Replication Data README

Hybrid Supervised–Unsupervised Modeling for Post-Hurricane Private Well Contamination Risk Score Using Empirical Validation and Community Ground-Truthing

1 Description

This repository provides the data required to reproduce all tables and figures presented in the manuscript and its Supporting Information.

The study develops and validates a hybrid supervised–unsupervised modeling framework to construct a composite private well contamination risk score in western North Carolina following Hurricane Helene. The dataset includes spatial indices for three modules—hazard, physical vulnerability, and social capacity—as well as the composite risk score and validation data from well testing.

2 Subject Area

Earth and Environmental Sciences

3 License

This dataset is released under the **Creative Commons CC0 1.0 Universal Public Domain Dedication (CC0 1.0)**.

The data may be freely used, modified, and distributed without restriction.

4 File Overview

All files are provided in **GeoPackage (.gpkg)** format.

Module-Level Indices

- `index_hazard.gpkg`
  Contains the constructed hazard index.
- `index_vulnerability.gpkg`
  Contains the constructed physical vulnerability index.

- `index_capacity.gpkg`
  Contains the constructed social capacity index.

Detailed Module Data (Variables + Principal Components)

- `index_hazard_detailed.gpkg`
  Contains original hazard-related variables, constructed principal components, and the hazard index.
- `index_vulnerability_detailed.gpkg`
  Contains original physical vulnerability variables, constructed principal components, and the physical vulnerability index.
- `index_capacity_detailed.gpkg`
  Contains original social capacity variables, constructed principal components, and the social capacity index.

Composite Risk Score

- `nc_grid_index.gpkg`
  Contains the composite contamination risk score derived from the three module-specific indices.

Supporting Gridded Data for Summary Statistics

- `nc_grid_vulnerability_929.gpkg`
  Used to compute descriptive statistics for physical vulnerability variables.
- `nc_grid_vulnerability_imputed.gpkg`
  Contains imputed physical vulnerability variables used in index construction and summary statistics.

Validation Data

- `well_tests.gpkg`
  Contains private well testing results used for empirical validation of the composite risk score.

5 Data Structure

All GeoPackage files include spatial polygon layers representing grid cells across the 25 western North Carolina counties included in the study.

Common fields may include:

- Grid identifiers
- County identifiers
- Module-specific principal components
- Module-specific indices

- Composite risk score (where applicable)

Coordinate reference system (CRS) information is embedded within each GeoPackage file.

6 Software Requirements

The data can be read using:

- **R** (packages: `sf`, `dplyr`, `terra`)
- **Python** (packages: `geopandas`, `fiona`)
- GIS software such as QGIS or ArcGIS

Example in R:

```
library(sf)
hazard <- st_read("index_hazard.gpkg")
```

7 Reproducibility

The data in this repository reproduce:

- All tables in the manuscript
- All figures in the manuscript
- All tables and figures in the Supporting Information

Associated code for reproducing results is available in the corresponding GitHub repository (see manuscript Data and Software Availability Statement).

8 Contact

Jilei Lin
Department of Statistics
The George Washington University