DEFRA 37 Pollutant Random Forest Model Report

Overview

This report summarises the Random Forest modelling results for DEFRA's 37 pollutant dataset. The expanded dataset tests whether the methodology validated on 6 regulatory pollutants generalises to VOCs with different chemical behaviours.

Dataset: 95 site-pollutant combinations across 37 pollutant types
Training samples: 17,036
Features: 1,188 (12 timesteps × 99 features)

Pollutant Categories

Category	Pollutants	Sites
Regulatory	NO2, PM2.5, PM10, O3, SO2, CO	40
Nitrogen	NO, NOx	26
Aromatic VOC	Benzene, Toluene, Xylenes, TMBs	8
Alkane	Ethane, Propane, Butanes, Pentanes, Hexanes	11
Alkene	Ethene, Propene, Butenes, Isoprene, Butadiene	9
Other VOC	Ethyne	1

Training Results

Model counts:

- Total trained: 95
- Valid models: 92
- Broken models: 3 (Tower Hamlets Roadside station failure)

Performance filtering based on Gilik, A., Ogrenci, A.S. and Ozmen, A. (2021) 'Air quality prediction using CNN+LSTM-based hybrid deep learning architecture':

- Useful models ($r^2 > 0.50$): 66
- Excluded models ($r^2 \leq 0.50$): 26

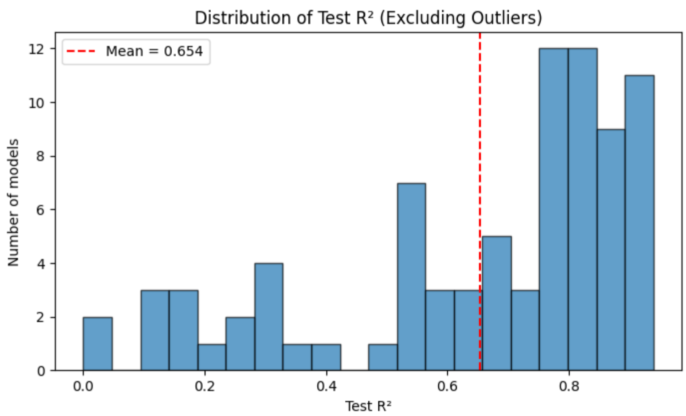
Performance Categorisation

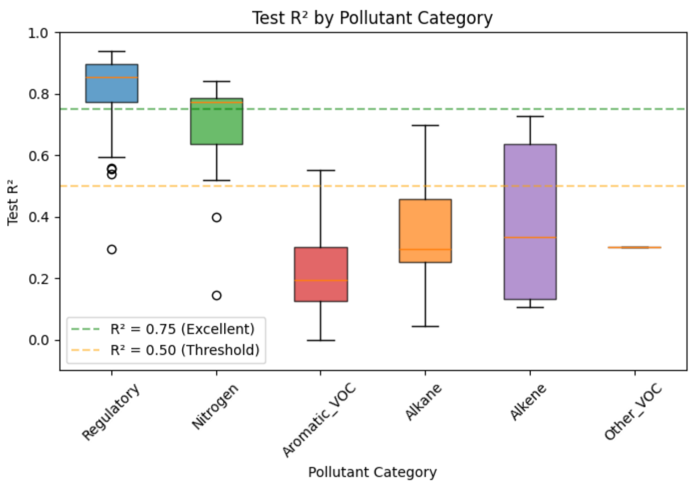
Performance thresholds derived from benchmarks:

Gilik et al. Results	r^2 Range	This Study Threshold
Best (PM10, O3)	0.88–0.92	Excellent ≥ 0.75
Typical (NOx, NO2)	0.55–0.74	Good ≥ 0.65
Lower (SO2)	0.46–0.62	Moderate > 0.50
Below range	< 0.46	Excluded ≤ 0.50

Results by category:

Performance	r^2 Range	Count	Categories Present
Excellent	≥ 0.75	44	Regulatory, Nitrogen
Good	0.65–0.75	8	Regulatory, Nitrogen, Alkane, Alkene
Moderate	0.50–0.65	14	All except Other_VOC
Excluded	≤ 0.50	26	Predominantly VOCs

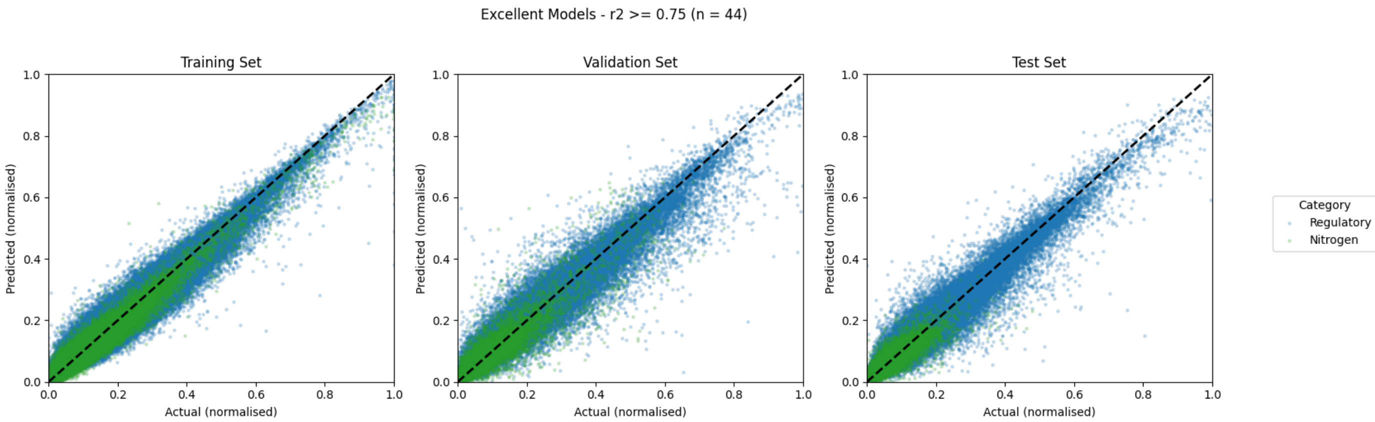




Findings by Performance Category

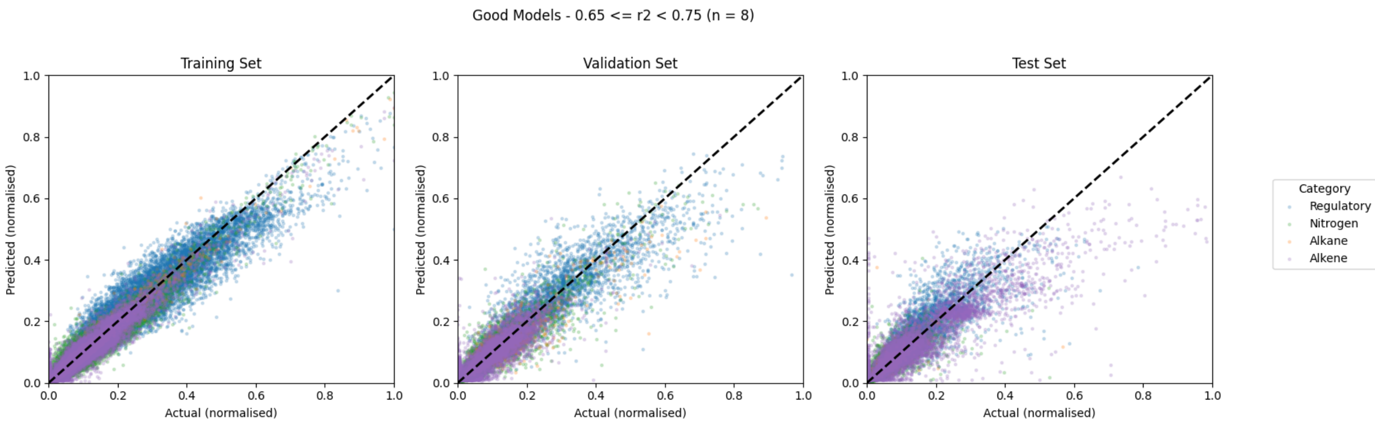
Excellent Models ($r^2 \geq 0.75$) 44 models

Only the Regulatory and Nitrogen categories achieve excellent performance. Tight clustering around the diagonal indicates accurate predictions across the full value range. Minimal train-to-test degradation confirms good generalisation.



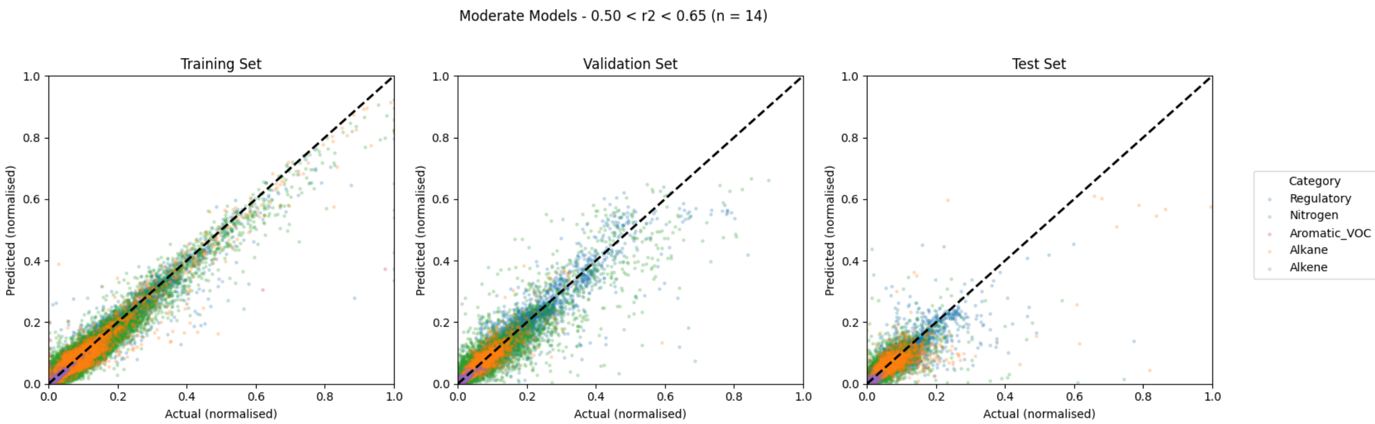
Good Models (0.65 ≤ r^2 < 0.75) 8 models

Four categories: Regulatory, Nitrogen, Alkane, Alkene. Wider scatter than excellent models, particularly at mid-range values. Includes biogenic Isoprene and stable Ethane.



Moderate Models (0.50 < r^2 < 0.65) 14 models

Five categories present with diverse compound types. Clear horizontal banding for some VOCs indicates mean-reversion predictions when true patterns can't be learned.



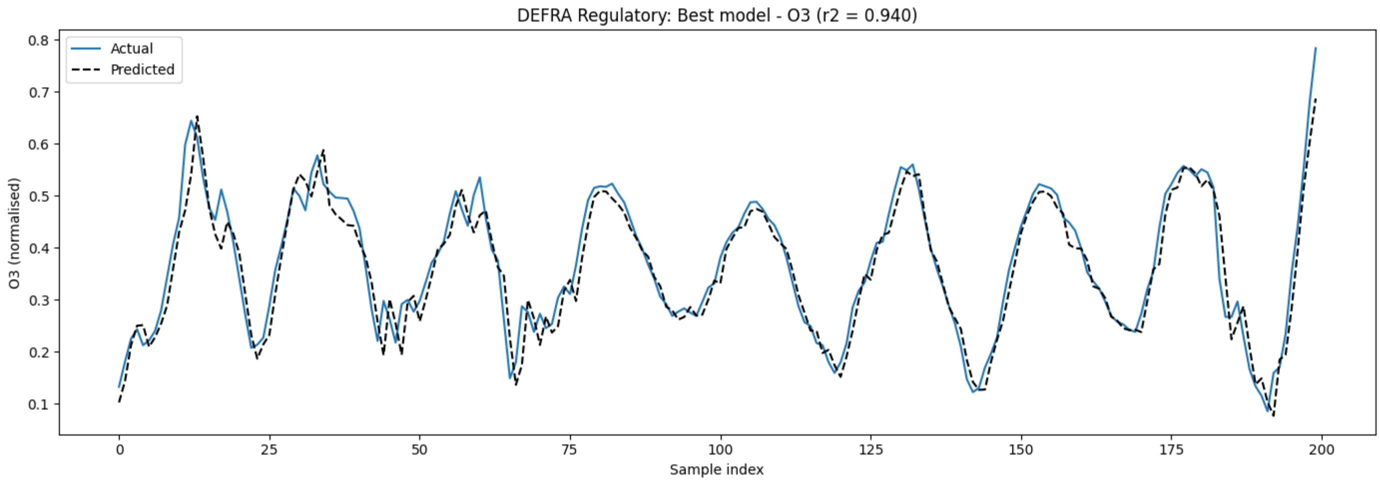
Best Model by Category

Category	Best Model	Pollutant	r²	Performance
Regulatory	London_Haringey_Priory_Park_South_O3	O3	0.940	Excellent
Nitrogen	London_N._Kensington_NO	NO	0.843	Excellent
Alkene	London_Marylebone_Road_Isoprene	Isoprene	0.726	Good
Alkane	London_Marylebone_Road_Ethane	Ethane	0.699	Good
Aromatic_VOC	London_Marylebone_Road_1_2_4_TMB	TMB	0.550	Moderate
Other_VOC	London_Marylebone_Road_Ethyne	Ethyne	0.302	Excluded

Time Series Analysis

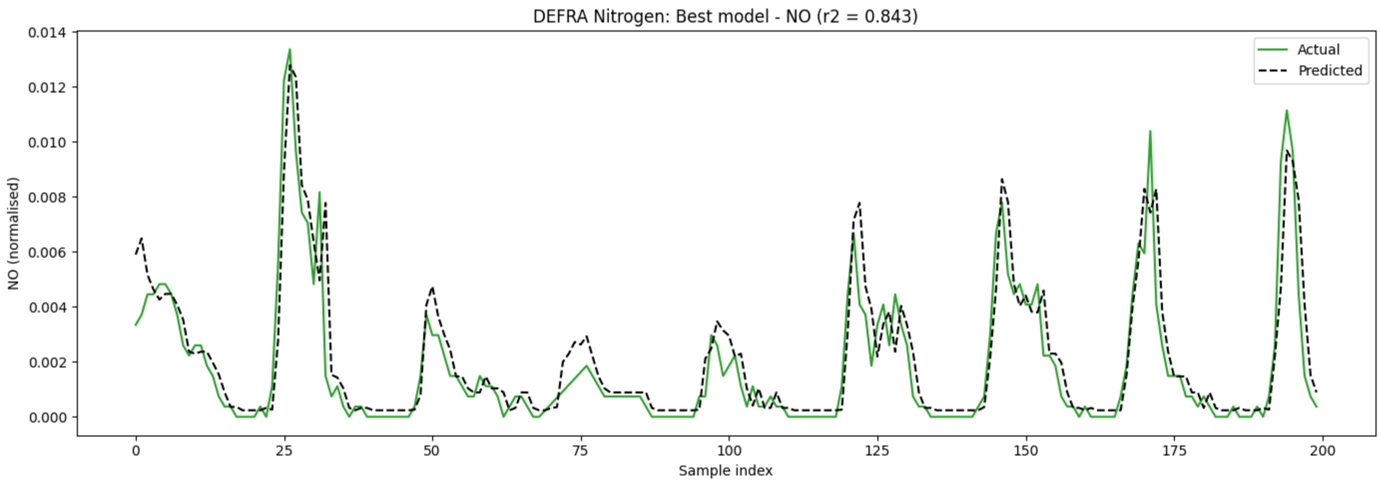
Regulatory — O3 (r^2 = 0.940)

Clear daily cycling with 24 hour periodicity. Predictions track actual values with minimal lag, like a shadow. Peak. Glady photochemical formation provides highly predictable patterns.



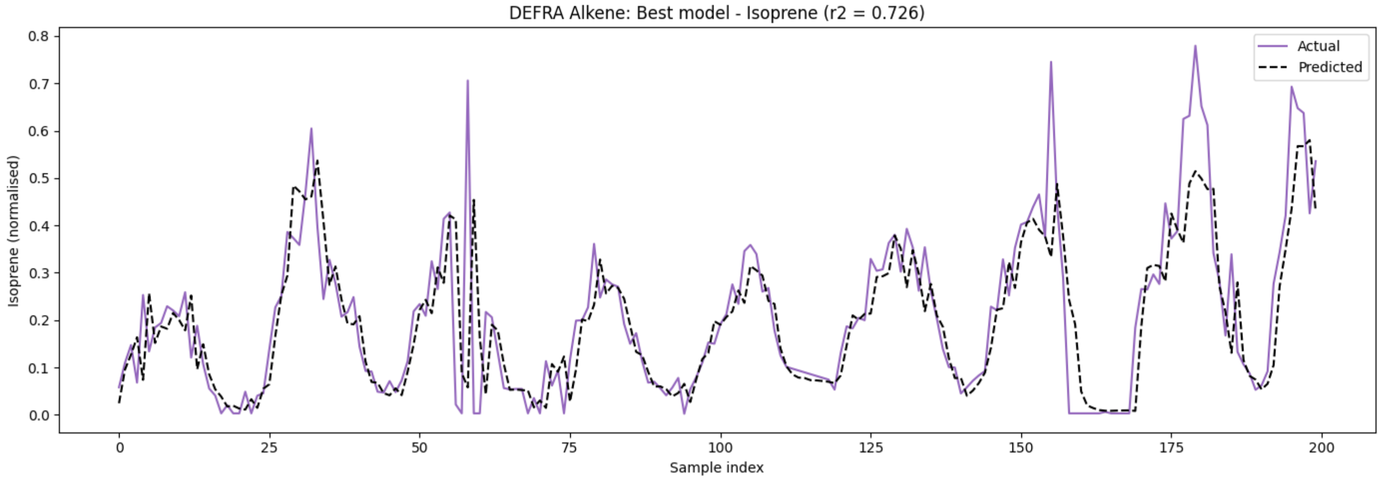
Nitrogen — NO (r^2 = 0.843)

Spiky pattern with sharp peaks reflecting traffic emissions. Low baseline with episodic spikes. Model captures spike timing but slightly underestimates peak magnitude.



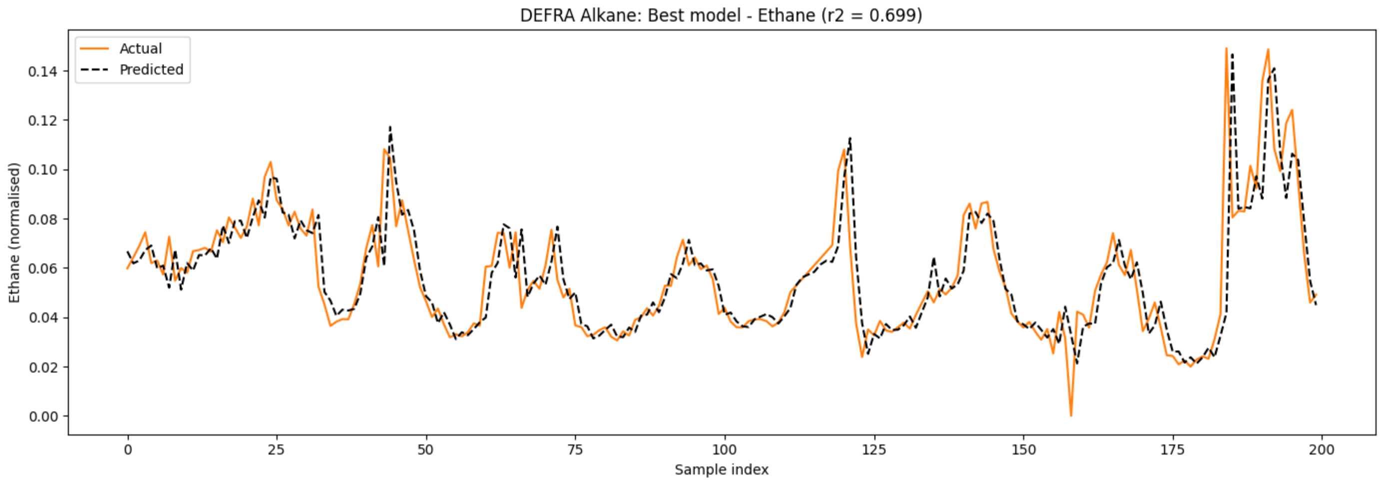
Alkene — Isoprene (r^2 = 0.726)

Strong diurnal pattern driven by temperature and sunlight. Biogenic emissions create predictable daytime peaks. More variable than regulatory pollutants due to vegetation response.



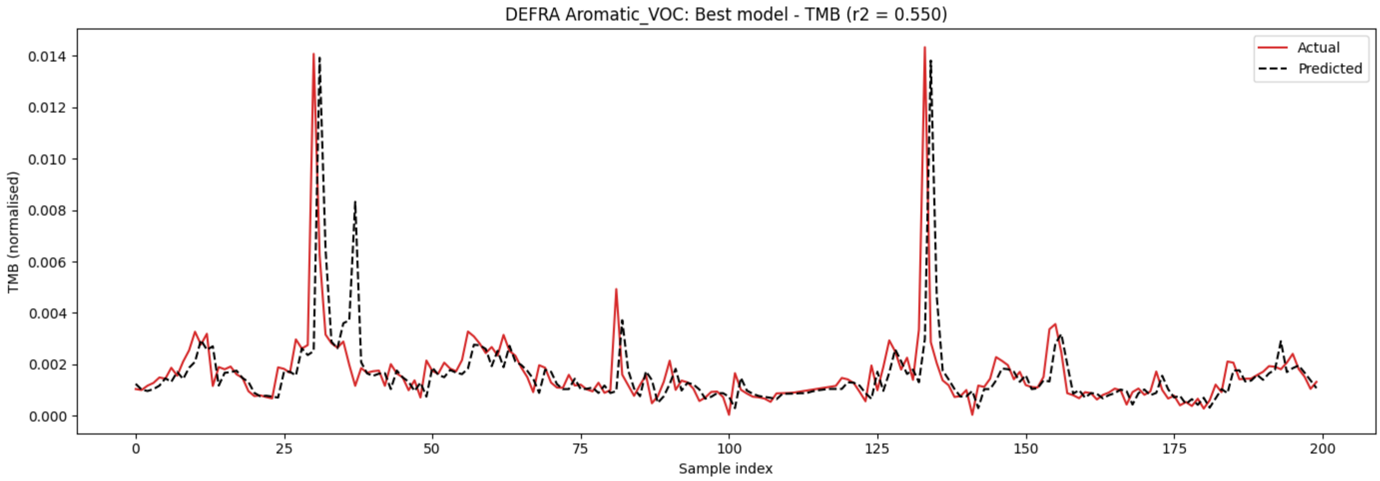
Alkane — Ethane ($r^2 = 0.699$)

Moderate variability with less noticeable cycling. Natural gas leakage provides a relatively stable baseline. Single-station limitation reduces spatial representativeness.



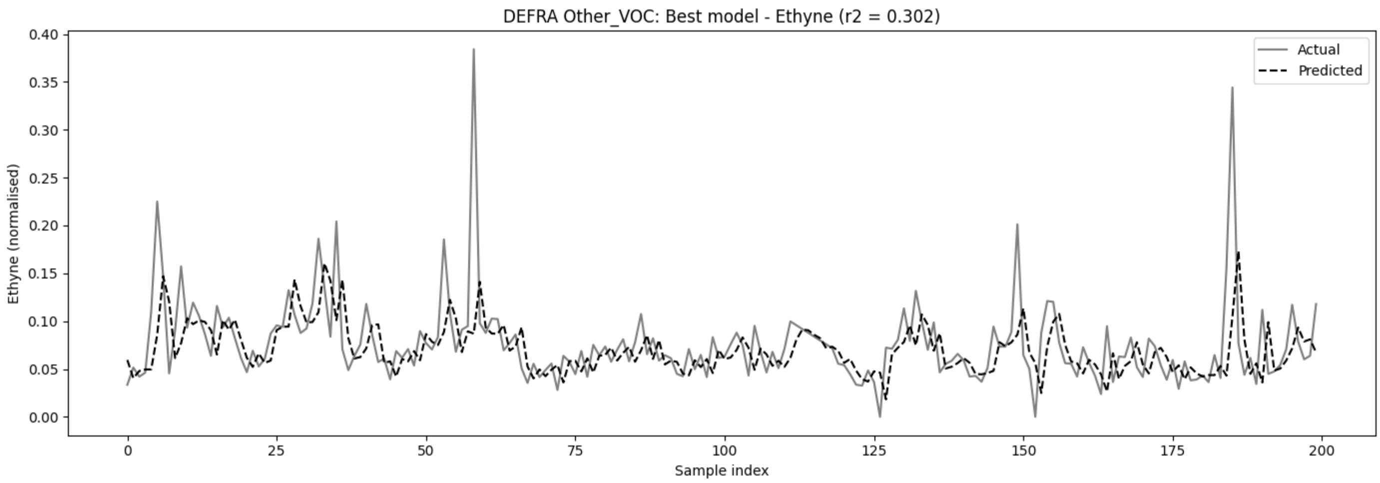
Aromatic_VOC — TMB ($r^2 = 0.550$)

Highly episodic with extreme spikes above a low baseline. The model captures major events but misses smaller variations. Traffic and solvent sources create unpredictable patterns.



Other_VOC — Ethyne ($r^2 = 0.302$)

Poor model performance, excluded from useful models. The combustion source is highly variable and localised. Single-station data insufficient for reliable prediction.



DEFRA vs LAQN Comparison (Regulatory Pollutants)

Pollutant	DEFRA Best Model	DEFRA r^2	LAQN Best Model	LAQN r^2	Better
O3	London_Haringey_Priory_Park_South_O3	0.940	HG4_O3	0.939	Similar
PM25	Borehamwood_Meadow_Park_PM25	0.920	HP1_PM25	0.899	DEFRA
CO	London_N._Kensington_CO	0.916	KC1_CO	0.823	DEFRA
SO2	London_Bloomsbury_SO2	0.906	BG1_SO2	0.915	Similar
NO2	Haringey_Roadside_NO2	0.865	TH2_NO2	0.878	LAQN
PM10	Borehamwood_Meadow_Park_PM10	0.861	CW3_PM10	0.854	Similar

DEFRA outperforms LAQN for CO and PM25. LAQN slightly better for NO2 due to denser station network. DEFRA's higher data quality (91.2% completeness) provides advantage for most pollutants.

Comparison with Gilik, A., Ogrenci, A.S. and Ozmen, A. (2021)

Metric	This Study (DEFRA)	Gilik, A., Ogrenci, A.S. and Ozmen, A. (2021)
Excellent models ($r^2 \geq 0.75$)	44 (48%)	~30% of results
Mean r^2 (useful models)	0.769	0.66
Best single model	O_3 ($r^2 = 0.940$)	PM10 ($r^2 = 0.92$)
Regulatory performance	Excellent	Good to Excellent

DEFRA results exceed Gilik, A., Ogrenci, A.S. and Ozmen, A. (2021) benchmarks for regulatory pollutants, validating the Random Forest methodology.

Key Findings

1. **Category determines performance ceiling.** Regulatory pollutants achieve excellent results ($r^2 \geq 0.75$) while VOCs from single stations rarely exceed moderate ($r^2 < 0.65$).
2. **Regulatory pollutants perform consistently.** Mean $r^2 = 0.788$ matches the 6 pollutant baseline. Adding VOC features does not degrade regulatory predictions.
3. **VOC limitation is data-driven, not methodological.** All VOC measurements come from Marylebone Road single station. Without spatial neighbours, models cannot learn from surrounding stations.
4. **Biogenic VOCs outperform traffic VOCs.** Isoprene ($r^2 = 0.726$) follows predictable temperature/light patterns. Traffic-related aromatics and alkenes show poor predictability due to episodic sources.
5. **Overfitting detected but manageable.** Mean train-validation gap of 0.21 indicates some memorisation. Regulatory pollutants show smallest gaps, VOCs show largest.

Conclusions

Random Forest effectively predicts regulatory pollutants and nitrogen species using 12-hour temporal features. VOC prediction requires expanded monitoring networks beyond single-station coverage.

The performance hierarchy reflects fundamental differences in pollutant behaviour:

- Regulatory pollutants: Strong temporal patterns, multiple stations, well-understood chemistry
- Nitrogen species: Traffic-related diurnal cycles despite rapid atmospheric reactions
- Biogenic VOCs: Temperature/light-driven emissions follow predictable patterns
- Traffic VOCs: Highly localised, episodic emissions resist prediction from temporal features alone

DEFRA's superior data quality (91.2% completeness) translates to better performance for most regulatory pollutants compared to LAQN.

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

Gilik, A., Ogrenci, A.S. and Ozmen, A. (2021) 'Air quality prediction using CNN+LSTM-based hybrid deep learning architecture', *Environmental Science and Pollution Research*, 29(8), pp. 11920–11938. doi:10.1007/s11356-021-16227-w.