Analysis Plan

1. Title Information

a. **Proposal Title:** Effects of Flood Risk and Climate Change on Census Tract-Level Health Outcomes

b. **Abbreviated Title:**

c. **Suggested key words:** **Keywords that are often used in publications**

2. Lead Author Name: Alvin Sheng

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

Name	Contact Information	Responsibilities
Kyle P. Messier		

5. Background/Rationale:

A brief literature review and the knowledge gap. You can think of this is a precursor to the first 1 or 2 paragraphs in a manuscript introduction.

Health effects (see Taylor's lit review)
Flood risk
BHM

6. Brief Overview:

A summary of your proposed analysis. This can build upon the background/rationale to include your analysis and some expected results. You can think of this as the penultimate or last paragraph in a manuscript introduction.

- Fjk
- P
- p

7. Research Questions & Hypotheses:

One to three aims or objectives of the manuscript. Since these aims are intended for one manuscript, then they are likely related but not necessarily dependent. One to four sentences each.

Aim #1: To investigate associations between flood risk and health outcomes.

Hypotheses: *We hypothesize that higher flood risk is associated with worse health outcomes. If longer term flood risk predictions are associated with higher health impacts,* *that would imply that long-term effects of climate change can have health impacts over the short term.*

<u>Aim #2:</u> To investigate whether the relationship between flood risk and health outcomes is modified by social vulnerability factors.

Hypothesis: *We hypothesize that the relationship between flood risk and health outcomes will be enhanced in areas of high social vulnerability.*

8. Data:

a. <u>Study domain and/or population:</u>

All census tracts in the conterminous United States.

Relevant data include the County Adjacency File provided by the Census Bureau (https://www.census.gov/programs-surveys/geography/library/reference/county-adjacency-file.html).

To-do: sort the states by environmental characteristics, i.e. whether it's inland or coastal (see plots in https://assets.firststreet.org/uploads/2020/06/first_street_foundation__first_national_flood_risk_assessment.pdf)

b. <u>Study years:</u>

- Outcome:
  - i. *Life Expectancy: 2014*
  - ii. *Age-specific mortality risk: 2014*
  - iii. *Deaths due to various CVD causes: 2016-2018*
- Exposures:
  - i. Flood Risk: 2020 (present) and 2050 (climate-adjusted future)
- Other Covariates:
  - i. CACES air pollution
  - ii. Smoking prevalence
- 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.

c. <u>Outcomes:</u>

Outcome Type	Variable Names and Description of Variable from Orig Source	Details from Orig Source	Analytical Treatment

Life Expectancy	“Life expectancy, 2014*” Data also available for 1980, 1985, 1990, 1995, 2000, 2005, and 2010 *Life expectancy at birth (years). Numbers in parentheses are 95% uncertainty intervals. “% Change in Life Expectancy, 1980-2014”	Results of the study were published in JAMA in May 2017 in "Inequalities in life expectancy among US counties, 1980-2014." http://ghdx.healthdata.org/record/ihme-data/united-states-life-expectancy-and-age-specific-mortality-risk-county-1980-2014 (see README_1.txt for suggested citation)	Uncertainty interval provided. Including uncertainty in outcome is probably standard in BHMs.
Age-specific Mortality Risk	“Mortality risk, 2014*” Available for age ranges 0-5, 5-25, 25-45, 45-65, and 65-85 y.o. Data also available for 1980, 1985, 1990, 1995, 2000, 2005, and 2010 *Probability of death, for given age range (%). Numbers in parentheses are 95% uncertainty intervals. “% Change in mortality risk, 1980-2014”	Mortality risk is the probability of death during the given age range conditional on being alive at the beginning of the age range. Same source as for life expectancy (see above).	Uncertainty interval provided. If I can find proportion of people alive at the start of the age ranges, I can recalculate the mortality risk for other age ranges.
Specific Health Outcomes	"Current asthma among adults aged >=18 years" "High blood pressure among adults aged >=18 years"	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	

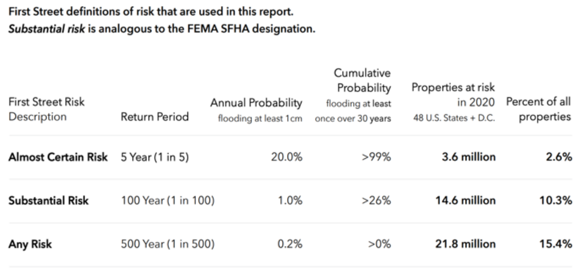
	"Mental health not good for >=14 days among adults aged >=18 years" "Coronary heart disease among adults aged >=18 years" "Physical health not good for >=14 days among adults aged >=18 years" And 8 other health outcomes	Data sources used to make dataset include BRFSS 2018 or 2017 data (just for HBP and cholesterol on left side), Census Bureau 2010 population data, and ACS 2014-2018 or 2013-2017 estimates	

d. Covariates:

- Exposure definition:

The main exposure of interest is flood risk as measured by the First Street Foundation (FSF) model. Source of dataset: https://registry.opendata.aws/fsf-flood-risk/. All details from original source can be found in https://assets.firststreet.org/uploads/2020/06/first_street_foundation__first_national_flood_risk_assessment.pdf.

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 county), 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.

<table>
<tr>
<td>Percent of First Street Properties at 3 levels of severity and 2 time points</td>
<td>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.</td>
<td>
First Street definitions of risk that are used in this report.
<em>Substantial risk</em> is analogous to the FEMA SFHA designation.
<table>
<tr><th>First Street Risk Description</th><th>Return Period</th><th>Annual Probability flooding at least 1cm</th><th>Cumulative Probability flooding at least once over 30 years</th><th>Properties at risk in 2020 48 U.S. States + D.C.</th><th>Percent of all properties</th></tr>
<tr><td><strong>Almost Certain Risk</strong></td><td>5 Year (1 in 5)</td><td>20.0%</td><td>&gt;99%</td><td><strong>3.6 million</strong></td><td><strong>2.6%</strong></td></tr>
<tr><td><strong>Substantial Risk</strong></td><td>100 Year (1 in 100)</td><td>1.0%</td><td>&gt;26%</td><td><strong>14.6 million</strong></td><td><strong>10.3%</strong></td></tr>
<tr><td><strong>Any Risk</strong></td><td>500 Year (1 in 500)</td><td>0.2%</td><td>&gt;0%</td><td><strong>21.8 million</strong></td><td><strong>15.4%</strong></td></tr>
</table>
According to environmental factors, there will be ~11% increase in flood risk over the next 30 years (to 2050).</td>
<td>Can subtract 2020 variable from 2050 variable to get percent change in properties at certain risk</td>
</tr>
<tr>
<td>Average Risk Score of Properties</td>
<td>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</td>
<td rowspan="2">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 year within the next 30 years have FF of 1 (minimal risk).</td>
<td></td>
</tr>
<tr>
<td>Percent of Properties with a given Flood Factor</td>
<td>pct_floodfactor1,
…,
pct_floodfactor10</td>
<td>Divide by count_property</td>
</tr>
</table>

- Confounders/other covariates

Other Covariate Type	**Variable Names and Description of Variable from Orig Source**	**Details from Orig Source**	**Analytical Treatment**
Smoking Prevalence			
CACES LUR Air Pollution (https://www.caces.us/data)	Population-weighted concentration based on block level centroid predictions for 6	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.;	Data is for the year 2015. There are other years available. Model estimates except for O3 are

	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	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 .""	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.

- Mediators: N/A

- Moderators:

These are variables that will interact with the covariates above. One can stratify on SVI (low/medium/high). Categorical and continuous variable interactions are more interpretable than continuous and continuous variable interactions. Can see https://jenfb.github.io/bkmr/overview.html for example model set ups and ggplot visualizations.

CDC Social Vulnerability Index (SVI) https://www.atsdr.cdc.gov/placeandhealth/svi/data_documentation_download.html (includes suggested citation). All variables are calculated from the 5-year American Community Survey (2014-2018 for the 2018 SVI version). Data also available for 2000, 2010, 2014, 2016. 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**
E_	estimate
M_	90% margins of error for the estimates. Can be incorporated in BHM.
EP_	Percentage of …
MP_	Margin of error for the percentage of… Can be incorporated in BHM.
EPL_	Percentile of the percentage of… (in Excel, calculated as PERCENTRANK.INC on EP_ variable)
SPL_	Sum of the EPL_ variables for the theme, suffixes are THEME1, THEME2, THEME3, THEME4
RPL_	Percentile ranking of SPL_ variable across counties (in Excel, calculated as PERCENTRANK.INC on SPL_ variable)

F_	Binary indicator where 1 means the county is in the top 10% (above 90th percentile) in a certain variable, and 0 means the county is not in the top 10%. Indicates high vulnerability

Aggregated Variable	Meaning
SPL_THEMES	sum of SPL_THEME1 + … + SPL_THEME4
RPL_THEMES	percentile ranking of SPL_THEMES across county
F_THEME#	Sum of flags for THEME1, … THEME4
F_TOTAL	Sum of flags for the four themes

SVI Type	Variable Names and Description of Variable from Orig Source	Details from Orig Source	Analytical Treatment
Description of County	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: **Missing, censored, or excluded data – both outcome and covariates.**

9. Statistical Analysis Plan and Methods:

Discuss the proposed statistical models – plain English explanation and some equations if applicable

a. Spatial Data Wrangling

These are the resources used to wrangle the spatial data, as in imported_data_wrangling.R and imported_data_wrangling_census_tract.R.

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
(unfortunately, it only allows me to download the shapefiles for one state at a time)

b. Data Pre-Processing

    a. Detecting Multicollinearity

Used vifstep:
Citation: 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.

c. Modeling

A Conditional Autoregressive model will be fitted. An adjacency matrix will be calculated, where 1 indicates pairs that are neighbors and 0 indicates pairs with only one census tract or census tracts that are not neighbors.

Can also try S.CARdissimilarity, if you have dissimilarity metrics handy

Can redo in rstan, for greater flexibility

Can use MVS.CARleroux() for multiple response variables

- List of expected or potential tables: *consult Dominici paper*
- List of expected or potential figures/graphics:

10. Anticipated pitfalls/challenges and limitations

- Challenges:
  - *Solution:*
- Limitations:

11. Manuscript Timeline

Goal for manuscript submission and other relevant subgoals (e.g. Methods, Results, Conference abstract).

12. References: