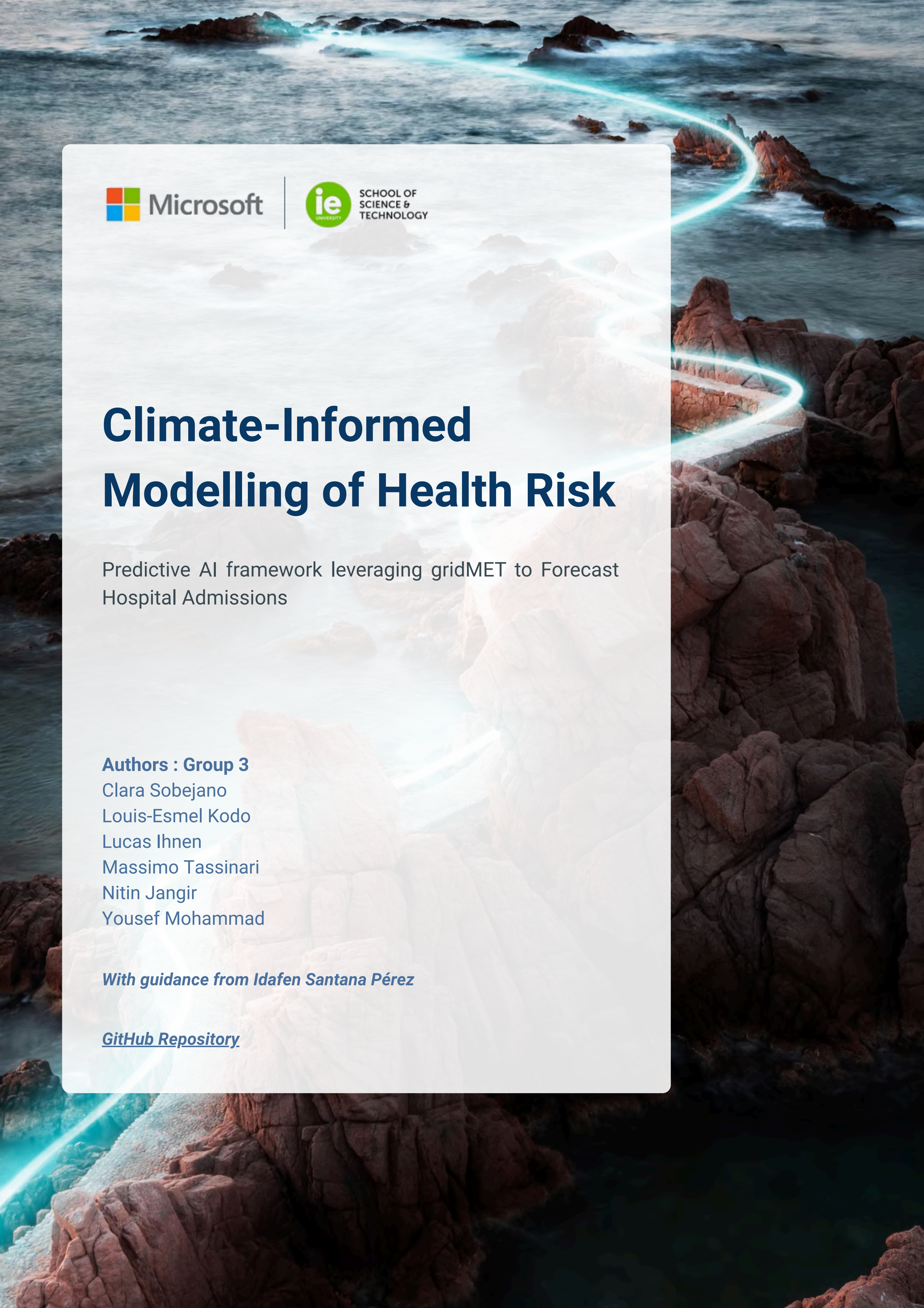


Climate-Informed Modelling of Health Risk

Predictive AI framework leveraging gridMET to Forecast Hospital Admissions

Authors : Group 3
Clara Sobejano
Louis-Esmel Kodo
Lucas Ihnen
Massimo Tassinari
Nitin Jangir
Yousef Mohammad

With guidance from Idafen Santana Pérez

GitHub Repository

Declaration

Data Ethics

All data used in this project is publicly available and anonymized. The hospitalization data was sourced from the CDC FluView Interactive platform and does not contain any personally identifiable information (PII). The environmental data was accessed through Microsoft's Planetary Computer, which provides open-access, high-resolution climate datasets (gridMET). No individual-level health records were used, and all data handling complied with terms of use and open-data policies.

Bias & Representativeness

Hospitalization rates and health outcomes can be influenced by a range of social determinants of health, including access to care, socioeconomic status, and race or ethnicity, that are not captured in environmental data alone. This model focuses only on climatic predictors, and we recognize that excluding such factors may lead to biased or incomplete representations of hospitalization risk, particularly across underserved communities. Therefore, any interpretations or applications of the model should be made with caution and awareness of these limitations.

Model Transparency & Reproducibility

The modeling process, including data preprocessing, feature engineering, algorithmic design, and evaluation, has been conducted in a transparent and reproducible manner. Time-aware data partitioning was employed to avoid information leakage and to simulate realistic forecasting scenarios. All code, methodology, and performance metrics (e.g., MAE, RMSE, R^2) are documented and available for replication or review.

Responsible Use

The goal of this project is to enhance understanding of the relationship between environmental change and public health, and to support anticipatory health planning. We do not intend or authorize this model to be used for surveillance, discrimination, or punitive action against any group or region. Model predictions should be interpreted as probabilistic estimates, not certainties, and should always be contextualized within broader epidemiological and social factors.

Contents

Objective

Microsoft’s Capstone Challenge tasked the team with demonstrating how open-access climate data from the Planetary Computer can be transformed into a commercially viable, revenue-generating product that delivers clear business value. In essence, the objective is to convert raw environmental datasets, specifically, gridMET, a comprehensive 40-year archive of daily meteorological variables such as temperature, humidity, wind, and precipitation at 4-kilometre resolution across the United States, into actionable insights for sectors willing to invest in climate intelligence.

Why focus on gridMET? Its fine spatial grid, long historical record, and daily temporal resolution make it one of the most versatile climate datasets available. Energy developers can size solar or wind assets with greater confidence; insurers can price flood, drought, or wildfire risk at property level; agriculture firms can adjust planting and irrigation plans; logistics companies can anticipate weather driven disruptions to supply chains; and public sector agencies can trigger early warnings for heat, smoke, or storm events.

In all these cases, the same core meteorological drivers serve as critical inputs into models that reduce uncertainty, improve operational efficiency, and/ or safeguard public health and infrastructure.

This project takes that broad mandate as its foundation. In the sections that follow, we describe our selection of a high-impact use case, the integration of gridMET with complementary public datasets, and the development of a prototype service. This end-to-end solution not only addresses the objectives of Microsoft’s challenge, but also illustrates the broader commercial potential of the Planetary Computer ecosystem.

Potential Applications

Our team brainstormed several ways to turn the gridMET meteorological dataset into viable business offerings. Ideas surfaced during cross-functional ideation workshops and were screened against six strategic dimensions :

- Business Impact & Monetisation Potential
- Specific Industry Knowledge required?
- External Datasets required?
- Temporal & Geographic resolution
- Applicable gridMET features
- Suitable Modelling Algorithm

The goal was to identify concepts that are technically feasible, data compatible, and commercially attractive either by lowering operational risk, optimising resources, or enabling better informed decisions for public or private sector clients. After scoring a long list of candidates, five directions emerged with the strongest combined score.

Climate Risk Modelling : By combining long term weather records with projection ensembles, this option would quantify the probability of hazards such as floods, droughts and heatwaves for specific assets and post codes. It would serve investors and facility managers who need to understand how climate change affects the physical resilience and valuation of their portfolios.

Wildfire Prediction : Here the focus was on forecasting the likelihood and potential spread of wildfires by tracking fuel moisture, wind patterns and antecedent rainfall. Emergency agencies and insurers could use the alerts to stage firefighting resources earlier, issue evacuation guidance and refine insurance pricing in areas of elevated fire danger.

Crop & Farming Advisory : The focus was on forecasting the likelihood & potential spread of wildfires by tracking fuel moisture, wind patterns and antecedent rainfall. Emergency agencies and insurers could use the alerts to stage firefighting resources earlier, issue evacuation guidance and refine insurance pricing in areas of elevated fire danger.

Renewable Energy Site Selection & Exploitation : Long-term averages of solar irradiance, wind speed and atmospheric stability would feed into a siting tool that ranks parcels of land for solar farms or wind turbines. Energy developers could screen locations faster, lower project-planning costs and de-risk future output estimates when pitching to investors.

Hospital Admissions Forecasting : Finally, the team considered if weather extremes and air-quality metrics could forecast surges in hospital admissions for heat-related illnesses, asthma, cardiovascular events and seasonal flu. Hospitals, public-health agencies and insurers stand to gain by staffing and resourcing proactively rather than reacting once emergency departments are already under pressure.

Chosen Application

The team ultimately chose to focus on *Forecasting Hospital Admissions* linked to climate conditions, **within the scope of influenza-related cases**. This use case stood out not only in our 6 aforementioned criteria, but also due to its clear societal relevance, data availability (via CDC FluView), and the potential to deliver tangible, high-impact outcomes for healthcare systems facing increasing pressure from climate-driven health events. It represented a well-defined and actionable problem, grounded in public health needs and supported by feasible modeling approaches.

Problem Statement

> Hospitals across the US face
> ***"unanticipated surges in patient volume",***
> many of which are triggered by environmental factors. These climate-driven spikes strain emergency departments, disrupt operations, and increase healthcare costs.

Extreme heat events, for example, have led to 3x increase in emergency room visits for heat-related illnesses in affected regions. Similarly, seasonal influenza causes up to 300,000 hospital admissions annually, each with an average cost of \$12,500, totalling \$3.75 billion in direct flu-related expenditures. Wildfire smoke and poor air quality further exacerbate respiratory and cardiovascular conditions, driving additional hospital visits.

These surges often occur with little warning, leading to:

- Overcrowded emergency departments
- Last-minute staffing shortages
- Medication and supply chain disruptions

The financial impact is substantial too :

\$2-4 Million in annual Efficiency Losses per large hospital

\$1 Billion added to Healthcare costs due to over 56,000 additional hospitalizations each U.S. summer

Looking ahead, climate change is expected to only intensify this trend & further drive up the wastage of crucial life-saving resources!

Data Sources

1 FluView

US CDC

To model virus-related hospitalization trends in response to environmental factors, this study utilizes data from the CDC FluView Interactive platform. FluView is maintained by the U.S. Centers for Disease Control and Prevention (CDC) and serves as the nation's primary surveillance system for tracking influenza activity, severity, and burden across the United States.

The specific dataset used in this project is drawn from the FluSurv-NET system within FluView, which provides standardized, laboratory-confirmed, and population-based weekly hospitalization rates for Influenza across select U.S. cities and states. These data are aggregated by,

- Age
- Season
- Region
- Weekly/ Annualy

FluSurv-NET is widely regarded as the gold standard for monitoring severe influenza outcomes and is frequently used to inform national influenza burden estimates, vaccine effectiveness studies, and public health preparedness efforts. The dataset provides a critical empirical foundation for training and validating predictive models that seek to link climatic conditions with seasonal fluctuations in influenza hospitalization rates.

As an official CDC resource, FluView data is subject to rigorous curation, validation, and continuous updates, ensuring reliability and consistency across seasons. Its city-level granularity and historical depth make it a robust and policy-relevant dataset for investigating the intersection of climate and public health at scale.

2 gridMET
Planetary Computer

The primary data source for this project is the gridMET (Gridded Surface Meteorological) dataset, made accessible through the Microsoft Planetary Computer. gridMET provides high-resolution, daily surface meteorological data across the United States, dating back to 1979 up to 2020. It combines the spatial granularity of gridded satellite data with the temporal accuracy of ground-based observations, making it highly suitable for climate analysis and environmental modeling.

The dataset covers over 12 billion data points, with spatial coverage at approximately 4 km x 4 km resolution (0.0417° lat/lon grid), including data from more than 800,000 sensor locations across 585 latitudes and 1,386 longitudes. Temporally, it includes more than 15,000 daily time points, ensuring extensive longitudinal analysis capabilities.

FEATURE	DESCRIPTION	DATA TYPE
time	date value (daily)	datetime64
lat	latitude	float64
long	longitude	float64
burning_index_g	Indicator of Fire Danger & intensity	float32
dead_fuel_moisture_1000hr	1000 hour fuel moisture (%)	float32
mean_vapor_pressure_deficit	mean vapor pressure deficit (kPa)	float32
precipitation_amount	Daily Accumulated Precipitation (mm)	float32
specific_humidity	Daily mean specific humidity (kg/kg)	float32
wind_speed	Daily Mean Wind Speed (m/s)	float32
wind_from_direction	Daily mean wind direction (degrees)	float32
surface_downwelling_shortwave_flux	Avg downward shortwave radiation/ day	float32

Data Access

The GRIDMET dataset is available via STAC (SpatioTemporal Asset Catalog) API, enabling efficient and scalable access through Python libraries such as *pystac-client*, *xarray*, & *zarr*. These tools facilitate data filtering, spatial subsetting, temporal slicing, & efficient memory usage for large-scale analysis. All processing and modeling in this project leveraged these open-access tools within a reproducible and modular pipeline.

Methodology

The methodology for this project was shaped by a set of practical and technical constraints identified early in the problem-solving process. Each constraint influenced the selection of tools, the design of data workflows, and the scope of analysis. This section outlines the constraints and how they guided our approach to implementing the solution.

Identified Constraints

The gridMET dataset offers significant analytical power due to its high-resolution meteorological coverage. However, from our team's assessment, it presents three key constraints that influenced our approach :

Temporal Coverage: The dataset spans from 1979 to 2020, making it ideal for analyzing historical climate trends, but less applicable to real-time or post-2020 forecasting use cases. Therefore, any supporting datasets needed to align with this time frame to ensure coherence.

Geographic Scope: GRIDMET only covers the United States. As such, any viable business case had to be geographically constrained within U.S. territory to ensure contextual relevance and regulatory compliance.

Data Volume: With over 12 billion data points, a single year of gridMET data can exceed 15 GB. This scale imposes computational and storage challenges, requiring efficient subsetting, filtering, and pipeline design to make the project feasible within the scope of a proof of concept. More on this in the “Data Extraction” section.

Implementation Approach

- First, we identified data sources that could circumvent these temporal & geographic constraints and restructured them to the same granularity.
- Next, we picked one location out of all the available options in FluView's database for validating our model before broader deployment. *Rochester, New York* was chosen due to its exposure to both extreme summer heat and intermittent wildfire smoke, providing diverse environmental conditions for model evaluation. Additionally, its city-level granularity ensures that the insights generated are operationally relevant for hospital systems.
- Finally, to test the adaptability & scalability of our solution, we implemented it across different Age Groups, Virus Types and the region of California.

Tools & Tech Stack

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

Development

A) Data Extraction

To explore and monetize the GRIDMET climate data, we first needed a way to extract clean, targeted data slices from a massive dataset containing over 12 billion records. This section describes how we did just that. Just for reference, everything mentioned here is within a notebook on the publicly available GitHub repository.

Tools & Tech Stack

To handle the extraction of regional slices from the GRIDMET climate dataset, we built our workflow using a simple but powerful Python-based stack:

pystac-client: To query and navigate Microsoft's STAC API from the Planetary Computer.

Planetary-computer: To sign and authenticate requests.

xarray: To work with multi-dimensional labeled arrays (ideal for time-latitude-longitude datasets).

zarr: A format that allows chunked, compressed storage for very large datasets.

Dask (via xarray): To handle large-scale computations lazily, avoiding memory overload on local machines.

1

Connecting to GRIDMET via Microsoft's STAC API

We began by tapping into Microsoft's Planetary Computer using their STAC API. This allowed us to programmatically browse and access the GRIDMET dataset without needing to manually download anything.

```
import pystac_client
import planetary_computer

catalog = pystac_client.Client.open(
    "https://planetarycomputer.microsoft.com/api/stac/v1",
    modifier=planetary_computer.sign_inplace,)
```

2

Opening the Dataset Efficiently

Once we found the GRIDMET collection, we loaded the data using xarray, a library designed for working with large climate datasets. Behind the scenes, this was powered by Zarr format, allowing us to load just the data we need, instead of the entire 40-year archive. This step made our work scalable from the start. Even working locally, we didn't need to compromise on speed or performance.

```
import xarray as xr

ds = xr.open_zarr( … )
```

3 Choosing a Region

To focus on specific locations rather than the entire US, we defined regions of interest such as Rochester, NY, using bounding boxes (latitude and longitude ranges).

```
Polygon_dict = {
 'rochester': {
        'lat_min': 43.075,
        'lat_max': 43.275,
        'lon_min': -77.745,
        'lon_max': -77.505
    }}
```

This modular approach let us extract only the relevant portion of the GRIDMET dataset for each area, keeping the workflow flexible. Using rectangular bounding boxes simplified filtering features by region.

4 Slicing Time and Space

After defining a region, we filtered the dataset by year (2015–2019) and the specified latitude/longitude bounds to create a focused “climate time-lapse” for our area of interest:

```
ds_polygon = ds.sel(
    time=slice("2015-01-01", "2019-12-31"),
    lat=slice(lat_max, lat_min), # Note: latitudes are descending
    lon=slice(lon_min, lon_max)
).compute()
```

5 Weekly resampling

Until this point, we connected to the Planetary Computer and filtered by date and region, but the data remained daily and spatially granular (one record per (lat, lon) per day). To make it manageable and compatible with external datasets, we aggregated to weekly values, specifying methods per feature. This produced a compact weekly representation for each variable.

```
agg_methods = {
    "air_temperature": "mean", # Average temperature per week
    "burning_index_g": "max", # Use max for fire danger indicators
    "dead_fuel_moisture_1000hr": "mean",
 . . .
}

ds_weekly = xr.Dataset({
    var: getattr(ds_subset[var].resample(time="1W"), method)()
    for var, method in agg_methods.items()
})
```

6

Averaging for the region

After resampling, we still had weekly data per (lat, lon). To create region-level records comparable to FluView, we averaged all points for each week, producing a clean region-week dataset ready for merging.

7

Export data

After reducing millions of records to just a few hundred weekly rows, we exported the dataset for later analysis. This processing took ~20 minutes on local machines. The data was saved in both CSV and Parquet formats for flexibility.

```
csv_path = f"data/{sel_polygon}_weekly_avg.csv"
parquet_path = f"data/{sel_polygon}_weekly_avg.parquet"

df_weekly_avg.to_csv(csv_path, index=False)
df_weekly_avg.to_parquet(parquet_path, index=False)
--------
✅ Data saved to:
- data/rochester_weekly_avg.csv
- data/rochester_weekly_avg.parquet
```

B) Exploratory Data Analysis

In this section, we conducted an exploratory data analysis to identify patterns linking climatic factors to influenza hospitalization rates in **Rochester, NY**. Specifically, we examined:

- Trends in weekly hospitalization rates over time.
- Seasonal and demographic patterns.
- Distributions and interrelationships among meteorological variables.
- Features with potential predictive value for modeling flu incidence.

Note: The dataset only includes typical flu seasons (October–April), limiting visibility into off-season baseline hospitalization trends.

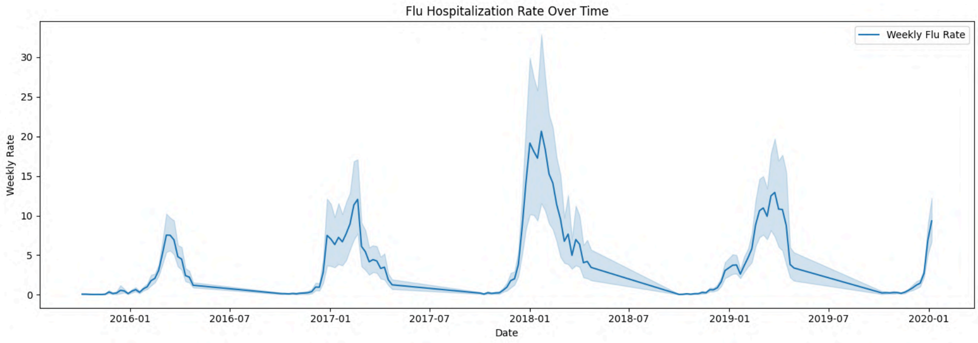
This process included:

DATA CLEANING AND INTEGRATION

The CDC FluView dataset provided weekly hospitalization rates, which we standardized by aligning all weeks to Mondays. The gridMET dataset contained daily meteorological measurements (temperature, precipitation, humidity, wind, and derived indicators). These records were filtered to the Rochester region, resampled to weekly aggregates, and merged with the hospitalization data by date. This produced a consolidated dataset combining weekly influenza admissions with regional climate indicators.

TIME SERIES VISUALIZATION

We plotted weekly influenza hospitalization rates over multiple seasons. These plots revealed clear seasonal peaks, typically between December and March each year, reflecting the known flu seasonality. Weeks outside this period consistently showed very low or zero admissions. The magnitude of these peaks varied by season, with some years — such as 2017–2018 — showing markedly higher hospitalization rates. This variability suggests the potential value of including climatic covariates to explain inter-season differences.



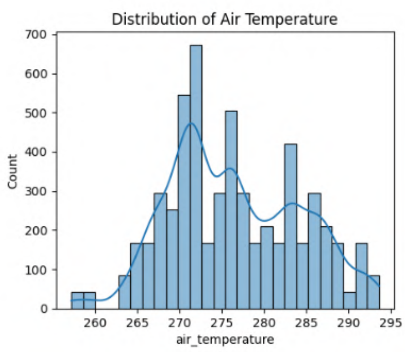
CLIMATE VARIABLE TRENDS

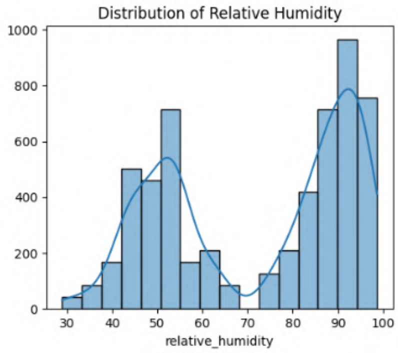
Time series plots were created for each meteorological variable.

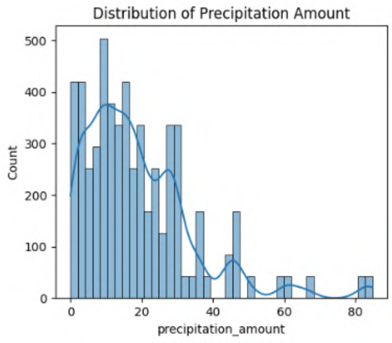
- **Air temperature** showed predictable annual cycles, lowest during peak flu months.
- **Specific humidity** and **precipitation** fluctuated more week to week.
- **Wind speed** and fire danger indicators (*burning index*) remained relatively stable but showed occasional spikes.

Histograms further revealed that **precipitation** and **vapor pressure deficit** exhibited right-skewed distributions, with frequent low values and occasional extremes. In contrast, **temperature** and **humidity** were more symmetrically distributed. These patterns helped identify potential outliers and informed decisions about scaling or transforming variables before modeling.

This analysis highlighted the need to create **lag features** (e.g., previous week's temperature) to capture potential delayed effects on flu incidence.



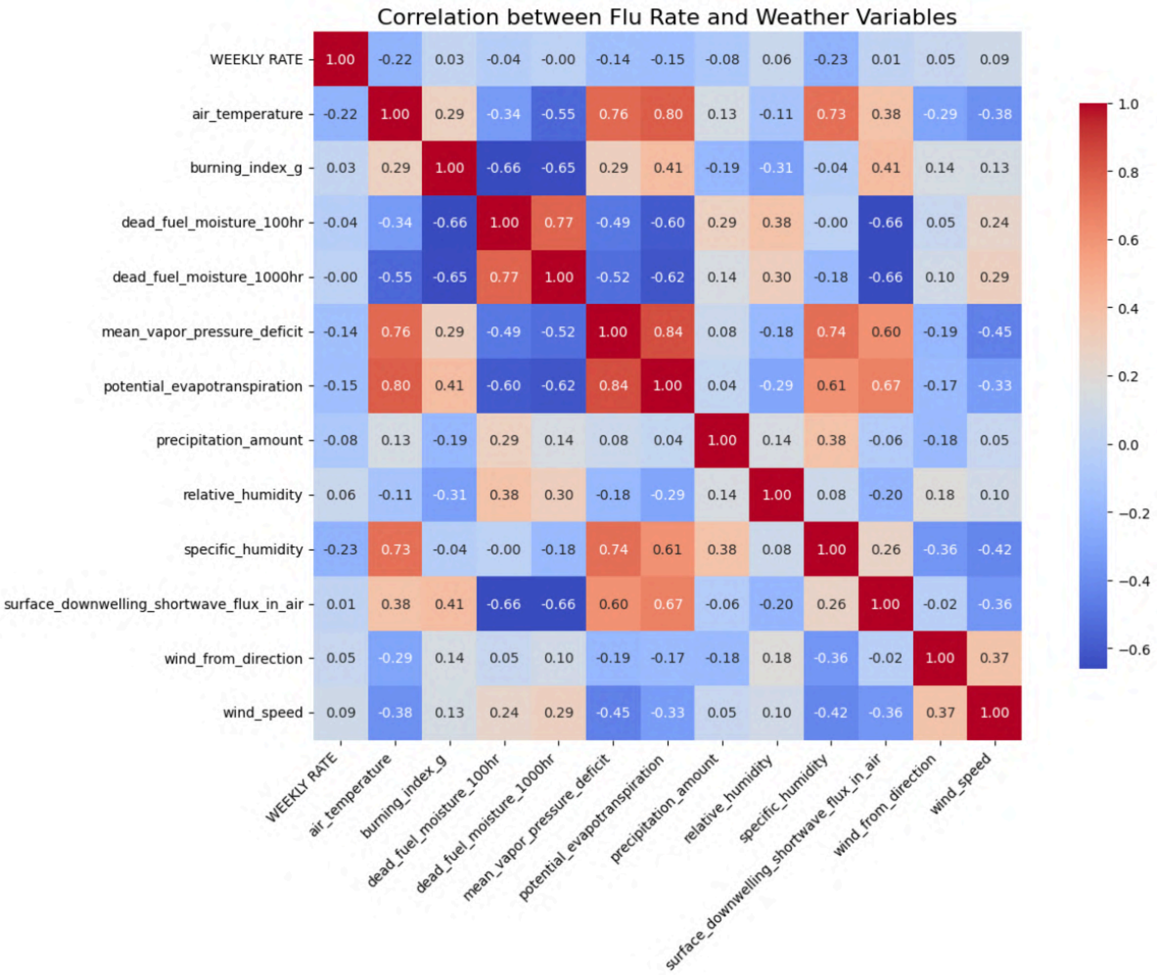




CORRELATION ANALYSIS

The correlation heatmap and tables revealed that **no single weather variable displayed a strong linear relationship** with weekly influenza hospitalization rates. The highest absolute correlation was modest (approximately −0.21) with potential evapotranspiration, suggesting that meteorological conditions alone are **insufficient** to reliably predict flu incidence. However, the analysis uncovered substantial intercorrelations among climate variables themselves—particularly between **mean vapor pressure deficit, air temperature**, and **specific humidity**—highlighting **clear underlying seasonal dynamics**. These strong associations between environmental factors imply redundancy that should be carefully managed during modeling, either through dimensionality reduction techniques or regularization. Overall, this finding reinforces that while weather contributes important contextual information, it is unlikely to drive accurate flu forecasting without integrating additional data sources such as viral surveillance or demographic trends.

While weather data alone may not robustly predict flu hospitalizations, it could still prove highly valuable if we broaden the focus to other climate-sensitive health outcomes, such as r**espiratory illnesses**, **dehydration during heat waves**, or **vector-borne diseases**.



	WEEKLY RATE	air_temperature	burning_index_g	dead_fuel_moisture_100hr	dead_fuel_moisture_1000hr	mean_vapor_pressure_deficit	potential_evapotranspiration	precipitation_amount	relative_humidity	specific_humidity	surface_downwelling_shortwave_flux_in_air	wind_from_direction	wind_speed
WEEKLY RATE	1.00	-0.22	0.03	-0.04	-0.00	-0.14	-0.15	-0.08	0.06	-0.23	0.01	0.05	0.09
air_temperature	-0.22	1.00	0.29	-0.34	-0.55	0.76	0.80	0.13	-0.11	0.73	0.38	-0.29	-0.38
burning_index_g	0.03	0.29	1.00	-0.66	-0.65	0.29	0.41	-0.19	-0.31	-0.04	0.41	0.14	0.13
dead_fuel_moisture_100hr	-0.04	-0.34	-0.66	1.00	0.77	-0.49	-0.60	0.29	0.38	-0.00	-0.66	0.05	0.24
dead_fuel_moisture_1000hr	-0.00	-0.55	-0.65	0.77	1.00	-0.52	-0.62	0.14	0.30	-0.18	-0.66	0.10	0.29
mean_vapor_pressure_deficit	-0.14	0.76	0.29	-0.49	-0.52	1.00	0.84	0.08	-0.18	0.74	0.60	-0.19	-0.45
potential_evapotranspiration	-0.15	0.80	0.41	-0.60	-0.62	0.84	1.00	0.04	-0.29	0.61	0.67	-0.17	-0.33
precipitation_amount	-0.08	0.13	-0.19	0.29	0.14	0.08	0.04	1.00	0.14	0.38	-0.06	-0.18	0.05
relative_humidity	0.06	-0.11	-0.31	0.38	0.30	-0.18	-0.29	0.14	1.00	0.08	-0.20	0.18	0.10
specific_humidity	-0.23	0.73	-0.04	-0.00	-0.18	0.74	0.61	0.38	0.08	1.00	0.26	-0.36	-0.42
surface_downwelling_shortwave_flux_in_air	0.01	0.38	0.41	-0.66	-0.66	0.60	0.67	-0.06	-0.20	0.26	1.00	-0.02	-0.36
wind_from_direction	0.05	-0.29	0.14	0.05	0.10	-0.19	-0.17	-0.18	0.18	-0.36	-0.02	1.00	0.37
wind_speed	0.09	-0.38	0.13	0.24	0.29	-0.45	-0.33	0.05	0.10	-0.42	-0.36	0.37	1.00

VIRUS TYPE CONTRIBUTIONS

To better understand drivers of flu-related hospital burden, we analyzed weekly hospitalization rates by virus type and age group.

Virus Type:

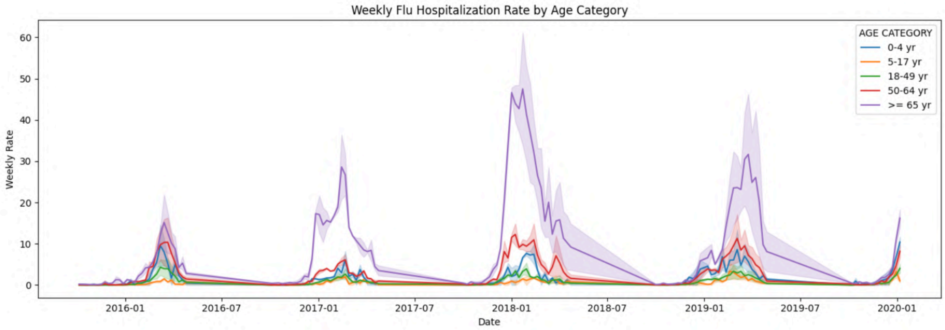
Time series plots show that **Influenza A is consistently the dominant contributor**, causing the highest peaks each winter. In contrast, **Influenza B results in lower rates overall** and often peaks slightly later. The bar chart of mean weekly rates highlights that Influenza A accounts for nearly **four times** the average burden compared to Influenza B. This distinction underscores the need to track virus strains separately when assessing severity and planning interventions.

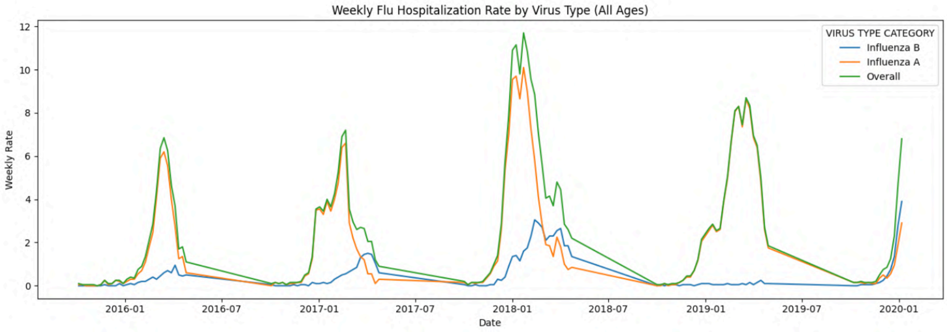
Age Groups:

Hospitalization rates differ substantially by age. The **≥65 years group is by far the most affected**, followed by those aged 50–64 years. Young children (0–4 years) also experience notable seasonal peaks. Meanwhile, rates among individuals aged 5–17 and 18–49 remain relatively low. These findings indicate that **age is a major factor in flu severity**, suggesting that models should include age or stratify predictions accordingly.

Key Insight:

Public health planning can be improved by targeting resources—such as ICU capacity and vaccination campaigns—toward older adults during peak flu seasons, and by monitoring virus-specific trends to anticipate surges.





C) Feature Engineering

To help the model understand patterns over time, we transformed the raw climate and flu data into a set of meaningful features. This step focused on capturing recent trends, seasonal effects, and interactions between climate conditions. The main techniques were:

Lag Features
We created features that represent the value of each climate variable from 1 to 8 weeks ago. This helps the model learn how past weather conditions may influence flu cases with a delay.

Rolling Statistics
For each variable, we calculated rolling averages, standard deviations, maximums, and minimums over windows of 2 to 8 weeks. These features capture recent trends and variability. For example, whether temperatures have been steadily rising or fluctuating.

Interaction Terms
We combined related climate variables into new features (e.g., humidity × temperature, wind × rainfall) to represent conditions that often occur together and may affect flu transmission, such as damp and windy weather.

Seasonality Features
Since flu is seasonal, we extracted the week of the year and applied sine and cosine transformations. This allows the model to recognize cyclical patterns, such as flu peaks in winter, without confusing week 1 and week 52 as being far apart.

Lagged Target Variable
To provide short-term context, we included the flu rate from the previous 1 to 3 weeks. This helps the model pick up on recent changes in flu activity.

Finally, we removed any rows with missing values caused by shifting or rolling operations. Together, these engineered features give the model a well-rounded view of both recent trends and seasonal context, key for forecasting flu rates based on environmental data.

D) Model

Model History

In this section, we conducted an exploratory data analysis to identify patterns linking climatic factors to influenza hospitalization rates in **Rochester, NY**. Specifically, we examined:

- Trends in weekly hospitalization rates over time.
- Seasonal and demographic patterns.
- Distributions and interrelationships among meteorological variables.
- Features with potential predictive value for modeling flu incidence.

Note: The dataset only includes typical flu seasons (October–April), limiting visibility into off-season baseline hospitalization trends.

Version	Model Type	Features	Tuning	RMSE	R2	Motivation
v1.0	XGBoost	All engineered features (lags, rolling, interactions, seasonality)	None	1.73	0.637	Initial full-featured model to set a baseline for performance
v2.0	XGBoost	Lag features + week_sin/cos (no climate)	Manual (n_estimators=200, learning_rate=0.1)	2.43	0.467	Test predictive power of autoregressive + seasonal features only
v3.0	XGBoost	All features	Randomized SearchCV	1.65	0.72	Improve performance through hyperparameter tuning
v4.0	XGBoost	Top 30 from feature importance	None	1.59	0.742	Reduce overfitting and training time by selecting top features

Version	Model Type	Features	Tuning	RMSE	R2	Motivation
v4.1	XGBoost	Top 30	Randomized SearchCV	1.55	0.758	Combine tuning with top features for better generalization
v5.0	XGBoost (Residual Correction)	Top 30 features (predicting residuals)	Randomized SearchCV	1.48	0.775	Use lag1 as baseline and predict only residual error
v5.1	XGBoost (Residual Correction)	Top 30	Optuna (30 trials)	1.42	0.789	Tune residual correction model for better hybrid performance
v5.2	XGBoost (Residual Correction)	Top 30	Optuna + SHAP	1.37	0.799	Add model explainability via SHAP to top-performing hybrid model
v6.0	XGBoost (Residual Correction)	Top 30	Optuna + diagnostics	1.43	0.781	Introduce diagnostics, reproducibility, and model saving
v6.1	XGBoost (Residual Correction)	Top 10 most important	Optuna	1.41	0.784	Test simplified version using top 10 features only
v6.2	XGBoost (Residual Correction)	Top 10 + correlation pruning	Optuna	1.35	0.821	Prune correlated features for a minimal, high-performing model

Selected Model

After iterating through multiple versions, we ultimately selected **v6.2**, an **XGBoost residual correction model** using the top 10 most important features with correlation pruning and Optuna tuning. This model achieved an **RMSE of 1.35** and an **R^2 of 0.821**, representing the highest accuracy & evaluation perfomance among all tested configurations.

Results

The Demonstrated Impact : Rochester (NY)

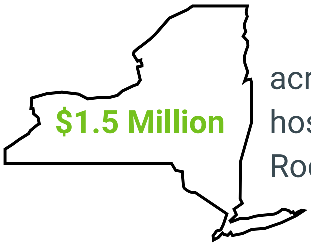
Baseline Model
RMSE = 2.43
$R^2 = 0.467$

Hybrid Model
RMSE = 1.34
$R^2 = 0.823$

The hybrid model significantly outperforms the baseline in predicting weekly influenza admission rates. This difference results in a 44.9% reduction in error and a 35.6 percentage point improvement in explained variance.

Since the RMSE is measured in admission rates per 100,000 population, and Rochester's 2020 population was 210 thousand, this translates to a reduction from approximately 5.14 to 2.83 unplanned admissions per week. Over the course of a year, this equates to roughly 120 fewer unplanned admissions.

Given that each influenza related hospital admission costs an average of $12,500, the improved accuracy of the hybrid model represents a potential annual cost saving of,



across all hospitals of Rochester (NY)

$125K per hospital

12 available hospitals within the city limits (American Hospital Association)

Market Scalability

According to the American Health Association's annual survey, there are more than 5,000 hospitals within the US, implying enough market size to scale the solution further. Furthermore, through the use of partnership with State and local public health departments, hospital networks or insurance providers it is very possible that we can have a 1% market penetration per year, which would mean 50 clients per year.

Solution Scalability

While our initial results were derived from the city-level dataset of Rochester (NY), we successfully applied the same modeling pipeline to state-level data from California with minimal adaptation. This demonstrates the modularity and robustness of our approach: regardless of geographic scale, the pipeline can be reused end to end.

In California, the global model trained on GRIDMET weather and temporal features achieved an RMSE of 0.45 and R^2 of 0.93, comparable to the performance observed in Rochester. This confirms that the model generalizes well across regions, as long as consistent weather and hospitalization data are available.

Monetisation Strategy

We propose offering the solution as a Model-as-a-Service (MaaS) product, delivered through a secure API layer that enables individual hospitals to access forecasts while ensuring proper handling of sensitive data and compliance with security standards.

The product will follow a subscription-based licensing model, priced per hospital, allowing for scalable adoption across institutions. Below is an outline of the proposed subscription tiers and the features included in each plan:

Product Tier	Product Offering	No. of Hospitals	Annual Price	Target Customer	Product ROI*
Pro	Forecast API & Dashboards	1	$100,000	Medium hospitals, regional clinics	125%
Enterprise	Multi-state forecasting, limited fine-tuning	5	$450,000	Large health systems, insurers, agencies	139%
Custom	custom models and fine tuning for different diseases	Unlimited	$2,000,000	Local Government agencies, national insurers	188% (for 30 hospitals)

* Product ROI calculated using the potential savings explained previously

Limitations & Future Improvements

To ensure both scientific soundness and business value, we critically evaluated the limitations of our current approach and outlined directions for future development. Below are the key areas we identified:

1 Regional Data Representation

Current Limitation:

At present, we reduce the high-resolution GRIDMET data (daily latitude-longitude points) to a single time series per region using a simple average across all grid cells within our selected polygon. While this method is efficient, it gives equal weight to all areas regardless of population density or land use.

This means that regions with low or zero population (e.g., lakes, forests) contribute as much as urban centers, potentially skewing results for our weekly regional representation.

Proposed Improvement:

Introduce **population-weighted aggregation**, using publicly available population density maps to weigh the influence of each point. This would provide a more accurate climate exposure signal, particularly when modeling effects on human health or infrastructure as this project intends to. The biggest restriction to achieve this is the computational cost of the process, which is going to be expanded on next.

2 Computational Constraints

Current Limitation:

Due to the size of the GRIDMET dataset (billions of points) and the limitation of working with our local machines, the Data Extraction pipeline prioritizes memory efficiency and minimal transformations. This forced us to generalize certain steps and **limit spatial representations** for the regions we wanted to extract data from to rectangular zones, instead of using complex polygons.

Proposed Improvement:

Deploy the pipeline on **cloud infrastructure** (e.g., Azure, AWS) to increase processing power and enable:

- Higher resolution extractions
- Region-specific tuning
- On-demand scalability

This would also allow real-time/ near-real-time data ingestion in production-grade systems.

3 Public Health Data Granularity

Current Limitation:

Our chosen business case focuses on modeling health outcomes (e.g., hospital admissions). However, available public datasets in the U.S. are highly aggregated due to **strict privacy regulations**. Data often comes in monthly or annual frequencies, and only at the state or county level, limiting modeling precision and insight.

Proposed Improvement:

To increase the model's predictive power, we propose sourcing **higher-frequency**, **de-identified admissions data**, ideally at the weekly and ZIP-code level. One promising way to achieve this is through the **Healthcare Cost & Utilization Project (HCUP)**, which offers granular hospitalization datasets via paid access.

4 Static Snapshot of gridMET

Current Limitation:

While the GRIDMET dataset on Microsoft's Planetary Computer was central to this challenge, it currently stops at 2020. As we operate in 2025, our entire model is effectively trained and evaluated on **historical data**, disconnected from today's climate reality.

Proposed Improvement:

To keep up with accelerating climate trends, we propose updating the dataset or **integrating newer sources** to reflect post-2020 conditions. This is crucial for ensuring that predictions and insights are grounded in today's world, not a "frozen in time" snapshot.

5 Modeling Limitations and Forecast Behavior

Current Limitation:

Our models favored interpretability and speed over complexity. We applied basic tree-based regressors without hyperparameter tuning and excluded important drivers like **vaccination rates** or **pandemic shocks**. No interaction models (e.g., Age 0–4 × Influenza A) were trained due to data gaps. As a result, forecasts are smoother than actual trends and often **miss sharp spikes**, especially in **low-sample groups like infants or Influenza B cases**.

Proposed Improvement:

Future work could adopt more advanced techniques such as **gradient boosting** or **neural networks** to better capture temporal dynamics. **Hyperparameter optimization** would likely improve accuracy and reduce overfitting. **Incorporating social and behavioral factors** (e.g., mobility data, public health interventions) and explicitly modeling interactions between age and virus type could further enhance performance. Finally, **probabilistic models or methods** designed to detect sudden surges may yield more responsive forecasts.

Conclusion

This project demonstrates the real-world potential of integrating open-access climate data with public health surveillance to forecast influenza-related hospital admissions. By leveraging gridMET's high-resolution meteorological indicators and the CDC's FluView dataset, we designed a **predictive framework capable of delivering actionable insights to healthcare systems** anticipating seasonal surges.

The model's performance, reducing prediction error by over 44% in Rochester and achieving an R^2 of 0.93 in California, validates the technical soundness and transferability of our approach across geographies. These improvements translate into tangible financial savings: **up to $1.5 million annually** for a single mid-sized city. When scaled nationally, the economic and operational impact for hospitals, insurers, and public health agencies becomes substantial.

Beyond forecasting, this solution opens the door to a broader class of climate-health intelligence products. Its modular architecture, grounded in reproducible data science practices, makes it **well-suited for commercial deployment** in the form of a Model-as-a-Service (MaaS) offering. As climate volatility intensifies and healthcare systems operate under increasing strain, our project illustrates a scalable path forward: where anticipatory data tools can **reduce inefficiencies**, **optimize resource allocation**, **and ultimately save lives**.

References

GitHub Repository - https://github.com/lucasihnen/microsoft_capstone_mbd_sept24_g3

- Agency for Healthcare Research and Quality. (n.d.). HCUP-US: Healthcare Cost and Utilization Project. Retrieved July 6, 2025, from https://hcup-us.ahrq.gov/
- American Hospital Association. (n.d.). AHA annual survey database. Retrieved July 6, 2025, from https://www.ahadata.com/aha-annual-survey-database
- Center for American Progress. (n.d.-a). Center for American Progress. Retrieved July 6, 2025, from https://www.americanprogress.org/
- Center for American Progress. (n.d.-b). Center for American Progress. Retrieved July 6, 2025, from https://www.americanprogress.org/
- Centers for Disease Control and Prevention. (n.d.). Flu hospitalization rates – CDC GRASP FluView. Retrieved July 6, 2025, from https://gis.cdc.gov/GRASP/Fluview/FluHospRates.html
- Centers for Disease Control and Prevention. (n.d.). Influenza disease burden: 2022–2023 data visualization. Retrieved July 6, 2025, from https://www.cdc.gov/flu-burden/php/data-vis/2022-2023.html
- Hu, T., Miles, A. C., Pond, T., Boikos, C., Maleki, F., Alfred, T., Lopez, S. M. C., & McGrath, L. (2024). Economic burden and secondary complications of influenza-related hospitalization among adults in the US: A retrospective cohort study. Journal of Medical Economics, 27(1), 324–336. https://doi.org/10.1080/13696998.2024.2314429
- ihnen, L. (2024). microsoft_capstone_mbd_sept24_g3 [GitHub repository]. Retrieved July 6, 2025, from https://github.com/lucasihnen/microsoft_capstone_mbd_sept24_g3
- National Library of Medicine. (n.d.). PubMed Central (PMC). U.S. National Institutes of Health. Retrieved July 6, 2025, from https://pmc.ncbi.nlm.nih.gov/
- New York State Department of Health. (n.d.). Influenza surveillance. Retrieved July 6, 2025, from https://health.ny.gov/
- Pines, J. M., Batt, R. J., Hilton, J. A., & Terwiesch, C. (2011). The financial consequences of lost demand and reducing boarding in hospital emergency departments. Annals of Emergency Medicine, 58(3), 331–340. https://doi.org/10.1016/j.annemergmed.2011.03.004
- University of Rochester Medical Center. (n.d.). URMC. Retrieved July 6, 2025, from https://www.urmc.rochester.edu