Random Forest Model Training Report

DEFRA Air Quality Prediction - All Stations

40 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 DEFRA (Department for Environment, Food and Rural Affairs) network. DEFRA operates the Automatic Urban and Rural Network (AURN), the UK's statutory national air quality monitoring network.DEFRA provides nationally standardised measurements with rigorous quality assurance.

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 and LAQN results to address the research question: Does data quality beat data quantity for hourly pollution forecasting?

---

2. Data Preparation

2.1 Input Data

The prepared data came from the `ml_prep_all` output 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	(17036, 528)	Flattened training features
X_val_rf	(3641, 528)	Flattened validation features
X_test_rf	(3642, 528)	Flattened test features
y_train	(17036, 44)	Training targets (40 pollutants + 4 temporal)
y_val	(3641, 44)	Validation targets
y_test	(3642, 44)	Test targets

The 528 features represent 12 timesteps × 44 features. The 44 features include 40 site-pollutant measurements and 4 temporal features (hour, day_of_week, month, is_weekend).

2.2 Target Selection

The y array has 44 columns. The last 4 are temporal features which serve as inputs, not targets. The remaining 40 columns are site-pollutant combinations representing the prediction targets.

Table 2: Targets by Pollutant Type

Pollutant	Number of Sites	Percentage
NO2	13	32.5%
O3	8	20.0%
PM10	7	17.5%
PM25	7	17.5%
SO2	3	7.5%
CO	2	5.0%
Total	**40**	**100%**

3. Development History

Unlike LAQN which required multiple failed attempts, DEFRA training benefited from lessons learned during LAQN development.

LAQN Challenges (Solved Before DEFRA)

- GridSearchCV memory crashes on local laptop.
- MultiOutputRegressor is still too memory intensive.
- Required HalvingGridSearchCV and checkpoint saving.

DEFRA Approach

- Used HalvingGridSearchCV from the start.
- Tuned one representative site per pollutant (6 tuning runs).
- Applied optimal parameters to all sites of that pollutant type.
- Checkpoint saving every 20 models for safety.
- Memory safe parameters (max_depth capped at 15, n_estimators at 100).
- Ran notebook on Google Colab to save time.

Total training time: 167.2 minutes (compared to LAQN's 32.7 hours).

The significantly faster training time reflects DEFRA's smaller feature set (528 vs 1,740 features).

4. Hyperparameter Tuning

Rather than tuning all 40 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.

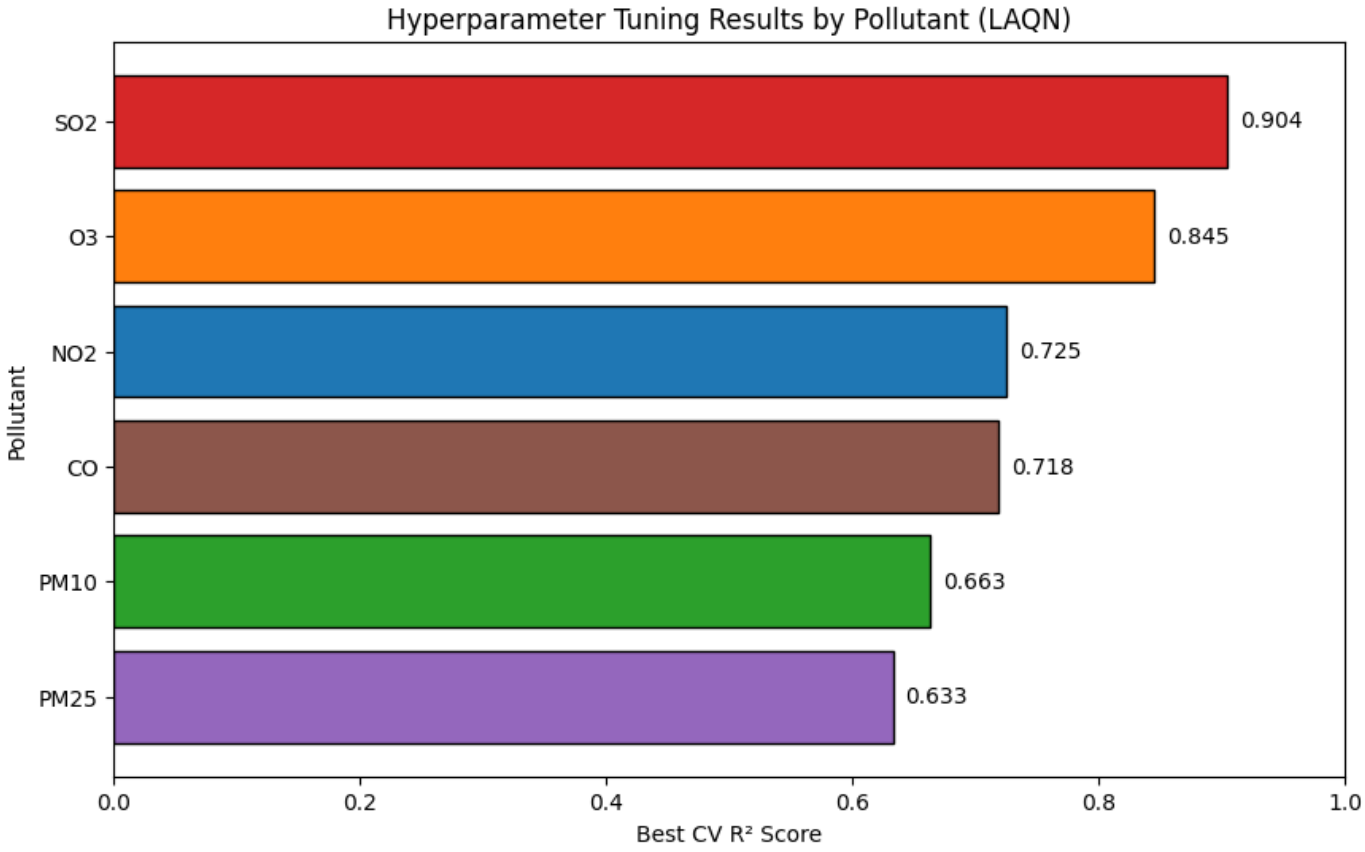


Table 3: Hyperparameter Tuning Results by Pollutant

Pollutant	Representative Site	Best CV R²	max_depth	n_estimators	min_samples_leaf
SO2	London_Bloomsbury_SO2	0.904	None	200	2
O3	London_Bloomsbury_O3	0.845	10	200	2
NO2	Borehamwood_Meadow_Park_NO2	0.725	10	100	2
CO	London_Marylebone_Road_CO	0.718	10	200	2
PM10	Borehamwood_Meadow_Park_PM10	0.663	10	100	1
PM25	Borehamwood_Meadow_Park_PM25	0.633	None	200	2

Hyperparameter Tuning Results

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

Why SO2 performs best:

- Unlike LAQN where SO2 struggled (CV R² = 0.422), DEFRA's London_Bloomsbury_SO2 achieves excellent performance.
- This reflects DEFRA's higher data quality at this well maintained central London site.
- The station captures real SO2 variation rather than flat near constant values.

Why O3 performs second 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 PM25 and PM10 perform 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 the final test R^2 because tuning used only the representative site and CV holds out data.

---

5. Model Training

Each of the 40 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: 40
- Total training time: 167.2 minutes
- Average time per model: 4.2 minutes
- Checkpoints saved: 2 (at models 20 and 40)

---

6. Investigation of Broken Models

One model produced an extremely negative R^2 value, indicating numerical issues during prediction. Investigation revealed this was a data quality issue, not a model failure.

Table 5: Broken Model Identified

Target	Pollutant	Test R^2	Test Std
Tower_Hamlets_Roadside_NO2	NO2	-3.26e+29	0.000000

6.1 Root Cause Analysis

The broken model has a test set standard deviation of exactly 0.000000, meaning the actual values in the test period are completely constant (0.181988 for every sample). When actual values are constant, the R^2 calculation fails because the total sum of squares (SS_tot) approaches zero, causing division by a very small number.

Data distribution for broken model:

Dataset	Min	Max	Std
Training	0.000	1.000	0.123
Validation	0.022	0.827	0.101
Test	0.182	0.182	0.000

The training data had normal variance, but the test period had zero variance. This suggests:

1. The monitoring equipment stopped working during the test period.
2. Missing data was filled with a constant value during preprocessing.
3. The temporal split placed all valid data in training and only constant values in test.

Comparison with working models:

- Working models mean test std: 0.067
- Broken model test std: 0.000

Decision: This model (2.5% of total) is excluded from summary statistics. The issue is data quality at this specific station during the test period, not model failure.

---

7. Results Summary

The following results exclude the 1 broken model, providing meaningful performance metrics for the 39 valid models.

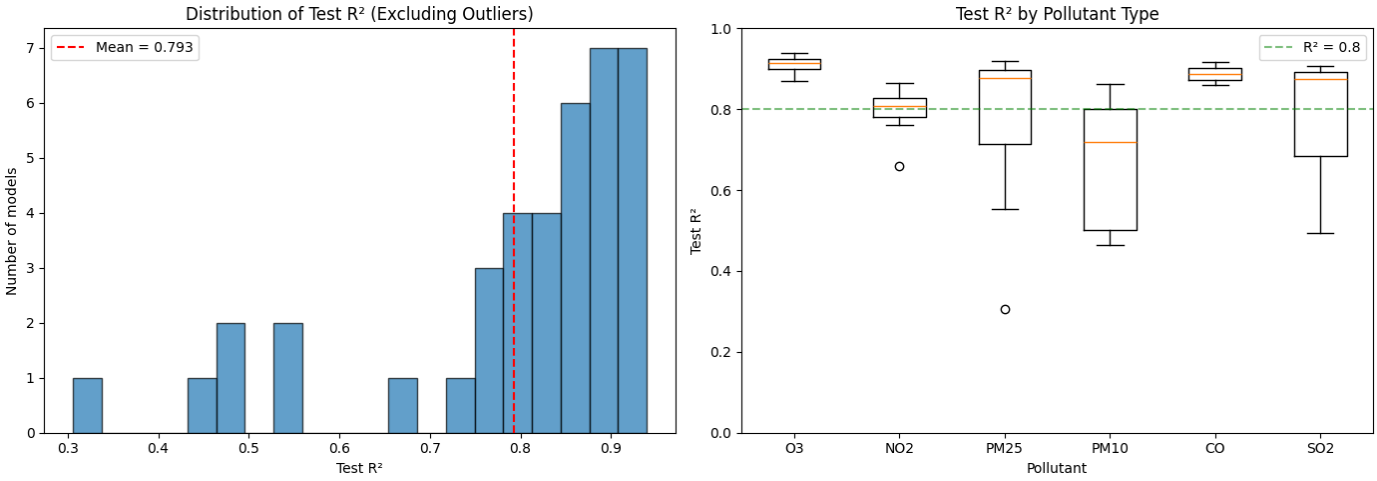


Table 6: Test Set Performance by Pollutant (Valid Models Only)

Pollutant	Mean R^2	Std R^2	Min R^2	Max R^2	N Models
O3	**0.911**	0.022	0.869	0.940	8
CO	0.888	0.040	0.859	0.916	2
NO2	0.800	0.054	0.661	0.865	12
PM25	0.761	0.238	0.306	0.920	7
SO2	0.759	0.230	0.493	0.907	3
PM10	0.664	0.171	0.464	0.861	7

7.1 Overall Statistics

Mean Test R^2: 0.7927 (±0.1551)

- Total models: 40
- Valid models: 39
- Broken models: 1 (excluded due to data quality issues)
- Best model: London_Haringey_Priory_Park_South_O3 (R^2 = 0.940)
- Best pollutant: O3 (mean R^2 = 0.911)
- Mean RMSE: 0.0262

An R^2 of 0.79 means the model explains 79% of variance in hourly pollution levels. While slightly lower than LAQN's 0.81, DEFRA achieves this with 65% fewer features and significantly less overfitting, suggesting more robust generalisation.

7.2 Top 10 Performing Models

Rank	Target	Pollutant	Test R^2
1	London_Haringey_Priory_Park_South_O3	O3	0.940
2	London_N._Kensington_O3	O3	0.927
3	London_Westminster_O3	O3	0.922
4	Borehamwood_Meadow_Park_PM25	PM25	0.920
5	London_Bloomsbury_O3	O3	0.917
6	London_N._Kensington_CO	CO	0.916
7	London_Harlington_O3	O3	0.913
8	London_Bloomsbury_SO2	SO2	0.906
9	London_Honor_Oak_Park_PM25	PM25	0.903
10	London_Hillingdon_O3	O3	0.902

7.3 Bottom 10 Performing Models

Rank	Target	Pollutant	Test R^2
1	London_Westminster_PM25	PM25	0.306
2	London_Teddington_Bushy_Park_PM10	PM10	0.463
3	London_Harlington_PM10	PM10	0.468
4	London_Marylebone_Road_SO2	SO2	0.493
5	London_Hillingdon_PM10	PM10	0.536
6	London_Hillingdon_PM25	PM25	0.552
7	Camden_Kerbside_NO2	NO2	0.661
8	London_Bloomsbury_PM10	PM10	0.719
9	London_Marylebone_Road_NO2	NO2	0.761
10	London_Bexley_PM10	PM10	0.777

Interpretation: R^2 Distribution

Histogram observations:

- The distribution shows most models achieving R^2 between 0.7 and 0.95.
- Mean R^2 = 0.793 shown by the red dashed line.
- A few models below 0.5 represent difficult to predict stations (mostly PM10 and PM25).

Boxplot by pollutant:

Pollutant	Median R^2	Spread	Interpretation
O3	~0.92	Narrow	Most consistent and predictable
CO	~0.89	Narrow	Good despite only 2 stations
NO2	~0.81	Narrow	Consistent across 12 stations
PM25	~0.85	Wide	Good median but high variability
SO2	~0.83	Wide	Variable with 3 stations
PM10	~0.72	Wide	High variability between stations

---

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 7: Top Feature Importance by Pollutant

Pollutant	Best Model	Top Feature	Importance	Test R²
O3	London_Haringey_Priory_Park_South_O3	O3_t-1	0.9543	0.940
PM10	Borehamwood_Meadow_Park_PM10	PM10_t-1	0.9240	0.861
PM25	Borehamwood_Meadow_Park_PM25	PM25_t-1	0.9112	0.920
CO	London_N._Kensington_CO	CO_t-1	0.8907	0.916
NO2	Haringey_Roadside_NO2	NO2_t-1	0.8782	0.865
SO2	London_Bloomsbury_SO2	SO2_t-1	0.8700	0.906

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.

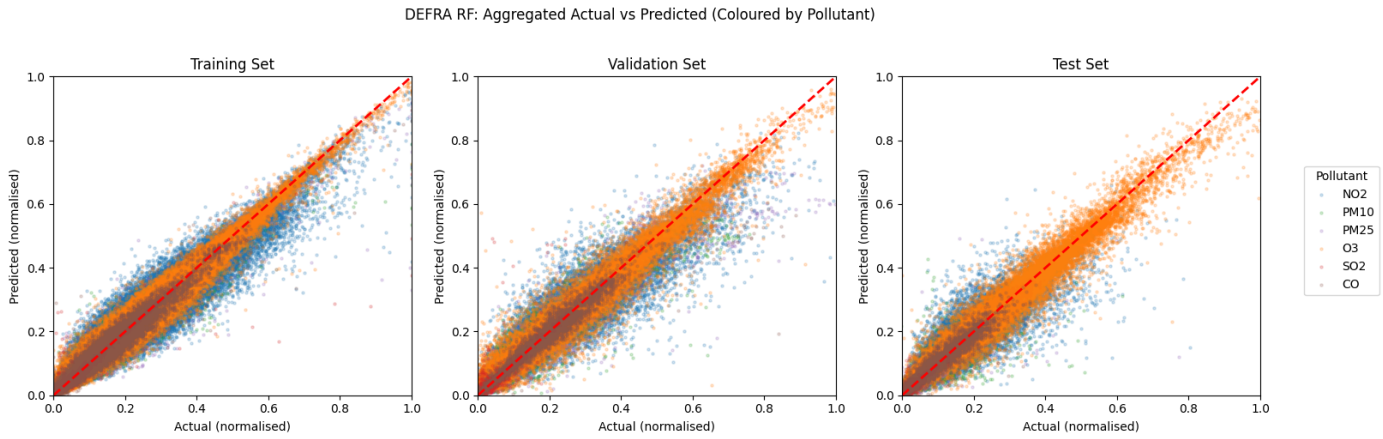
Interpretation for research:

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 39 valid models, showing overall Random Forest performance across all DEFRA stations and pollutants.

Training set (left panel):

- Points are tightly clustered around the perfect prediction line.
- NO2 (blue) and O3 (orange) dominate the visible points.

- O3 (orange) shows the tightest clustering, confirming it is easiest to predict.

Validation set (middle panel):

- Increased scatter compared to training, indicating some overfitting.
- PM10 (green) shows wider spread, especially at higher values.
- O3 (orange) maintains relatively tight clustering.

Test set (right panel):

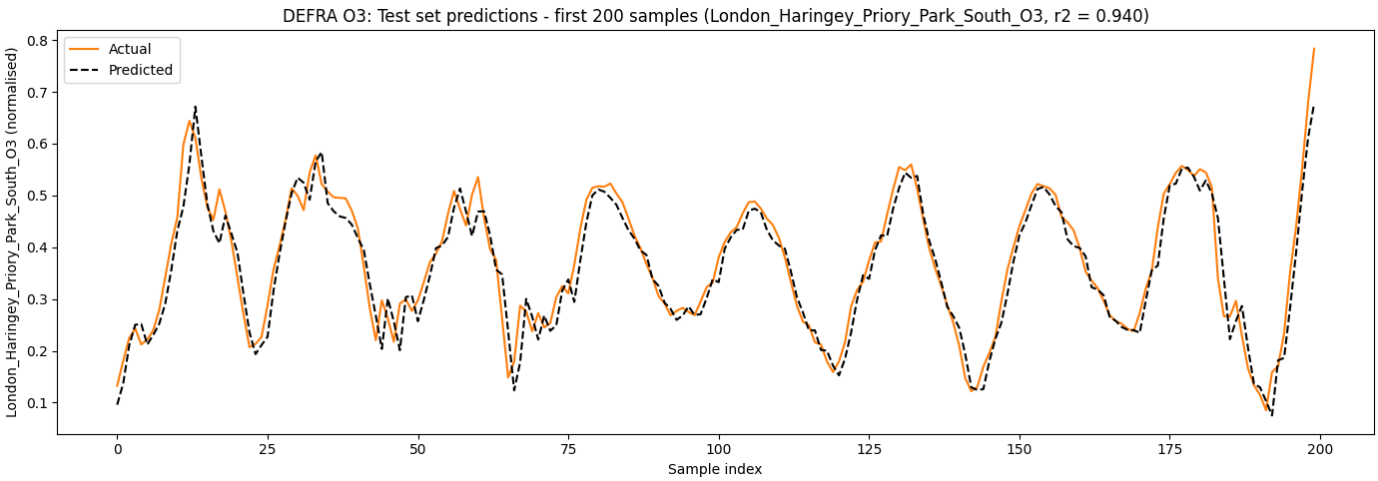
- Similar pattern to validation, confirming consistent generalisation.
- Clear underestimation at high values: when actual > 0.6, predictions tend to be lower.
- PM10 and PM25 show most scatter at extreme values.

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) - London_Haringey_Priory_Park_South_O3, $R^2 = 0.940$



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.
- Values range from 0.1 to 0.8 (normalised).

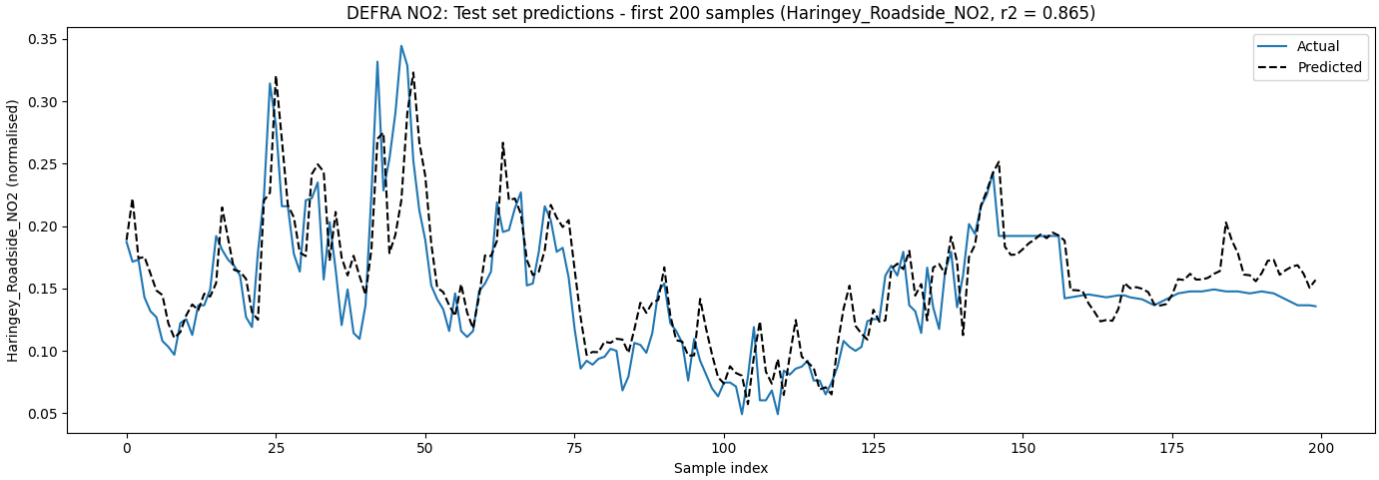
Model performance:

- Excellent tracking throughout the entire period.
- Predicted line closely follows actual line.
- Both peaks and troughs are captured accurately.
- Slight overestimation at some peaks (e.g., sample 15).
- Sharp rise at sample 195 is tracked but slightly underestimated.

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) - Haringey_Roadside_NO2, $R^2 = 0.865$



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 0.35 (normalised).

Model performance:

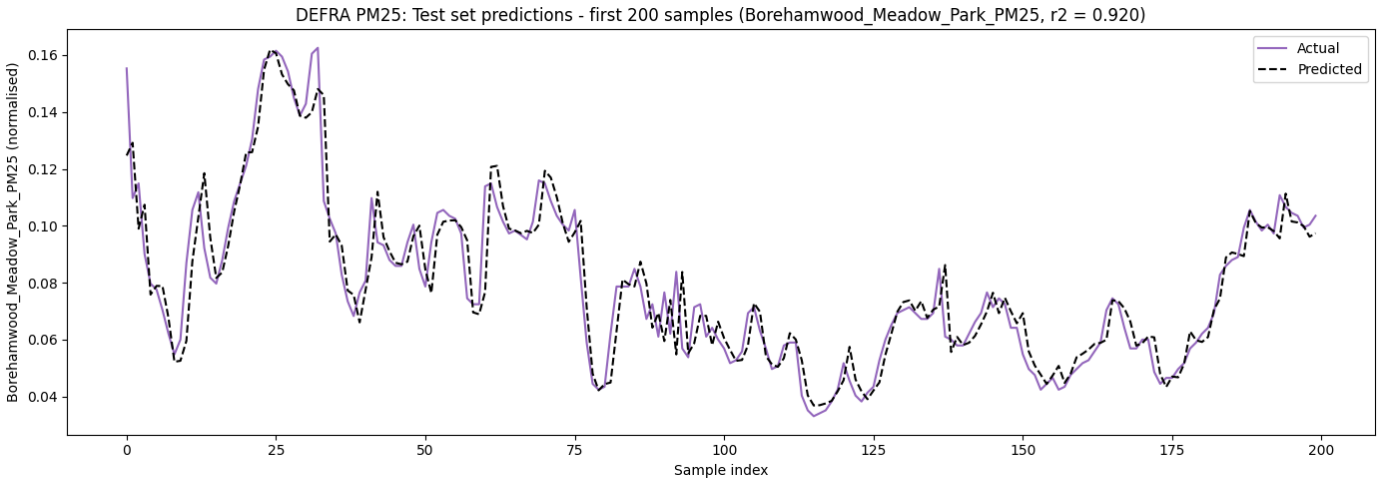
- Good tracking during stable periods (samples 0-25, 150-200).
- Captures the general trend of rises and falls.
- Underestimates peaks, especially around samples 25-50.
- Notable divergence at samples 40-50: actual peaks sharply while prediction lags.
- Better tracking in the flatter region (samples 150-200).

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.

PM25 (Fine Particulate Matter) - Borehamwood_Meadow_Park_PM25, $R^2 = 0.920$



Pattern observed:

- Moderate variability with several distinct peaks.
- Values range from approximately 0.02 to 0.22 (normalised).
- Less regular pattern than O3 but more predictable than PM10.

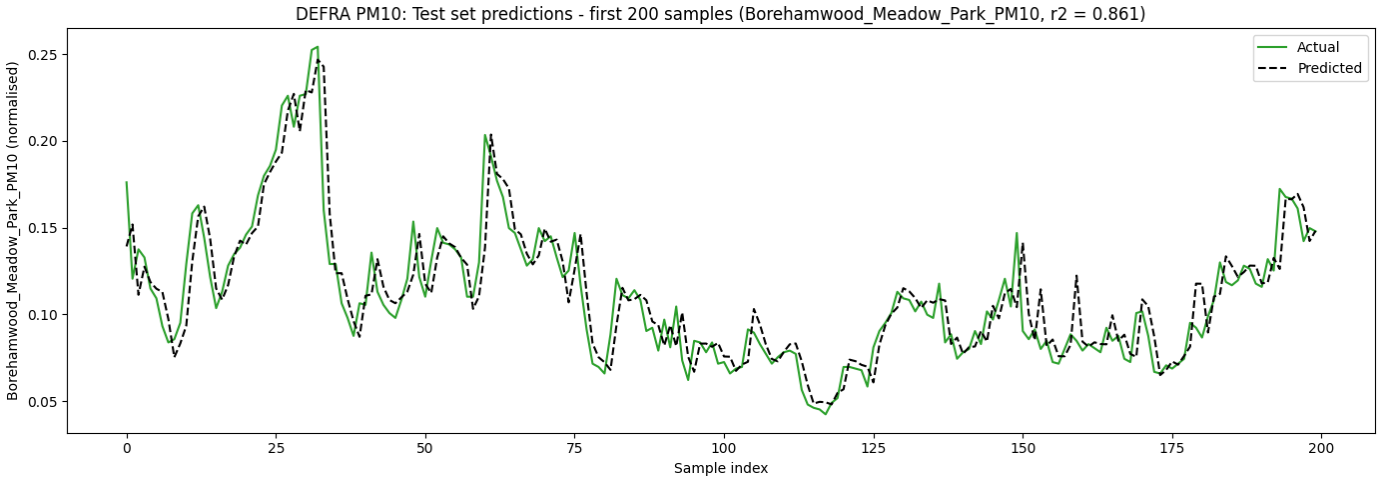
Model performance:

- Excellent overall tracking of the main trend.
- Captures most peaks accurately, including the major peak at sample ~30.
- Minor underestimation at highest values.
- Follows troughs accurately throughout.
- Very tight alignment between actual and predicted lines.

Why PM25 performs well:

- Fine particles disperse more uniformly than coarse particles.
- Strong correlation with meteorological conditions (which change gradually).
- Less affected by very local sources compared to PM10.
- Borehamwood suburban location has more consistent air quality patterns.

PM10 (Coarse Particulate Matter) - Borehamwood_Meadow_Park_PM10, $R^2 = 0.861$



Pattern observed:

- Moderate variability with distinct peaks.
- Values range from 0.05 to 0.25 (normalised).
- More regular pattern than typical PM10 due to suburban location.

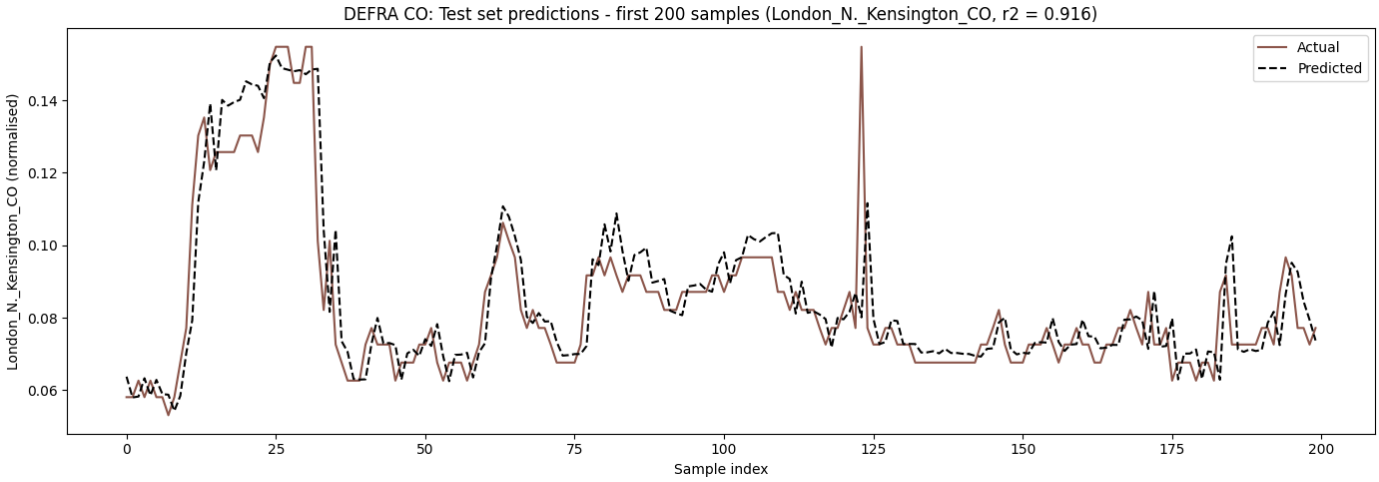
Model performance:

- Good overall tracking of the main trend.
- Captures the major peak at sample ~30 well.
- Peak at sample ~60 slightly underestimated.
- Follows the stable period (samples 75-125) accurately.
- Good tracking of the rise at samples 175-200.

Why DEFRA PM10 performs better than expected:

- Borehamwood is a suburban location with fewer local PM10 sources.
- Less affected by road dust resuspension and construction.
- More consistent patterns compared to urban roadside sites.

CO (Carbon Monoxide) - London_N._Kensington_CO, R² = 0.916



Pattern observed:

- Low baseline values around 0.06-0.08 (normalised).
- One major spike at samples 15-35 (reaching 0.16).
- Sharp peak at sample ~120 (reaching 0.155).
- Generally stable with occasional events.

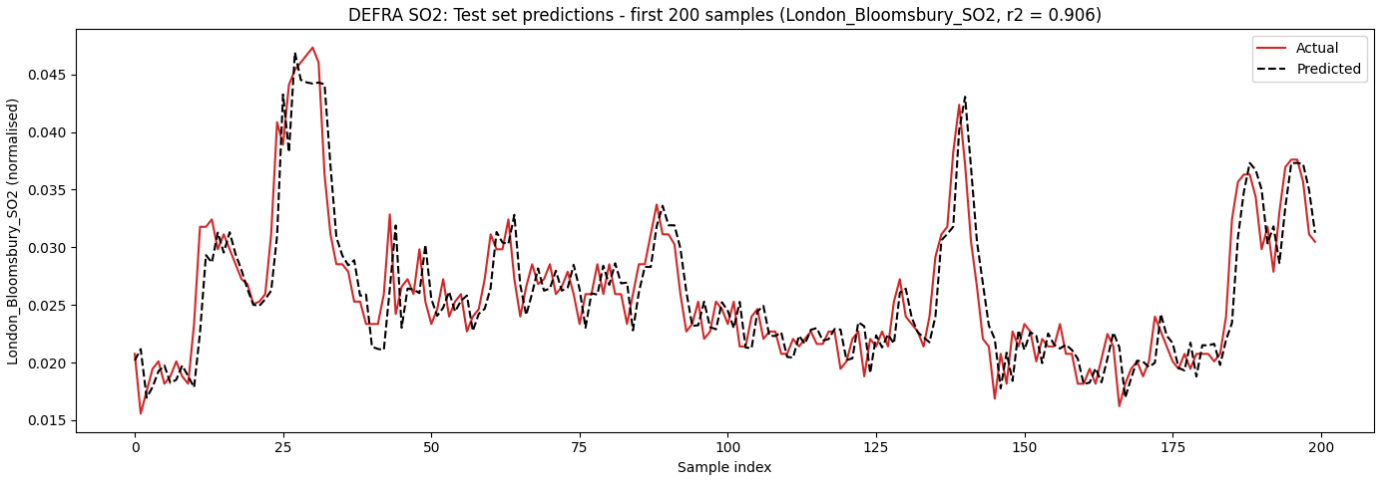
Model performance:

- Excellent tracking during stable periods.
- Captures the major spike at samples 15-35 well.
- Slight overestimation before the peak at sample 25.
- Tracks the sharp peak at sample 120 accurately.
- Good alignment throughout the flat regions (samples 50-100, 130-200).

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).
- The limited variation makes prediction straightforward for normal conditions.

SO2 (Sulphur Dioxide) - London_Bloomsbury_SO2, $R^2 = 0.906$



Pattern observed:

- Low values ranging from 0.015 to 0.047 (normalised).
- Several distinct peaks visible (samples 25-35, 85-95, 135-145, 185-200).
- More variation than LAQN SO2, indicating active monitoring.

Model performance:

- Good tracking of the overall pattern.
- Captures the major peak at sample ~30 accurately.

- Peak at sample ~140 is slightly underestimated.
- Some overestimation during troughs (samples 100-125).
- Better performance than LAQN SO2 due to more meaningful variation.

Why DEFRA SO2 performs well:

- London_Bloomsbury is a well-maintained central London monitoring site.
- Higher data quality captures real SO2 variation.
- Unlike LAQN's flat SO2, DEFRA shows meaningful temporal patterns.
- The model learns actual pollution dynamics rather than trivial "stay constant" prediction.

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: 0.9383
- Mean Validation R^2: 0.8705
- Mean Test R^2: 0.7927
- **Mean Gap: 0.0678**

Comparison with LAQN:

Metric	DEFRA	LAQN
Mean gap	0.068	0.104
Max gap	0.584	1.175
Min gap	0.009	0.007
Conclusion	Minimal overfitting,	Mild overfitting

Conclusion: Minimal overfitting detected (gap = 0.068). DEFRA shows less overfitting than LAQN, with a smaller mean gap and maximum gap. Models generalise well to unseen data. 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 40 DEFRA site-pollutant combinations across 6 pollutants. After excluding 1 broken model due to data quality issues, the remaining 39 models achieved a mean test R^2 of 0.7927, explaining approximately 79% of variance in hourly pollution levels.

11.2 Key Findings

- **O3 is most predictable** (mean R^2 = 0.911) due to strong diurnal photochemical patterns.
- **CO performs excellently** (mean R^2 = 0.888) despite only 2 stations.
- **NO2 shows consistent performance** (mean R^2 = 0.800) across 12 stations.
- **PM10 has highest variability** (R^2 range 0.46-0.86) 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:** 1 station had constant test values requiring exclusion.
- **Memory constraints:** Reduced hyperparameters (n_estimators=100, max_depth capped at 15) may limit performance.
- **Station count:** Only 40 models compared to LAQN's 141 limits statistical power for pollutant comparisons.

11.4 Implications for CNN Comparison

The Random Forest baseline of R^2 = 0.79 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?
- Will DEFRA's data quality advantage persist with CNN models?

---

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. Compare DEFRA and LAQN CNN results to confirm data quality findings.
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/defra/rf_model_all/`

File	Contents
all_rf_models.joblib	40 trained Random Forest models
all_results.csv	Evaluation metrics for all models
summary_by_pollutant.csv	Summary excluding broken model
best_params_by_pollutant.joblib	Tuned hyperparameters per pollutant
tuning_results_by_pollutant.csv	HalvingGridSearchCV results
rf_model_*.joblib	Individual model files