Forecasting High Speed 2 Construction Environmental Impacts

Xu Feng[1], Khuong An Nguyen[1], and Matthew Tye[2]

[1] Department of Computer Science, Royal Holloway University of London, Egham, Surrey TW20 0EX, United Kingdom
Xu.Feng@rhul.ac.uk; Khuong.Nguyen@rhul.ac.uk

[2] High Speed 2, Snow Hill Queensway, Birmingham B4 6GA, United Kingdom
Matthew.Tye@hs2.org.uk

Abstract. High Speed Two (HS2) is Britain's current largest rail infrastructure project. However, its delivery has faced significant operational challenges, leading to lengthy delays. The ability to forecast environmental impacts could support more effective site management, resource allocation, and mitigation strategies. Thus, we investigate whether air quality and noise monitoring data from HS2 construction sites can be predicted ahead of time. We implemented several state-of-the-art time-series models and benchmarked them against statistical models. To address volatile environmental conditions, we proposed a Conformal Prediction based time-series approach to provide statistically valid prediction intervals. The results demonstrated that background noise was the most forecastable target, predictable one-hour ahead achieving RMSE 1.35 dB(A). Surprisingly, we found no correlation between air and noise measures. To the best of our knowledge, this was the first work to apply Machine Learning to forecast HS2 construction environmental impacts.

Keywords: High Speed 2 · air quality · noise · time series forecasting

1 Introduction

HS2 promised to bridge the UK North-South divide by producing long-term benefits for employment, regional connectivity and economic growth. However, given the scale of this construction project, environmental impacts including air quality and noise require careful management. As construction activity intensifies, the increased presence of people and equipment on site may lead to higher levels of noise and particulate matter. Under the Government's Environmental Minimum requirements Act[1], HS2 are required to measure these environmental effects and ensure the limits are not exceeded.

However, the difficulty with current environmental monitoring is that it only reports exceedances after they have occurred. This limits managers' ability to

[1] https://www.gov.uk/government/publications/environmental-minimum-requirements - last accessed on July 1st 2026.

allocate resources whilst remaining compliant with regulations. If future conditions could be anticipated in advance, managers could identify sites capable of additional construction activities or take measures to maintain safety thresholds.

More importantly, air quality and noise measures are highly volatile in construction sites. The environmental conditions may change rapidly due to weather, traffic, and other external factors. As a result, traditional Machine Learning (ML) approaches may struggle to provide reliable estimations.

Therefore, to examine the environmental forecast accuracy and uncertainty, we compared statistical modelling baselines against advanced ML models. Each model was evaluated across multiple air and noise quality targets monitored around HS2 construction sites. We also proposed a Conformal Prediction (CP) time-series approach to generate prediction intervals with guaranteed coverage. Figure 1 illustrates the result of our method on forecasting HS2 noise level, showcasing the prediction interval that guaranteed to include the true noise measure.

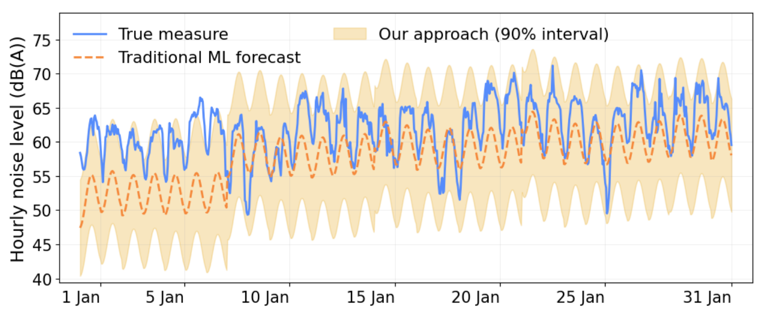

Fig. 1: An example of our approach for predicting hourly noise level on HS2 construction site. While traditional ML models struggle to predict the noise measures, our approach correctly captured the true measure 90% of the time, for a 90% prediction interval.

Specifically, our paper makes the following contributions:

- We compare statistical baseline methods against ML models, and apply Conformal Prediction to quantify forecasting uncertainty.
- We evaluate the predictability of multiple environmental indicators, identifying which air quality and noise targets can be predicted reliably.
- We present the first study, to the best of our knowledge, that applies ML for forecasting HS2 construction conditions, demonstrating how such predictions may support operational planning and environmental compliance.

The remaining of the paper is structured as follows. Section 2 reviews the related work on air quality and noise prediction. Section 3 describes the HS2 environmental data. Section 4 defines our forecasting framework. Section 5 presents the experimental results. Lastly, Section 6 concludes our work.

2 Related work

This section compares existing work on ML approaches for air quality and noise forecasting (see Table 1). Crucially, existing studies do not provide reliable predictions with statistical coverage guarantees for construction-related environmental impacts, which our work addresses.

Table 1: Summary of the related work reviewed in this study.

Paper Reference	**ML Model Employed**	**Performance**	**Notes**
[1]	DTR, LNR, RFR.	DTR AQI MAE 2.02%, RMSE 9.94%.	Cloud computing reduced model execution time.
[3]	SWR, SVR.	SVR $PM_{2.5}$ R^2 0.78-0.97, RMSE 5.62-22.21.	Suggests we can use data from surrounding stations to predict a station's own $PM_{2.5}$.
[7]	MLP, KNN, RF, XGB.	RF PM_{10} $R^2 = 0.953$, RMSE 0.032, MAE 0.019.	All struggled to predict black carbon PM_{10}. The best $R^2 =$ 0.585 by RF.
[9]	R, SVR, RFR, ETR, XGB.	All $PM_{2.5}$ $MAE < 2$, $RMSE < 2.5$.	As more input features were added, the performance improved.
[11]	GBM, MLR.	GBM Noise $RMSE < 1dB(A)$, $R^2 > 0.7$.	Historical features were more informative than time features.
[12]	KNN, RF, NB.	RF overall accuracy 0.95, AUC 1 for the "Good" PM_{10} concentration range	The most impactful features were proximity to transport, urban and industrial activity.
[14]	LTSM, RW, SAE, SVM.	LTSM Noise 10 min interval - RMSE 0.48-0.93, MAE 0.31-0.69.	Larger intervals generally improve accuracy as randomness is averaged out.

*M*odel: AQI=Air Quality Index; DTR=Decision Tree Regression; ETR=Extra Trees Regressor; XGB=Extreme Gradient Boosting; GBM=Gradient Boosting; KNN=K-Nearest Neighbours ; LNR=Linear Regression; LSTM=Long Short Term Memory Network; MLP=Multi-layer Perceptron; MLR=Multiple Linear Regression; NB=Naïve Bayes; RF=Random Forest; RFR=Random Forest Regressor; RW=Random Walk; R=Ridge Regression; SAE=Stacked Autoencoder; SWR=Stepwise Regression; SVM=Support Vector Machine; SVR=Support Vector Regression.
Metrics: MAE=Mean Absolute Error; RMSE=Root Mean Square Error; AUC=Area Under the Curve.

3 HS2 environmental monitoring data

This section describes the HS2 environmental monitoring data, its transformation from heterogeneous raw workbooks into a unified dataset, and the preprocessing applied before Machine Learning modelling.

3.1 Air quality and noise monitoring

To assess construction impacts and support compliance with environmental requirements, HS2 collects air-quality and noise data around active construction sites. These measurements are recorded by fixed monitoring instruments installed at specific locations near the worksites. Each instrument is treated as one monitor and provides an independent environmental time series.

In this study, six monitors from three HS2 construction areas were selected to represent contrasting environments. The London Borough of Camden (LBC) and Birmingham City Council (BCC) were chosen as major urban areas with intensive construction activity, while Buckinghamshire Council (BC) provided a more rural setting (see Figure 2).

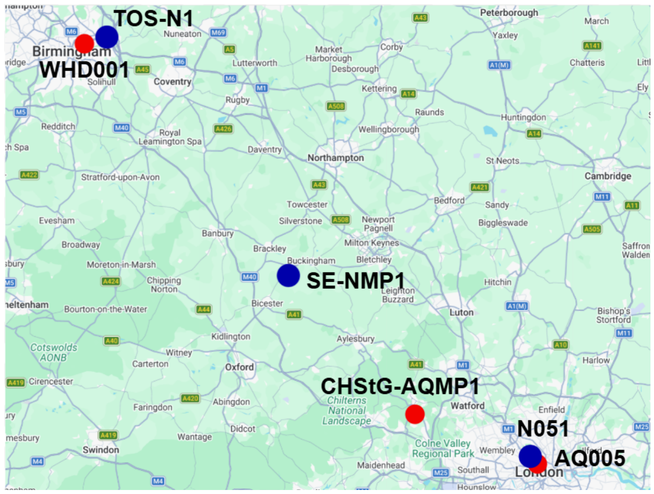

Fig. 2: Locations of the six HS2 monitors used in this study. Red markers indicate air quality monitors and blue markers indicate noise monitors. Camden sites comprise AQ005 (Stephenson Way) and N051 (Adelaide Road Ventilation Shaft). Buckinghamshire sites are CHStG-AQMP1 (Hobbs Hole Cottage) and SE-NMP1 (School End, Chetwode). Birmingham sites consist of WHD001 (Washwood Heath Depot) and TOS-N1 (Birmingham Road, Water Orton).

Air monitors are AQ005, CHStG-AQMP1, and WHD001. Each has one prediction target PM_{10}, the particulate matter PM_{10} concentration in micrograms per cubic metre $\mu g/m^3$.

Noise monitors are fitted with a windshield protected microphone and are installed at approximately 4 m above ground level in a free field area. Its calibration is checked at the start and end of each measurement. Each monitor has three prediction targets, all measured in A-weighted decibels dB(A), which represents sound pressure levels adjusted to reflect human hearing. $L_{\mathrm{pA90},T}$ is the level exceeded for 90% of the period (i.e. the background noise), $L_{\mathrm{pAeq},T}$ is the continuous sound pressure level over the period (i.e. the average), $L_{\mathrm{pAF,max}}$ is the maximum fast time sound pressure level (i.e. short time high noise events).

The final audit checked metadata, schemas, timestamps, chronological order, missing measurements, and duplicated records (see Table 2). The final dataset contains 6 monitor keys, 26,546 rows, and 52,976 finite environmental readings, with an overall capture rate of 99.88%. The audit also identified 66 missing source values and 48 additional duplicate monitor-timestamp records.

Table 2: Overview of the HS2 environmental dataset used.

Feature	Value
Monitor-level files	6
Councils included	Buckinghamshire (BC), Birmingham (BCC), and Camden (LBC)
Number of monitors by council	BC: 2; BCC: 2; LBC: 2
Physical data records	26,546
Air-quality measurement type	PM10
Noise measurement types	LpAeq, T, $L_{\mathrm{pA90},T}$, and $L_{\mathrm{pAF,max}}$
Finite environmental readings	52,976
Nominal monitoring duration	184 calendar days
Nominal sampling interval	1 h

4 Reliable HS2 environmental forecasting

This section outlines the forecasting framework used to predict HS2 environmental data, particularly, the baseline models, machine learning models, the evaluation metrics and the conformal prediction framework.

Throughout, y_t symbolises the true value of the target at hour t, $\hat{y}_t$ is the models predicted value for that hour, and n is the number of predictions being evaluated. The forecast horizon h is the number of hours ahead being predicted, where $h = 1$ is adopted. Moreover, $\mathbf{x}$ is the input feature vector for a single prediction i.e. the flattened lag window of recent target values together with the calendar features. $\mathbf{x}_i$, y_i, $\hat{y}_i$ are the feature vector, true value, and prediction for the i-th sample. The i index represents training samples and t is time.

4.1 Baseline models

Two baselines were evaluated on the January test set to provide a reference point for comparing against more complex models.

- **Persistence (lag)** - The value h hours ago is used as the prediction $\hat{y}_t = y_{t-h}$, where y_{t-h} is the observed value h hours before t. This naive forecast is used as it is strong and hard to beat for short term forecasts. [4].
- **Rolling mean** - The mean of the previous window size w is used $\hat{y}_t = \frac{1}{w}\sum_{i=1}^{w} y_{t-h-i}$, where w is the number of past hours i.e. window size and y_{t-i} is the observed value i hours before t. This is included as a smoothed model capturing the average while ignoring momentary spikes.

4.2 Machine Learning models

Four ML models were evaluated in this study.

- **Linear regression** predicts the target as a weighted combination of the input features, $\hat{y} = \beta^\top \mathbf{x} + b$, where β contains the learned coefficients and b is the intercept. It provides a simple and interpretable model of linear relationships but can be affected by multicollinearity. [8].

- **Gradient boosting** combines regression trees sequentially according to $F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \nu h_m(\mathbf{x})$, where h_m is the tree added at iteration m and ν is the learning rate. It can capture nonlinear relationships and interactions between temporal and environmental features. [2].
- **ARIMA** represents the differenced time series using autoregressive and moving-average terms. An ARIMA(p, d, q) model uses p autoregressive lags, differences the series d times, and includes q lagged forecast errors. The model was repeatedly refitted using all observations available before each forecast cut-off and then used to generate multi-step forecasts. [6].
- **ARIMAX** extends ARIMA by estimating with exogenous inputs. The term $\boldsymbol{\beta}^\top \mathbf{x}_t$ is added, where $\mathbf{x}_t$ is the feature at hour t. [5].

4.3 Conformal Prediction

With traditional ML models, a single predicted value does not show how certain the result is. This is a serious concern for a large and complex project as HS2, where unreliable forecasts could lead to poor site management, inefficient resource allocation, and greater disruptions.

Therefore, we adopted conformal prediction (CP) to complement each forecast with a prediction interval and provide an explicit assessment of reliability. CP is a model-agnostic method that uses calibration errors to quantify forecast uncertainty. It provides finite-sample marginal coverage without assuming a specific error distribution [10].

For HS2 environmental data, we used a rolling window of recent out-of-sample errors, allowing interval widths to adapt over time [13]. The required reliability is defined by the nominal coverage level $1 - \alpha$, where α is the significance level. For example, $\alpha = 0.10$ gives a 90% prediction interval, meaning that the target is expected to fall inside the interval, 90% of the time.

For a calibrated forecast at hour t, the absolute residual is used as the nonconformity score $e_t = |y_t - \hat{y}_t|$. A larger score means less accurate forecast. For a rolling calibration window of W hours, the scores before forecasting hour t are:

$$\mathcal{E}_t = \{e_i : t - W \leq i < t\} \tag{1}$$

Let $m_t = |\mathcal{E}_t|$ be the number of available calibration scores. These scores are then sorted. For nominal coverage $1 - \alpha$, the selected position is:

$$k_t = \min\{m_t, \lceil (m_t + 1)(1 - \alpha) \rceil\} \tag{2}$$

This rank selects an error threshold large enough to cover approximately the required proportion of previous calibration errors. If $e_{t,(1)} \leq \cdots \leq e_{t,(m_t)}$ are the ordered scores, the conformal error quantile is $q_{t,1-\alpha} = e_{t,(k_t)}$.

The value $q_{t,1-\alpha}$ is used as the half-width of the prediction interval. The symmetric interval for hour t is therefore:

$$C_{t,1-\alpha} = [\hat{y}_t - q_{t,1-\alpha}, \hat{y}_t + q_{t,1-\alpha}] \tag{3}$$

Under exchangeability, CP guarantees that the true value y_t falls within $C_{t,1-\alpha}$ with marginal probability of at least $1 - \alpha$.

5 Experimental results

This section presents our experimental results, which were organised around the following research questions:

- **Can air and noise targets be forecast from historical monitoring data and calendar features?** It is believed that targets dominated by regular daily cycles will be forecastable, whereas targets dominated by random events will not.
- **Which of the targets are predictable?** We expect air, background noise and average noise to be the most predictable whilst the loud short event noises will be random.
- **Are the air and noise monitors at the same area correlated?** In one instance, they are both in Camden Borough so we expect some correlation.
- **How reliable and informative are our proposed prediction intervals?** It is expected that conformal prediction intervals will achieve stated coverage levels while remaining sufficiently narrow.

5.1 Baseline results

To evaluate the performance of the ML models, we employed Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and symmetric Mean Absolute Percentage Error (sMAPE). MAE measures the average magnitude of the prediction errors, while RMSE gives greater weight to larger errors and is therefore more sensitive to substantial deviations. sMAPE expresses the prediction error as a symmetric percentage of the observed and predicted values, enabling comparisons across measurements with different scales. For all three metrics, lower values indicate better forecasting performance.

Across all three air quality monitors, the rolling mean typically outperforms the seasonal persistence baseline at both forecast horizons (see Table 3). For the 168-hour window, the rolling mean achieves lower MAE and RMSE overall. Similarly, for the 720-hour window, the rolling mean records lower MAE, RMSE, and sMAPE overall than the monthly seasonal persistence baseline.

Table 3: 168-hour and 720-hour window PM_{10} baseline forecasting performance of each monitor (all MAE and RMSE values in $\mu g/m^3$).

Baseline	AQ005			CHStG-AQMP1			WHD001		
	MAE	**RMSE**	**sMAPE**	**MAE**	**RMSE**	**sMAPE**	**MAE**	**RMSE**	**sMAPE**
168-hour rolling mean	**5.97**	**10.10**	**44.36**	**3.54**	**5.05**	**56.86**	**1.30**	**1.86**	106.08
Weekly seasonal persistence	8.19	12.13	61.09	5.73	10.58	69.07	2.47	5.85	**100.70**
720-hour rolling mean	**6.84**	**10.87**	**50.05**	**3.36**	**4.66**	**56.50**	**0.87**	**0.95**	**88.12**
Monthly seasonal persistence	9.38	13.48	61.12	6.18	11.46	71.09	2.25	5.40	97.50

Furthermore, the three noise monitors mostly suggest $L_{\mathrm{pAF,max}}$ as the hardest target to forecast, with the highest absolute errors in both baselines and both horizons (see Table 4). At TOS-N1 in Birmingham, the 168-hour rolling mean generally dominates across all targets. Contrastingly, at SE-NMP1 in Buckinghamshire, the weekly seasonal persistence is overall the best baseline. Interestingly, at N051 in Camden, the results are mixed where on the whole, 168-hour rolling mean is best for $L_{\mathrm{pAeq},T}$ and $L_{\mathrm{pAF,max}}$ but weekly seasonal persistence is better for$L_{\mathrm{pA90},T}$.

Table 4: 168-hour and 720-hour noise forecasting baseline performance for monitors SE-NMP1, TOS-N1, and N051 (all MAE and RMSE values in dB(A)).

Monitor	Baseline	$L_{\mathrm{pAeq},T}$			$L_{\mathrm{pA90},T}$			$L_{\mathrm{pAF,max}}$		
		MAE	**RMSE**	**sMAPE**	**MAE**	**RMSE**	**sMAPE**	**MAE**	**RMSE**	**sMAPE**
`SE-NMP1`	Weekly seasonal persistence	**5.32**	**7.26**	**12.19**	**4.08**	5.17	**11.33**	**6.88**	**8.96**	**11.26**
	Monthly seasonal persistence	7.44	9.80	16.32	4.25	5.55	11.77	8.87	11.18	14.08
	168-hour rolling mean	8.60	10.04	19.26	4.25	5.08	12.13	10.49	12.09	16.56
	720-hour rolling mean	8.53	10.18	19.10	4.20	**5.07**	12.00	10.35	12.26	16.34
`TOS-N1`	Weekly seasonal persistence	3.39	4.17	5.41	3.31	4.15	5.55	6.26	8.42	7.84
	Monthly seasonal persistence	4.92	6.01	7.78	4.60	5.59	7.60	8.90	11.16	11.20
	168-hour rolling mean	**2.98**	**3.63**	**4.62**	**3.20**	**3.88**	**5.22**	**4.63**	**5.82**	**5.64**
	720-hour rolling mean	3.42	4.25	5.31	3.50	4.28	5.70	5.99	6.96	7.31
`N051`	Weekly seasonal persistence	2.53	3.36	**3.77**	**2.77**	**3.64**	**5.68**	7.81	9.99	8.58
	Monthly seasonal persistence	2.75	3.54	4.11	4.24	5.32	8.58	8.18	10.44	8.98
	168-hour rolling mean	**2.51**	**3.31**	3.78	4.91	5.82	9.97	**6.90**	**7.96**	**7.64**
	720-hour rolling mean	2.54	3.33	3.83	5.16	6.09	10.46	6.92	7.96	7.66

5.2 Air forecasting

At the one-hour horizon, gradient boosting and linear regression are generally the best models for PM_{10} forecasting whilst ARIMA and ARIMAX mostly underperform (see Table 5). Overall, gradient boosting achieves the best MAE, RMSE, sMAPE at both AQ005 and WHD001. Linear regression is typically the best performing at CHStG-AQMP1 where it achieves the lowest MAE and RMSE of any model across the three monitors. For the most part, ARIMA and ARIMAX both produce similarly higher errors. Thus altogether, incorporating exogenous features in ARIMAX does not provide meaningful improvement over ARIMA.

Table 5: One-hour-ahead PM_{10} forecasting performance of each monitor (all MAE and RMSE values in μg/m^3).

Model	`AQ005`			`CHStG-AQMP1`			`WHD001`		
	MAE	**RMSE**	**sMAPE**	**MAE**	**RMSE**	**sMAPE**	**MAE**	**RMSE**	**sMAPE**
Linear Regression	3.11	6.38	23.82	**1.63**	**2.83**	**26.27**	2.41	4.75	96.33
Gradient Boosting	**2.56**	**5.04**	**19.45**	3.37	8.76	37.37	**1.83**	**4.42**	**79.82**
ARIMA	6.46	10.84	47.80	4.59	10.46	53.58	2.31	4.89	97.00
ARIMAX	6.27	10.57	46.80	4.68	10.44	59.58	3.54	5.95	107.54

5.3 Noise forecasting

Gradient boosting is generally the best model at the one-hour horizon across all three monitors and noise targets (see Table 6). Linear regression is usually consistently second for $L_{\mathrm{pA90},T}$ and $L_{\mathrm{pAeq},T}$. Overall, at every monitor $L_{\mathrm{pAF,max}}$ is the toughest target with high absolute errors that are similar across models. ARIMA typically produced the poorest performance, particularly at SE-NMP1, although integrating exogenous features in ARIMAX improved accuracy at this site, this was not the case at TOS-N1 where instead the performance reduced.

Table 6: One-hour-ahead noise forecasting performance for monitors N051, SE-NMP1, and TOS-N1 (all MAE and RMSE values in dB(A)).

Monitor	Model	$L_{\mathrm{pAeq},T}$			$L_{\mathrm{pA90},T}$			$L_{\mathrm{pAF,max}}$		
		MAE	**RMSE**	**sMAPE**	**MAE**	**RMSE**	**sMAPE**	**MAE**	**RMSE**	**sMAPE**
N051	Linear Regression	1.94	3.31	2.94	1.25	2.12	2.58	6.79	9.50	7.52
	Gradient Boosting	**1.62**	**2.21**	**2.41**	**1.20**	**1.57**	**2.45**	6.24	7.40	6.90
	ARIMA	2.52	3.31	3.79	4.36	5.44	8.88	6.42	7.50	7.10
	ARIMAX	2.05	2.66	3.07	3.04	3.86	6.08	**6.16**	**7.40**	**6.80**
SE-NMP1	Linear Regression	2.88	4.63	7.11	1.69	2.86	4.87	5.12	8.46	9.08
	Gradient Boosting	**2.31**	**3.04**	**5.28**	**1.44**	**1.97**	**3.99**	**4.18**	**5.46**	**6.90**
	ARIMA	8.37	9.69	18.73	4.10	4.96	11.67	9.97	11.53	15.73
	ARIMAX	5.11	6.29	11.51	3.06	4.00	8.63	6.30	7.92	10.16
TOS-N1	Linear Regression	1.39	2.37	2.18	1.04	1.76	1.71	4.33	6.73	5.35
	Gradient Boosting	**1.22**	**1.61**	**1.89**	**1.02**	**1.35**	**1.66**	**3.67**	**4.65**	**4.46**
	ARIMA	2.87	3.57	4.43	3.52	4.30	5.74	5.01	6.40	6.14
	ARIMAX	3.14	3.84	4.97	4.47	5.31	7.53	4.27	5.45	5.21

Figure 3 shows the ARIMAX prediction results for $L_{\mathrm{pAeq},T}$ at TOS-N1 throughout 1/2026. The predicted pattern broadly captures the true daily measure. Despite this, it generally underestimates the true values in the first week.

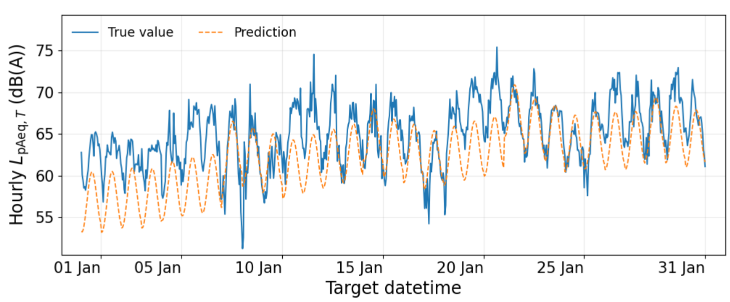

Fig. 3: Actual and ARIMAX predicted $L_{\mathrm{pAeq},T}$ noise levels for Birmingham City Council monitor TOS-N1.

5.4 Air and noise correlation

The two monitors assessed were in Camden. Thus, we initially expected air and noise to correlate on hourly and weekly scales which would have shown up as shared peak periods. However, the peaks did not overlap. PM_{10} peaks during 7:00-17:00 whereas noise peaks from 5:00 to late in the evening. Additionally, the weekday and weekend contrast was strong for air but weak for noise. Furthermore, PM_{10} against $L_{\mathrm{pA90},T}$, $L_{\mathrm{pAeq},T}$ and $L_{\mathrm{pAF,max}}$ give Spearman's correlation rank results ≤ 0.15 (see Figure 4). Thus, knowing the PM_{10} value at a given hour indicated nothing about the noise levels at that hour and vice versa.

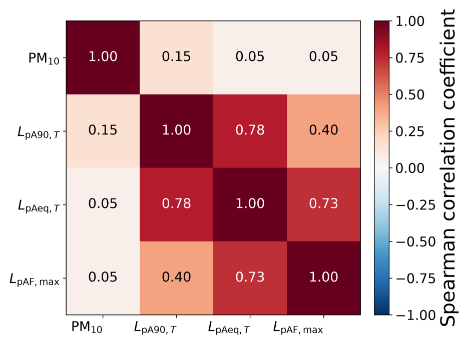

Fig. 4: Spearman correlation between air quality and noise levels (AQ005 & N051, 8/2025-1/2026).

5.5 Evaluation of conformal prediction intervals

Traditional ML models provide a single predicted value for each future time point, without any uncertainty associated with each prediction. Thus, we evaluate our proposed Conformal Prediction intervals to assess both their reliability and efficiency.

A rolling 30-day calibration window was used to estimate predictive uncertainty from recent forecasting errors. Nineteen miscoverage levels were evaluated:

$$\mathcal{A} = \{0.05, 0.10, \ldots, 0.95\} \tag{4}$$

where $\alpha \in \mathcal{A}$ corresponds to a nominal coverage level of $1 - \alpha$, ranging from 5% to 95%. The prediction interval for hour t at nominal coverage level $1 - \alpha$ is denoted by $C_{t,1-\alpha}$.

Interval reliability was evaluated using the prediction interval coverage probability (PICP):

$$\mathrm{PICP}_{1-\alpha} = \frac{1}{n} \sum_{t=1}^{n} \mathbf{1}\left\{y_t \in C_{t,1-\alpha}\right\} \tag{5}$$

where n is the total number of predictions being evaluated, t indexes the prediction hours, and $\mathbf{1}\{\cdot\}$ equals one when the observed value lies within the corresponding prediction interval and zero otherwise. A reliable prediction interval should achieve a $\mathrm{PICP}_{1-\alpha}$ close to the nominal coverage level $1-\alpha$.

Interval efficiency was evaluated using the mean prediction interval width (MPIW):

$$\mathrm{MPIW}_{1-\alpha} = \frac{1}{n}\sum_{t=1}^{n}\left(U_{t,1-\alpha} - L_{t,1-\alpha}\right) \tag{6}$$

where $L_{t,1-\alpha}$ and $U_{t,1-\alpha}$ denote the lower and upper bounds of the prediction interval for hour t. A smaller $\mathrm{MPIW}_{1-\alpha}$ indicates a more informative interval, provided that the required coverage level is maintained. Together, $\mathrm{PICP}_{1-\alpha}$ and $\mathrm{MPIW}_{1-\alpha}$ were used to evaluate the reliability and efficiency of the conformal prediction intervals generated from the rolling calibration residuals.

The evaluation included all six monitors mentioned above. Four forecasting models were evaluated: linear regression (LNR), gradient boosting (GB), ARIMA, and ARIMAX. The target measurements included PM_{10}, $L_{\mathrm{pAeq},T}$, $L_{\mathrm{pA90},T}$, and $L_{\mathrm{pAF,max}}$.

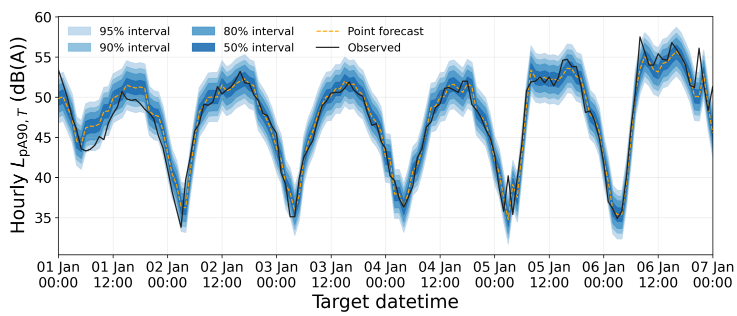

Fig. 5: Nested conformal prediction intervals for hourly $\mathrm{L}_{\mathrm{pA90},T}$ forecasts at monitor `N051` in London Camden from 1–7 January 2026. The gradient boosting model captured the daily noise pattern, with 90% and 95% PICP values of 88.84% and 95.03%, respectively.

The one-week example in Figure 5 provides a clear view of the nested intervals around the hourly forecasts. A clear coverage–width trade-off is demonstrated, with interval widths increasing consistently as the nominal coverage level rises.

The complete PICP and MPIW results are reported in Table 7. It is observed that the noise intervals generally achieved coverage close to or above their nominal levels. For example, GB achieved 90% and 95% PICP values of 90.32% and 95.30% for $L_{\mathrm{pAeq},T}$ at `SE-NMP1`, and 90.05% and 95.97% at `TOS-N1`. Air-quality intervals, on the other hand, showed greater under-coverage, particularly at the 90% level. However, air-quality coverage consistently increased for the 95% intervals. Overall, ARIMAX provided the strongest coverage, whereas LNR and GB achieved a better balance between coverage and interval width.

Table 7: Conformal prediction intervals for noise monitors N051, SE-NMP1, and TOS-N1 in January 2026. PICP is reported as a percentage and MPIW in dB(A).

Monitor	Model	$L_{\mathrm{pAeq},T}$				$L_{\mathrm{pA90},T}$				$L_{\mathrm{pAF,max}}$			
		90% interval		**95% interval**		**90% interval**		**95% interval**		**90% interval**		**95% interval**	
		PICP	**MPIW**	**PICP**	**MPIW**	**PICP**	**MPIW**	**PICP**	**MPIW**	**PICP**	**MPIW**	**PICP**	**MPIW**
N051	LNR	89.38	7.80	93.95	10.89	**90.19**	**4.87**	**95.70**	7.06	89.92	24.84	94.35	30.15
	GB	90.19	**6.77**	94.09	**9.27**	88.84	4.91	95.03	**6.28**	**90.19**	23.38	95.03	**26.89**
	ARIMA	**90.59**	11.22	94.76	13.45	88.31	18.14	94.49	21.13	88.98	**22.85**	**95.16**	27.00
	ARIMAX	89.52	8.45	**95.16**	10.77	87.50	12.50	94.76	15.24	90.05	23.52	94.09	27.65
SE-NMP1	LNR	92.34	12.87	95.83	16.32	90.46	6.93	95.03	9.58	92.20	21.76	95.97	27.49
	GB	90.32	**10.45**	95.30	**13.26**	88.58	**6.18**	94.62	**7.71**	93.68	**20.14**	97.18	**24.45**
	ARIMA	83.74	26.92	91.53	31.25	88.58	15.52	94.49	18.23	88.71	34.62	93.01	39.11
	ARIMAX	**95.30**	23.81	**97.04**	27.82	**92.47**	14.23	**96.10**	17.65	**93.95**	28.96	**97.85**	35.58
TOS-N1	LNR	90.99	5.61	95.70	7.09	88.17	**3.68**	94.62	5.13	94.09	**19.50**	96.37	24.32
	GB	90.05	**5.14**	95.97	**6.65**	86.69	3.94	93.82	**5.07**	96.64	20.05	98.39	**23.44**
	ARIMA	94.49	13.82	98.66	17.57	88.04	13.36	94.62	15.93	97.18	28.72	98.66	32.22
	ARIMAX	**96.91**	15.14	**99.73**	18.20	**89.52**	17.51	**94.89**	20.20	**98.12**	25.87	**99.46**	30.11

5.6 Discussion

A consistent finding from all our initial hypotheses is that the predictability of a target is determined by how cyclical or event driven it is. Thus, we find that background noise $L_{\mathrm{pA90},T}$ is the most forecastable target, reaching $R^2 = 0.73$ with gradient boosting, because it is stable and strongly cyclical so its historical value holds real meaning. The average $L_{\mathrm{pAeq},T}$ is only partially predictable and the loud short event is essentially unpredictable since they are random in real-world construction sites. Similarly, the best model for PM_{10} barely matches the mean and its largest errors trace to a pollution episode rather than construction. Furthermore, the 1 hour persistence baseline is very strong for the cyclical targets PM_{10} $R^2 = 0.82$ and $L_{\mathrm{pA90},T}$ $R^2 = 0.84$ however, a 1-hour lead time is not useful in practice. Surprisingly, the air and noise data were not correlated.

Overall, the results indicate that meaningful short-term forecasting is attainable. However, predictability varies between targets. For noise, RMSE scores of 1.61 for $L_{\mathrm{pAeq},T}$ and 1.35 for $L_{\mathrm{pA90},T}$ were achieved with gradient boosting at TOS-N1. Overall, this suggests that predicting $L_{\mathrm{pAeq},T}$ and $L_{\mathrm{pA90},T}$ provides reliable results. Contrastingly, $L_{\mathrm{pAF,max}}$ is typically the toughest noise target to forecast, its overall best RMSE 4.65 with gradient boosting at TOS-N1. Moreover, we found that PM_{10} reaches its best RMSE of 2.83 at CHStG-AQMP1 with linear regression. Despite this, even at its best performance across all monitors with gradient boosting and linear regression, PM_{10} sMAPE values range 19.45-79.82 %. Where as across all noise targets sMAPE values range 1.66-6.90 %. Overall, it is evident that gradient boosting is largely the strongest model, followed by linear regression. Surprisingly, the air and noise data were generally not correlated.

6 Conclusion

We set out to determine whether HS2 construction-period air quality and noise monitoring data can be forecast. This was motivated by the limitation of current environmental monitoring only reporting exceedances after they have occurred. Thus, we compared statistical baselines with ML models to assess the predictability of different air quality and noise indicators. Strikingly, we found no holistic correlation between air and noise measures. Additionally, we provided reliable predictions with statistical coverage guarantees. This work will potentially enable HS2 construction site teams to initiate mitigation strategies to meet environmental obligations. Future work may be directed by incorporating additional pollutants such as $PM_{2.5}$, NO_2, NO_x, and adaptive conformal methods across more sites and longer monitoring periods.

Acknowledgements

Khuong An Nguyen is supported by the UKRI DAFNI Fellowship grant, and Innovate UK grant for the 'TunnelGuard: A Ground-Air Robotic Safety Twin for HS2 Construction' project.

References

1. Al-Eidi, S., Amsaad, F., Darwish, O., Tashtoush, Y., Alqahtani, A., Niveshitha, N.: Comparative analysis study for air quality prediction in smart cities using regression techniques. IEEE Access **11**, 115140–115149 (2023). https://doi.org/10.1109/ACCESS.2023.3323447
2. Anghel, A., Papandreou, N., Parnell, T.P., Palma, A.D., Pozidis, H.: Benchmarking and optimization of gradient boosting decision tree algorithms. arXiv preprint arXiv:1809.04559 (2018), https://arxiv.org/abs/1809.04559
3. di He, H., Li, M., li Wang, W., yong Wang, Z., Xue, Y.: Prediction of pm2.5 concentration based on the similarity in air quality monitoring network. Building and Environment **137**, 11–17 (2018). https://doi.org/https://doi.org/10.1016/j.buildenv.2018.03.058
4. Hyndman, R.J., Athanasopoulos, G.: Forecasting: Principles and Practice. OTexts, Melbourne, Australia, 3rd edn. (2021), https://otexts.com/fpp3/
5. Islam, F., Imteaz, M.A.: Use of teleconnections to predict western australian seasonal rainfall using arimax model. Hydrology **7**(3) (2020), https://www.mdpi.com/2306-5338/7/3/52
6. Khan, S., Alghulaiakh, H.: Arima model for accurate time series stocks forecasting. International Journal of Advanced Computer Science and Applications **11**(7) (2020). https://doi.org/10.14569/IJACSA.2020.0110765
7. Liu, K., Xu, J., Zhang, F., Meng, Z., Hu, X., Wang, Z.: Coupled impacts of road construction and vehicle emissions on spatiotemporal dynamics of roadside particulate matter pollution. Water, Air, & Soil Pollution **237**(8), 495 (2026). https://doi.org/10.1007/s11270-025-09068-7
8. Roustaei, N.: Application and interpretation of linear-regression analysis. Medical Hypothesis, Discovery & Innovation in Ophthalmology **13**(3), 151–159 (2024). https://doi.org/10.51329/mehdiophthal1506

9. Samad, A., Garuda, S., Vogt, U., Yang, B.: Air pollution prediction using machine learning techniques – an approach to replace existing monitoring stations with virtual monitoring stations. Atmospheric Environment **310**, 119987 (2023). https://doi.org/https://doi.org/10.1016/j.atmosenv.2023.119987
10. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World. Springer, New York (2005). https://doi.org/10.1007/b106715
11. Wen, P.J., Huang, C.: Noise prediction using machine learning with measurements analysis. Applied Sciences **10**(18) (2020). https://doi.org/10.3390/app10186619
12. Yasin, K.H., Yasin, M.I., Iguala, A.D., Gelete, T.B., Tulu, D., Kebede, E.: Predictive machine learning and geospatial modeling reveal pm10 hotspots and guide targeted air pollution interventions in addis ababa, ethiopia. Discover Applied Sciences **7**(4), 263 (2025). https://doi.org/10.1007/s42452-025-06723-w
13. Zaffran, M., Féron, O., Goude, Y., Josse, J., Dieuleveut, A.: Adaptive conformal predictions for time series. In: Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 25834–25866. PMLR (2022), https://proceedings.mlr.press/v162/zaffran22a.html
14. Zhang, X., Zhao, M., Dong, R.: Time-series prediction of environmental noise for urban iot based on long short-term memory recurrent neural network. Applied Sciences **10**(3) (2020). https://doi.org/10.3390/app10031144