Group 3 Studies: AQI Data Analysis

1. Context

Our team supports air-quality monitoring and decision-making for provincial and municipal authorities responsible for advisories, resource allocation, and mitigation design. Prior analysis demonstrates that AQI dynamics across Ontario are systematically structured along two dimensions. At the daily level, lagged traffic and meteorological variables explain a stable portion of next-day variation. At the intraday level, pollution follows a consistent temporal pattern—morning peak (6:00–8:00 AM), midday trough, and evening rise—with higher weekday levels and strong seasonal amplification, particularly in winter and summer. Hierarchical and nonlinear models confirm that these effects remain statistically significant after accounting for station heterogeneity, indicating stable structural drivers rather than random noise. Building on this evidence, the present analysis extends from explanation to operational forecasting. We evaluate whether validated traffic, meteorological, and temporal effects can support reliable one-step-ahead AQI prediction and which modeling approach best balances predictive performance, interpretability, and deployability. Forecasting is thus framed as an operational extension of structural analysis, supporting precision, time-targeted environmental management and evidence-based policy intervention.

2. Research Questions and Business Meaning

Q1 (Integrated Daily Effects)

Research Question

Across Ontario monitoring locations, how are daily AQI levels associated with daily traffic volume after controlling for meteorological factors (temperature, wind, humidity), and does this association vary by season?

Business Meaning

Understanding the magnitude and seasonal stability of the traffic–pollution relationship informs whether traffic management policies (e.g., congestion pricing, remote-work incentives, peak-hour regulation) can generate measurable next-day AQI improvements. If the association is systematic and stable, it provides an empirical foundation for forecast-driven traffic intervention strategies.

Q2 (Temporal Structure and Heterogeneity)

Research Question

How do AQI concentrations vary by hour of day and weekday versus weekend, and do these temporal patterns differ across seasons or stations?

Business Meaning

Identifying intraday and seasonal exposure windows enables time-targeted environmental management. If pollution consistently peaks during specific hours or seasons, advisories and mitigation measures can be deployed with greater precision rather than uniformly across time. Understanding station-level heterogeneity further supports geographically adaptive policy design.

Forecasting Extension

Research Question

Given the empirically validated daily and temporal structure of AQI dynamics, to what extent can these relationships support a reliable one-step-ahead forecasting framework?

Business Meaning

If structural effects identified in Q1 and Q2 are sufficiently stable, they can be leveraged to build operational next-day forecasting systems. The policy decision then becomes whether interpretable linear models are adequate for deployment or whether additional model complexity meaningfully improves predictive performance without sacrificing stability and transparency.

3. Executive Summary

3.1 Structural and Predictive Evidence

This report integrates next-day forecasting with hierarchical time-structure analysis. Results show that AQI across Ontario is systematically structured rather than random.

Lagged traffic and meteorological variables explain a stable portion of next-day variation. Linear models achieve the lowest out-of-sample error, while nonlinear approaches provide no meaningful improvement. Under the current feature set, the dominant daily signal is approximately linear.

Diurnal and seasonal patterns remain strong, including a morning peak (6:00–8:00 AM), a midday trough, an evening rise, higher weekday levels, and elevated winter risk. Traffic activity and atmospheric conditions jointly shape pollution dynamics.

3.2 Model Constraints

Forecast errors are aligned across model families and concentrated on extreme-event days, suggesting missing environmental drivers rather than algorithmic limitations. Model performance is therefore constrained more by feature completeness than by model complexity.

3.3 Operational Implications

Linear models (OLS or ElasticNet) provide competitive accuracy, interpretability, and stability, making them suitable as production baselines. Forecast-driven, time-specific interventions—particularly during winter weekday mornings—can enhance advisory precision and traffic-related mitigation strategies.

Further predictive gains depend primarily on expanding environmental feature coverage rather than increasing algorithmic complexity.

4. Statistical Framework

4.1 Data Construction and Integration

The analysis focuses on Ontario monitoring stations to ensure relevance to our operational scope. We restrict the sample to April 26, 2022 – September 26, 2024, the period with consistent traffic coverage. Including dates outside this window would substantially increase missing data and weaken the reliability of the results.

Air quality readings are summarized at the daily level to align with traffic and weather data and to reduce short-term volatility that is less relevant for policy planning. Key meteorological variables—such as temperature, precipitation, and wind conditions—are included because weather plays a central role in how pollutants disperse. Traffic and weather

information are linked to each air quality station using nearby monitoring sources within a 10 km radius, balancing local relevance with data availability.

To maintain data quality, variables with severe missingness were excluded, and remaining gaps were handled in a way that preserves the overall sample size. Persistent differences across monitoring stations are accounted for in the modeling process. In forecasting exercises, only past information is used to predict future AQI values, ensuring realistic out-of-sample evaluation.

4.2 Model Fitting

4.2.1 Models for Q1: Integrated Daily Effects

For our first research question, we implemented a one-step-ahead forecasting framework using lagged daily predictors. A time-based split was used to prevent look-ahead bias, and hyperparameters were selected using rolling cross-validation (TimeSeriesSplit). Station ID was encoded categorically to capture persistent cross-station baseline differences analogous to fixed effects. We compared linear models (OLS, ElasticNet, Neural Networks) against nonlinear learners (XGBoost) to evaluate whether additional flexibility improves out-of-sample performance.

4.2.2 Models for Q2: Temporal Structure and Heterogeneity

For our second research question, we adopted a hierarchical modeling strategy to isolate structural time effects. A mixed-effects model incorporated station as a random intercept to quantify between-station heterogeneity while estimating fixed effects for season and weekday. To capture nonlinear diurnal structure, a Generalized Additive Model (GAM) was fitted with smooth hourly terms and hour–season interactions. Finally, XGBoost was used to assess short-term predictive capability under nonlinear feature interactions. This multi-model approach allows us to distinguish linear structural relationships from nonlinear seasonal modulation while evaluating predictive robustness.

5 Results

5.1 Daily Traffic–Pollution Association

Across five held-out stations and a three-month test horizon, the linear OLS model achieved the lowest RMSE and MAE among all candidate models. ElasticNet performed nearly identically, while XGBoost, and Neural Networks failed to deliver meaningful improvements in out-of-sample error.

	model	RMSE_test	MAE_test	n_eval
0	OLS	11.958300	9.057869	3633
1	ElasticNet	12.023960	9.102588	3633
2	XGBoost	13.537576	10.273238	3633
3	Naive (y_hat = y_lag1)	13.753437	9.804844	3633
4	MLP	14.058493	10.394399	3633

This finding is analytically important: despite allowing for nonlinear interactions and high model flexibility, complex models did not outperform a linear specification. This indicates that, conditional on lagged traffic and meteorological variables, the dominant next-day AQI signal is approximately linear. The bias–variance trade-off favors lower-variance linear estimators under the current feature set. Note that In prior exploratory work, we also tested using a first-difference target (ΔAQI) instead of level AQI, the qualitative ranking and conclusions remained similar.

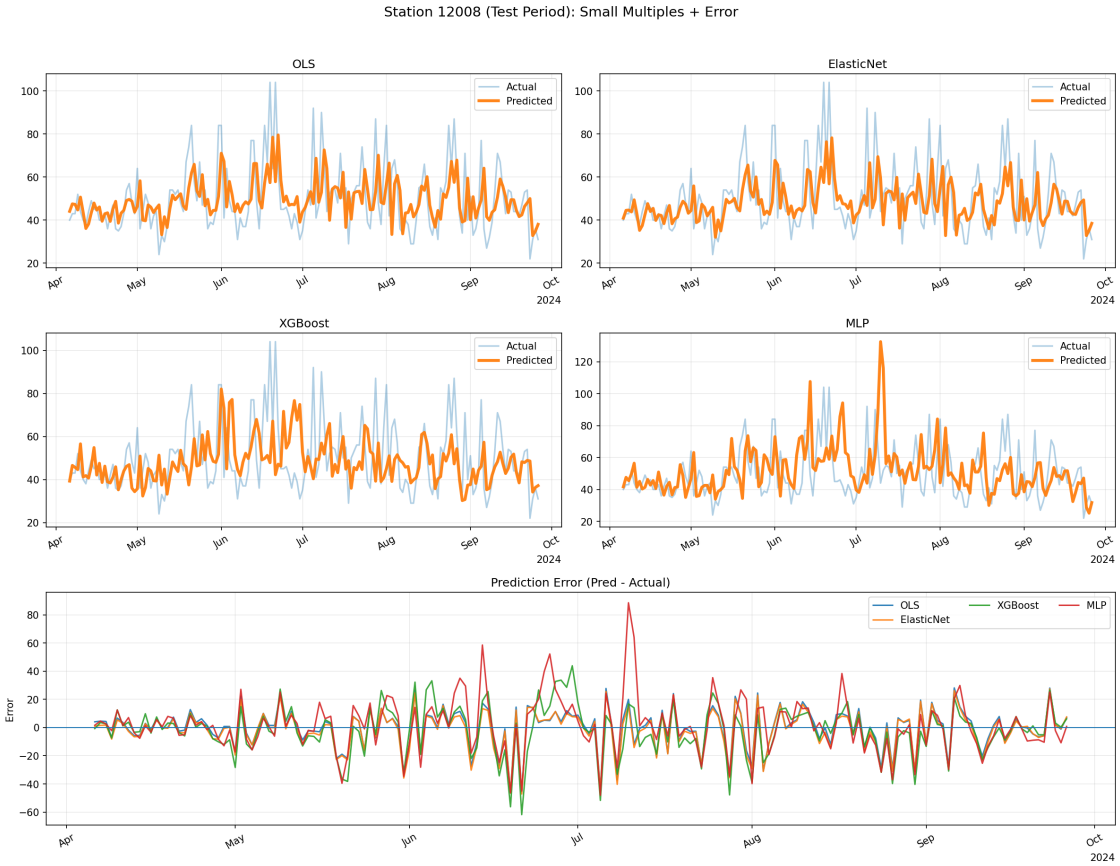

Figure 1: Station 12008: forecasts vs actual (top) and prediction errors (bottom) over the test period.

Station-level diagnostics reinforce this conclusion. In this plot we picked one of the stations and showed the predicted vs actual AQI. We can see that prediction errors across model families exhibit strong temporal alignment, with error spikes occurring simultaneously. This pattern implies shared structural blind spots—likely extreme-event days or unobserved transport phenomena—rather than model mis-specification. Increasing algorithmic complexity therefore cannot compensate for missing explanatory drivers.

From a policy perspective, the results provide empirical evidence that:

- The next-day AQI is highly correlated with the current-day AQI. We can see even our best model (OLS) do behave not much better than the naive model.
- There are strong linear relationship between the current-day weather & traffic vs next-day AQI. And complex models like NN and XGBoost capture that relationship worse than simple models like OLS.

This strengthens the case for deploying linear forecasting models in operational settings.

5.2 Intraday, Weekly, and Seasonal Effects

Hourly AQI demonstrates a highly structured diurnal cycle consistent across 39 monitoring stations:

- A pronounced morning peak (6:00–8:00 AM)
- A midday trough (12:00–15:00)
- A secondary evening rise (18:00–21:00)

The morning peak represents the largest intraday deviation from baseline and aligns temporally with commuting activity.

The mixed-effects model confirms these patterns remain statistically significant after controlling for station heterogeneity (between-station variance 4.85). Relative to fall:

- Spring AQI is 0.75 units lower ($p < 0.001$)
- Summer AQI is 1.23 units higher ($p < 0.001$)
- Winter AQI is 0.66 units higher ($p < 0.001$)

Weekends are associated with a 1.05-unit reduction relative to weekdays ($p < 0.001$).

These are not marginal differences—they are systematic, highly significant, and persistent after spatial adjustment. The weekday effect directly supports the interpretation that anthropogenic activity, particularly traffic, materially elevates pollution levels.

The GAM further reveals strong nonlinear hourly effects and significant hour–season interactions. In practical terms, the shape and amplitude of the daily AQI curve differ meaningfully across seasons. Winter mornings exhibit both higher levels and steeper gradients, suggesting reduced dispersion and amplified emission impact.

Predictive modeling at the hourly level achieved an out-of-sample RMSE of approximately 3.5 AQI units, demonstrating that nonlinear interactions can be exploited for operational forecasting.

Taken together, the evidence shows:

1. AQI dynamics are structurally nonlinear over the day.
2. Weekday commuting materially elevates pollution.
3. Seasonal atmospheric conditions amplify baseline risk.
4. Spatial heterogeneity affects magnitude but not structural pattern.

These findings strongly corroborate and complement the Q1 daily results: traffic effects are observable both in next-day forecasting and within-day diurnal structure.

6 Risk and Limitations

The analysis focuses on April 2022 to September 2024, the period with reliable traffic coverage. While this ensures consistency across datasets, the findings reflect conditions during this specific window and may not capture longer-term shifts. Traffic and weather data are linked to air quality stations using nearby sources within a 10 km radius. This improves local relevance but remains an approximation, introducing some unavoidable measurement noise.

Also, pollutants such as CO and SO_2 exhibit structural missingness and were excluded from temporal modeling, limiting pollutant diversity. Second, extreme-event days produce synchronized forecast errors across models, indicating missing drivers such as wildfire smoke transport or regional inflow. Third, residual heteroskedasticity suggests that volatility increases during peak seasons, potentially understating uncertainty under linear assumptions.

Finally, model performance is constrained more by feature completeness than algorithm choice. Without incorporating additional environmental or transport covariates, predictive gains will remain incremental.

7 Recommendations

7.1 Standardize a Production Forecasting Model

Deploy OLS/ElasticNet as the official next-day AQI forecasting baseline. It delivers the best out-of-sample performance with full interpretability and low operational complexity. The production forecasting model should be operationalize the forecasting model through automated daily retraining with a rolling window, continuous performance monitoring (e.g., RMSE drift and residual checks), and predefined forecast thresholds that directly trigger air-quality advisories and targeted mitigation actions. This establishes a stable analytics foundation for municipal and provincial advisory systems.

7.2 Target High-Risk Time Windows

Winter weekday mornings (6:00–8:00 AM) represent the most structurally elevated pollution window. Interventions should be time-specific rather than broad and seasonal. Interventions should be forecast-driven: trigger targeted advisories when predicted AQI exceeds defined risk bands, coordinate short-term traffic mitigation on anticipated peak days, and pilot staggered work-hour programs in high-density corridors to reduce morning congestion impacts. This aligns intervention intensity with statistically identified risk periods.

7.3 Integrate Traffic Policy with Environmental Impact Modeling

The persistent weekday effect confirms measurable anthropogenic influence. Analytics teams should support scenario modeling to quantify AQI impact under traffic reductions (e.g., 5–10% peak-hour decrease). This enables evidence-based evaluation of congestion pricing, remote-work incentives and traffic flow optimization. Policy discussions should be supported by modeled air-quality impact estimates rather than descriptive trends.

7.4 Strengthen Extreme-Event Detection

Forecast errors cluster around unobserved episodic events. Integrating wildfire smoke indicators and regional transport metrics should be prioritized to improve advisory precision during high-volatility periods.

Overall, the evidence supports a shift toward precision, time-targeted, and seasonally adaptive air-quality management. Structured forecasting combined with traffic-linked intervention modeling will enable data-driven environmental decision-making at both municipal and provincial levels.