Climate-Informed Modelling of Health Risk

Authors: Group 3
Clara Sobejano, Louis-Esmel Kodo, Lucas Ihnen, Massimo Tassinari, Nitin Jangir, Yousef Mohammad
Mentor: Idafen Santana Pérez

Microsoft | ie SCHOOL OF SCIENCE & TECHNOLOGY

Our Repository

Introduction

Project Objective
Microsoft's Capstone Challenge asked us to turn GRIDMET, a 40-year archive of high-resolution climate data, into a revenue-generating service with clear real-world value.

Why GRIDMET?
With daily weather data across the U.S. at 4 km resolution, GRIDMET supports decisions in energy, agriculture, insurance, logistics, and public health. These variables help reduce risk, improve planning, and protect lives.

Our Focus
We chose to forecast hospital admissions linked to climate-driven events like heatwaves, flu seasons, and wildfire smoke. These surges often overwhelm hospitals, causing:

- Emergency department overcrowding
- Staffing and supply shortages
- Significant cost increases

The stakes are high:

- $3.75B/year in flu-related admissions
- $1B in extra costs from summer heat
- $2–4M/year in efficiency losses per hospital

Proof of Concept on Rochester (NY)

Methodology

Data Sources

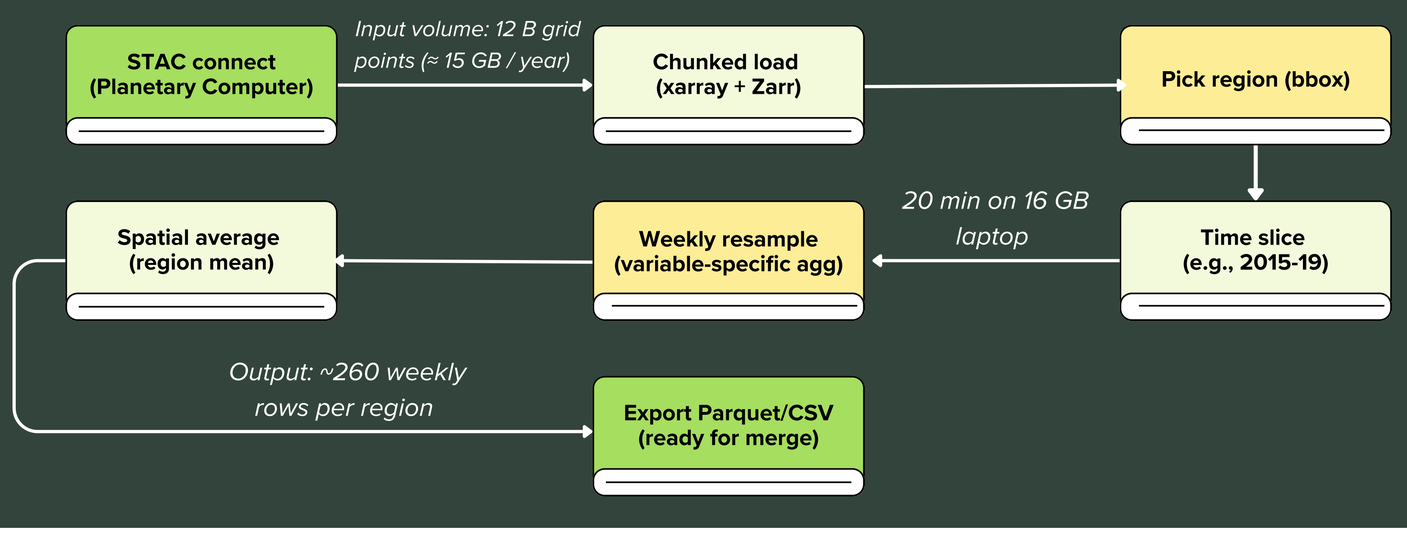
FluView

- Coverage from 2009 refreshed in near-real time
- City & State level confirmed weekly influenza hospitalisations
- Age group, Virus Type, Race & Gender granularity

GridMET Planetary Computer

- 40 years of daily U.S. weather, 4 km grid
- 10+ variables (temp, humidity, precip, wind, fire danger...)
- > 12 billion records in daily granularity and (lat, long) pairs

Data Extraction Pipeline
A modular, cloud-ready pipeline to transform massive climate data into structured health inputs



Feature Engineering
To capture trends and improve predictive power:

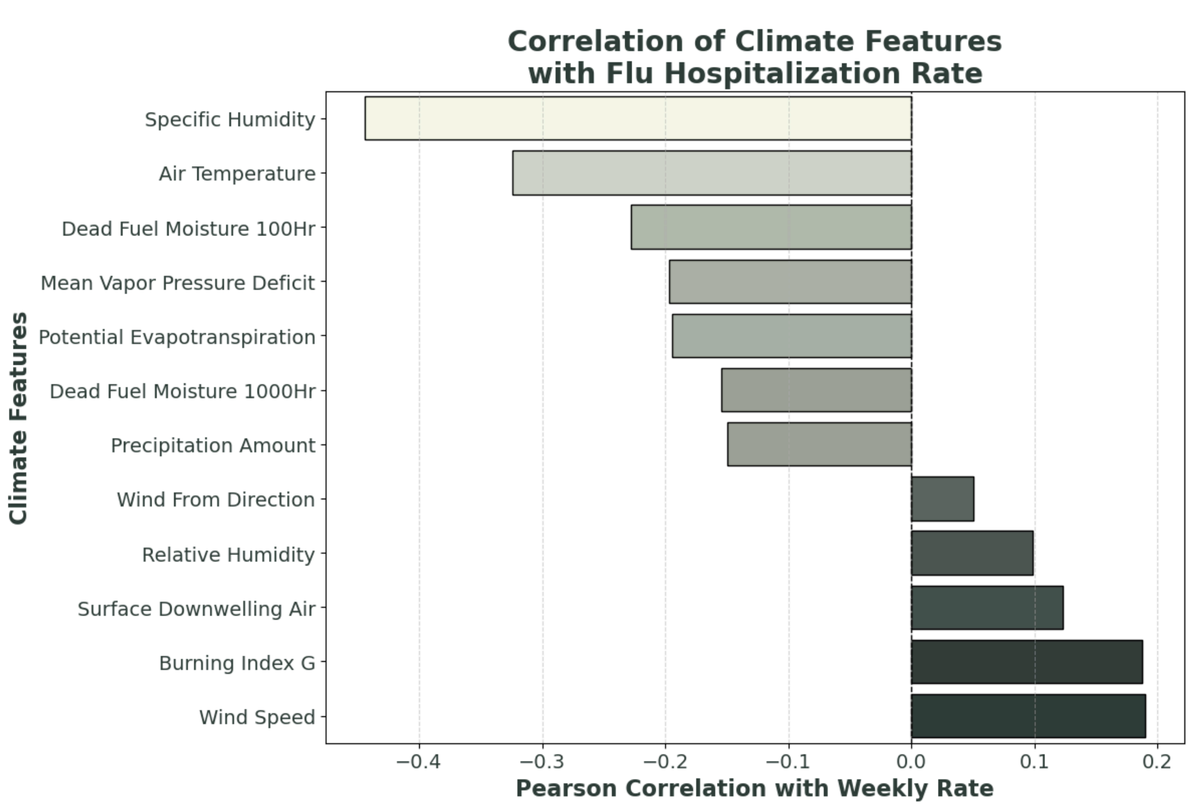
Lag Features: 1–8 weeks of past climate data

Rolling Statistics: Weekly averages, min/max, variability

Interaction Terms: E.g., humidity × temperature

Seasonality Features: Captured recurring temporal patterns relevant to flu cycles

Lagged Target Variable: Flu rates from the past 1–3 weeks



Model Development & Selection

- 10+ XGBoost models tested with progressive enhancements
- Improvements through tuning, feature selection, and hybrid design
- Final model (v6.2): predicts correction over lag-1 baseline
- Optimized via time-series CV + Optuna (30 trials)

XGBoost + Optuna

R^2 : 0.82
RMSE : 1.35

best among 10+ tested

Tools & Tech Stack

Data Manipulation: pandas, NumPy
Data Extraction: STAC SpatioTemporal Asset Catalog, Zarr, xarray
EDA: matplotlib, seaborn, plotly
Machine Learning: scikit-learn

Results

MODEL PERFORMANCE

Hybrid model outperforms baseline:

- **RMSE**: 2.43 → **1.34** (**−44.9% error**)
- **R^2**: 0.467 → **0.823** (+35.6 pts explained variance)

OPERATIONAL IMPACT

- **~120 fewer admissions/year** in Rochester
- At **$12,500 per flu admission**, that's **$1.5M saved**
- Across **12 hospitals: $125K annual savings per hospital**

MARKET OPPORTUNITY

- 5,000 hospitals in the U.S.
- Just **1% penetration = 50 clients/year**
- Product ROI: **125%–188%**, depending on tier

Conclusion

This project shows how open climate data and public health surveillance can be combined to forecast hospital admissions with high accuracy.

By leveraging GRIDMET and CDC FluView, we built a model that reduced prediction error by over 44% in Rochester and reached an R^2 of 0.93 in California. These improvements translate to up to $1.5M in annual savings for a mid-sized city.