Plan of Work

1. Title: Effects of Flood Risk and Climate Change on U.S. Census Tract-Level Health Outcomes

2. Lead Author Name: Alvin Sheng

3. Co-authors, Contact Information, and Responsibilities:

Name	Contact Information	Responsibilities
Kyle P. Messier	kyle.messier@nih.gov	Preceptor at NIEHS
Brian J. Reich	bjreich@ncsu.edu	Preceptor at NCSU

5. Background/Rationale:

Floods have been linked to various health outcomes such as mental disorders and chronic diseases. This is likely due to psychosocial and post-traumatic stress caused by natural disasters and inadequate responses to them[1].

6. Research Questions & Hypotheses:

Aim #1: To investigate associations between flood risk and health outcomes, including prevalence of coronary heart disease, asthma, high blood pressure, and poor mental health.

> ***Hypotheses:*** *We hypothesize that higher flood risk is associated with worse health outcomes.*

7. Data:

a. Study domain and/or population:

All census tracts in the conterminous United States.

b. Study years:

- Outcomes: Prevalence of
    i. *Coronary heart disease (2018)*
    ii. *Asthma (2018)*
    iii. *High blood pressure (2017)*
    iv. *Poor mental health (2018)*
- Exposures:
    i. Flood Risk: 2020 (present) and 2050 (climate-adjusted future)
- Other Covariates:
    i. CACES air pollution (2015)
    ii. Smoking prevalence (2018)
- Mediators, Moderators, etc.
    i. CDC SVI: 2018

There may be a mismatch of years between the outcome and exposures. The assumption is that the exposure doesn't change drastically over the short term. An example of a study that has used data with similarly mismatched years is Wu et al. (2020)[2], which is a U.S. county-level cross-sectional study examining the association between air pollution and COVID-19 mortality.

c. Outcomes[3]:

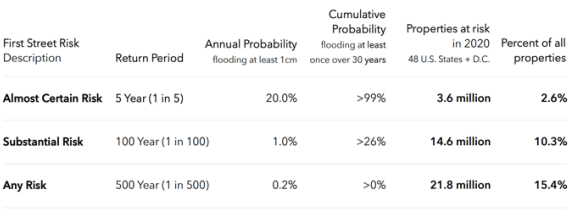
Outcome Type	**Variable Names and Description of Variable from Orig Source**	**Details from Orig Source**	**Analytical Treatment**
Specific Health Outcomes	"Current asthma among adults aged >=18 years" "High blood pressure among adults aged >=18 years" "Mental health not good for >=14 days among adults aged >=18 years" "Coronary heart disease among adults aged >=18 years" "Physical health not good for >=14 days among adults aged >=18 years" And 8 other health outcomes	In addition to the health outcomes of interest on the left, there are 5 chronic disease-related unhealthy behaviors, and 10 on use of preventative services. https://chronicdata.cdc.gov/500-Cities-Places/PLACES-Local-Data-for-Better-Health-Census-Tract-D/cwsq-ngmh Data sources used to make dataset include BRFSS 2018 data (2017 for HBP and cholesterol on left side), Census Bureau 2010 population data, and ACS 2014-2018 or 2013-2017 estimates	

d. Covariates:

- Property flood risk[4]:

The First Street Foundation (FSF) model calculates the flood risk at every property in the contiguous United States. Source of dataset: https://registry.opendata.aws/fsf-flood-risk/. The "details from original source" below can be found in https://assets.firststreet.org/uploads/2020/06/first_street_foundation__first_national_flood_risk_assessment.pdf.

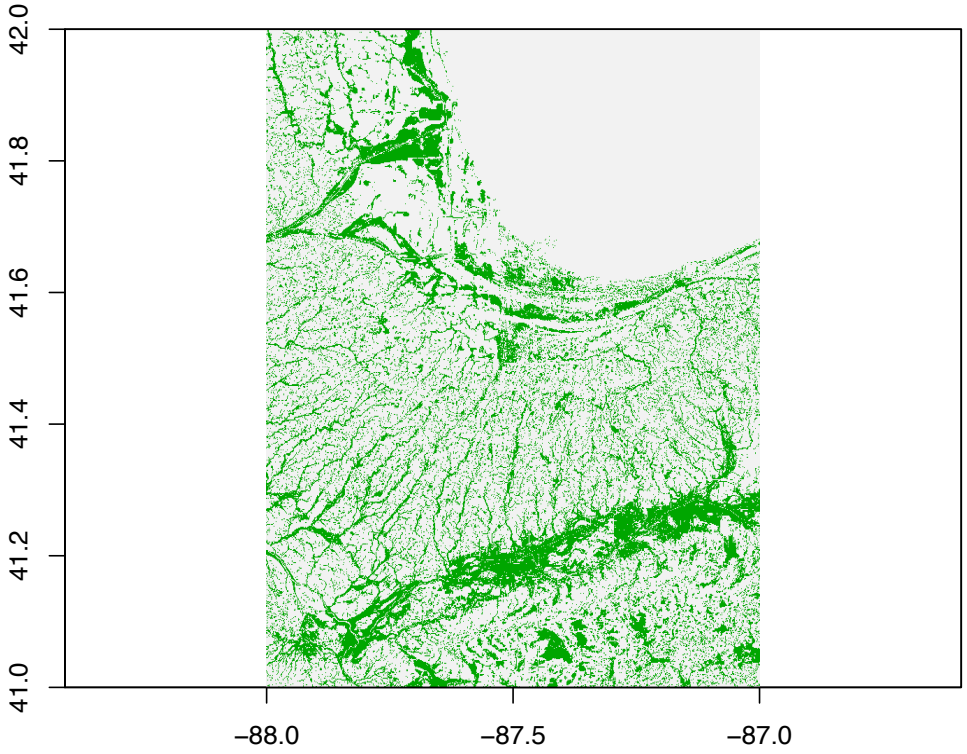
Because the data is given by zip codes, I use the 2010 ZCTA to Census Tract Relationship File Layout (https://www.census.gov/programs-surveys/geography/technical-documentation/records-layout/2010-zcta-record-layout.html#par_textimage_3) to merge the property flood risk data with the rest of the data by census tract.

Flood Risk Type	**Variable Names and Description of Variable from Orig Source**	**Details from Orig Source**	**Analytical Treatment**
Comparison with Federal Emergency Management Agency (FEMA) Special Flood Hazard Areas (SFHA)	count_property (the number of First Street properties in the census tract), count_fema_sfha (number of properties in FEMA SHFA), pct_fema_sfha (percent of properties in FEMA SFHA), pct_fs_fema_difference_2020 (percent difference between number of First Street properties and FEMA properties at risk in 2020)	FEMA classifies 8.7 M properties as having substantial risk (1% annual), i.e. within SFHAs. By contrast, the FSF classifies 14.6 M properties with same level of risk. Discrepancy is due to FSF using current climate data, mapping precip as a stand-alone risk, and includes areas FEMA doesn't (https://firststreet.org/mission/)	These variables will probably not be directly used in the model. They may be used to process other variables.
Percent of First Street Properties at 3 levels of severity and 2 time points	pct_fs_risk_2020_5, pct_fs_risk_2050_5, pct_fs_risk_2020_100, pct_fs_risk_2050_100, pct_fs_risk_2020_500, pct_fs_risk_2050_500. 2020 refers to present-time, and 2050 refers to the climate adjusted future. See right for the 5, 100, 500.	First Street definitions of risk that are used in this report. ***Substantial risk*** **is analogous to the FEMA SFHA designation.**  According to environmental factors, there will be ~11% increase in flood risk over the next 30 years (to 2050).	Can subtract 2020 variable from 2050 variable to get percent change in properties at certain risk
Average Risk Score of Properties	avg_risk_score_all, avg_risk_score_2_10, avg_risk_fsf_2020_100, avg_risk_fsf_2020_500, avg_risk_score_sfha, avg_risk_score_no_sfha	The Flood Factor (FF) is an indicator of a property's practical flood risk from 1 to 10. High flood factors correspond to being more likely to flood and/or more likely to experience high floods. FF is determined by the property's likelihood of flooding and the potential depth of that flood. Flood risks accumulate over time, so FF	

Percent of Properties with a given Flood Factor	pct_floodfactor1, …, pct_floodfactor10	specifically looks at the likelihood of water reaching the building/center of empty lot at least once within the next 30 years. Properties with less than 0.2% chance of experiencing any depth of flooding in any year within the next 30 years have FF of 1 (minimal risk).	Divide by count_property

- Raster flood risk[4]:

The FSF model also predicts flood depth at a 3-meter level resolution. Each pixel represents flood depth in centimeters.



Raster for the Chicago region. Green pixels have flood depths; gray pixels have no calculated flood depths.

Flood risk raster data are available for selected combinations of the below factors. The percentile refers to the quartiles of a distribution of environmental outcomes based on simulations using the RCP 4.5 curve and the GCM ensemble

(https://help.floodfactor.com/hc/en-us/articles/360049241313-How-is-future-environmental-change-incorporated-into-Flood-Factor-). The return periods refer to floods of a given probability. For instance, 1 in 5 refer to floods with a 20% probability of occurring in a given year. Rarer floods tend to be more severe with higher flood depths (https://floodfactor.com/methodology).

Factor	Levels
Year	2020, 2035, 2050
Percentile	0p25, 0p50, 0p75
Return Period	1 in 2, 1 in 5, 1 in 20, 1 in 100, 1 in 250, 1 in 500

Summary statistics of flood depth values (mean, percentiles, etc.) can be extracted for each census tract to be used in the model (Section 8).

- Confounders/other covariates

Other Covariate Type	Variable Names and Description of Variable from Orig Source	Details from Orig Source	Analytical Treatment
Smoking Prevalence[2]	"Current smoking among adults aged >=18 years"	Same data source as the Outcomes (see table in section 7c.), i.e. PLACES Local Data for Better Health	
CACES LUR Air Pollution[5] (https://www.caces.us/data)	Population-weighted concentration based on block level centroid predictions for 6 pollutants: co (ppm), no2 (ppb), o3 (ppb), pm10 (µg/m^3), pm25 (µg/m^3), so2 (ppb) Population-weighted latitude and longitude based on block	Citation: "This article includes concentration estimates developed by the Center for Air, Climate and Energy Solutions using v1 empirical models as described in Kim S.-Y.; Bechle, M.; Hankey, S.; Sheppard, L.; Szpiro, A. A.; Marshall, J. D. 2020. "Concentrations of criteria pollutants in the contiguous U.S., 1979 – 2015: Role of prediction model parsimony in integrated empirical geographic regression." PLoS ONE 15(2), e0228535. DOI: 10.1371/journal.pone.0228535 ."	Data is for the year 2015. There are other years available. Model estimates except for O3 are annual-average values. Ozone model estimates are the average during May-Sept of the daily maximum 8-hr moving average. Either way, only for years

	level centroid: lat/lon		with available monitoring data.

- Moderators[6]:

CDC Social Vulnerability Index (SVI) https://www.atsdr.cdc.gov/placeandhealth/svi/data_documentation_download.html. All variables are calculated from the 5-year American Community Survey (2014-2018 for the 2018 SVI version).
There are four themes of social vulnerability: socioeconomic, household composition/disability, minority status/language, housing type/transportation. The EPL_ variables (see below) are percentile ranks for each of the variables, ordered by census tract. Higher values of the EPL_ variables indicate higher social vulnerability.

There are several prefixes that can go before each variable listed in the next table.

Prefix	**Meaning**
EP_	Percentage of …
MP_	Margin of error for the percentage of… Can be incorporated in BHM.

SVI Type	**Variable Names and Description of Variable from Orig Source**	**Details from Orig Source**	**Analytical Treatment**
Description of Census Tract	TOTPOP (population), HU (# housing units), HH (# households)		These variables will not be directly used in the model. They may be used to process other variables.
Socioeconomic	POV (below poverty) UNEMP (16+ unemployed) PCI (per capita income) NOHSDP (25+ no high school)	E_PCI/EP_PCI and M_PCI/MP_PCI are the same	
Household Composition/Disability	AGE65 (≥ 65 y.o.) AGE17 (≤ 17 y.o.) DISABL (civilian noninstitutionalized w/ disability) SNGPNT (single parent household with < 18 y.o. children)		

Minority Status/Language	MINRTY (all except white non-hispanic) LIMENG (≥5 y.o. speak English "less than well")		
Housing Type/Transportation	MUNIT (housing in structures w/ ≥ 10 units) MOBILE (mobile homes) CROWD (household level, more people than rooms) NOVEH (households with no vehicles) GROUPQ (persons in group quarters)		
Other Variables	UNINSUR (those w/o health insurance in the total civilian noninstitutionalized population) E_DAYPOP (estimated daytime population)	UNINSUR has E_, M_, EP_, MP_ versions These variables are excluded from the SVI rankings	

e. Missingness/ Exclusion criteria:

The outcomes, flood risk variables, and CDC SVI have a few NAs (maximum is ~0.6% values missing in a column).

8. Statistical Analysis Plan and Methods:

a. Spatial Data Wrangling

a. Census Tracts

Because the property flood risk data is aggregated by zip code rather than census tract, the flood risk data was merged with the census-tract level data according to the 2010 Zip Code Tabulation Area (ZCTA) Relationship File (https://www.census.gov/programs-surveys/geography/technical-documentation/records-layout/2010-zcta-record-layout.html). Each census tract was assigned an average of the flood risk values for the zip codes covered by the census tract. The average was weighted by the proportion of housing units in the census tract that is covered by each zip code.

I use the census tract adjacency file provided by the Diversity and Disparities project (https://s4.ad.brown.edu/projects/diversity/index.htm) to construct the census tract adjacency matrix needed for the Bayesian hierarchical model as described later. The adjacency matrix is binary, where 1

indicates pairs that are neighbors and 0 indicates pairs involving the same census tract or census tracts that are not neighbors.

For constructing maps, I use the 2010 TIGER/Line Shapefiles to get the boundaries of the census tracts corresponding to the 2010 census: https://www.census.gov/cgi-bin/geo/shapefiles/index.php?year=2010&layergroup=Census+Tracts.

b. Data Pre-Processing

a. Dimensionality Reduction

There are a couple of unsupervised methods to reduce the large number of flood risk variables to a small subset of interest. In particular, the property flood risk variables and summary statistics derived from flood risk raster data would be highly correlated. One method is the variance inflation factor (VIF), which indicates how much multicollinearity there is in a group of covariates. The function `vifstep` in the package `usdm` can be used to iteratively exclude highly correlated variables through a stepwise procedure[7].

Another way is to use factor analysis to find a set of latent flood risk factors that explain the other flood risk variables.

b. Modeling

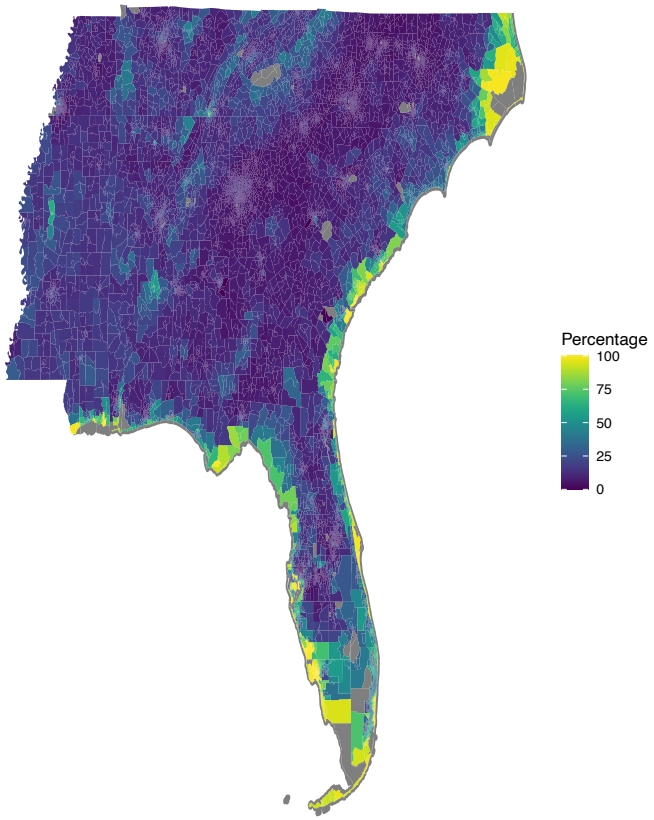
We will fit a Bayesian hierarchical model (BHM) where the outcome variables at a census tract are linear functions of the covariates discussed in section 7. To account for the spatial correlation among census tracts, we will use a Gaussian multivariate conditional autoregressive (MCAR) prior.

Missing outcome values will be treated as additional unknown parameters that are updated via data augmentation. Missing covariate values will be mean imputed.

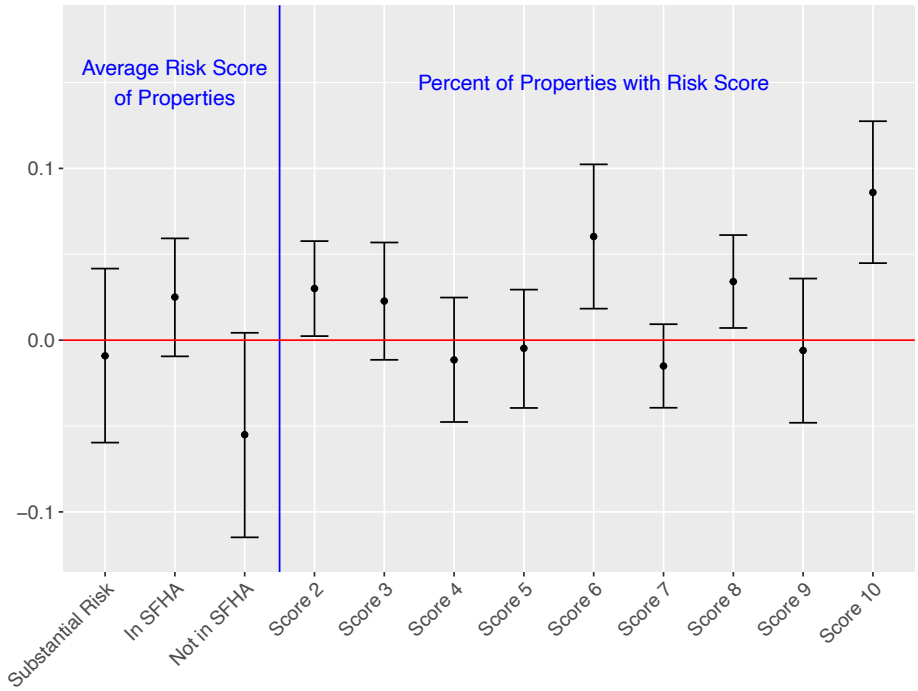
Uncertainty intervals (95% and 50%) are available for the outcomes and most covariates. Thus, measurement error for a variable can be incorporated in the model by assuming that the measurement error is normal with a known deviation derived from the uncertainty interval (https://mc-stan.org/docs/2_21/stan-users-guide/bayesian-measurement-error-model.html).

- List of expected or potential figures/graphics:

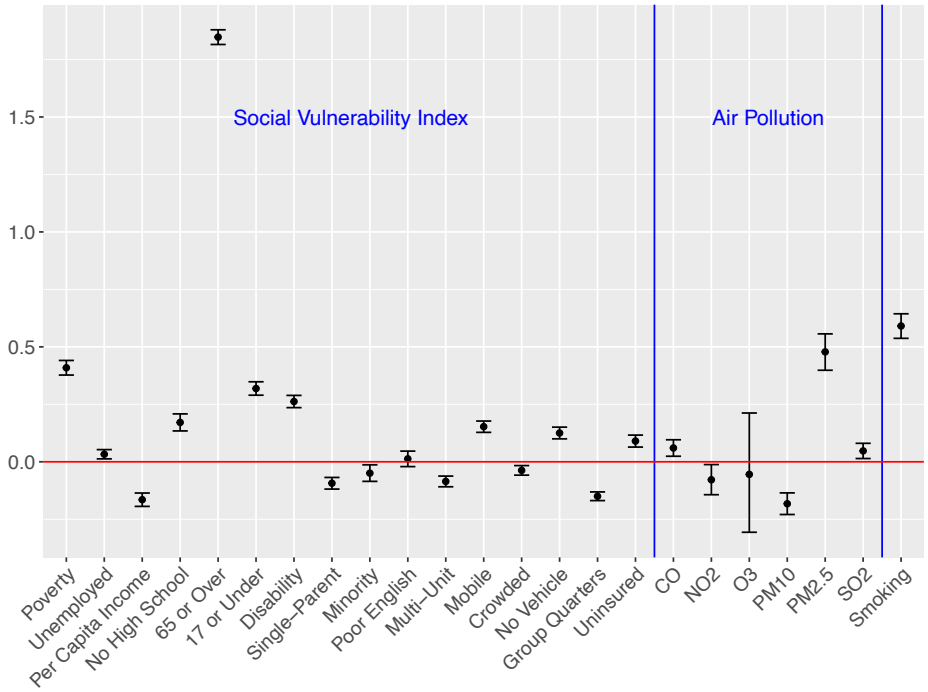
The below figures and graphics are only for seven states in the Southeastern US: North Carolina, South Carolina, Tennessee, Georgia, Alabama, Mississippi, and Florida. These figures and graphics will be extended to the entire contiguous United States.



Percent of Properties with Any Risk of Flooding in each Census Tract.



95% Credible Intervals of Coefficients for Flood Risk Standardized Covariates



95% Credible Intervals of Coefficients for Other Standardized Covariates

9. Anticipated pitfalls/challenges and limitations

- Limitations: Assumptions needed to conduct causal inference may not be valid for the above model. Thus, associations between flood risk and health outcomes should not be interpreted causally.

10. Manuscript Timeline

Goal: Finish a rough draft of the manuscript by mid-February.

11. References:

1. Hsin-I Shih, Tzu-Yuan Chao, Yi-Ting Huang, Yi-Fang Tu, Tzu-Ching Sung, Jung-Der Wang, and Chia-Ming Chang. Increased medical visits and mortality among adults with cardiovascular diseases in severely affected areas after Typhoon Morakot. International Journal of Environmental Research and Public Health, 17(18), September 2020.
2. Xiao Wu, Rachel C Nethery, M Benjamin Sabath, Danielle Braun, and FrancescaDominici.Exposure to air pollution and covid-19 mortality in the united states.medRxiv, April 2020.

3. Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, and Division of Population Health. Places: Local data for better health, census tract data 2020 release. https://chronicdata.cdc.gov/500-Cities-Places/PLACES-Local-Data-for-Better-Health-Census-Tract-D/cwsq-ngmh, January 2021.
4. First Street. First street foundation flood risk summary statistics. https://registry.opendata.aws/fsf-flood-risk/, May 2021.
5. Sun-Young Kim, Matthew Bechle, Steve Hankey, Lianne Sheppard, Adam A. Szpiro, and Julian D. Marshall. Concentrations of criteria pollutants in the contiguous U.S., 1979 – 2015: Role of prediction model parsimony in integrated empirical geographic regression. PLoS ONE, 15(2), 2020.
6. Centers for Disease Control and Prevention, Agency for Toxic Substances and Disease Registry, and Geospatial Research, Analysis, and Services Program. CDC/ATSDR Social Vulnerability Index 2018 Database US. https://www.atsdr.cdc.gov/placeandhealth/svi/data documentation download.html, April 2021.
7. Naimi, B., Hamm, N.A.S., Groen, T.A., Skidmore, A.K., and Toxopeus, A.G. 2014. Where is positional uncertainty a problem for species distribution modelling?, Ecography 37 (2): 191-203.
8. Duncan Lee. CARBayes version 5.2.3: An R Package for Spatial Areal Unit Modelling with Conditional Autoregressive Priors. University of Glasgow.