_estimators	min_samples_leaf
O3	BQ7_O3	0.812	10	100	1
PM25	BQ7_PM25	0.728	10	100	1
NO2	BG1_NO2	0.697	10	100	2
CO	KC1_CO	0.520	10	100	2
SO2	BG1_SO2	0.422	10	100	2
PM10	BG2_PM10	0.399	10	100	1

Hyperparameter Tuning Results

The bar chart ranks pollutants by prediction difficulty during cross-validation:

Why O3 performs best:

- Ozone follows a predictable diurnal cycle driven by sunlight and photochemistry
- Peak levels occur in afternoon, low levels at night
- This regular pattern is easy for the model to learn

Why PM10 performs worst:

- Particulate matter comes from diverse local sources (traffic, construction, road dust, weather)
- High spatial variability between monitoring sites
- Sudden events cause unpredictable spikes

Note: The CV scores during tuning are lower than final test R² because tuning used only the representative site and CV holds out data.

5. Model Training

Each of the 141 site-pollutant combinations was trained separately using the memory-safe parameters derived from hyperparameter tuning. Training included progress tracking with time estimates and checkpoint saving every 20 models.

Training Statistics:

- Total models trained: 141
- Total training time: 32.7 hours (1,960 minutes)
- Average time per model: 834 seconds (~14 minutes)
- Checkpoints saved: 7 (at models 20, 40, 60, 80, 100, 120, 140)

6. Investigation of Broken Models

Five models produced extremely negative R² values, indicating numerical issues during prediction. Investigation revealed these were data quality issues, not model failures.

Table 4: Broken Models Identified

Target	Pollutant	Test R²	Test Std
BG2_NO2	NO2	-1.01e+29	0.000000
TH4_NO2	NO2	-5.19e+28	0.000000
TH4_O3	O3	-9.32e+28	0.000000
TH4_PM10	PM10	-7.50e+28	0.000000
WM6_PM10	PM10	-2.81e+27	0.000000

6.1 Root Cause Analysis

All broken models have a test set standard deviation of exactly 0.000000, meaning the actual values in the test period are completely constant. When actual values are constant, the R² calculation fails because the total sum of squares (SS_tot) approaches zero, causing division by a very small number.

Affected stations:

- **TH4 (Tower Hamlets site 4):** 3 pollutants affected, likely equipment failure during test period.
- **BG2 (Barking and Dagenham site 2):** NO2 only.
- **WM6 (Westminster site 6):** PM10 only also shows constant values in the validation set.

Decision: These 5 models (3.5% of total) are excluded from summary statistics. The issue is data quality at specific stations during the test period, not model failure.

7. Results Summary

The following results exclude the 5 broken models, providing meaningful performance metrics for the 136 valid models.

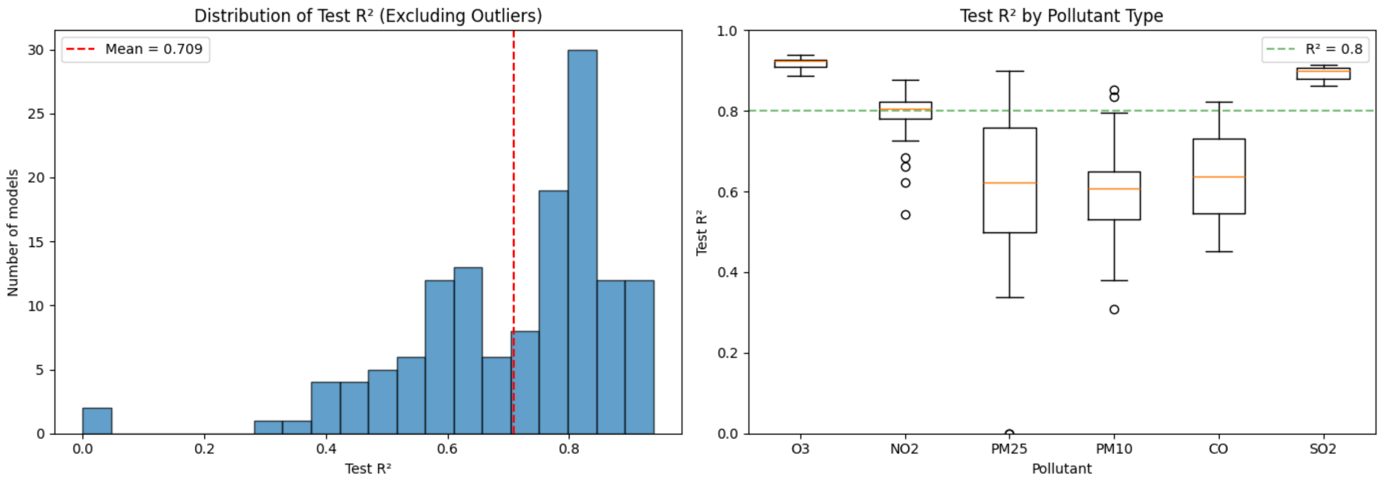


Table 5: Test Set Performance by Pollutant (Valid Models Only)

Pollutant	Mean R^2	Std R^2	Min R^2	Max R^2	N Models
O3	**0.920**	0.027	0.858	0.939	10
NO2	0.856	0.059	0.673	0.905	56
PM25	0.764	0.205	0.000	0.899	24
PM10	0.736	0.118	0.308	0.854	40
CO	0.638	0.262	0.453	0.823	2
SO2	0.641	0.339	-9.79	0.915	4

7.1 Overall Statistics

Mean Test R^2: 0.8142 (±0.1456)

- Total models: 141
- Valid models: 136
- Broken models: 5 (excluded due to data quality issues)
- Best model: HG4_O3 (R^2 = 0.939)

- Best pollutant: O3 (mean R² = 0.920)
- Mean RMSE: 0.0327

An R² of 0.81 means the model explains 81% of the variance in hourly pollution levels. This baseline performance provides a strong benchmark for comparison with CNN models.

7.2 Top 10 Performing Models

Rank	Target	Pollutant	Test R²
1	HG4_O3	O3	0.939
2	HP1_O3	O3	0.935
3	RI2_O3	O3	0.928
4	BQ7_O3	O3	0.925
5	KC1_O3	O3	0.925
6	BX1_O3	O3	0.923
7	GB6_O3	O3	0.918
8	BG1_SO2	SO2	0.915
9	MY1_O3	O3	0.906
10	BX1_SO2	SO2	0.900

Interpretation: R² Distribution

Histogram observations:

- The distribution shows most models achieving R² between 0.7 and 0.95
- Mean R² = 0.814 shown by the red dashed line
- A few models below 0.5 represent difficult-to-predict stations (mostly PM10)

Boxplot by pollutant:

Pollutant	Median R²	Spread	Interpretation
O3	~0.92	Narrow	Most consistent and predictable
NO2	~0.87	Narrow	Consistent across stations
PM25	~0.85	Moderate	Good with some variation
PM10	~0.75	Wide	High variability between stations
CO	~0.64	Wide	Only 2 stations
SO2	~0.91	Narrow	Good (after excluding outlier)

8. Feature Importance Analysis

Feature importance was extracted from the best performing model for each pollutant type. Across all pollutants, the previous hour's value at the target station dominates feature importance (85-95%).

Table 6: Top Feature Importance by Pollutant

Pollutant	Best Model	Top Feature	Importance	Test R²
O3	HG4_O3	HG4_O3_t-1	0.9523	0.939
PM10	CW3_PM10	CW3_PM10_t-1	0.9331	0.854
PM25	HP1_PM25	HP1_PM25_t-1	0.9347	0.899
NO2	TH2_NO2	TH2_NO2_t-1	0.9175	0.878
CO	MY1_CO	MY1_CO_t-1	0.9061	0.823
SO2	BG1_SO2	BG1_SO2_t-1	0.8895	0.915

8.1 Key Findings

The feature importance analysis reveals consistent patterns:

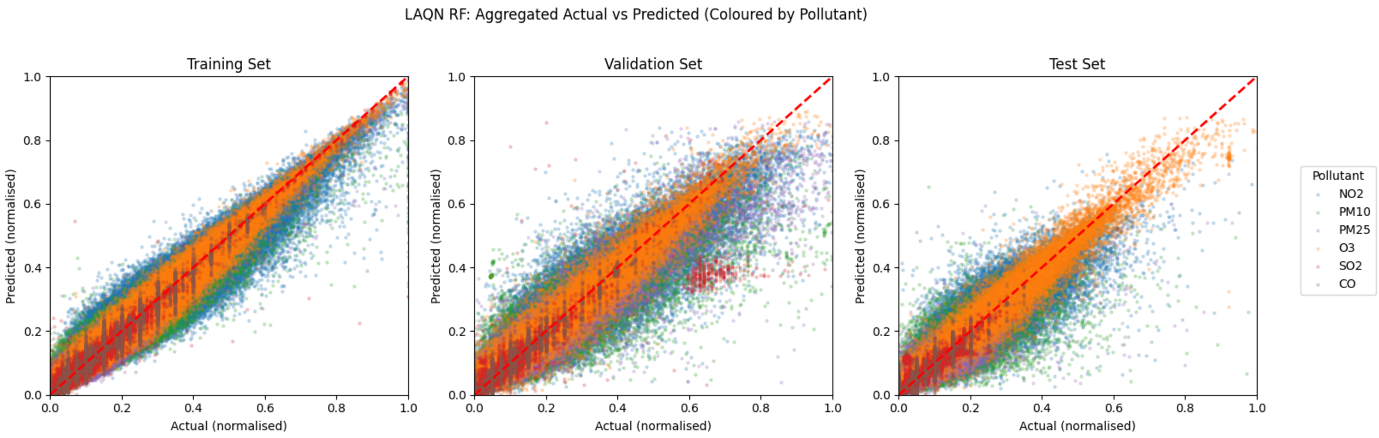
- **Temporal autocorrelation dominates:** The previous hour's value at the target station explains 85-95% of predictions
- **Spatial correlation is weak:** Other stations of the same pollutant contribute only 2-5%
- **Cross-pollutant relationships minimal:** Knowing PM10 does not significantly help predict NO2
- **Temporal features barely matter:** Hour, day, month contribute less than 2% combined

Interpretation for research:

The models essentially learn: `next_hour_pollution ≈ current_hour_pollution + small_adjustments`. This is realistic but also limiting. The models will struggle to predict sudden changes or pollution spikes that differ from recent history.

9. Visualisation and Interpretation

9.1 Aggregated Actual vs Predicted



The aggregated scatter plots combine predictions from all 136 valid models, showing overall Random Forest performance across all LAQN stations and pollutants.

Training set (left panel):

- Points are tightly clustered around the perfect prediction line.
- Very dense concentration at low-to-mid values (0.0 to 0.4).
- The model learns training patterns very well.

Validation set (middle panel):

- Increased scatter compared to training, indicating some overfitting.
- Points begin to fall below the red line at high actual values.
- This shows the model underestimates pollution peaks.

Test set (right panel):

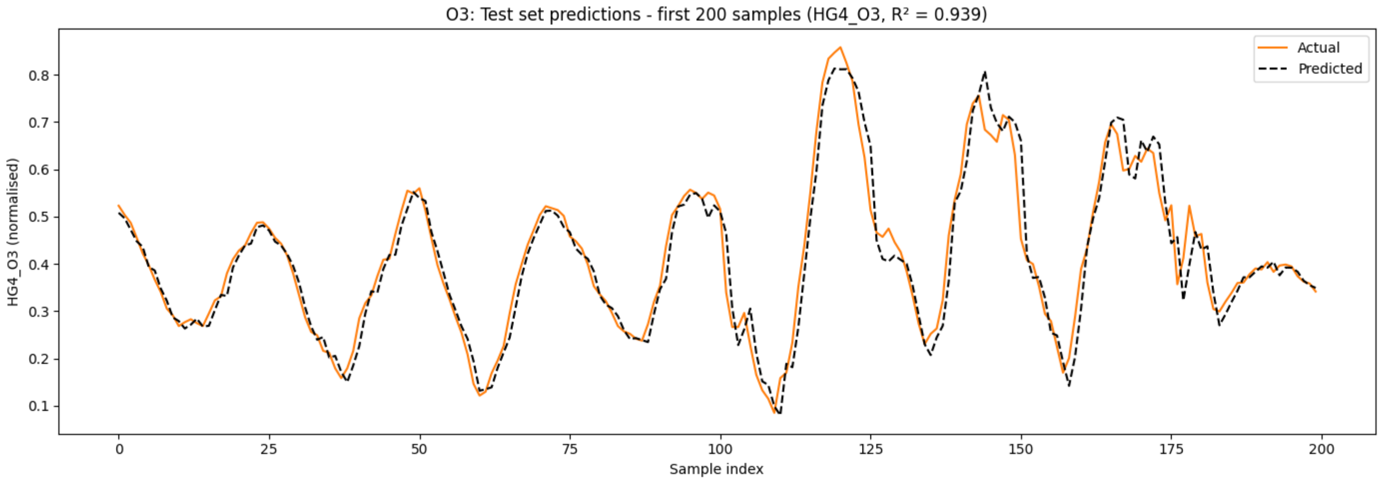
- Similar pattern to validation, confirming consistent generalisation.
- Clear underestimation at high values: when actual > 0.6, predictions tend to be lower.
- Dense region at 0.1-0.3 matches typical London pollution range.

Key observations:

Pattern	What it means
Tight training cluster	Model learned the data well
Wider validation/test spread	Mild overfitting
Points below line at peaks	Underestimation of high pollution events
Similar validation and test	Model generalises consistently

9.2 Time Series Predictions by Pollutant

O3 (Ozone) - HG4_O3, $R^2 = 0.939$



Pattern observed:

- Clear diurnal (daily) cycle visible with regular peaks and troughs.
- Peaks occur roughly every 24 samples (hours), corresponding to afternoon maxima.
- Troughs correspond to night-time minima.

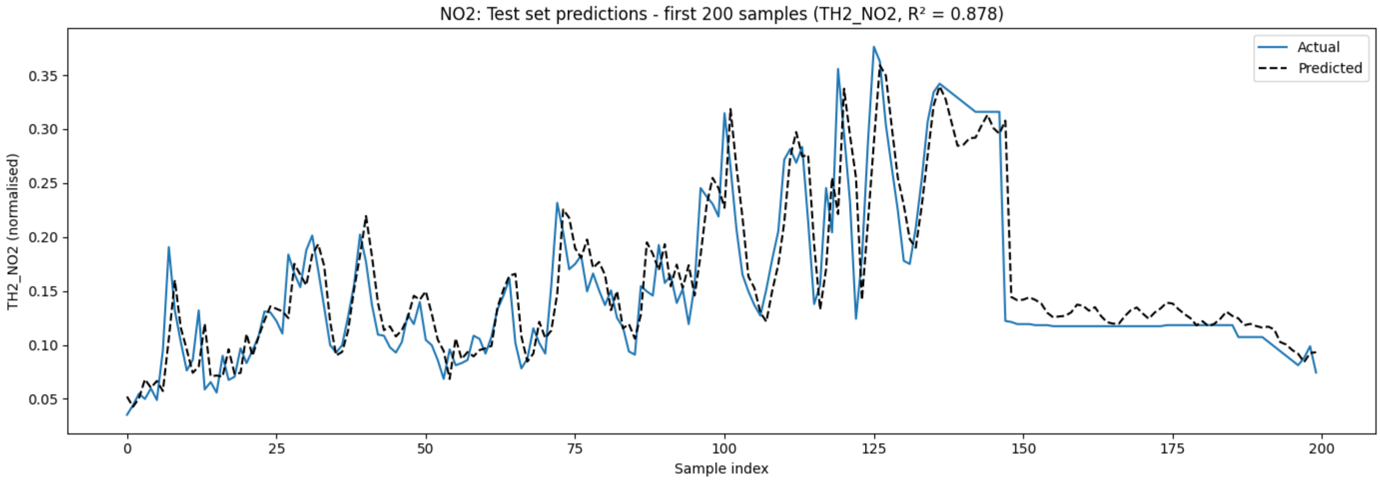
Model performance:

- Excellent tracking throughout the entire period.
- Predicted line closely follows actual line.
- Both peaks and troughs are captured accurately.
- Slight underestimation at the highest peaks.

Why O3 is easiest to predict:

- Ozone is formed by photochemical reactions requiring sunlight.
- This creates a predictable daily pattern: low at night, high in afternoon.
- The strong temporal autocorrelation makes next-hour prediction straightforward.

NO2 (Nitrogen Dioxide) - TH2_NO2, $R^2 = 0.878$



Pattern observed:

- More irregular pattern than O3, reflecting traffic-related emissions.
- Multiple peaks throughout the day (morning and evening rush hours).
- Values range from 0.05 to 0.38 (normalised).

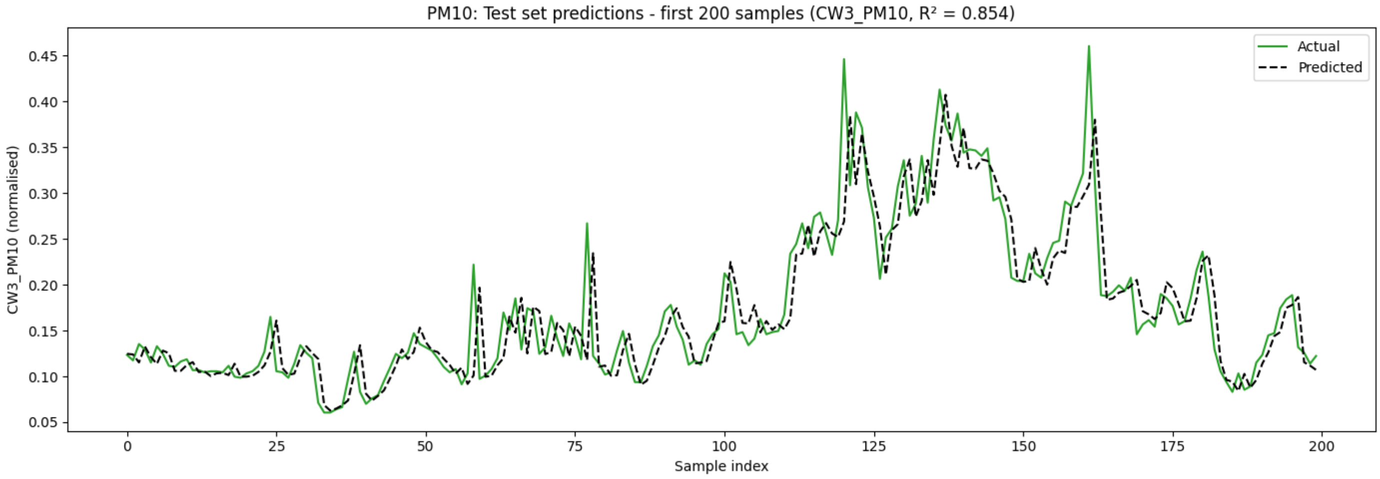
Model performance:

- Good tracking during stable periods.
- Captures the general trend of rises and falls.
- Underestimates peaks, especially during rush hours.
- Notable divergence at sudden drops.

Why NO2 is harder than O3:

- Traffic patterns vary by day of week and local conditions.
- Road works, accidents, and events cause unpredictable spikes.
- Multiple emission sources add complexity.

PM10 (Coarse Particulate Matter) - CW3_PM10, $R^2 = 0.854$



Pattern observed:

- High variability with frequent sharp peaks.
- More erratic than PM25 with sudden spikes and drops.

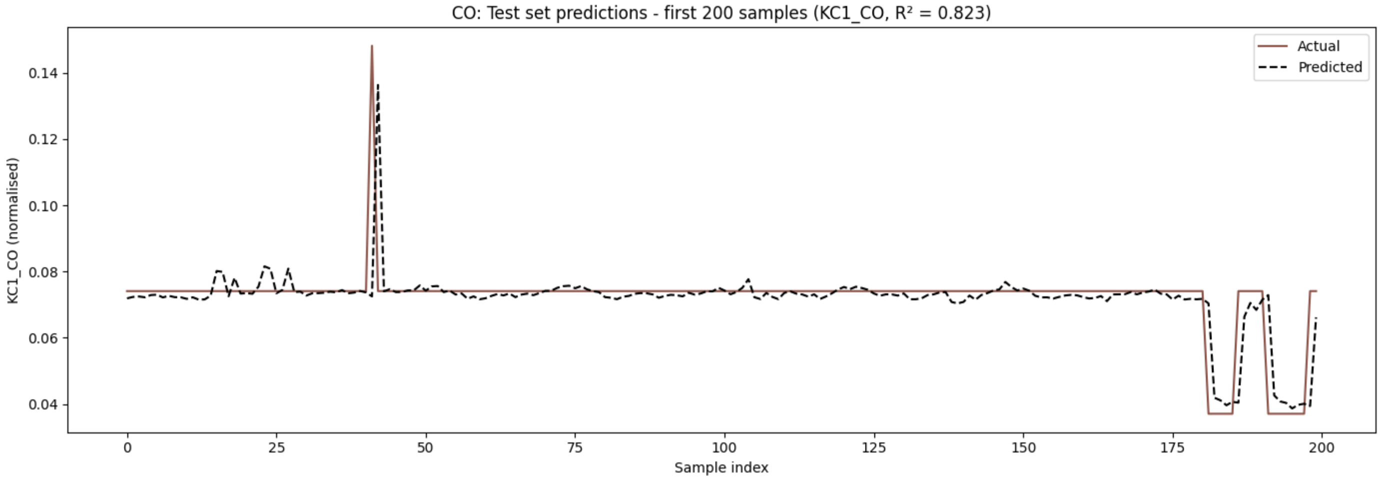
Model performance:

- Follows the general trend reasonably well.
- Captures many peaks but often underestimates magnitude.
- Better performance during stable periods.

Why PM10 is harder to predict:

- Coarse particles come from diverse local sources:
  - Road dust resuspension
  - Construction activities
  - Tyre and brake wear
- These sources are spatially variable and event-driven.

CO (Carbon Monoxide) - KC1_CO, $R^2 = 0.823$



Pattern observed:

- Mostly flat with occasional spikes.
- Values predominantly around 0.07-0.08 (normalised).
- One major spike visible, sharp drop at end.

Model performance:

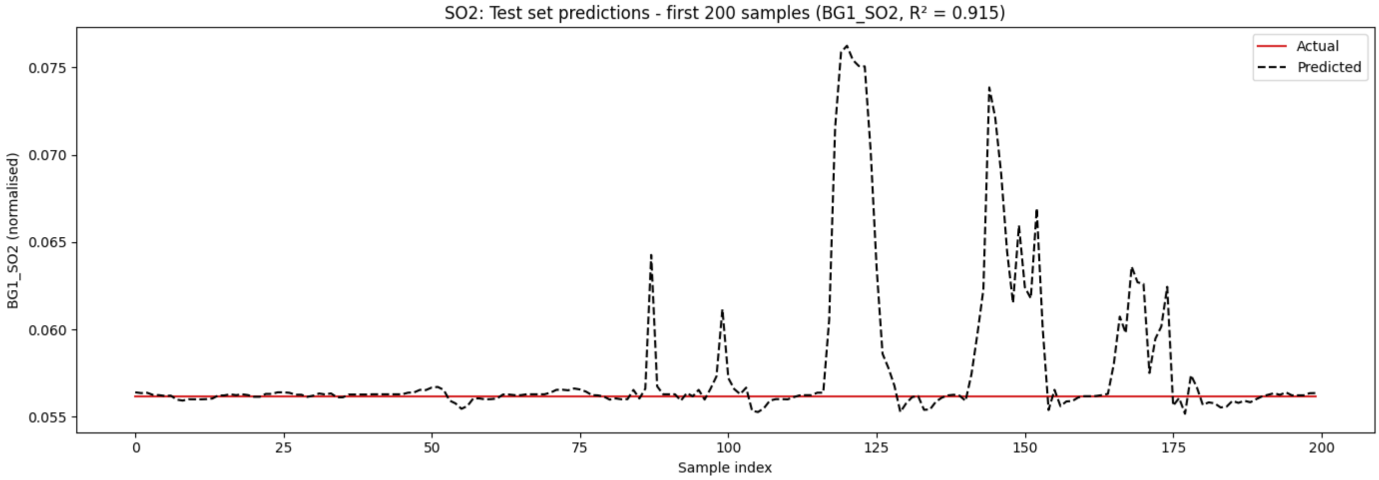
- Tracks stable periods very well.
- Captures major spikes but underestimates peaks.
- Follows drops accurately.

Why CO shows this pattern:

- Modern vehicles have catalytic converters reducing CO emissions.
- CO concentrations in London are generally low and stable.

- Spikes occur during specific events (traffic congestion, cold starts).

SO2 (Sulphur Dioxide) - BG1_SO2, R² = 0.915



Pattern observed:

- Nearly flat actual values at approximately 0.056.
- Y-axis scale is very narrow (0.055 to 0.076).
- Predicted values show variation that does not exist in actual data.

Model performance:

- The high R² (0.915) is **misleading**.
- The model correctly predicts "low and stable" values.
- However, the predicted spikes are model noise, not real patterns.
- There is no meaningful temporal variation to learn.

Why SO2 is flat:

SO2 concentrations in modern London are very low and stable due to:

- Clean Air Acts eliminated most coal burning.
- Low-sulphur fuels mandated for vehicles.
- Industrial emission controls.
- No significant local SO2 sources remain.

The model essentially learns: "SO2 tomorrow ≈ SO2 today ≈ very low". This is correct but trivial.

9.3 Common Findings Across All Pollutants

1. **Shadow effect:** All models show predictions trailing behind rapid changes due to reliance on lag-1 features.
2. **Peak underestimation:** Sudden pollution spikes are consistently underestimated across all pollutant types.

3. **Stable period accuracy:** All models perform well during periods of gradual change or stable values.
4. **Autocorrelation dominance:** The previous hour's value is the strongest predictor for all pollutants, explaining 85-95% of feature importance.
5. **Practical implication:** Random Forest models are reliable for general trend forecasting but should not be relied upon for predicting high pollution episodes that trigger health warnings.

10. Overfitting Analysis

The gap between training and validation/test performance indicates the degree of overfitting.

Overfitting metrics:

- Mean Training R^2: 0.9528
- Mean Validation R^2: 0.8232
- Mean Test R^2: 0.8142
- **Mean Gap: 0.1036**

Conclusion: Mild overfitting detected (gap = 0.1036). The gap indicates some memorisation of training data, but models still generalise reasonably well. This level of overfitting is acceptable for the comparative study and does not require additional regularisation.

11. Conclusions

11.1 Summary of Results

Random Forest models were successfully trained for 141 LAQN site-pollutant combinations across 6 pollutants. After excluding 5 broken models due to data quality issues, the remaining 136 models achieved a mean test R^2 of 0.8142, explaining approximately 81% of variance in hourly pollution levels.

11.2 Key Findings

- **O3 is most predictable** (mean R^2 = 0.92) due to strong diurnal photochemical patterns.
- **NO2 shows consistent performance** (mean R^2 = 0.86) across 56 stations despite traffic variability.
- **PM10 has highest variability** (R^2 range 0.31-0.85) due to diverse local emission sources
- **Temporal autocorrelation dominates** with t-1 feature importance of 85-95% across all pollutants.
- **Peak underestimation is consistent** across all pollutant types due to lag-based prediction.

11.3 Limitations

- **Peak prediction:** Models consistently underestimate high pollution events.
- **Data quality:** 5 stations had constant test values requiring exclusion.
- **Memory constraints:** Reduced hyperparameters (n_estimators=100, max_depth=10) may limit performance.

- **SO2 trivial prediction:** Near-constant concentrations make meaningful prediction impossible.

11.4 Implications for CNN Comparison

The Random Forest baseline of $R^2 = 0.81$ provides a solid benchmark for CNN comparison. Key questions for the neural network approach include:

- Can CNN better capture peak pollution events?
- Do spatial features provide more value with convolutional layers?
- Can the shadow effect be reduced through sequence modelling?

12. Next Steps

1. Train CNN models using same data splits for fair comparison.
2. Compare performance metrics (RMSE, MAE, R^2) across both approaches.
3. Analyse whether CNN captures peak events better than Random Forest.
4. Replicate analysis on DEFRA dataset for cross-network comparison.
5. Document findings for the dissertation methodology chapter.

References

Géron, A. (2023) *Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow*. 3rd edn. O'Reilly Media.

HalvingGridSearchCV (no date) scikit-learn. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.HalvingGridSearchCV.html

RandomForestRegressor (no date) scikit-learn. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

r2_score (no date) scikit-learn. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html

Appendix: Output Files

All outputs saved to: `/data/laqn/rf_model_all/`

File	Contents
all_rf_models.joblib	141 trained Random Forest models
all_results.csv	Evaluation metrics for all models
summary_by_pollutant_corrected.csv	Summary excluding broken models
best_params_by_pollutant.joblib	Tuned hyperparameters per pollutant
tuning_results_by_pollutant.csv	HalvingGridSearchCV results
dissertation_plots/	All visualisation PNG files