Air Quality Analysis

Findings Reveal Vaccination and Healthcare Spending as Key Drivers of Mortality Reduction

Junbo Li, Yikai Ma, Yuhan Fu, Ruolong Zhang,

February 13, 2026

123

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 become 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 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, and communications, thereby increasing disruption 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, 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 station or 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 by season and region. 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, subject to request limits.

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 scripts 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 steps included timestamp standardization, removal or recoding of invalid values, and harmonization of spatial identifiers to support station-level and region-level analysis. The resulting dataset provides a consistent foundation for evaluating next-day PM2.5 forecasting models and downstream alerting performance.

2.2 Cleaning and transformation

2.3 EDA

3 Model

3.1 LASSO Baseline Model

We implemented a time-aware, region-wise LASSO regression as a linear baseline for PM2.5 prediction. The response variable was daily PM2.5 concentration (pm25_label). To account for spatial heterogeneity and avoid information sharing across geographically distinct areas, separate models were trained for each region.

For each region, observations were ordered chronologically by date. Non-predictive identifiers, geographic coordinates, and the response variable were excluded from the feature set, and only numeric predictors were retained. Rows with missing values in either predictors or the response were removed to ensure a valid design matrix. All predictors were standardized using z-score normalization prior to model fitting. Standardization was performed within a scikit-learn pipeline to ensure that scaling parameters were learned exclusively from the training data, thereby preventing data leakage.

A strict time-based split was applied within each region, with the first 80% of observations used for training and the remaining 20% reserved for testing. Model estimation was performed using LASSO regression with the regularization parameter selected via five-fold cross-validation on the training set. The L1 penalty induces sparsity in the coefficient vector, enabling automatic feature selection while retaining interpretability. The optimal regularization strength varied across regions, reflecting differences in signal complexity and predictor relevance.

4 Reults

4.1 LASSO

Model performance was evaluated on the held-out test data using metrics computed on the original PM2.5 scale, including mean absolute error (MAE), root mean squared error (RMSE), coefficient of determination (R^2), mean absolute percentage error (MAPE), symmetric MAPE (sMAPE), and the Pearson correlation coefficient between predictions and observations.

Across the 26 regions, the LASSO baseline achieved a mean test R^2 of approximately 0.25 and a median R^2 of 0.22, indicating that the linear model explains a modest fraction of the variability in future PM2.5 concentrations. Average test RMSE and MAE were approximately 2.9 $\mu g/m^3$ and 2.2 $\mu g/m^3$, respectively, while the mean Pearson correlation between predicted and observed values was approximately 0.55, suggesting moderate agreement in temporal variation.

Predictive performance varied substantially across regions. Some regions exhibited relatively strong linear relationships, with test R^2 values exceeding 0.45, whereas others showed limited predictive skill, with R^2 values below 0.10. This variability highlights pronounced regional differences in PM2.5 dynamics, likely driven by differences in local emission sources, meteorological conditions, and spatial coverage of monitoring stations.

The LASSO models produced sparse solutions, selecting on average several dozen predictors per region, with the exact number depending on the chosen regularization strength. Predictors related to wind dynamics, precipitation, seasonal patterns, and regional PM2.5 aggregates were among the most consistently selected across regions. While the sparsity of the models supports interpretability and facilitates physical interpretation, the overall performance indicates that

important nonlinear relationships and temporal dependencies are not fully captured by a linear framework.

In summary, the region-wise LASSO provides a transparent and leakage-safe baseline that captures key linear and seasonal components of PM2.5 variation. Its limited explanatory power relative to more flexible models motivates the use of nonlinear and sequence-based approaches, such as the LSTM, for improved predictive performance.

5 Discussion

123

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*.