1) Which U.S. state?

- **New York was selected for this study on air pollution, weather, and health outcomes because it provides a clear and data-rich example of the connection between environmental conditions and population health. The June 2023 Canadian wildfire smoke event produced an approximately 82% statewide surge in asthma emergency-department visits, offering a well-documented natural experiment for analyzing pollution spikes, lag effects, and forecasting respiratory outcomes. The state's robust surveillance infrastructure, including detailed datasets from the CDC Tracking Network and New York's own health and environmental monitoring systems, ensures data quality and feasibility. Furthermore, New York's geographic and climatic diversity allows examination of regional and seasonal variations in exposure–response patterns, making it an ideal setting for studying the intersection of air quality, weather, and health.**

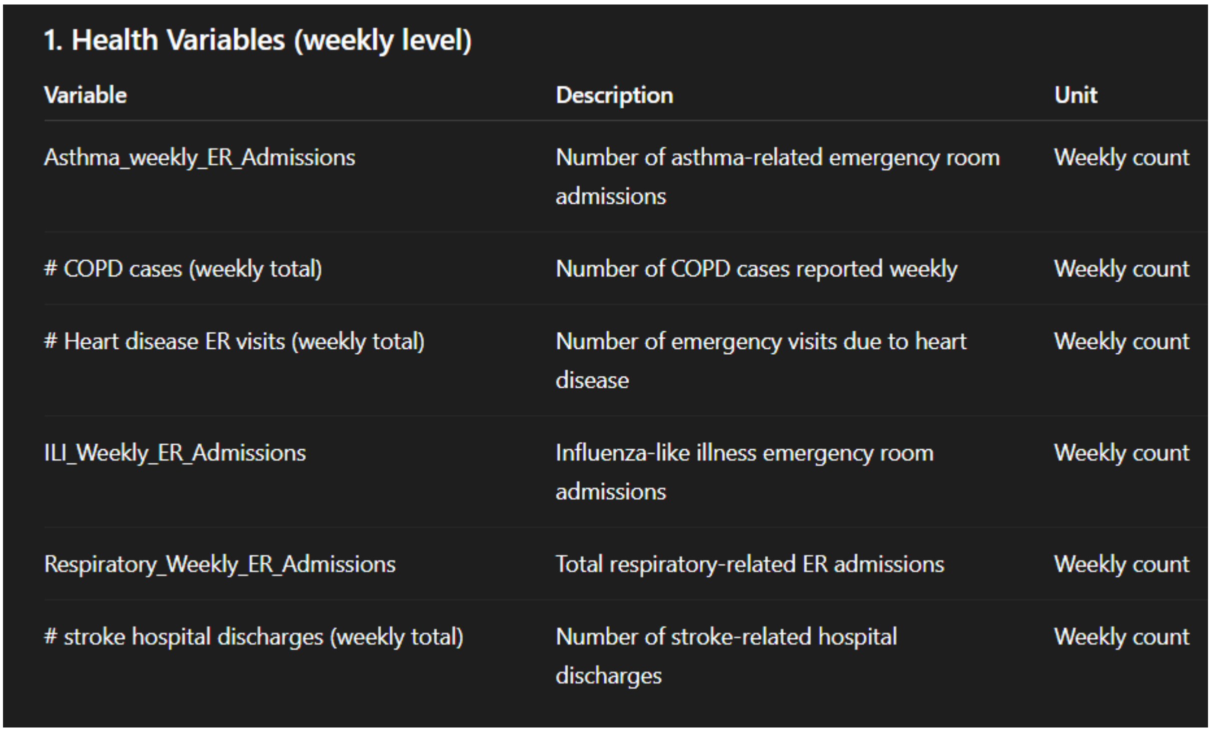
2) Which variables were included?

The air quality and weather variables in this study were originally available at the daily level and were subsequently aggregated into weekly averages or totals to align with the temporal frequency of the health indicators. Daily observations from 2015 through 2024 were grouped into 521 weeks, representing ten complete years of data. To maintain consistent weekly cycles, leap days (February 29) were removed from all years, and the final two days of the ten-year period were excluded to ensure that each week contained an exact seven-day interval (since 3650 days % 7 = 3).

For continuous variables such as pollutant concentrations and temperature, weekly means were computed to represent the average exposure over each week. For accumulation-based variables like precipitation and snowfall, weekly totals were calculated to capture the total amount within each seven-day period. This ensured that both short-term variation and seasonal trends were preserved while maintaining a uniform temporal structure across datasets.

All health outcomes—including asthma, COPD, heart disease, stroke, influenza-like illness (ILI), and respiratory ER admissions—were available directly at the weekly level, eliminating the need for additional temporal aggregation. Aligning all datasets to a consistent weekly timescale (n = 521) allowed for robust modeling of short-term relationships and lag effects between environmental exposures and health outcomes.

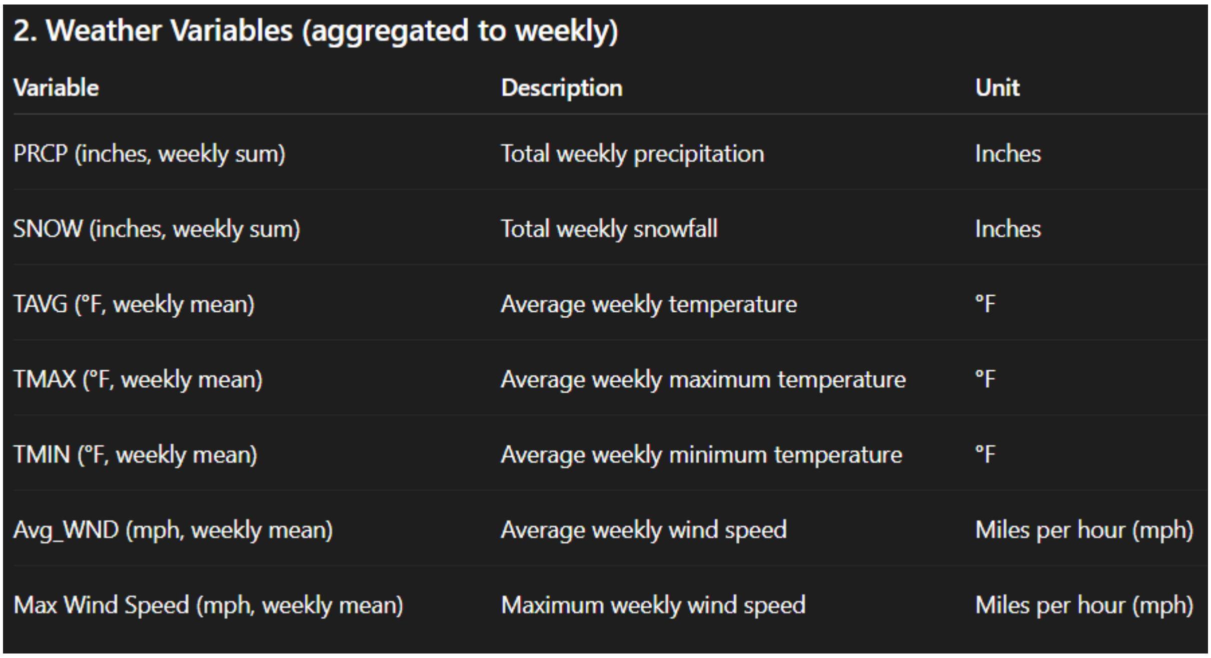
1. **Health variables**



1. Health Variables (weekly level)

Variable	Description	Unit
Asthma_weekly_ER_Admissions	Number of asthma-related emergency room admissions	Weekly count
# COPD cases (weekly total)	Number of COPD cases reported weekly	Weekly count
# Heart disease ER visits (weekly total)	Number of emergency visits due to heart disease	Weekly count
ILI_Weekly_ER_Admissions	Influenza-like illness emergency room admissions	Weekly count
Respiratory_Weekly_ER_Admissions	Total respiratory-related ER admissions	Weekly count
# stroke hospital discharges (weekly total)	Number of stroke-related hospital discharges	Weekly count

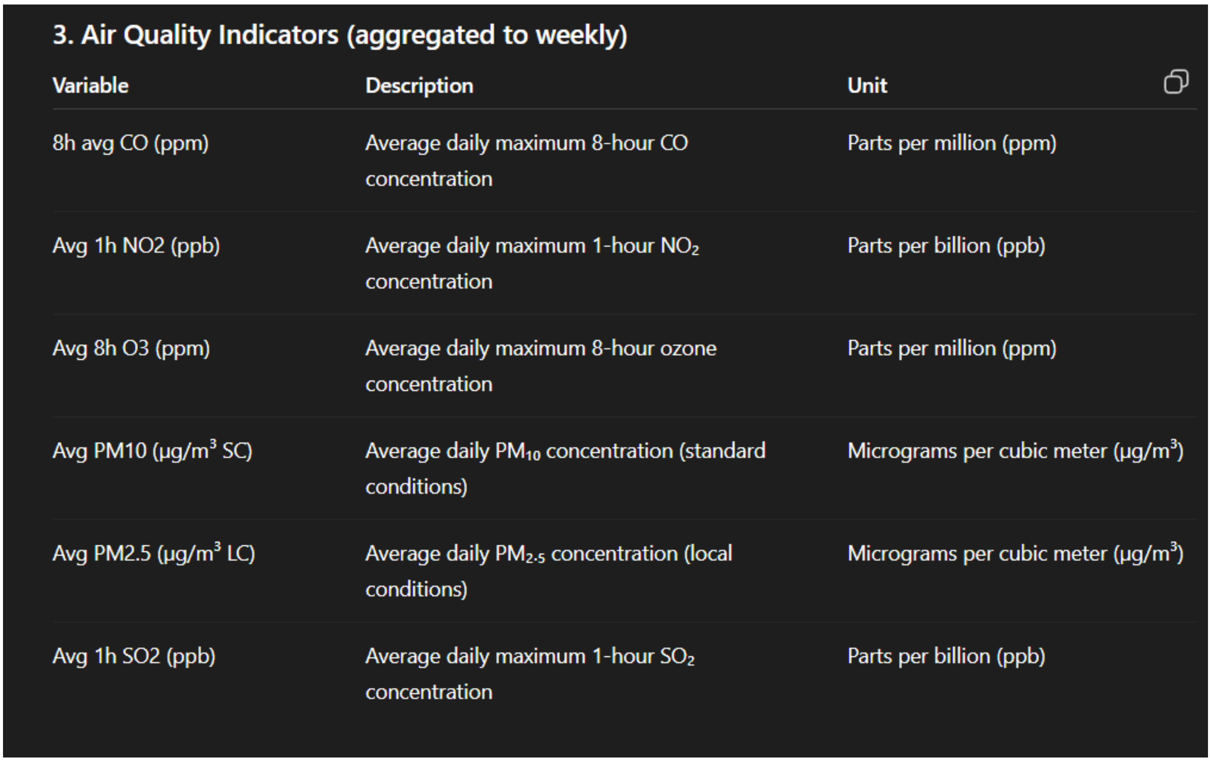
2. **Weather variables**



2. Weather Variables (aggregated to weekly)

Variable	Description	Unit
PRCP (inches, weekly sum)	Total weekly precipitation	Inches
SNOW (inches, weekly sum)	Total weekly snowfall	Inches
TAVG (°F, weekly mean)	Average weekly temperature	°F
TMAX (°F, weekly mean)	Average weekly maximum temperature	°F
TMIN (°F, weekly mean)	Average weekly minimum temperature	°F
Avg_WND (mph, weekly mean)	Average weekly wind speed	Miles per hour (mph)
Max Wind Speed (mph, weekly mean)	Maximum weekly wind speed	Miles per hour (mph)

3. **Air quality variables**



3. Air Quality Indicators (aggregated to weekly)

Variable	Description	Unit
8h avg CO (ppm)	Average daily maximum 8-hour CO concentration	Parts per million (ppm)
Avg 1h NO2 (ppb)	Average daily maximum 1-hour NO_2 concentration	Parts per billion (ppb)
Avg 8h O3 (ppm)	Average daily maximum 8-hour ozone concentration	Parts per million (ppm)
Avg PM10 ($\mu g/m^3$ SC)	Average daily PM_{10} concentration (standard conditions)	Micrograms per cubic meter ($\mu g/m^3$)
Avg PM2.5 ($\mu g/m^3$ LC)	Average daily $PM_{2.5}$ concentration (local conditions)	Micrograms per cubic meter ($\mu g/m^3$)
Avg 1h SO2 (ppb)	Average daily maximum 1-hour SO_2 concentration	Parts per billion (ppb)

Sources of variables:
1) Air quality variables: https://ephtracking.cdc.gov/DataExplorer/

2) Weather variables: https://www.ncei.noaa.gov/cdo-web/datatools/findstation

3) Health variables:

a. **Asthma, COPD, Heart Disease-** https://health.ny.gov/environmental/public_health_tracking/
b. **Respiratory illnesses (bronchitis, pneumonia, and other upper respiratory tract infections), ILI (flu and like diseases)-**https://a816-health.nyc.gov/hdi/epiquery/visualizations?PageType=tsi&PopulationSource=Syndromic&Topic=All&Subtopic=All&Indicator=Influenza-like%20illness%20(ILI)&Year=2025
c. **Stroke-** https://www.health.ny.gov/diseases/cardiovascular/stroke/designation/data.htm

3) Literature review

1) Ravindra et al., 2019 — *Environment International* (methods review on GAM/time-series)

- **Summary: This methodological review outlines how generalized additive models (GAMs) and Poisson time-series frameworks are used to study associations between air pollutants, meteorological variables, and health outcomes. The paper provides detailed guidance on incorporating nonlinear splines, seasonality adjustments, and lag structures in time-series models. It uses a weekly modeling approach—specifically in controlling for seasonality (week-of-year effects) and including lagged pollutant exposures (0–3 weeks).**

2) Sidell et al., 2022 — *Environmental Research* (short- & long-term exposure, large cohort)

- **Summary: This large Southern California cohort study evaluates both short-term and long-term exposures to $PM_{2.5}$ and NO_2 in relation to COVID-19 incidence, accounting for temporal and meteorological confounders. The authors developed multi-window exposure metrics (weekly means, rolling multi-week averages) and simultaneous adjustment strategies. The methodology informs the design of multi-window exposure variables—distinguishing between background (chronic) and acute (short-term) pollution effects on weekly health outcomes.**

3) Meek et al., 2023 — *CDC MMWR* (NY wildfire smoke; abrupt asthma spike)

- **Summary: This surveillance report investigates the June 2023 Canadian wildfire smoke event in New York, documenting an ~82% statewide increase in asthma-related emergency-department (ED) visits on the peak $PM_{2.5}$ day, with strong regional $PM_{2.5}$–asthma correlations. The findings validate New York as a high-signal environment for studying pollution-health dynamics and motivate the creation of event-study indicators (e.g., “smoke week” flags) and maximum $PM_{2.5}$ variables.**

4) Kalashnikov et al., 2022 — *Science Advances* (co-occurring $PM_{2.5}$ + O_3 extremes)

- **Summary: This paper demonstrates a rising trend in co-occurring $PM_{2.5}$ and O_3 extremes across the western United States, driven by heatwaves and wildfire activity. The study emphasizes the health risks of compound exposure under worsening meteorological conditions. It provides justification for incorporating interaction terms ($PM_{2.5}$ × O_3) and joint-extreme indicators into predictive models. Even in the Northeast, such co-exposures were observed during the 2023 wildfire event.**

5) Al-Kindi et al., 2020 — *Nature Reviews Cardiology* (cardiovascular mechanisms & evidence)

- **Summary: This review synthesizes epidemiologic and mechanistic evidence linking air pollution exposure to cardiovascular morbidity and mortality, describing biological pathways involving inflammation, oxidative stress, and endothelial dysfunction. It strengthens the scientific rationale for including cardiovascular outcomes in environmental-health analyses.**

Research Question/Objective:

1. Short-Term Forecasting of Respiratory ED Visits Using Air Quality and Weather Variables

Approach:

- **Build one-month-ahead forecasting models using ensemble methods such as Random Forest, CatBoost, or XGBoost, comparing them with baseline GAM or Poisson regression models.**

- **Include lagged exposures (0–3 weeks) for pollutants ($PM_{2.5}$, O_3, NO_2) and temperature to capture delayed health effects.**

- **Evaluate predictive performance using metrics like RMSE, MAE, and R^2 with actual unseen data from 2025.**

- **Perform feature importance analysis (e.g., SHAP values) to interpret which environmental variables contribute most to short-term health risks.**

Novelty (vs. Ravindra et al. 2019; Sidell et al. 2022):
Prior studies focused on estimating associations rather than prediction accuracy. Your study emphasizes out-of-sample forecasting and interpretable ML—bridging statistical modeling and predictive analytics for public health preparedness.

Idea 2: Exploring Interactions and Joint Extremes Between Pollutants and Meteorology

Research Question:

> **Do co-occurring extremes in $PM_{2.5}$, O_3, and temperature jointly amplify respiratory health risks compared to individual exposures?**

Approach:

- **Construct joint-extreme indicators (e.g., $PM_{2.5}$ > 90th percentile & O_3 > 90th percentile & temperature > 85th percentile).**
- **Incorporate interaction terms ($PM_{2.5} \times O_3$, $PM_{2.5}$ × temperature) in both statistical models (GAMs) and nonlinear ML models (CatBoost/Neural Networks) to capture synergistic effects.**
- **Visualize results via partial dependence plots or SHAP interaction values.**
- **Compare model fit and forecast accuracy with and without interaction features.**

Novelty (vs. Kalashnikov et al. 2022):
Kalashnikov et al. demonstrated rising joint extremes regionally, but not their direct health implications or forecasting value. Your project advances this by quantifying incremental health risks from joint exposures and testing whether incorporating these features improves predictive accuracy.

Idea 3: Quantifying the Impact of the 2023 Canadian Wildfire Smoke Event

Research Question:

> **How did the June 2023 Canadian wildfire smoke episode alter pollutant–health dynamics in New York, and can we detect lasting changes in model behavior or pollutant sensitivity?**

Approach:

- **Use an event-study design: create a "smoke week" indicator (based on Meek et al., 2023) to compare predicted vs. observed asthma/respiratory ER rates before, during, and after the event.**

- **Fit separate models with and without the wildfire period to assess differences in parameter estimates and pollutant contributions.**

- **Quantify changes in $PM_{2.5}$–health sensitivity (slope coefficients, feature importances) during the event.**

- **Optionally, test for structural breaks in relationships using Chow tests or rolling-window ML forecasts.**

Novelty (vs. Meek et al. 2023):
While Meek et al. described the short-term surge descriptively, your analysis generalizes the event into a forecasting + model-interaction framework, revealing how pollutant-health relationships shift under extreme conditions.

4) Analyses (Can be found in the Google Drive link)

DS480 Honors Project Report

5) Results

Analysis 1:

The first analysis evaluated the predictive performance and interpretability of two modeling frameworks—Generalized Additive Models (GAMs) and CatBoost—in forecasting weekly respiratory and cardiovascular health outcomes from 2015 to 2024. GAMs served as the transparent, epidemiologically grounded baseline using spline terms to capture nonlinear seasonality and flexible pollutant–response shapes, while CatBoost represented a modern nonlinear machine-learning alternative capable of capturing interactions automatically.

Across outcomes, the initial untuned CatBoost model performed poorly, with negative R^2 and high RMSE for several conditions. After incorporating best-performing lags, autoregressive terms, and mild tuning, CatBoost improved but still only matched—not exceeded—the GAM's performance. This result highlights the structural nature of the data: most of the signal in these health outcomes comes from smooth seasonal patterns, temperature, and long-term trends, all of which GAMs naturally capture with splines. CatBoost's added flexibility had limited benefit because the underlying pollutant–health relationships were relatively smooth and low-dimensional. However, the machine-learning models proved valuable for uncovering nonlinear interactions (explored in Analysis 2). Overall, both GAM and CatBoost produced comparable forecasting accuracy, but GAMs remained superior for interpretation and mechanistic insight, while CatBoost demonstrated potential for capturing synergistic effects not explicitly coded into statistical models.

In Analysis 1, the model consistently underestimates cases because the forecasting period (late 2023–2024) lies in a new, higher-risk regime that is not represented in the training data. Most of the GAM is fit on 2015–2023, a period with lower baseline ER visits and weaker pollutant–health sensitivities. After the June 2023 wildfire event, however, respiratory and cardiovascular outcomes shifted upward, creating a structural break that the model cannot anticipate. The strong smoothing penalties used in the GAM, combined with the log–exponential transformation and the inclusion of smoothed AR terms, further dampen peaks and pull predictions toward historical averages. As a result, when the true system enters a more intense post-wildfire period with higher visit counts, the model—trained on a calmer past—systematically predicts values that are too low.**

These forecasting results show that weekly health outcomes are driven mostly by predictable seasonal cycles and weather patterns rather than highly chaotic or complex interactions. Because GAMs capture these smooth trends naturally, they performed just as well—and often more reliably—than CatBoost. This means that for real-world hospital planning or public-health surveillance, transparent statistical models remain highly dependable. CatBoost still adds value by revealing subtle nonlinear patterns and interactions, especially during unusual events like wildfires, but routine forecasting can be handled effectively with interpretable models.

Analysis 2:

The second analysis focused on uncovering whether co-occurring environmental extremes amplify health risks beyond the sum of individual exposures. Six scientifically motivated interaction terms were constructed:
$PM_{2.5}$ × Temperature, $PM_{2.5}$ × O_3, $PM_{2.5}$ × NO_2, Temperature × O_3, Temperature × Wind, and $PM_{2.5}$ × Wind.
These interactions were incorporated into multiple linear regression models for three outcomes that later showed evidence of structural breaks: ILI, Respiratory, and Stroke.

Adding interactions improved model fit to varying degrees (e.g., adjusted R^2 increased substantially for Respiratory and Stroke) and revealed distinct synergy patterns across outcomes. ILI showed strong amplification in $PM_{2.5}$ × Temperature, $PM_{2.5}$ × O_3, and Temperature × O_3, suggesting heightened vulnerability to combined smoke and meteorological stressors. Respiratory outcomes showed the largest and most volatile interaction shifts, particularly for $PM_{2.5}$ × O_3 and $PM_{2.5}$ × NO_2, reflecting altered atmospheric chemistry and increased airway reactivity. Stroke outcomes displayed more subtle but consistent increases in interactions such as Temperature × O_3 and $PM_{2.5}$ × Temperature, indicating cardiovascular sensitivity to combined thermal and particulate stress.

Across all three outcomes, the consistently important interactions were
(1) Temperature × O_3,
(2) $PM_{2.5}$ × Temperature, and
(3) $PM_{2.5}$ × O_3,
while wind-based and NO_2-based interactions were comparatively weak. These findings show that wildfire conditions alter not only individual pollutant effects but also the ways pollutants and meteorology interact—emphasizing the need to incorporate interaction terms in environmental-health models.

The interaction analysis highlights that people face multiple environmental stressors at once, and these combined exposures often amplify health risks beyond what each pollutant would cause alone. Heat, smoke, and ozone together were especially harmful for respiratory and cardiovascular outcomes. These results show that wildfire smoke does more than raise $PM_{2.5}$—it changes how pollutants and meteorology interact. In real-life terms, public-health warnings and hospital preparedness should prioritize periods when smoke, heat, and ozone coincide, because those weeks carry the greatest risk for vulnerable populations.

Analysis 3:

The third analysis used the June 2023 Canadian wildfire smoke episode as a natural experiment to test whether the pollutant–health relationship structurally changed after this extreme exposure event. For each outcome, three models were fitted:
(1) a pooled model excluding the wildfire weeks,
(2) a pre-wildfire model, and
(3) a post-wildfire model.
Comparing coefficients side-by-side provided initial evidence of slope changes, which were then formally tested using the Chow test for structural breaks.

Results showed strong and statistically significant structural breaks for all three outcomes. ILI displayed the most dramatic shifts, with large increases in the effects of $PM_{2.5}$, temperature, and their interactions after the wildfire. Respiratory outcomes exhibited substantial and directionally volatile coefficient changes, especially for ozone and $PM_{2.5}$ × O_3, suggesting severe alterations in atmospheric chemistry and respiratory vulnerability during and after the smoke event. Stroke outcomes, while more stable, still exhibited significant shifts in key interactions such as Temperature × O_3 and $PM_{2.5}$ × Temperature. Chow tests confirmed structural breaks for ILI ($p \approx 0.00008$), Respiratory ($p \approx 0.0147$), and Stroke ($p \approx 0.000000$), demonstrating that the wildfire introduced a new environmental regime with altered pollutant sensitivities and interaction dynamics.

Overall, Analysis 3 provides strong evidence that wildfire smoke fundamentally changed the pollutant–health relationship, validating the need to examine interactions (Analysis 2) and supporting the observation that nonlinear machine-learning models and traditional statistical models perform differently around extremes (Analysis 1). Collectively, the three analyses reveal a coherent narrative: normal pollutant–health dynamics break down during wildfire events, interactively amplify after extreme smoke episodes, and require new modeling strategies that capture synergy, thresholds, and regime shifts.

The wildfire analysis shows that June 2023 marked a turning point: after the smoke event, pollutants affected health outcomes differently than before. ILI, respiratory, and stroke sensitivities to $PM_{2.5}$, ozone, and temperature all shifted, meaning even moderate exposures may now trigger stronger health responses. This suggests that both population vulnerability and atmospheric chemistry changed following the wildfire. Practically, this means older forecasting models may underestimate current risks, and agencies should update public-health guidelines and early-warning systems to reflect this new post-wildfire risk environment.

6) Future recommendations

1. Explore Advanced Forecasting Models With Hybrid Statistical–ML Structures

Although this project compared GAM and CatBoost for weekly respiratory-visit prediction, future work should investigate hybrid models that explicitly combine the strengths of both.
Examples include:

- **GAM-LSTM hybrid models, where GAM handles long-term seasonality and smooth climate–pollution effects, and the LSTM captures short-term temporal dependencies and nonlinear wildfire-smoke shocks.**

- **ARIMAX-CatBoost ensembles, where ARIMA models the autoregressive structure while CatBoost models nonlinear interactions and high-dimensional lag effects.**

- **Bayesian GAMs that incorporate uncertainty in smoothing parameters, particularly important during extreme wildfire events.**

Because Ravindra et al. (2019) emphasizes that pollutant–health links are often nonlinear and lagged, a hybrid approach could better capture these structures while maintaining the interpretability of GAMs and the predictive power of gradient-boosted trees.

2. Model Interaction Effects and Structural Breaks Caused by the June 2023 Canadian Wildfires

The 2023 wildfire smoke intrusion represented the largest air-quality disruption in the 2015–2024 period and likely caused a structural break in both pollutant levels and their relationship with health outcomes. Future analyses should:

- **Perform a Chow test or Bai–Perron multiple-breakpoint test to formally detect structural breaks in the pollutant–health relationship.**

- **Explicitly model interaction terms between wildfire-related variables and pollutants (e.g., $PM_{2.5}$ × Fire_Density, or $PM_{2.5}$ × Wind_Direction).**

- **Use regime-switching models (e.g., Markov switching GAM or switching regression models) to allow the system to behave differently during smoke-intrusion vs. normal periods.**

- **Incorporate satellite-derived wildfire exposure metrics (AOD, plume height, plume transport pathways) for richer spatial–temporal coverage.**

This would expand upon the methodological guidance from Ravindra et al. (2019), who highlight the importance of multi-day lag structures and meteorological confounders when pollutants exhibit unusual episodic behavior.

3. Build a Multi-Outcome, Multi-Pollutant Causal Framework Using DLNM/GAMMs With Cross-Basis Functions

The current project modeled a single respiratory outcome at a time. Future research should build a multi-response model capable of simultaneously estimating effects across asthma, COPD, influenza-like illness, and total respiratory visits.

A strong direction is implementing:

- Distributed Lag Nonlinear Models (DLNM) with cross-basis terms for $PM_{2.5}$, ozone, temperature, and humidity.
- GAMMs (Generalized Additive Mixed Models) with random effects for region/county, enabling spatial sharing of information.
- Causal inference designs, such as:
    - Synthetic control (for wildfire weeks),
    - Difference-in-differences (pre- vs post-June 2023),
    - Augmented inverse-probability weighting for exposure.

DLNM/GAMMs directly follow the best practices summarized by Ravindra et al. (2019), who discuss spline-based approaches, nonlinear meteorological interactions, and the importance of lagged pollutant effects on health outcomes. Such future work would shift the analysis from pure prediction toward causal interpretation and public-health relevance.

Together, these extensions would allow future studies to move beyond baseline prediction models and toward a deeper understanding of nonlinear, lagged, and wildfire-specific pollutant–health relationships, while aligning closely with the methodological recommendations of Ravindra et al. (2019) and other contemporary air-quality time-series literature.

7) Appendix (Codes) (Can be found in the Google Drive link)

https://drive.google.com/drive/folders/1T8AWOfJhzHxbIbUtHoOMzbnHAKF2XKqx?usp=sharing