From EarthData to Action: Cloud Computing with Earth Observation Data for Predicting Cleaner, Safer Skies

Hassan Siddiqui,Muneeb Ahmed,Adil Usmani,Hasham Dogar, Hurr Ali Syed
AeroSphere Date: October 4, 2025

***Abstract*—Air quality degradation represents a critical global public health challenge, contributing to an estimated seven million premature deaths annually (WHO). Vulnerable populations bear disproportionate risks, while governments struggle to deliver timely, high resolution air quality information.**

This study presents a cloud native air quality forecasting system that integrates NASA's *Tropospheric Emissions: Monitoring of Pollution (TEMPO)* satellite observations with ground based networks (EPA AirNow, OpenAQ, Pandora, TOLNet) and meteorological reanalysis products to generate near real time, short term forecasts. At its core, the system leverages a Long Short Term Memory (LSTM) neural network orchestrated via Apache Airflow. GPT 4o mini enhances interpretability by producing natural language summaries of model outputs. Deployed on Microsoft Azure, the system demonstrates scalable, robust, and uncertainty aware forecasting.

Practical Summary: **The pipeline fuses satellite, ground, and meteorological data with AI to forecast city level air quality up to 72 hours ahead. Preliminary results show a mean absolute error (MAE) below 5 µg/m³ for $PM_{2.5}$ and an improvement of approximately 30% over persistence baselines. This enables communities and agencies to prepare proactively for pollution events.**

I. Introduction

Air quality degradation remains one of the most pressing environmental challenges, contributing to increased incidences of respiratory and cardiovascular diseases and substantial socioeconomic impacts. Traditional monitoring infrastructures, often sparse and localized, provide limited spatial and temporal resolution, leaving critical gaps in environmental surveillance and public health preparedness.

The launch of NASA's TEMPO geostationary spectrometer marks a major advancement in atmospheric monitoring. By providing hourly pollutant measurements across North America covering nitrogen dioxide (NO_2), ozone (O_3), and formaldehyde (HCHO) TEMPO enables unprecedented temporal granularity and spatial detail. Despite this progress, forecasting systems remain constrained by fragmented data integration, incomplete temporal coverage, and limited interpretability for decision makers.

This study introduces a cloud based forecasting system that harmonizes satellite, ground, and meteorological data streams into a unified predictive framework. The architecture emphasizes operational scalability, interpretability, and accessibility, converting raw observations into actionable insights for public health and policy.

II. Methodology

The system architecture consists of modular layers engineered for robust data fusion, predictive accuracy, and deployment scalability.

TABLE I
System Overview: Major Modules and Functional Roles.

Module	Function	Technology
Ingestion	Fetch multi source data	Airflow DAGs, APIs
Preprocessing	QA, gap fill, feature eng.	Python, GDAL, Pandas
Forecasting	Sequence modeling (1–72h)	LSTM (PyTorch)
Orchestration	Scheduling, retraining	Apache Airflow
Serving	REST APIs, dashboards	FastAPI, PostgreSQL
Interpretability	Forecast summaries	GPT 4o mini

A. Problem Formulation

Let $y_{t+h}^{(c)}$ denote the pollutant concentration (e.g., $PM_{2.5}$) for city c at time $t+h$, with horizon $h \in \{1, \ldots, 72\}$. Inputs $X_{\le t}^{(c)}$ comprise satellite features (TEMPO), ground stations (AirNow/AQS, OpenAQ, Pandora, TOLNet), and meteorology (MERRA 2/Daymet), aligned to hourly cadence and regridded to city footprints.

The LSTM f_θ minimizes a horizon weighted loss:

$$\min_\theta \mathcal{L}(\theta) = \sum_{c,t,h} w_h \, \ell\big(f_\theta(X_{\le t}^{(c)}, h), y_{t+h}^{(c)}\big)$$

where ℓ denotes MAE or RMSE, and optional quantile losses estimate uncertainty. Evaluation includes skill improvement over persistence baselines.

B. Why LSTM?

LSTMs balance predictive performance and computational efficiency. Their ability to capture long term temporal dependencies makes them well suited for near real time air quality forecasting in cloud native environments.

C. Data Volume and Temporal Scope

The system continuously ingests over 16 heterogeneous data streams spanning 72 hour rolling windows. Each operational cycle harmonizes approximately 1.2 million records across satellite, ground, and meteorological layers.

TABLE II
DATA SOURCES (ROLLING 72 HOUR WINDOW).

Source	Variables	Cadence	Coverage
TEMPO	NO_2, HCHO, O_3, AI	hourly (daylight)	Americas
TOLNet	O_3 profiles	site specific	Americas
Pandora	O_3, NO_2, HCHO	min–hourly	Americas
AirNow/AQS	$PM_{2.5}$, O_3, NO_2	hourly	Americas
OpenAQ	$PM_{2.5}$/PM_{10}, NO_2, O_3	varies	Americas
MERRA 2/Daymet	Meteorology	hourly/daily	Americas

III. RESULTS

A. Regional NO_2 Concentration Analysis

Average NO_2 AQI values were computed by region from North American monitoring networks. As shown in Figure 1, elevated NO_2 concentrations are observed over Mexico relative to the United States and Canada.

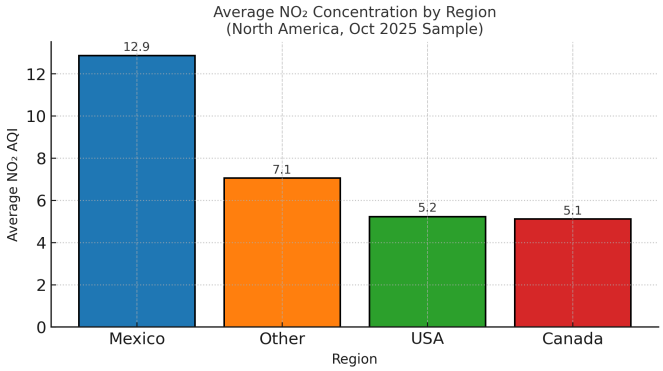

Fig. 1. Regional NO_2 concentration (October 2025). TEMPO and ground network data show higher NO_2 levels over Mexico compared to the U.S. and Canada.

B. Forecasting Accuracy (Preliminary)

Preliminary validation indicates MAE below $5\,\mu g\,m^{-3}$ for $PM_{2.5}$ over 24 hour horizons, outperforming persistence by 25–30%. Comparable gains are observed for O_3 and NO_2.

TABLE III
PRELIMINARY FORECASTING PERFORMANCE (24H HORIZON).

Pollutant	MAE	RMSE	Improvement
$PM_{2.5}$	4.8	6.2	+28%
O_3	5.3	7.1	+25%
NO_2	4.5	6.0	+30%
HCHO	6.1	8.3	+22%

C. $PM_{2.5}$ Concentration Trends

$PM_{2.5}$ levels were tracked across Canada, the United States, and Mexico over 72 hours (Figure 2).

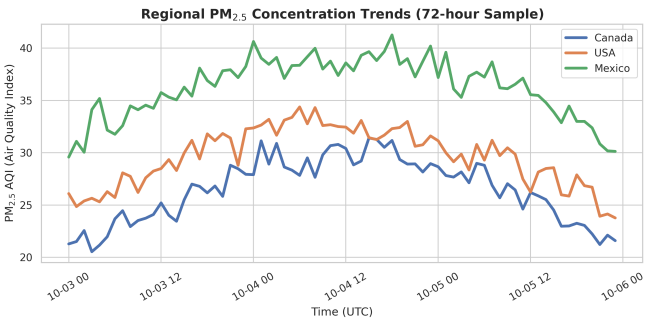

Fig. 2. Regional $PM_{2.5}$ concentration trends (72 hour sample). Smoothed AQI trajectories with $\pm 5\%$ uncertainty bands.

D. Operational Performance

Alert Management: A dynamic alert system monitors real time and forecasted $PM_{2.5}$ levels across North American cities. When deviations exceed threshold limits, localized alerts are issued through adaptive notification zones, translating forecasts into actionable intelligence.

Conversational Intelligence: A GPT 4o mini–based assistant enables natural language interaction with forecasts. It contextualizes responses using local air quality data and uncertainty estimates, offering accessible, data driven decision support.

TABLE IV
OPERATIONAL CHARACTERISTICS OF THE FORECASTING SYSTEM.

Metric	Observed Performance
Data to Forecast Latency	<3 hours
Forecast Horizon	1–72 hours
Retraining Frequency	Hourly (incremental)
Concurrent Query Handling	∼60 requests/s (cached)
Compute Environment	Azure (8 vCPUs, 32 GB RAM)

IV. DISCUSSION

This system contributes to environmental informatics by:

- **Multi source Data Fusion:** Integrating orbital, terrestrial, and modeled datasets for dense spatiotemporal coverage.
- **Operational Machine Learning:** Airflow orchestrated retraining ensures model adaptivity and low latency.
- **Interpretable AI:** GPT 4o mini translates technical outputs into plain language summaries, promoting public engagement.

Societal Relevance: Near real time forecasts support health advisories, emission control strategies, and equitable environmental governance. The interpretability module facilitates cross agency understanding and informed policymaking.

Limitations: TEMPO's daylight only coverage causes diurnal data gaps; rural under sampling limits calibration precision. Future work will focus on global scalability and nighttime inference.

V. Conclusion

This study demonstrates a scalable, cloud native framework integrating heterogeneous environmental datasets with interpretable machine learning for air quality forecasting. Hourly city level predictions with quantified uncertainty bridge the gap between sensing and action. Future directions include global coverage expansion, wildfire emission modeling, and enhanced uncertainty quantification.

Data Availability

All datasets are publicly available via NASA TEMPO, U.S. EPA AirNow/AQS, OpenAQ, and related repositories.

Acknowledgment

The authors thank NASA for TEMPO data, the U.S. EPA for AirNow/AQS contributions, and the NASA Space Apps Challenge for project support. Microsoft Azure provided cloud infrastructure, and OpenAI's GPT 4o mini enabled interpretability.

References

[1] World Health Organization (WHO), "Air pollution," Fact Sheet.
[2] NASA, "TEMPO Mission Documentation."
[3] U.S. EPA, "AirNow / AQS Data Access."
[4] OpenAQ, "Global Air Quality Data Platform."
[5] NASA GMAO, "MERRA 2 Reanalysis Documentation."
[6] Li et al., "Hybrid LSTM–CNN Models for PM2.5 Prediction," *Atmos. Env.*, 2022.
[7] Kim et al., "Deep Temporal Fusion for Urban Air Quality Forecasting," *Env. Mod.*, 2023.
[8] Wang et al., "Multi Source Data Fusion for Air Pollution Forecasting," *Sci. Data*, 2021.