Projects

Team Formation and Project Selection Instructions

We have **14 students** participating in the program. To ensure a balanced and productive collaboration experience:

- Each team will have **2 or 3 members**
  - **3 members** is the optimal team size
  - **2 members** is also manageable
- A total of **5 project teams** will be formed:
  - **4 teams** will have 3 members
  - **1 team** will have 2 members

Project Preferences and Assignment

To match students with projects that best fit their interests:

- You will receive a link to a **Google Sheet** where you will rank the 4 project descriptions in order of preference:
  - Rank from **1 (top choice)** to **4 (least preferred)**
- Since there are 5 teams and only 4 projects, **two teams will be assigned to the same project topic**.
  - However, the project descriptions are intentionally broad and customizable, so each team will likely pursue a unique direction.

Mentorship, Timeline, and Deliverables

- Each team will be assigned a **faculty mentor** and a **graduate student mentor (TA)**.
- The **faculty mentor** will help your team:

  - Define a **timeline**
  - Establish **intermediate deliverables**
  - Guide the **scope** of the project, which may evolve over time
- A **GitHub repository** will be created for each team to:
  - Organize and document all work
  - Facilitate communication with mentors
  - Track progress and contributions

Expectations and Logistics

- **All efforts should be well documented**, including code, analysis, meeting notes, and decisions.
- Teams will meet during the scheduled **project meeting time: 4:00–5:00 PM**.
- Additional **lab time will be dedicated** to interaction with mentors and TAs.

Timeline Summary

- **Tuesday (end of day):** Submit your ranked project preferences
- **Wednesday:** Receive your team and project assignment
- **Post-assignment:** Coordinate a kickoff meeting with your team and mentors

Overview

This set of four customizable research projects explores the intersection of flood hazards, health outcomes, and environmental equity. While each project has a unique focus—ranging from predictive modeling and spatial risk analysis to the use of AI for extracting impact narratives—they are united by a shared goal: to generate actionable insights that support climate resilience and health equity.

Most projects concentrate on Washington, D.C., but the study area extends to neighboring counties in Maryland and Virginia to increase sample size and enhance analytical depth, particularly for historical flood events. Each project will be supervised by a faculty mentor and supported by a graduate student mentor, creating a collaborative learning environment for undergraduate researchers.

Together, these projects address several overarching themes:

- **Flood Risk Assessment:** Using meteorological, topographic, and spatial data to identify areas vulnerable to different types of flooding.

- **Health and Social Vulnerability:** Investigating how chronic diseases (like asthma and diabetes) and socioeconomic indicators (such as income, race, and housing conditions) intersect with flood exposure.
- **Environmental Justice:** Identifying disparities in exposure and resilience, especially in under-resourced communities.
- **Innovative Data Science Tools:** Applying methods such as random forest modeling, spatial regression, and large language models (LLMs) to structured and unstructured data.

Each project provides a clear direction and starting point but can be adapted to suit student interests and available data. The emphasis is on hands-on experience, interdisciplinary thinking, and the creation of impactful research products that can inform policy, planning, and public awareness.

Final Project Deliverables and Review Process

By the **morning of June 13**, each team is expected to submit a written document summarizing your progress toward your project goals. This **technical report** should include the following sections:

- **Introduction**
- **Methods**
- **Results**
- **Discussion**

The report must use clear technical language. Results and outcomes should be thoroughly described and supported by tables, graphs, and maps. This constitutes your **Tier 1 deliverable**.

Tier 2: Publishable Technical Report

Teams are encouraged to expand their work into a **publishable technical report**. This may include deeper analysis, refined writing, and enhanced visualizations. Faculty mentors will continue to support teams after the bootcamp.

Tier 3: Final Presentation

A formal **presentation will be held on September 7**, summarizing your project's motivation, methods, results, and conclusions. This presentation builds on the Tier 2 report and is intended for a broader audience.

Additional Deliverables

Some projects may also produce alternative outputs such as:

- Interactive dashboards

- Web-based maps
- Data products

All materials will be shared via **GitHub** or other publicly accessible platforms.

External Review

All projects will be reviewed by a panel of **community partners**, including representatives from **D.C. government, industry, and academia**. Their feedback will help assess the relevance, technical quality, and community impact of each project.

1 Classifying Flood Events by Impact in the Mid-Atlantic Region Using Random Forest Models

Objective

This project aims to develop a classification model that distinguishes flood events with recorded damages or injuries from those without, using historical NOAA storm event data (1996–present) for Virginia (VA), Maryland (MD), and Washington, D.C. The model will leverage event characteristics, precipitation levels, and sociodemographic data to understand the conditions under which flood events result in harm.

Data Sources

1. **NOAA Storm Events Database (1996–Present)**
   Focused on flood-related event types: *Flood*, *Flash Flood*, *Coastal Flood*. Key variables include:
   - Event date and time
   - Event type and location (county or subdivision)
   - Property and crop damage estimates
   - Injuries and fatalities
2. **Precipitation Data (Daymet or PRISM)**
   Daily precipitation values corresponding to the location and dates of each flood event.
3. **Sociodemographic Data (ACS or Census APIs)**
   Retrieved at county or finer spatial resolution. Variables include:
   - Median household income
   - Poverty rates
   - Population density
   - Race/ethnicity and age distribution

Methodology

1. Data Integration

Merge storm events with precipitation and demographic data by geolocation and event date. Label each event as:

- **Class 1:** Flood events *with* recorded damage or injury
- **Class 0:** Flood events *without* recorded damage or injury

2. Feature Engineering

- Event characteristics: type, season, duration, precipitation amount
- Location context: urban/rural designation, elevation (optional)
- Demographics: income, population vulnerability indicators

3. Modeling Approach

- Use a **Random Forest Classifier** implemented in Python (`scikit-learn`)
- Apply cross-validation and assess feature importance
- Evaluate using accuracy, precision, recall, F1 score, ROC AUC

4. Tools and Technologies

- **Language:** Python
- **Libraries:** `pandas`, `numpy`, `scikit-learn`, `matplotlib`, `geopandas`, `xarray`
- **APIs and Data Portals:** NOAA Storm Events, Daymet, U.S. Census Bureau

Expected Outcomes

- A predictive model classifying flood events based on recorded damage/injury
- Identification of key predictors such as precipitation and demographics
- Visualizations: feature importance plots, impact maps
- A framework extensible to other regions or disaster types

Applications

- Emergency management: Identify high-risk flood conditions
- Environmental justice: Understand links between harm and community vulnerability
- Regional resilience planning

2 Environmental Equity, Health Vulnerability, and Topographic Flood Susceptibility in Washington, D.C.: A Spatial Analysis of Risk and Resilience

Project Objective

This project aims to assess the intersection of **topographic flood susceptibility**, **social vulnerability**, and **health vulnerability** (specifically asthma and diabetes prevalence) in Washington, D.C., with the goal of identifying communities facing compounded risks and informing equitable climate resilience strategies.

Research Questions

1. Are socially and medically vulnerable populations (e.g., low-income, racial minorities, renters, people with asthma or diabetes) more likely to live in topographically flood-prone areas?
2. Does proximity to major water bodies increase exposure among these populations?
3. Where do topographic flood risk, social vulnerability, and health burden overlap, and what are the implications for environmental justice?

Study Area

- Washington, D.C.
- Primary hydrological features: *Anacostia River*, *Potomac River*, *Rock Creek*
- Flood-prone areas assessed using topographic indicators and FEMA flood zones

Data Sources

- FEMA Flood Hazard Maps
- USGS Digital Elevation Models (DEMs): elevation, slope, and flow accumulation
- CDC PLACES data: census tract–level estimates of asthma and diabetes prevalence
- American Community Survey: indicators of social vulnerability (income, race/ethnicity, housing tenure, age)
- DC GIS: river networks and land use data

Methodology

1. Topographic Susceptibility Mapping

Use DEMs to compute elevation, slope, and hydrological flow paths. Identify low-elevation, high-flow accumulation zones as flood-susceptible areas.

2. Hydrological Exposure

Calculate Euclidean (straight-line) distance from residential areas to the nearest river. Proximity is treated as an additional flood exposure factor.

3. Health Vulnerability Mapping

Integrate CDC PLACES estimates of asthma and diabetes. Normalize and map these variables to examine their spatial distribution relative to flood susceptibility.

4. Spatial Overlay and Analysis

Overlay topographic flood risk, social vulnerability, and health indicators to identify spatial clusters where risks converge.

5. Statistical Modeling

Apply regression models to quantify the relationships between vulnerability factors and flood susceptibility:

$$\text{FloodSusceptibility} = \beta_0 + \beta_1\text{Income} + \beta_2\text{Race} + \beta_3\text{Asthma} + \beta_4\text{Diabetes} + \beta_5\text{Elevation} + \beta_6\text{DistanceToWater} +$$

Expected Outcome

- A composite vulnerability map of Washington, D.C. highlighting neighborhoods with overlapping flood risk, chronic disease burden, and social disadvantage.
- Evidence-based recommendations for equity-centered, health-aware flood resilience planning.

3 Mapping Vulnerability Hotspots in Washington, D.C.: Interactions Between the Built Environment, Flood Risk, and Health Equity

Objective

To identify and map vulnerability hotspots in Washington, D.C. where disparities in the *built environment* interact with *flood risk* to *modify* the relationship between place-based conditions and adverse health outcomes, particularly chronic diseases such as asthma and diabetes.

Rather than simply adding up risk factors, the project tests how flood exposure *intensifies or alters* the effects of built environment conditions on health, providing evidence to support the Sustainable DC 2.0 goal of reducing place-based health disparities by 15% by 2032.

Research Questions

- Which areas of D.C. show compounded vulnerability due to poor built environment features and high flood risk?
- Does flood exposure *modify* the relationship between features of the built environment (e.g., green space, housing density, walkability) and health outcomes (e.g., asthma, diabetes)?
- Can we define and track a composite vulnerability index tied to built environment disparities, climate resilience, and health equity over time?

Conceptual Innovation

- Moves beyond additive models by testing *effect modification*, e.g., is the impact of poor housing on asthma stronger in flood-prone areas?
- Frames health inequity as a place-based phenomenon with dynamic risk interactions.
- Supports long-term policy tracking aligned with DC's 2032 equity goals.

Key Variables

- **Health Outcomes:** Census tract–level asthma and diabetes prevalence (CDC PLACES).
- **Built Environment Indicators:**
  - Housing quality proxies (e.g., age of housing, overcrowding)
  - Access to parks and green space
  - Street connectivity and walkability
  - Proximity to public transit
  - Heat island exposure
- **Flood Risk:** Overlap with FEMA flood zones via spatial analysis.
- **Sociodemographics:** Percent below poverty, percent minority, percent uninsured, percent without vehicle (ACS).

Methodology

- **Interaction Modeling:** Linear or spatial regression models with interaction terms (e.g., housing quality $\times$ flood risk).
- **Hotspot Mapping:** Create a composite vulnerability index and map high-burden tracts.
- **Scenario Projection:** Outline a framework for evaluating progress toward the 2032 target by tracking composite vulnerability over time.

Deliverables

- Vulnerability Hotspot Map (static and interactive)
- Visualizations of interaction effects (e.g., stratified associations)
- Policy recommendations for neighborhood-level intervention and resilience planning

4 Narrative Intelligence and Impact Signatures of Flash Flood Events in DC, Maryland, and Virginia

Project Objectives

1. Download and filter flash flood event data from 1996 to present for Washington, D.C., and neighboring counties in Maryland and Virginia, using only events that contain narrative descriptions.
2. Apply large language models (LLMs) to extract structured impact tags from unstructured event narratives.
3. Quantify flood severity by computing a composite **Flood Impact Score (FIS)** for each event.
4. Explore contextual patterns through co-occurrence of impact types, spatial clustering, and temporal trends.
5. Develop a data-driven understanding of high-impact flood typologies to inform resilience and emergency planning.

LLM-Based Narrative Tagging

Each narrative is processed with an LLM to classify presence or absence of the following impact indicators:

Tag	**Description**
`death`	Human fatalities
`injury`	Physical injuries
`evacuation`	Any form of population displacement
`rescue`	Emergency or civilian water rescues
`car_crash`	Explicit vehicle collisions or crashes
`home_damage`	Flooded or damaged residences
`infrastructure_damage`	Roads, bridges, water plants, etc.
`school_damage`	Impacts on schools or educational buildings
`hospital_damage`	Impacts on hospitals or clinics
`road_closure`	Blocked roads or intersections

`power_outage`	Loss of electricity or power infrastructure

Quantifying Combined Impact: Flood Impact Score (FIS)

Each event receives a **Flood Impact Score (FIS)** — a weighted sum of the above tags that quantifies overall severity based on:

- Human impact
- Physical infrastructure disruption
- Service system breakdowns

Weights may be adjusted based on expert judgment, statistical learning, or downstream modeling needs.

Descriptive and Exploratory Analysis Plan

Tag Frequency Analysis

- Distribution of impact types across all events.
- Cross-tabulations by state and decade.

Tag Co-Occurrence Networks

- Heatmap and network visualizations showing which tags appear together most often.
- Identification of characteristic combinations (e.g., `rescue + car_crash + home_damage`).

Spatial Analysis

- Mapping average and maximum FIS by county (VA/MD) and Washington, D.C.
- Detecting spatial clusters of high-impact events using local spatial statistics.

Temporal Trends

- Time series of flood counts and average FIS.
- Analysis of changes over time in narrative complexity or dominant flood typologies.
- Examination of seasonal patterns (e.g., spring snowmelt vs. summer storms).

Advanced Contextual Analyses

Flood Typology Clustering

- Use of LLM-labeled tags to cluster floods into categories such as:
  - "High-disruption, low-casualty"
  - "Urban flash floods with rescues"
  - "Infrastructure-heavy winter floods"

Risk Prediction Modeling (Optional)

- Predicting FIS using metadata: date, location, damage estimates.

Expected Deliverables

- A structured event-level dataset with LLM-labeled impacts.
- A derived Flood Impact Score for each event.
- Visual analytics: tag frequency plots, spatial heatmaps, co-occurrence networks.
- A replicable method to use LLMs for extracting disaster intelligence from text.