Air Quality Analysis

Supporting Proactive Air-Quality Decisions in Ontario Schools

Junbo Li, Yikai Ma, Yuhan Fu, Rilong Zhang, Yiming Shen

February 13, 2026

Climate-driven increases in wildfire activity have led to more frequent short-term spikes in fine particulate matter (PM2.5), posing growing risks to children's health and challenging the ability of education systems to respond proactively. This study evaluates whether next-day (24-hour mean) PM2.5 concentrations in Ontario can be predicted using routinely available meteorological and traffic-related data. We construct a leakage-safe, region-level dataset combining PM2.5 measurements with cleaned and spatially aligned weather observations. A transparent region-wise LASSO baseline is compared with a nonlinear Long Short-Term Memory (LSTM) model under a strict time-aware evaluation framework. While the LASSO captures linear and seasonal effects (mean test R^2 0.25), the LSTM substantially improves predictive performance ($R^2 = 0.36$), outperforming persistence baselines. Results demonstrate the feasibility of policy-ready next-day PM2.5 forecasting to support proactive air-quality decision-making in schools.

Table of contents

1 Introduction

Over the past several decades, advances in environmental regulation, emissions controls, and monitoring technologies have led to substantial improvements in air quality across many regions. In high-income countries, long-term declines in fine particulate matter (PM2.5) concentrations have contributed to reductions in pollution-related morbidity and mortality, representing a major public-health achievement, particularly for children (World Health Organization 2021).

Despite this progress, recent years have seen stagnation and episodic reversals in air-quality improvements. Since approximately 2015, climate-driven factors—most notably increasingly severe and frequent wildfires—have emerged as a dominant source of extreme PM2.5 exposure (Reid et al. 2016). In Canada, wildfire smoke has repeatedly affected population centres far from fire origins, including the Greater Toronto Area and Ottawa–Gatineau, resulting in prolonged periods of hazardous air quality that strain existing public-health response systems.

The health and operational impacts of poor air quality are unevenly distributed. Children are particularly vulnerable due to developing respiratory systems and higher relative inhalation rates, and exposure to elevated PM2.5 is strongly linked to adverse respiratory and cardiovascular outcomes (Brook et al. 2010). At the institutional level, preparedness also varies substantially across regions. Many education systems continue to rely on same-day or reactive air-quality advisories, leaving limited time for schools to adjust outdoor activities, ventilation strategies, transportation logistics, and communications. This reactive posture increases disruption and limits the effectiveness of mitigation during air-quality emergencies.

Short-term air quality is shaped by a combination of atmospheric conditions and human activity. Meteorological variables such as wind, temperature, humidity, and precipitation influence pollutant dispersion and accumulation, while traffic remains a major local contributor to PM2.5 concentrations. Prior work suggests that incorporating meteorological and traffic information can improve short-horizon air-quality forecasts beyond persistence and classical time-series approaches (Cheng, Li, et al. 2021). However, the policy relevance of these improvements—particularly for next-day decision-making in education systems—remains insufficiently examined.

In this study, we assess whether next-day PM2.5 (24-hour mean) can be reliably predicted using routinely available traffic and weather data, and whether these forecasts can be translated

into actionable next-day air-quality alerts. The primary stakeholder for this work is Provincial Ministries of Education, particularly policy, analytics, and student well-being units responsible for issuing guidance to school boards during environmental health events. We find that incorporating traffic and meteorological predictors improves forecasting performance relative to persistence and classical baselines and supports the construction of transparent alerting systems that balance recall and precision.

The estimand of this study is the effect of lagged air-quality measures, meteorological variables, and traffic indicators on next-day PM2.5 concentrations at the regional level. By quantifying these relationships, this analysis evaluates the feasibility of a short-horizon, policy-ready forecasting and alerting framework to support proactive and equitable decision-making during air-quality emergencies affecting schools.

The remainder of this paper is organized as follows. Section 2 describes the data sources, spatial alignment, and preprocessing steps. Section 3 outlines the modeling and validation strategy. Section 4 presents forecasting and alerting performance results. Finally, Section 5 discusses policy implications, limitations, and directions for future work.

2 Data

2.1 Overview

This study integrates air quality and meteorological data for Ontario to construct an analysis-ready dataset for next-day PM2.5 forecasting. All datasets are publicly available and provided by provincial or federal agencies under open data licenses. Data access and preprocessing were designed to be reproducible, with scripted workflows used wherever possible.

Hourly PM2.5 concentration data were obtained from Air Quality Ontario, operated by the Ontario Ministry of the Environment, Conservation and Parks (MECP) (Government of Ontario 2025). The dataset contains historical PM2.5 measurements collected at fixed ambient air monitoring stations across the province. Data are accessible through a web-based interface that supports CSV downloads by station and time range.

Meteorological data were obtained using scripted downloads from publicly available weather monitoring sources corresponding to Ontario stations (Environment and Climate Change Canada 2025). Variables include temperature, wind speed and direction, precipitation, and other atmospheric measures commonly used in air-quality modeling. Automated pipelines were used to retrieve multi-year weather data, standardize timestamp formats, and validate schema consistency across files.

Weather observations were temporally aligned with PM2.5 measurements at the daily level and spatially matched to air-quality monitoring locations or aggregated regions using geographic coordinates. Across both datasets, preprocessing included timestamp standardization, removal

or recoding of invalid values, and harmonization of spatial identifiers to support region-level analysis.

2.2 Weather Data Cleaning

Daily weather station data from 2020–2025 were cleaned into an analysis-ready station–day table and a station-level missingness summary. Cleaning steps included standardized column naming, resolution of duplicate fields, normalization of station identifiers, numeric coercion with invalid entries set to missing, and engineering of quality indicators from metadata flags. Physical plausibility bounds were applied to detect and flag outliers, and records were deduplicated to enforce one observation per station per day. A station-level missing-rate table was constructed to support data quality screening.

2.3 PM2.5 Data Cleaning

Hourly PM2.5 data from January 1, 2020 to December 31, 2024 were consolidated into a unified dataset with standardized formats. Daily PM2.5 concentrations were computed using available hourly measurements; days with entirely missing hourly values were retained as missing rather than imputed. Each monitoring station was mapped to its corresponding Ontario administrative region using official metadata, enabling aggregation from station-level readings to region-day averages.

The PM2.5 and weather datasets were merged by region and date. All region-day PM2.5 observations were retained; when weather data were unavailable for a given region-day, weather variables were recorded as missing rather than dropping the observation. This design ensures that pollution events are not systematically excluded during integration.

2.4 Exploratory Data Analysis

The integrated dataset contains 51,830 region-day observations with strong coverage; PM2.5 exhibits only 0.4% missingness. The distribution of daily PM2.5 is highly right-skewed, with occasional extreme values exceeding 400 $\mu g/m^3$, indicating that air-quality risk is driven by rare but severe events. Regional heterogeneity is pronounced, with southern and industrialized regions exhibiting higher average concentrations.

Temporal diagnostics reveal strong short-term persistence, with lag-1 autocorrelation around 0.72, and a moderate correlation between current and next-day PM2.5 ($r \approx 0.57$). Meteorological variables show meaningful associations: wind speed and precipitation are negatively correlated with PM2.5, while temperature is positively correlated. An Augmented Dickey–Fuller test rejects the unit-root hypothesis ($p < 0.001$), suggesting no long-term explosive trend.

These patterns motivate models that incorporate temporal dependence, seasonality, and meteorological drivers while remaining robust to heavy-tailed outcomes.

2.5 Feature Engineering

To capture temporal structure while strictly avoiding data leakage, all features were constructed using past information only. Lagged PM2.5 values and rolling statistics were generated to capture persistence and short-term trends. Aggregated and lagged weather variables were created to reflect delayed meteorological effects, and seasonal cycles were encoded using cyclical time features. Rows with incomplete feature histories were removed to preserve temporal integrity.

3 Model

3.1 LASSO Baseline Model

We implemented a time-aware, region-wise LASSO regression as a transparent linear baseline. Separate models were trained for each region to account for spatial heterogeneity. Predictors were standardized within a pipeline, and non-predictive identifiers were excluded.

A strict chronological split was applied within each region, with the first 80% of observations used for training and the remaining 20% reserved for testing. The regularization parameter was selected via five-fold cross-validation on the training set. The L1 penalty induces sparsity, enabling automatic feature selection and interpretability.

3.2 LSTM Model Training and Evaluation

A sequence-based LSTM model was trained using fixed-length historical windows (14 days) to capture temporal dependencies and nonlinear interactions. Data were split chronologically into training, validation, and test sets to prevent leakage. The model uses a two-layer recurrent encoder with dropout, optimized using AdamW with early stopping. The target variable was log-transformed during training to stabilize variance, and all metrics were computed on the original PM2.5 scale.

Baseline comparators—including persistence and window-mean predictors—were evaluated on the same test split to establish meaningful performance thresholds.

4 Results

4.1 LASSO Results

Across 26 regions, the LASSO baseline achieved a mean test R^2 of approximately 0.25 and a median R^2 of 0.22, indicating modest explanatory power. Average RMSE and MAE were approximately 2.9 µg/m³ and 2.2 µg/m³, respectively, with a mean Pearson correlation of 0.55. Performance varied substantially across regions, reflecting heterogeneous pollution dynamics.

The model produced sparse solutions, typically selecting several dozen predictors per region. Wind-related variables, precipitation, seasonal indicators, and lagged PM2.5 measures were consistently selected, aligning with known physical mechanisms. However, the limited predictive power suggests that linear models are insufficient to capture complex nonlinear and temporal interactions.

4.2 LSTM Results

The LSTM substantially outperformed all baselines, achieving MAE 1.96, RMSE 2.71, R^2 0.356, and Pearson r 0.598 on the test set. This represents a 12–19% reduction in MAE relative to persistence and window-mean baselines. Gains from extending the input window beyond 7 days were modest, indicating that most predictive information lies in recent history.

Permutation importance analysis shows that recent PM2.5 history dominates predictive performance, with meaningful secondary contributions from temperature, precipitation, wind, and seasonal features. Performance improvements are especially pronounced during elevated pollution periods, supporting the model's relevance for alerting applications.

5 Discussion

This study demonstrates that next-day PM2.5 concentrations can be predicted with meaningful accuracy using routinely available meteorological and traffic-related data, particularly when temporal dependence and nonlinear structure are explicitly modeled. While a region-wise LASSO baseline captures important linear and seasonal effects, its limited explanatory power underscores the necessity of sequence-based models for short-horizon forecasting.

From a policy perspective, the LSTM's improvement over persistence is critical. Reactive systems that rely solely on same-day measurements leave little time for institutional response. Even modest next-day forecasting skill enables proactive decisions regarding outdoor activities, ventilation strategies, transportation planning, and communication with families—especially in high-risk regions.

The results also highlight important equity considerations. Regions with consistently higher pollution levels may benefit disproportionately from proactive alerting systems, suggesting that forecasting tools should be integrated into differentiated preparedness strategies rather than uniform province-wide policies.

Several limitations warrant discussion. Forecast uncertainty increases during abrupt wildfire events and long-range smoke transport, where local meteorology alone may be insufficient. Additionally, the absence of explicit wildfire or satellite-based aerosol indicators likely constrains peak-event performance. Future work should integrate remote-sensing products, chemical transport model outputs, or wildfire plume indicators to improve tail-risk prediction.

Methodologically, further gains may be achieved by incorporating explicit region or station identifiers via embeddings, adopting loss functions that emphasize high-pollution days, and reporting stratified performance metrics across pollution regimes. Despite these limitations, the proposed framework demonstrates that policy-ready, leakage-safe next-day PM2.5 forecasting is feasible and can meaningfully support proactive decision-making in education systems.

References

Brook, Robert D., Sanjay Rajagopalan, C. Arden Pope, et al. 2010. "Particulate Matter Air Pollution and Cardiovascular Disease." *Circulation* 121 (21): 2331–78.

Cheng, Yun, Xia Li, et al. 2021. "Deep Learning for Air Quality Forecasting: A Review." *Atmospheric Environment* 262: 118600.

Environment and Climate Change Canada. 2025. "Historical Climate Data." Government of Canada. https://climate.weather.gc.ca/.

Government of Ontario. 2025. "Air Quality Ontario: Historical Air Quality Data." Ministry of the Environment, Conservation; Parks. https://www.airqualityontario.com/history/index.php.

Reid, Colleen E., Michael Brauer, Fay H. Johnston, Michael Jerrett, John R. Balmes, and Catherine T. Elliott. 2016. "Critical Review of Health Impacts of Wildfire Smoke Exposure." *Environmental Health Perspectives* 124 (9): 1334–43.

World Health Organization. 2021. "WHO Global Air Quality Guidelines." *WHO Guidelines*.