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 72539 census tracts (based on the 2010 census) 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. GRIDMET temperature and relative humidity (2005-2020)
    ii. CACES air pollution (2000-2015)
    iii. 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]:

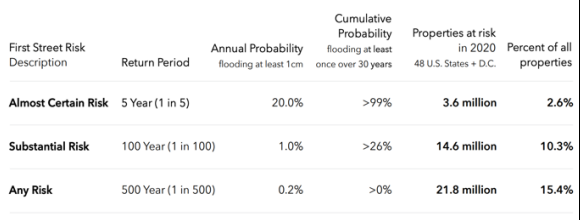
Outcome Type	**Variable Names and Description of Variable from Orig Source**	**Details from Orig Source**	**Analytical Treatment**
Specific Health Outcomes	"Current asthma among adults aged >=18 years" "High blood pressure among adults aged >=18 years" "Mental health not good for >=14 days among adults aged >=18 years" "Coronary heart disease among adults aged >=18 years"	In addition to the health outcomes of interest on the left, there are 9 other health outcomes, 5 chronic disease-related unhealthy behaviors, and 10 variables 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), 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 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	
Percent of Properties with a given Flood Factor	pct_floodfactor1, …, pct_floodfactor10		Divide by count_property

First Street Risk Description	Return Period	Annual Probability flooding at least 1cm	Cumulative Probability flooding at least once over 30 years	Properties at risk in 2020 48 U.S. States + D.C.	Percent of all properties
Almost Certain Risk	5 Year (1 in 5)	20.0%	>99%	**3.6 million**	**2.6%**
Substantial Risk	100 Year (1 in 100)	1.0%	>26%	**14.6 million**	**10.3%**
Any Risk	500 Year (1 in 500)	0.2%	>0%	**21.8 million**	**15.4%**

		year within the next 30 years have FF of 1 (minimal risk).	

- 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	
Center for Air, Climate and Energy Solutions (CACES) Land Use Regression (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 level centroid: lat/lon	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 averaged over 16 years 2000-2015. 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 with available monitoring data.
Gridded Surface Meteorological (GRIDMET) Temperature and	4 km x 4 km maximum temperature and maximum relative humidity predictions, summer (June-September) and	Citation: Abatzoglou J. T., Development of gridded surface meteorological data for ecological applications and modelling, International Journal of Climatology. (2012) doi:10.1002/joc.3413 Data acquired on October 14, 2021.	Data is averaged over 16 years 2005-2020 and averaged over each census tract.

Relative Humidity	winter (December-March)		

- 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, GRIDMET variables, and CDC SVI have some missing values. Outcome variables are available for 71825 out of 72539 total census tracts in the conterminous U.S. (< 1% missing).

The flood risk variable Average Risk Score of Special Flood Hazard Areas has an exceptionally large number of missing values, so it will be omitted from the analysis.

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. There are 2 census tracts not included in the adjacency file, so they are omitted from the analysis.

To get the list of U.S. census tracts and to construct 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. However, for Virginia and South Dakota, I use 2019 census tract boundaries, because of county fip code changes that occurred in the 2010s (https://www.ddorn.net/data/FIPS_County_Code_Changes.pdf).

b. Analysis

a. Dimensionality Reduction

To reduce the large number of flood risk variables to a small set of variables, we will use principal components analysis. We will include the first few principal components such that at least 80% of the variance is accounted for.

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 conditional autoregressive (CAR) prior.

Missing outcome values will be treated as additional unknown parameters that are updated via data augmentation. Missing covariate values will be mean imputed. Thus, all 72537 census tracts in the conterminous U.S. will be considered (2 census tracts are omitted due to not being included in the adjacency file).

The eigendecomposition of an adjacency matrix for 72537 census tracts, necessary for estimating the spatial smoothing parameter ρ, is intractable. We will circumvent this issue by fixing ρ at 1, thus fitting an intrinsic CAR model.

c. Sensitivity Analyses

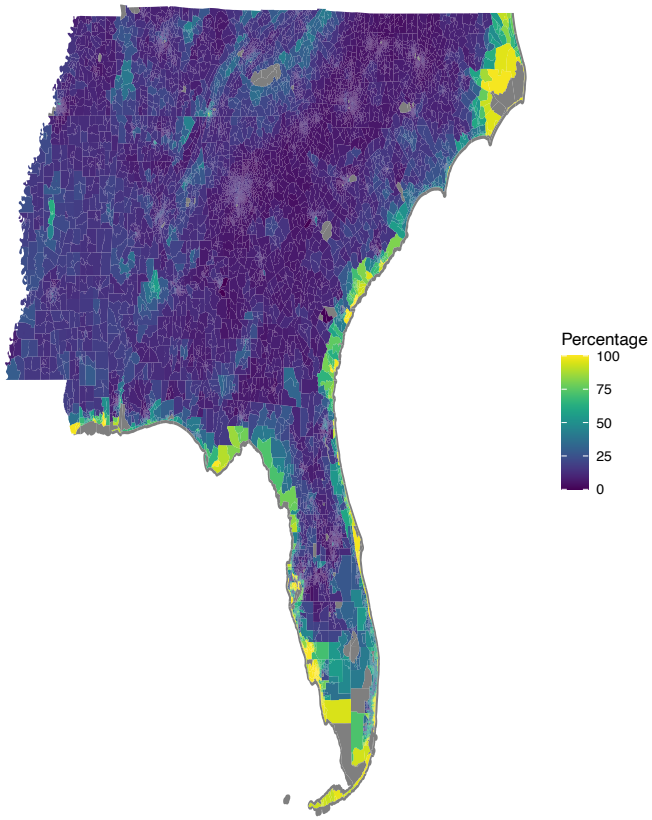
To assess the robustness of our results to the data and our modeling choices, we will conduct several sensitivity analyses. One adjustment is using more principal components such that 90% of the variance is accounted for. Another adjustment is to estimate ρ through divide and conquer: fit separate CAR models to 10 connected subregions of the U.S., take the average of the 10 estimated ρ parameters, and use the average ρ as a fixed parameter for the CAR model fitted to the entire contiguous U.S.

We can also apply the spatial causal inference technique of incorporating a two-stage propensity score adjustment to the model.

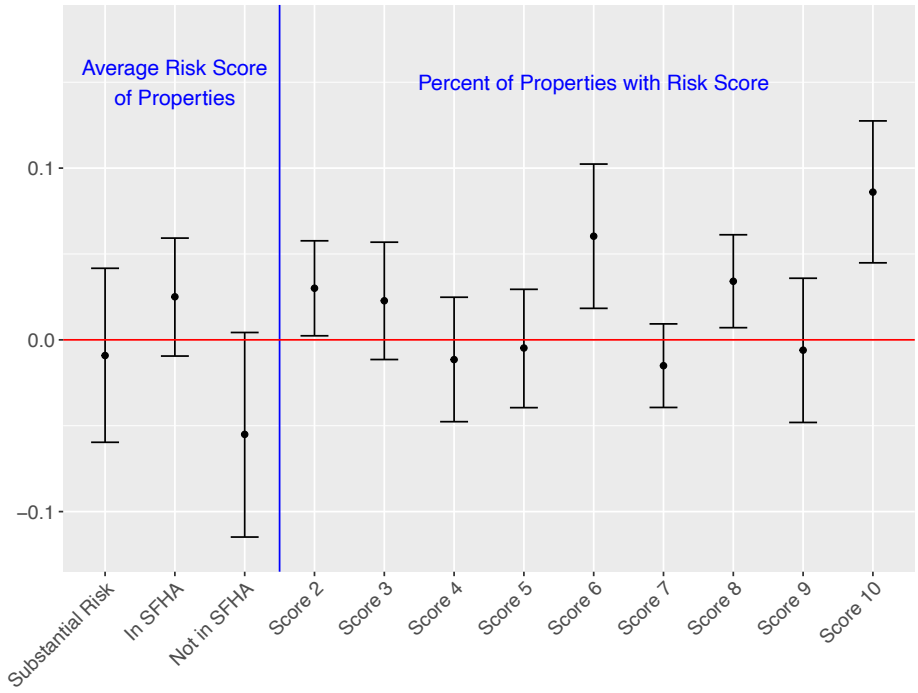
Uncertainty intervals (95% and 50%) are available for the outcomes and most covariates. Thus, measurement error for a variable can be incorporated in the model by assuming that the measurement error is normal with a known deviation derived from the uncertainty interval (https://mc-stan.org/docs/2_21/stan-users-guide/bayesian-measurement-error-model.html).

- List of expected or potential figures/graphics:

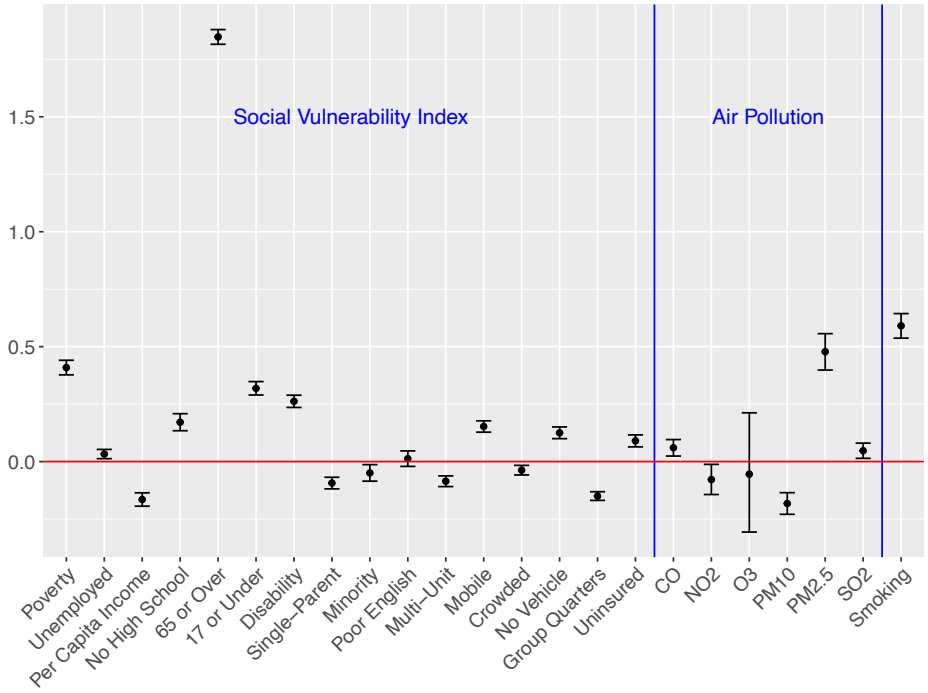
The below figures and graphics are only for seven states in the Southeastern US: North Carolina, South Carolina, Tennessee, Georgia, Alabama, Mississippi, and Florida. These figures and graphics will be extended to the entire contiguous United States.



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, 2022.

After the manuscript is complete, we will polish and bundle the code for this project into an R package. We aim to complete the R package at the end of this project in May 2022.

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.