Application of machine learning for integrated flood risk assessment: Case study of Hurricane Harvey in Houston, Texas

Behrang Bidadian[1], Aaron E. Maxwell[2], Michael P. Strager[3]

[1]School of Design & Community Development, West Virginia University, Morgantown, WV 26506, USA
[2]Department of Geology and Geography, West Virginia University, Morgantown, WV 26505, USA
[3]School of Natural Resources, West Virginia University, Morgantown, WV 26506, USA

Correspondence to: Behrang Bidadian (behrang.bidadian@mail.wvu.edu)

Abstract. Flood risk, encompassing hazard, exposure, and vulnerability is defined concerning potential losses. Machine learning techniques have gained traction among researchers to address the complexities of multi-variable flood risk assessment models and overcome issues associated with non-linear relationships. However, the focus has primarily been on flood hazard prediction rather than comprehensive risk assessment and damage estimations. Therefore, there is a need for experiments that combine risk elements using such methods. To address this need, this study utilized the Random Forest algorithm to analyze the correlations between the physical flood damage caused by Hurricane Harvey in 2017 in Houston, Texas and certain hazard, exposure, and vulnerability-related variables. The study identified poorly drained soils as the primary contributor to the losses, followed by population density and the ratio of developed lands with medium intensity. The study's findings also explored the reasons for the unexpectedly low importance of social vulnerability factors compared to the environmental justice concept. These findings and conclusions can provide insights to planners and stakeholders enhancing their understanding of the underlying causes contributing to flood risk. Future research can expand upon this study's methodology and findings by incorporating additional factors related to climate change.

1 Introduction

Flood risk as the combination of inundation hazard, human or physical exposure, and vulnerability of the exposed elements (Birkmann and Welle, 2015; Crichton, 2002; Hallegatte, 2014) is defined in relation to potential damages or losses (Dewan, 2013; Hallegatte, 2014; Hammond et al., 2015; Kubal et al., 2009). The above definition considers flood source (e.g., heavy rainfall) and pathway (e.g., floodways and floodplains) as the hazard, the receptors (e.g., at-risk population, buildings, and infrastructure) as the exposure elements, the susceptibility of receptors as the vulnerability, and the probable consequent damages as the risk (Jha et al., 2012; Kaźmierczak and Cavan, 2011; Taramelli et al., 2022).

In a conventional integrated flood risk analysis, the floodplains or inundation zones should be mapped first, then the intersections of the population and assets with those delineated hazard zones should be extracted considering their vulnerability characteristics (Kaźmierczak and Cavan, 2011; Kubal et al., 2009; Merz et al., 2010; Tate et al., 2021). After identifying the exposed elements and their vulnerability levels, predictive impact assessment models can be developed to determine the estimated monetary damages and human losses. The flood consequences can be divided into direct (or immediate) human and physical losses and indirect (also known as higher-order) damages such as long-run business interruptions, economic consequences at the regional scale, or long-term social impacts (Hallegatte, 2014; Hammond et al., 2015; Merz et al., 2010; Rose, 2004).

The conventional models for direct loss estimation typically rely on depth-damage functions considering flood depth and property use type to produce fractions of the structure replacement cost as the potential damage percentages (Hammond et al., 2015; Merz et al., 2004; Thieken et al., 2008; Wagenaar et al., 2017). However, many of those models exhibit substantial uncertainties due to weak correlations between the losses and flood depth values (Hammond et al., 2015; Merz et al., 2004; Thieken et al., 2008; Wagenaar et al., 2017). The main reason is that flood risk and damage are influenced by various factors beyond just the water depth (Hammond et al., 2015; Merz et al., 2004; Thieken et al., 2008; Wagenaar et al., 2017).

In recent years, a shift has emerged towards multi-variable flood risk models that can assist researchers and practitioners in conducting more accurate comprehensive flood loss analyses with a better understanding of the effective factors. In addition to the inundation depth, such models consider several other influential variables, such as environmental factors, characteristics of exposed structures, and socioeconomic conditions (Kubal et al., 2009; Merz et al., 2004, 2010; Thieken et al., 2008; Wagenaar et al., 2017). However, multi-variable flood risk prediction is a complicated process comprising non-linear relationships with many variables (Merz et al., 2013; Tehrany et al., 2013; Wang et al., 2015).

The emergence of artificial intelligence (AI) has brought about transformative changes in flood risk prediction and analysis methods (Mosavi et al., 2018; Tehrany et al., 2013). AI refers to the ability of machines, such as computers and robotic systems, to perform tasks that traditionally require human intelligence (Poole and Mackworth, 2017). Machine learning (ML) techniques, as essential branches of AI, can train computers to process spatial and non-spatial big data and discover patterns or correlations for more accurate predictions (Murphy, 2012; Wagenaar et al., 2017, 2020). Additionally, they can address the complexities and non-linear relationships among multiple influential factors associated with flood risk (Alipour et al., 2020; Kalaycıoğlu et al., 2023; Mosavi et al., 2018; Tehrany et al., 2013; Wagenaar et al., 2020). Researchers increasingly apply automated ML techniques to flood prediction and mapping processes instead of conventional methods (Tehrany et al., 2013).

More attention has been paid to flood hazard prediction than comprehensive risk assessment and damage estimations (Knighton et al., 2020; Merz et al., 2010; Mohanty and Simonovic, 2021). A cause for this trend is the presence of uncertainties in the currently available flood zone maps.

Understanding the vulnerability of exposed populations and assets is critical for appropriate flood risk mitigation and communication efforts (Knighton et al., 2020; Merz et al., 2010). Socioeconomic or demographic attributes such as income, race, ethnicity, age, household composition, and housing characteristics can contribute to social vulnerability, affecting people’s ability to anticipate, respond to, and recover from natural disasters (Birkmann and Welle, 2015; Collins et al., 2018; Cutter and Finch, 2008; Cutter et al., 2003; Flanagan et al., 2011; Kaźmierczak and Cavan, 2011). Furthermore, incorporating demographic and socioeconomic data can potentially enhance the estimation accuracy of flood risk and damage models (Knighton et al., 2020). Despite the potential of machine learning in integrating social vulnerability into flood risk and loss studies while addressing non-linearity and complexity, only a few recent studies such as those conducted by Alipour et al. (2020), Kalaycıoğlu et al. (2023), and Knighton et al. (2020) employed socioeconomic data to investigate the vulnerability in their models.

The parameters used in multi-variable flood risk models, also referred to as conditioning factors, are primarily location-based. The above statement pertains to the environmental or geomorphological factors, some of which may hold significant influence in one location but may not be effective in another study area (Tehrany et al., 2013; Wang et al., 2015). Similarly, the same principle applies to socioeconomic variables that determine human exposure and vulnerability, as they are contingent on the specific social context of a given area (Cutter et al., 2003; Dewan, 2013). Thus, the conditioning factors in flood risk models may differ in terms of importance degrees based on the region. The decision about these levels is a challenging task. For example, Cutter et al. (2003) adopted an equal contribution level for the factors included in their development of the social vulnerability index (SoVI) due to the absence of a robust method for assigning weights. They mentioned the development of a reasonable weighting scheme as a future need (Cutter et al., 2003). Machine learning models have the capability to determine the importance degree of each variable used in regression or prediction processes (Debeer and Strobl, 2020; Kalaycıoğlu et al., 2023; Wagenaar et al., 2017). This ability can aid in understanding and interpreting the results of risk models more accurately (Kalaycıoğlu et al., 2023). By identifying the relative contributions of different variables, researchers and stakeholders can gain a deeper understanding of the factors driving risk and make informed decisions based on the model outputs.

Many experts involved in social vulnerability analyses hold the view that there is an unfair distribution of natural disaster risk resulting from environmental inequity (Birkmann, 2013; Collins et al., 2018; Kaźmierczak and Cavan, 2011). According to several authors, socially disadvantaged groups, for example the poor and racial minorities, not only face higher vulnerability but also experience a greater likelihood of exposure to natural hazards like floods (Collins et al., 2018; Dewan, 2013; Kaźmierczak and Cavan, 2011; Maldonado et al., 2016; Tate et al., 2021). However, some recent analyses in the United States such as those conducted by Cutter et al. (2018), Hale et al. (2018), and Knighton et al. (2020) indicated contrary findings, not in harmony with the above environmental justice notion. Their experiments revealed that racial minorities or economically disadvantaged people were less likely to reside in hazardous flood zones. Therefore, further studies employing ML techniques are still needed to assess and validate the environmental equity concept in different locations.

This study aimed to fill a current research gap in the area of comprehensive multi-variable flood loss analyses and assessment of the environmental justice notion in order to better understand the factors that significantly influence flood risk levels in the study area and determine their respective contribution levels in relation to the above concept. Specifically, it addressed the question: What role do environmental and socioeconomic characteristics play in shaping flood risk in urban areas? As the primary goal, the experiment tested this hypothesis in Houston, Texas as the study area: Flood risk correlates with the socioeconomic attributes of urban communities as well as the environmental characteristics. Appropriate ML methods in conjunction with geospatial technologies were employed for regression analysis of the data to uncover relationships between environmental factors, development patterns, human characteristics, and flood loss. We considered the monetary damages caused by Hurricane Harvey in 2017 declared in the flood insurance claims published by the Federal Emergency Management

Agency (FEMA) as an indicator of flood risk. We examined conditioning or predictor variables and their significance level in the risk prediction process categorized into three main groups: hazard-, exposure-, and vulnerability-related factors.

The remainder of this paper is organized as follows. The next section describes the study area and the spatial units of analysis, selection of response and predictor variables, data collection and processing, and the ML modeling process. Next, the results obtained from the machine learning model are presented. Then, the discussion section offers an interpretation of the outcomes considering their significance along with the limitations of the processes. Finally, the conclusions section summarizes the most critical findings, discusses their implications and contributions to the field, and suggests potential directions for future research.

2 Material and methods

2.1 Study area and units

Houston, one of the major cities in the United States, has gained a reputation for experiencing recurrent instances of flooding that significantly impact its urban area. The most recent catastrophic event in Houston was the flooding triggered by Hurricane Harvey in 2017 resulting in the tragic loss of 68 lives and causing approximately 125 billion dollars in physical damage in the state. On August 25, 2017, Hurricane Harvey made landfall in the vicinity of Rockport, Texas unleashing winds surpassing 240 kilometers per hour. As the storm progressed inland, its pace decelerated, leading to the accumulation of unprecedented rainfall in southeastern Texas. In certain locations, the rainfall persisted for eight consecutive days, surpassing 1520 millimeters (60 inches) in magnitude. This quantity exceeded the average annual rainfall levels for eastern Texas and the coast by approximately 380 millimeters (15 inches) (Watson et al., 2018). The impact of Hurricane Harvey on Houston was primarily attributed to the devastating flooding caused by record-breaking rainfall rather than wind damage or storm surge (Bedient et al., 2017).

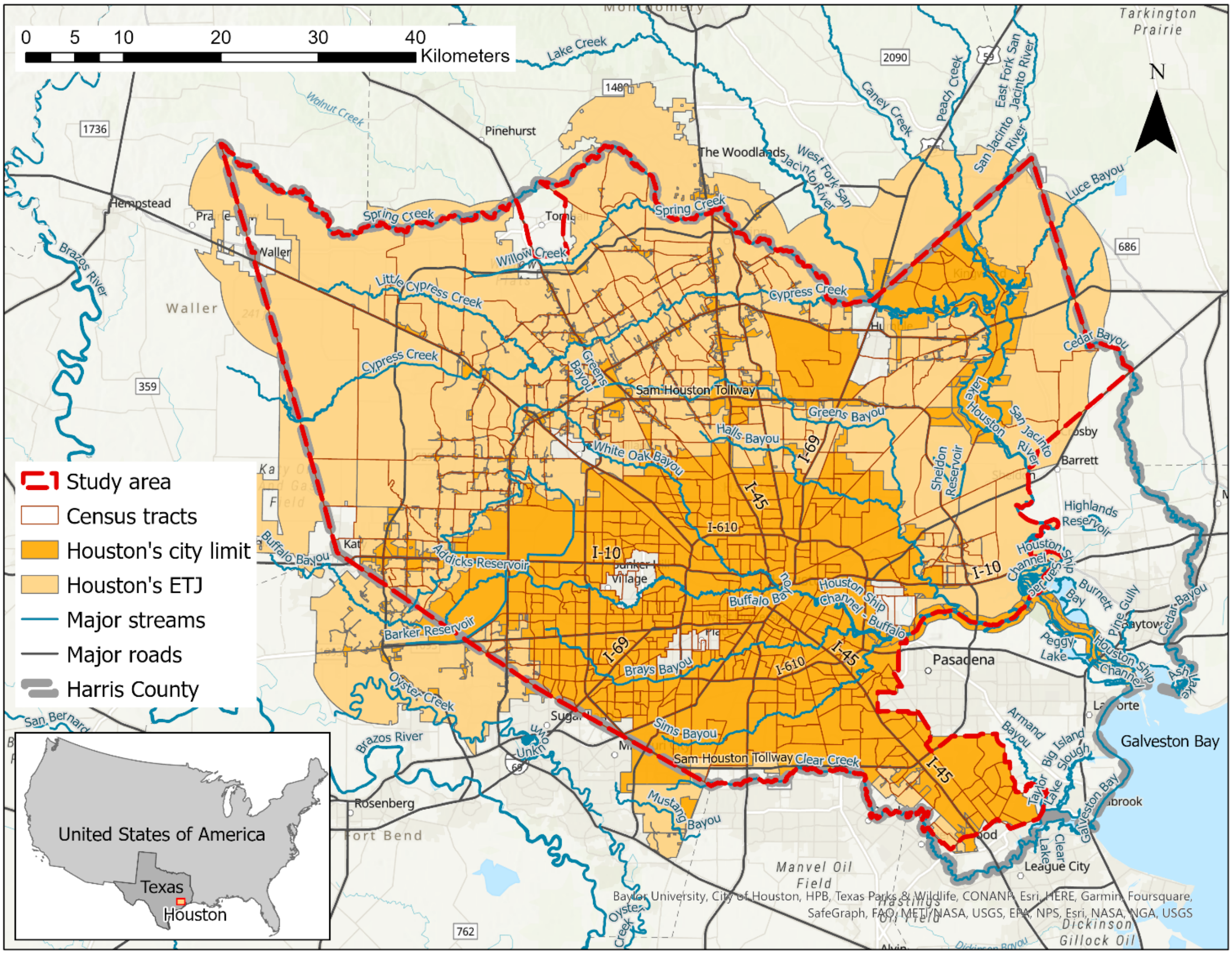
Houston is located in the southeastern region of Texas in close proximity to Galveston Bay and approximately 80 kilometers away from the Gulf of Mexico. The city of Houston is situated at the convergence of major interstate highways, specifically I-10, I-45, and I-69. In terms of racial and ethnic composition, Houston stands out as one of the most diverse cities in the United States. According to the 2020 decennial census (DEC), its demographic makeup includes a population with 32% identifying as white, 23% as African-American, 7% as Asian, 1% as American Indian or belonging to other Native groups, 21% as belonging to other races, and 16% as individuals of mixed race heritage (United States Census Bureau, 2020). From a topographical standpoint, Houston falls in the Coastal Plain physiographic province and has been developed in a predominantly flat, low-altitude region. The urban area within the city limits encompasses approximately 1,554 square kilometers. Additionally, the municipality holds certain rights and responsibilities in the adjacent metropolitan area known as the extraterritorial jurisdiction (ETJ). The ETJ is an area extending eight kilometers (five miles) beyond the city's primary limits except where it intersects with another municipality or jurisdiction (City of Houston, 2022).

The bayous in Houston can be tranquil under normal weather conditions. However, during the occurrence of Hurricane Harvey in 2017, all 22 bayous in the region surpassed their capacity, resulting in extensive flooding (Bedient et al., 2017). Among

https://doi.org/10.5194/nhess-2023-113
Preprint. Discussion started: 25 August 2023

these waterways, Buffalo Bayou stands as the largest, flowing through the city until it meets the San Jacinto River at Galveston Bay, which then connects to the Houston Ship Channel. There are two large reservoirs in the western part of the area called Addicks and Barker constructed by the U.S. Army Corps of Engineers to mitigate flooding. Both reservoirs reached their emergency spillway levels during Hurricane Harvey (Bedient et al., 2017).

For this study, census tracts of 2017 provided by the U.S. Census Bureau were the units of analysis and model development (U.S. Census Bureau, Geography Division, 2017). Census tracts serve as subdivisions within the United States counties for which detailed demographic data collected by the Census Bureau are publicly accessible. These tracts are typically delineated based on the assumption of demographic homogeneity, making them suitable units for various planning and government purposes (Flanagan et al., 2011). The study area encompassed the portions of the city limit and ETJ situated in Harris County. Some adjustments were made at the edges to align with the boundaries of the census tracts, ensuring consistency and accuracy in the area. For that purpose, the tracts that partially intersected the designated area were retained for analysis only if their centroids fell within the boundaries of the study area (Fig. 1). Certain census tracts, primarily those encompassing large non-residential lots such as the airports and universities, were excluded from the study due to incomplete availability of demographic data. Finally, a total of 678 census tracts remained as the study units, covering an area of 3,760 square kilometers. Based on the 2017 American Community Survey (ACS), the population residing in this region amounted to 3,994,164 individuals during that year (U.S. Census Bureau, 2017).



Baylor University, City of Houston, HPB, Texas Parks & Wildlife, CONANP, Esri, HERE, Garmin, Foursquare, SafeGraph, FAO, METI/NASA, USGS, EPA, NPS, Esri, NASA, NGA, USGS

Figure 1: Houston study area, census tracts, city limit, and extraterritorial jurisdiction (ETJ).

2.2 Variable selection

In this study, we investigated the damage ratios of flood insurance claims made through the National Flood Insurance Program (NFIP) in response to Hurricane Harvey in 2017 in Houston. The above ratios were specifically limited to the direct physical losses incurred by the buildings. The damage data, which were published by FEMA, served as the response or dependent variable to be modeled against the potentially influential factors. FEMA has redacted the physical addresses and coordinates in the data to protect individual privacy. Therefore, it was not possible to accurately identify the specific locations of the loss claims at the property level. However, the corresponding census tracts were recognizable through the use of the provided Census Bureau's 11-digit codes. Using the above dataset, we calculated the damage percentage for each claim by dividing the amount of building damage by the value of the building property as estimated in the claim dataset. Then, the average damage ratio was computed for each census tract to be utilized as the dependent variable in our analysis.

This research expanded upon the previous study conducted by Knighton et al. (2020) by adopting a more comprehensive approach to analyze the contribution of various factors to the damages stated in flood insurance claims. We examined the correlation between the average damage ratios and 17 predictor variables, which were categorized into three main groups. The first group consisted of environmental or geomorphological hazard-related factors that were identified in the literature as having a significant impact on flooding. Within this group, we considered variables such as elevation, land surface curvature, the path to the streams affected by slope or flood source cost distance, and the presence of poorly drained soils. More details of these parameters are explained in the data collection and processing subsection.

The second group of predictor variables focused on flood exposure. We investigated population density in the census tracts as a measure of human exposure, and we also took into account the intensity of developed land cover as an indicator of physical exposure. Population density correlates with the distribution of housing units (Ferguson and Ashley, 2017; Figueiredo and Martina, 2016). Consequently, it can also represent the physical exposure of residential buildings.

Land use/land cover (LULC) data are often appropriate for flood exposure assessments on the macro scale, particularly in the context of cities and metropolitan regions (Kaźmierczak and Cavan, 2011; Merz et al., 2010). The National Land Cover Database (NLCD) includes four classes of developed lands. First, developed, open space encompasses parks, golf courses, single-family housing units in large lots, and green spaces in developed environments for recreational or aesthetic purposes and erosion control. Impervious surfaces such as roads, buildings, and paved areas, occupy less than 20% of the total land cover in this class. Second, developed, low-intensity comprises areas characterized by a combination of constructed materials and vegetation wherein impervious surfaces account for 20% to 49% of the total land cover. These areas are mostly associated with single-family housing units, striking a balance between constructed elements and natural vegetation. Third, developed, medium-intensity denotes areas characterized by a blend of constructed materials and vegetation, where impervious surfaces constitute 50% to 79% of the total land cover. These areas are typically associated with the presence of single-family housing units, reflecting a higher proportion of built environment in relation to natural elements. Fourth, developed, high-intensity refers to areas with a dense concentration of structures where a significant number of people reside or work. This category includes apartment complexes, row houses, and commercial or industrial establishments. Impervious surfaces occupy 80% to 100% of the total land within these areas, indicating minimal natural vegetation (Multi-Resolution Land Characteristics Consortium (MRLC), 2023). In this study, we excluded the developed open spaces from consideration due to the absence of a significant number of primary insurable structures in those parts.

As the third group of independent variables, we incorporated vulnerability-related socioeconomic factors that were derived from the literature. The 2017 American Community Survey (ACS) was the source for collecting the data relevant to poverty, unemployment, race, ethnicity, citizenship status, age, education, and housing characteristics such as median housing value and the ratio of renter-occupied residential units (U.S. Census Bureau, 2017). Table 1 presents all variables that were utilized in the spatial and correlation analyses categorized into the aforementioned groups.

Table 1: Dependent and predictor variables used in the study and their grouping.

Group	Variable	Previous authors stating importance of the field
Response (dependent)	Average damage ratios	Cutter et al. (2018); Knighton et al. (2020)
Hazard-related predictors	Median elevation	Tehrany et al. (2013); Wang et al. (2015)
	Median curvature	Tehrany et al. (2013)
	Median cost distance	Brivio et al. (2002); Joy et al. (2019)
	Percentage of poorly drained soils	Lin and Billa (2021)
Exposure-related predictors	Population density	Alipour et al. (2020); Knighton et al. (2020)
	Percentage of low-intensity developed	Cutter et al. (2003, 2018)
	Percentage of medium-intensity developed	
	Percentage of high-intensity developed	
Vulnerability-related predictors	Percent below poverty level	Collins et al. (2018); Cutter et al. (2003)
	Unemployment rate	Flanagan et al. (2011); Kalaycıoğlu et al. (2023)
	African-American ratio	Collins et al. (2018); Cutter et al. (2003)
	Hispanic or Latino ratio	Cutter et al. (2003); Flanagan et al. (2011)
	Non-citizen ratio	Collins et al. (2018); Flores et al. (2020)
	Median age	Collins et al. (2018); Cutter and Finch (2008)
	No high school diploma ratio	Cutter et al. (2003); Flanagan et al. (2011)
	Median housing value	Knighton et al. (2020); Merz et al. (2013)
	Renter-occupied ratio	Cutter et al. (2003); Flores et al. (2020)

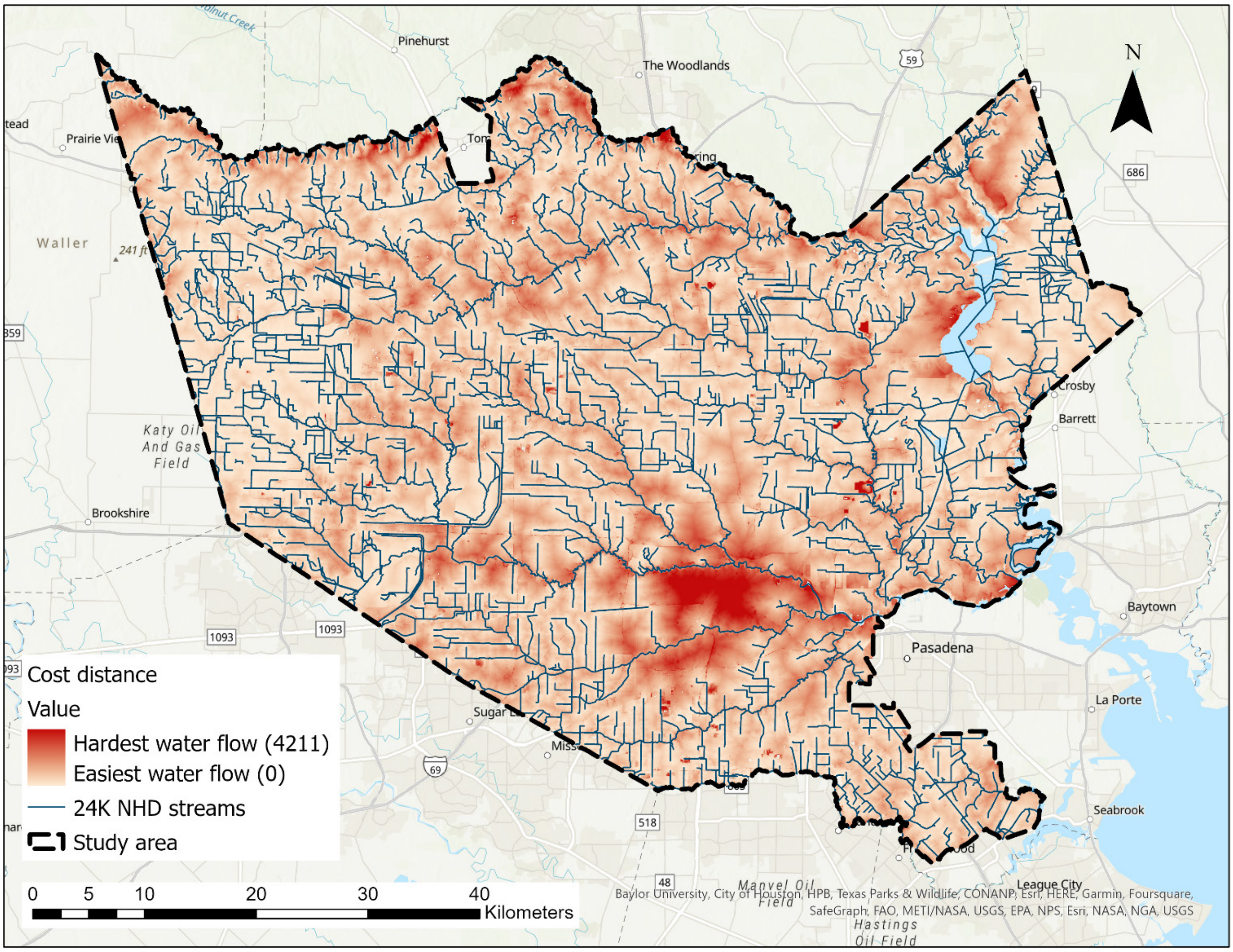
2.3 Data collection and processing

The FEMA NFIP Redacted Claims (v2) were downloaded from OpenFEMA database (FEMA, 2023). This version of the dataset, including over 2.5 million flood insurance claims transactions nationwide, has been derived from the National Flood Insurance Program (NFIP) record system to list comprehensive information on claims filed from 1978 up until May 2023. We employed the data analysis software of R (R-4.2.3) (R Core Team, 2023) to extract the relevant records pertaining to Hurricane Harvey within the designated study area.

National, state, and local sources were used to collect the required data for the hazard-related variables. The 10 m digital elevation model (DEM) utilized in the elevation processing was obtained from the Houston–Galveston Area Council (H-GAC) website (H-GAC, 2019). We produced raster grids of slope and curvature with a 10 m cell size based on the above DEM using ArcGIS Pro version 3.1 (Environment Systems Research Institute (ESRI), 2023). The curvature grid assigned a numerical value to each cell based on the surface's level of concavity, convexity, or flatness. In order to map the streams, we relied on the 1:24,000 (24k) National Hydrography Dataset (NHD) downloaded from the United States Geological Survey (USGS) database (USGS, 2023). Instead of a Euclidean distance measure, we created a raster grid of cost distance to the streams, weighted by slope, employing ArcGIS Pro (ESRI, 2023), to account for terrain-related conditions that may impact floodwater movement. The integration of distance and slope in this variable was aimed at offering a more holistic assessment of the risk effects arising from both stream proximity and topographic characteristics. Figure 2 displays the developed cost distance grid.

https://doi.org/10.5194/nhess-2023-113
Preprint. Discussion started: 25 August 2023

To acquire data on soil types and their drainage characteristics, we accessed the Soil Survey Geographic Database (SSURGO) provided by the Natural Resources Conservation Service (NRCS) under the United States Department of Agriculture (USDA). We used ESRI's SSURGO Downloader web-based application to download the relevant data (Esri and USDA NRCS, 2022).



Baylor University, City of Houston, HPB, Texas Parks & Wildlife, CONANP, Esri, HERE, Garmin, Foursquare, SafeGraph, FAO, METI/NASA, USGS, EPA, NPS, Esri, NASA, NGA, USGS

Figure 2: Cost distance raster grid developed in relation to 1:24,000 (24k) National Hydrography Dataset (NHD) streams.

The 2016 30 m National Land Cover Database (NLCD), acquired from the Multi-Resolution Land Characteristics Consortium (MRLC) website, was the source for extracting the developed areas in three categories of high, medium, and low intensity (MRLC, 2016). The selection of the 2016 dataset was approximately aligned with the occurrence time of Hurricane Harvey, ensuring temporal relevance for our analysis. Population density for each census tract was determined by dividing the total population, as per the 2017 American Community Survey (ACS), by the area of the tract measured in hectares (U.S. Census Bureau, 2017). The same data source was used to access the needed demographic and socioeconomic data for the development of the social vulnerability variables described earlier. The 2017 ACS was preferred among the datasets due to its temporal alignment with Hurricane Harvey.

To identify the values of predictor variables on insurable buildings, we utilized the building footprints (v2.0) published by Microsoft in 2018. This is an open-source dataset comprising more than 125 million building footprints in the United States and was created using AI-assisted mapping techniques developed on high spatial resolution aerial images. The data are publicly available on the associated GitHub website (Microsoft Corporation, 2018). We selected the footprints within the defined study area and mapped their centroid points. Then, we extracted the values of hazard- and exposure-related variables for each centroid point by referencing the corresponding raster grids. Finally, we removed the points with null values such as those slightly outside or very close to the study area boundary. We were left with a total of 1,189,569 remaining centroid points, each assigned with the relevant predictor values in 678 tracts.

2.4 Machine learning model development

In this comprehensive study, we decided to employ suitable machine learning techniques. This choice was driven by the capability of ML to effectively handle the anticipated complexity and non-linear relationships among the variables utilized in our analysis. Machine learning algorithms are computer programs that can be trained to discover data patterns to conduct predictions (Wagenaar et al., 2020). Different algorithms may be applied in ML models based on the purpose and data. Random Forest (RF) is a type of algorithm widely used in flood-related studies. For example, Alipour et al. (2020), Knighton et al. (2020) and Wang et al. (2015) employed it in their ML models. Random Forest, which works based on creating several decision trees is an appropriate method for extracting correlations or regressions as well as classification. As a major concern of working with powerful ML algorithms, overfitting can occur when a learning algorithm generates a hypothesis that fits the training data extremely well but fails to generalize to unseen data. The overfitted model has learned the training data so precisely that it may incorporate noise or irrelevant patterns leading to poor performance when applied to new, unseen data (Maxwell et al., 2018). One advantage of RF is its inherent capability to mitigate the overfitting problem by employing an ensemble of multiple decision trees each trained independently on a randomly selected subset of the training data. It also only allows for a subset of the available predictor variables to be selected from at each decision node. The goal of using a subset of the training data in each tree in the ensemble and a subset of the variables at each decision node is to reduce the correlation between the trees in the ensemble and potentially reduce overfitting and improve generalization (Wang et al., 2015). Another benefit of Random Forest is its high ability to estimate the contribution of each variable used in the model, often referred to as variable importance (Debeer and Strobl, 2020; Maxwell et al., 2018; Wang et al., 2015). Moreover, it generally shows an acceptable tolerance to outliers and noise in data as well as the correlations among the predictor variables (Wang et al., 2015). The RF algorithm is highly applicable to flood hazard risk assessment as it effectively addresses highly interactive multi-variable and non-linear correlations, making it a valuable tool for capturing complex relationships and providing robust predictions (Wang et al., 2015).

We developed the machine learning model using RF algorithm on the 678 census tracts in R (R-4.2.3) (R Core Team, 2023). First, for each census tract, the values of elevation, curvature, and cost distance extracted for the centroid points in that tract were summarized by calculating their median values. Then, the proportions of points within each tract were calculated to determine the distribution across various development intensities, specifically low-, medium-, and high-intensity developed lands. Additionally, the percentages of points situated on poorly drained soils, including the dominant condition classes of very poorly, poorly, and somewhat poorly drained soils, were also calculated and assigned to their respective census tracts. Furthermore, we incorporated the collected data on vulnerability-related variables and population density for each census tract. Before constructing the machine learning model, a random selection process was employed to divide the census tracts into two, non-overlapping groups for training and validation. The training dataset comprised 70 percent of the total census tracts, while the remaining 30 percent constituted the validation set. Following the completion of model training, an evaluation process was conducted to assess its performance using the validation dataset. Relevant metrics were generated to gauge the performance which served as quantitative measures of how well the model generalized to the unseen data.

Finally, we carried out an importance analysis to assess the contribution level of each predictor variable while being included with the other ones to the damage ratios. The variable importance measures generated by Random Forest algorithms can demonstrate to exhibit bias when dealing with highly correlated variables (Debeer and Strobl, 2020; Maxwell et al., 2020). To investigate the above issue, we performed a Spearman's correlation test (Spearman, 2010), a measure of non-linear, monotonic correlation. According to the findings illustrated in Fig. 3, significant correlations were found between certain variables. For instance, a positive relationship was observed between the ratio of Hispanic or Latino population and the ratio of individuals without a high school diploma (p-value = 0.883), as well as between the percentage of high-intensity developed lands and the ratio of renter-occupied households (p-value = 0.658). There were also instances of high negative correlation such as between the percentages of developed lands of high and low intensities (p-value = –0.580) or between the median age and the ratio of the Hispanic or Latino population (p-value = –0.540). Therefore, we applied a conditional permutation importance (CPI) analysis utilizing the "permimp" package in R that considers correlation in the importance calculation (Debeer et al., 2021). The method is based on comparing the prediction accuracy before and after permuting each variable to assess its impact on the model's performance. In conjunction with the Random Forest model, the package conducts importance analysis iteratively for each variable, evaluating at each tree of the model and subsequently averaging the results across all trees (Debeer and Strobl, 2020; Kalaycıoğlu et al., 2023).

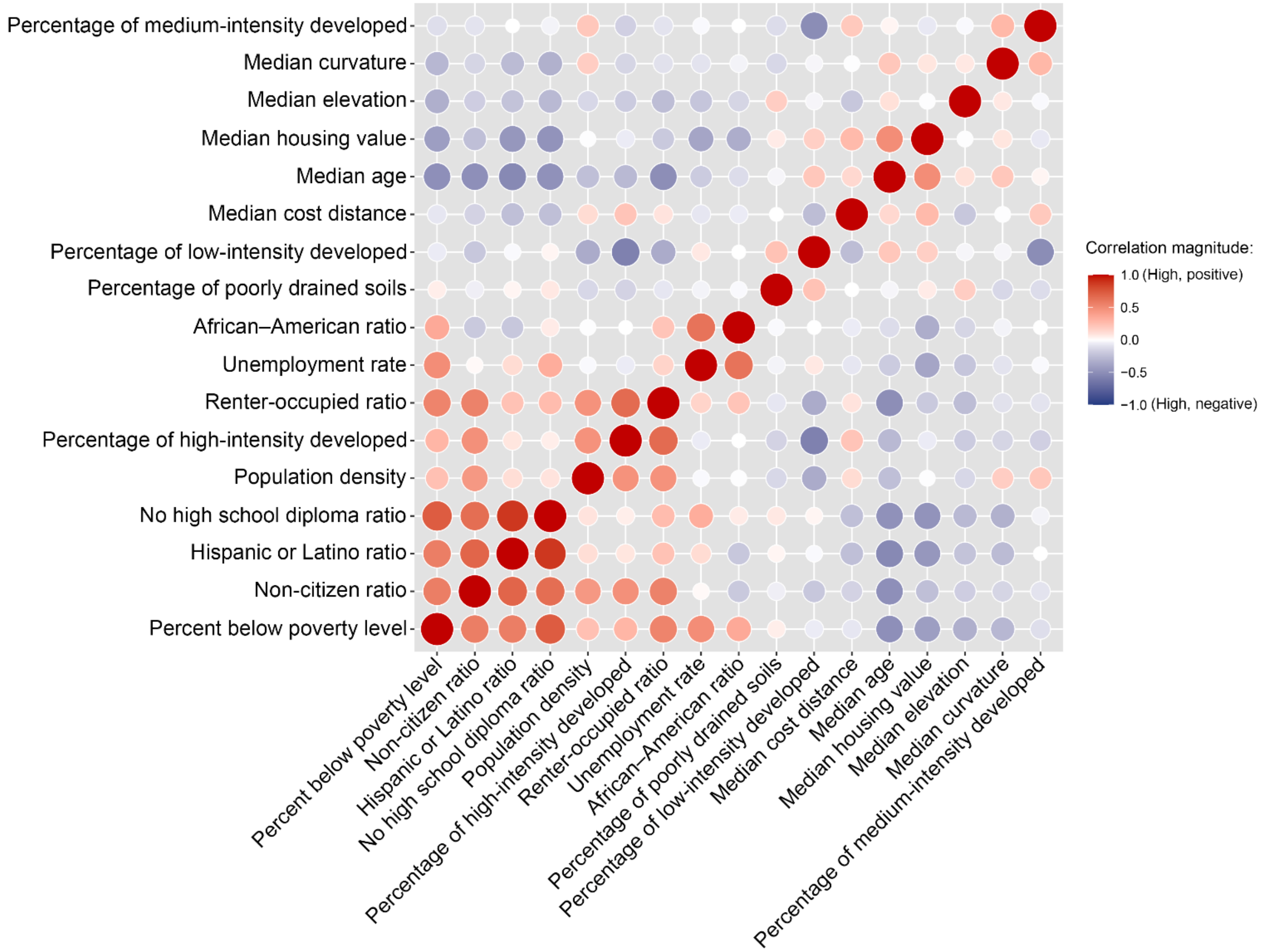


Figure 3: Result of correlation analysis among the predictor variables.

3 Results

The analysis of the NFIP flood insurance records revealed that Hurricane Harvey resulted in building damages amounting to a total of 4.33 billion dollars claimed in the designated study area. The studied census tracts experienced average damage ratios ranging from zero to 86.3 percent. Figure 4 illustrates the distribution of these loss ratios, classified into five distinct categories using the natural breaks method. The natural breaks method is a statistical technique employed to derive meaningful categories from the data, ensuring that the variance in each grouping is minimized while maximizing the variance between different groupings.

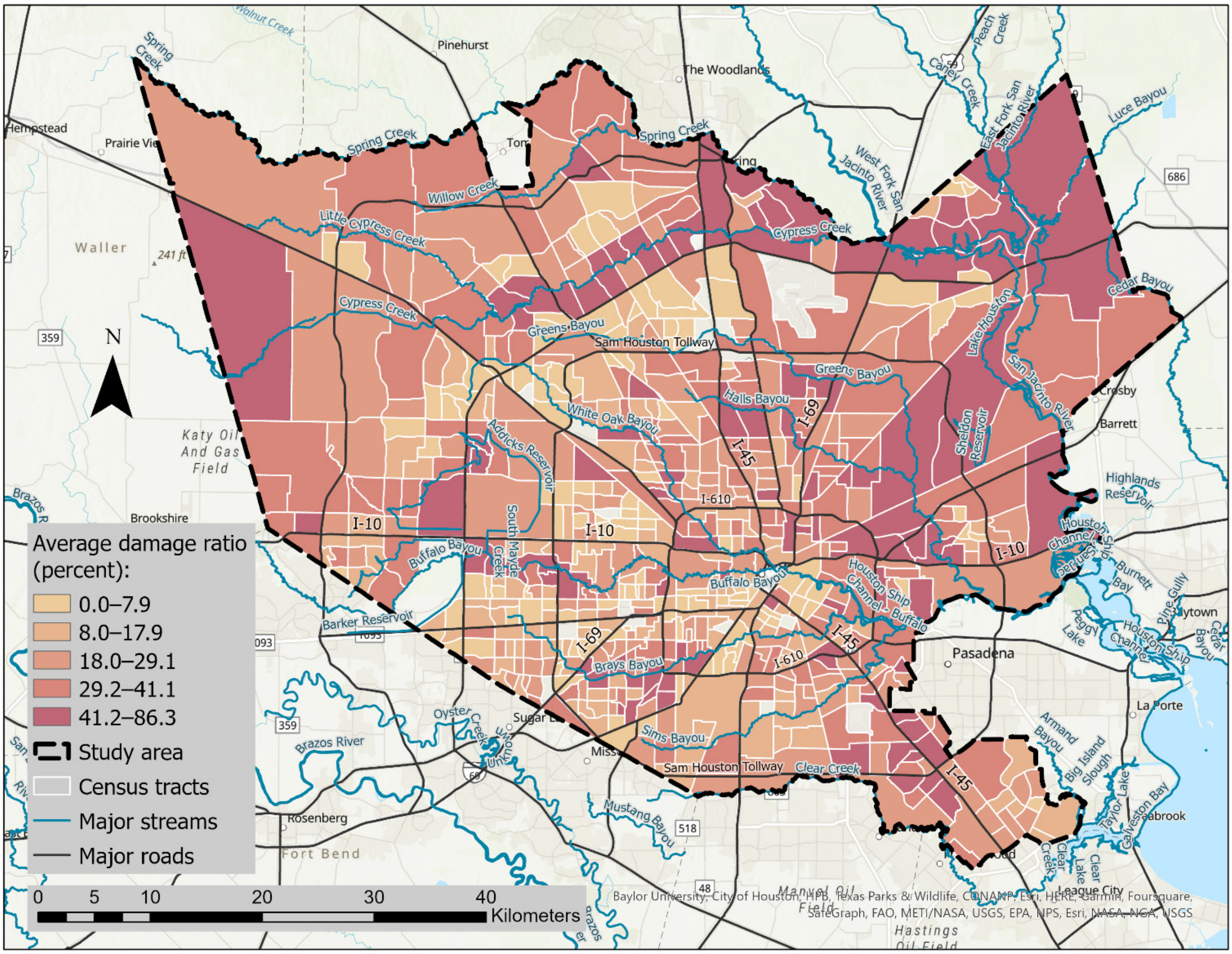
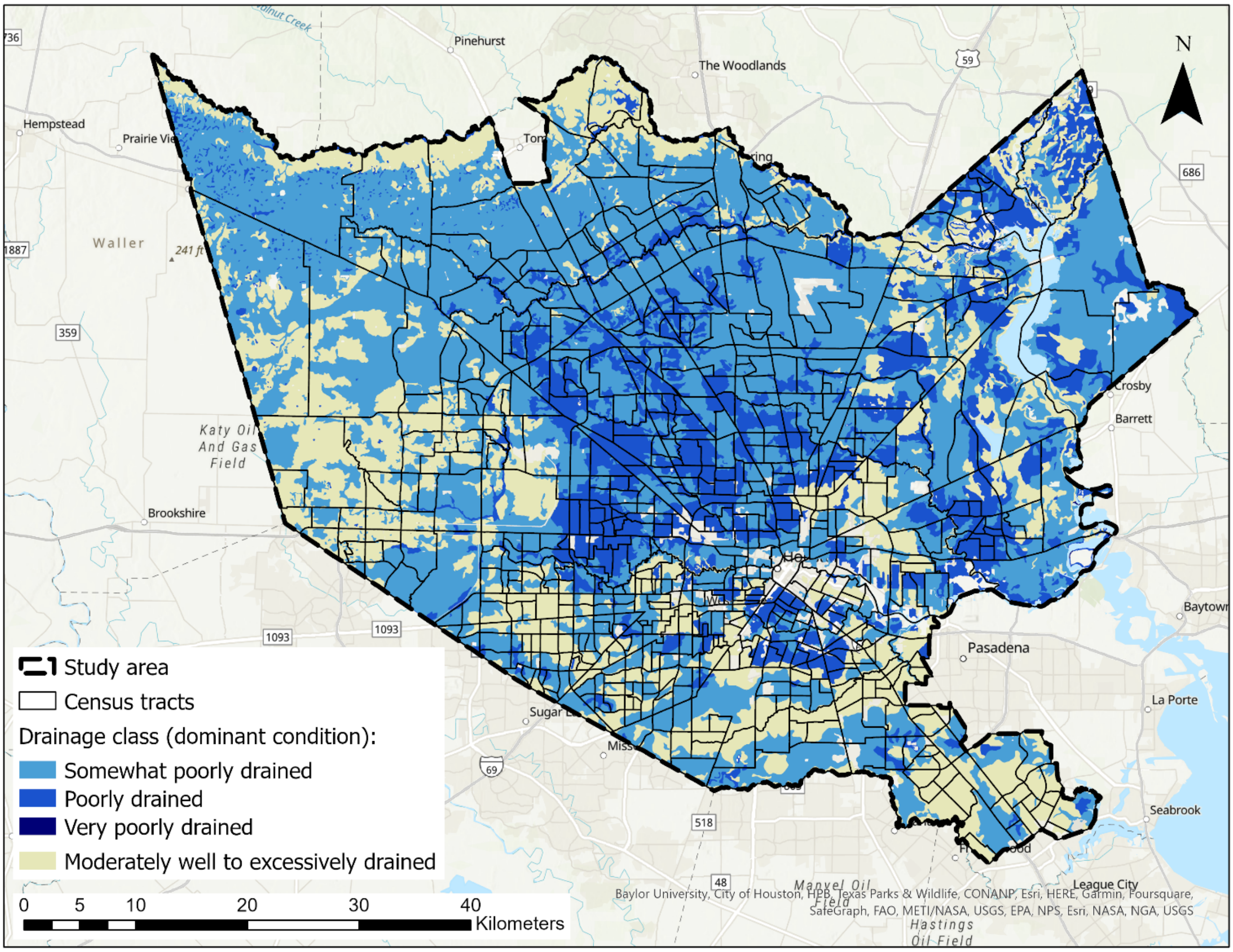


Figure 4: Average damage ratio by census tract based on the National Flood Insurance Program (NFIP) redacted claims.

3.1 Hazard-related variables

290 The DEM analysis showed that the elevation varied between -7.3 and 99.0 meters in the entire study area indicating a range of about 106 meters. Specifically, the median elevation values of the studied census tracts derived from the building footprint centroids ranged from 5.2 to 67.9 meters. The minimum median values were observed in the southeastern region while the highest elevations were found in the northwestern part of the city. The produced curvature grid affirmed the prevalence of flat terrain in the tracts, with the exception of minor man-made modifications in the urban area and along streams. The flood source 295 cost distance grid revealed that the most challenging movement of floodwater, as indicated by the highest values, could occur in the census tracts in the downtown region between Buffalo and Brays Bayous. According to the conducted soil type study, it was found that within 250 out of the total 678 census tracts (37%), all the building centroids were situated on a category of

poorly drained soils. Additionally, in 244 tracts, 50 to 99 percent of the centroids were situated on soils categorized as the above drainage class. Figure 5 represents the distribution of soils across the census tracts in the study area, categorized based on their respective drainage abilities.



Baylor University, City of Houston, HPB, Texas Parks & Wildlife, CONANP, Esri, HERE, Garmin, Foursquare, SafeGraph, FAO, METI/NASA, USGS, EPA, NPS, Esri, NASA, NGA, USGS

Figure 5: Dominant conditions of soil drainage in census tracts of the study area based on the Soil Survey Geographic Database (SSURGO) by the Natural Resources Conservation Service (NRCS) (2022).

3.2 Exposure-related variables

The population density across the census tracts ranged from zero (less than one person) to 134 persons per hectare. The highest values were found in the downtown area and the southwestern part of the city, as illustrated based on the natural breaks in Fig. 6. The variation in population density across the census tracts reflects the diverse spatial distribution of residents, with some locations exhibiting more urbanization and human activity while others remain less densely populated. The median population density in the entire study area was 18 persons per hectare.

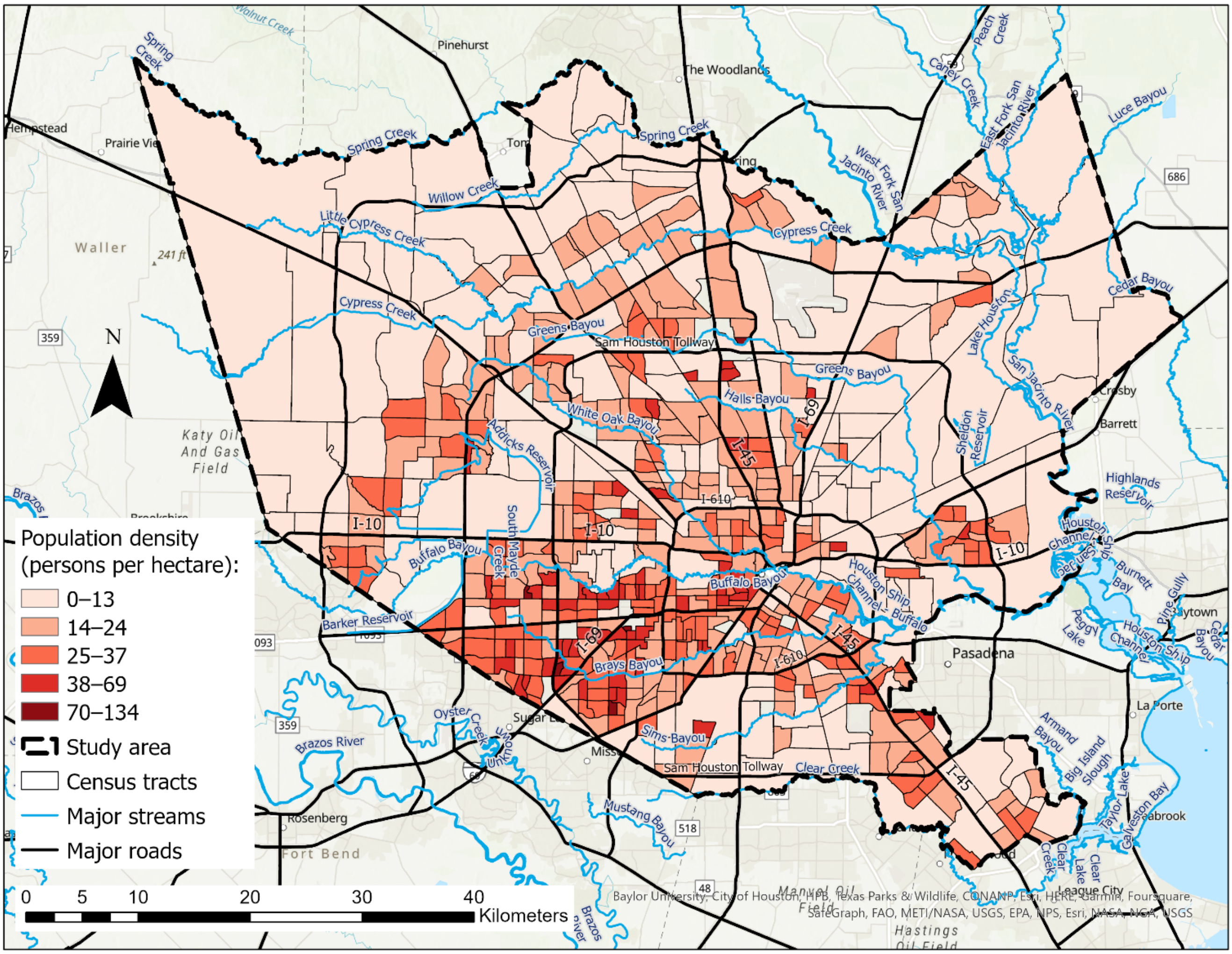
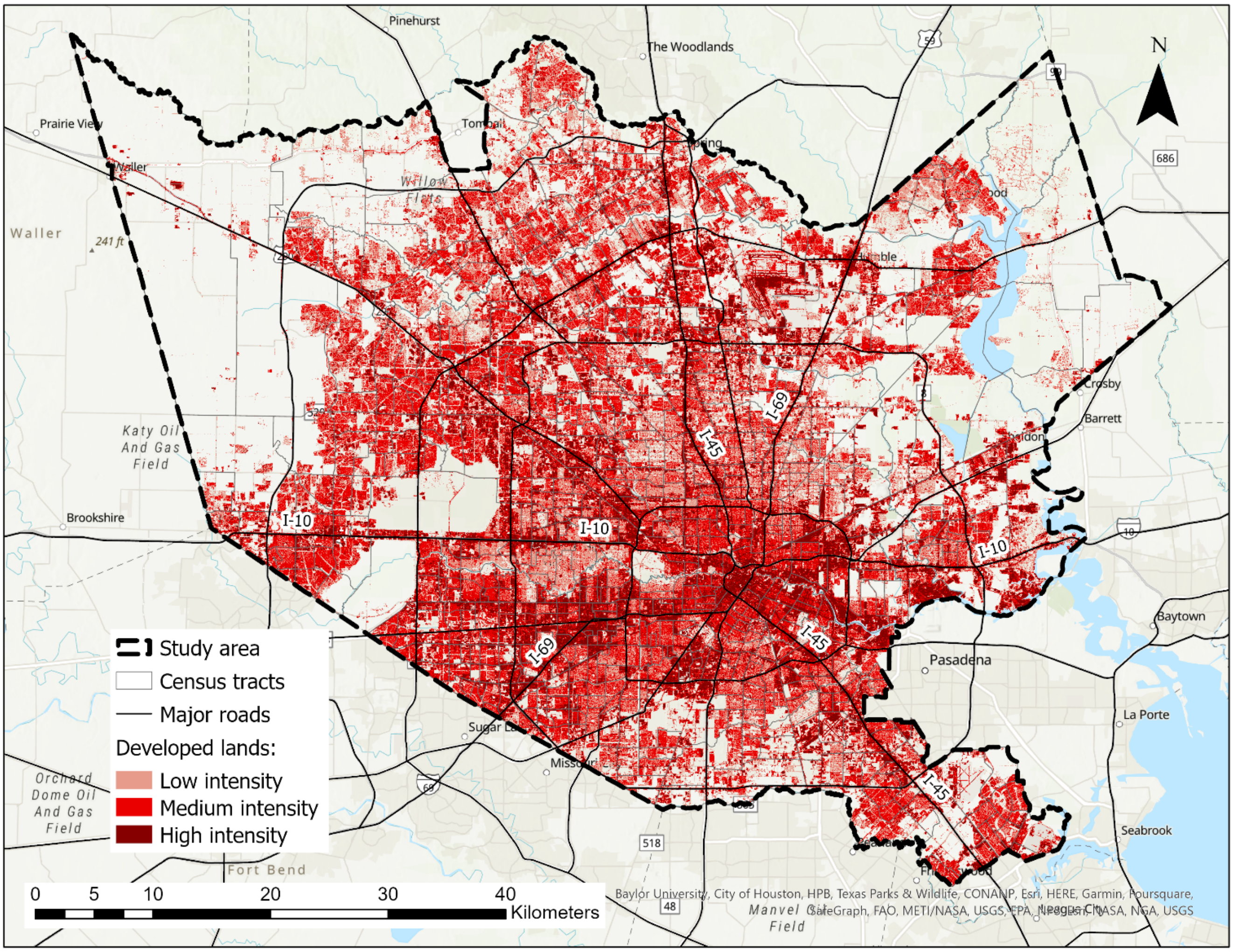


Figure 6: Population density (persons per hectare) in the studied census tracts based on the 2017 American Community Survey (ACS) by the U.S. Census Bureau (2017).

Regarding physical development, the census tracts displayed a range of zero to 72 percent for the ratio of low-intensity developed lands. The percentage of medium-intensity development spanned from two to 87% while high-intensity developed lands covered zero to 86% of the tract areas. Notably, a significant concentration of high-intensity development was observed in the downtown area and along the major interstate highways, indicating a spatial preference for such urbanization patterns (Fig. 7).



Baylor University, City of Houston, HPB, Texas Parks & Wildlife, CONANP, Esri, HERE, Garmin, Foursquare, SafeGraph, FAO, METI/NASA, USGS, EPA, NPS, NRCan, NASA, NGA, USGS

Figure 7: Distribution of the 2016 developed land cover categories in the study area based on the 2016 National Land Cover Database (NLCD) published by Multi-Resolution Land Characteristics Consortium (MRLC) (2016).

3.3 Vulnerability-related variables

Based on the 2017 American Community Survey (ACS), the percentage of individuals living below the poverty line within the census tracts varied between 0.4% and 66.0%. The spatial pattern of this ratio across the study area is depicted using the natural breaks method in Fig. 8. A predominant concentration of census tracts with high poverty ratios can be found in the central part of the urban area, encircled by the Sam Houston Tollway loop.

https://doi.org/10.5194/nhess-2023-113
Preprint. Discussion started: 25 August 2023

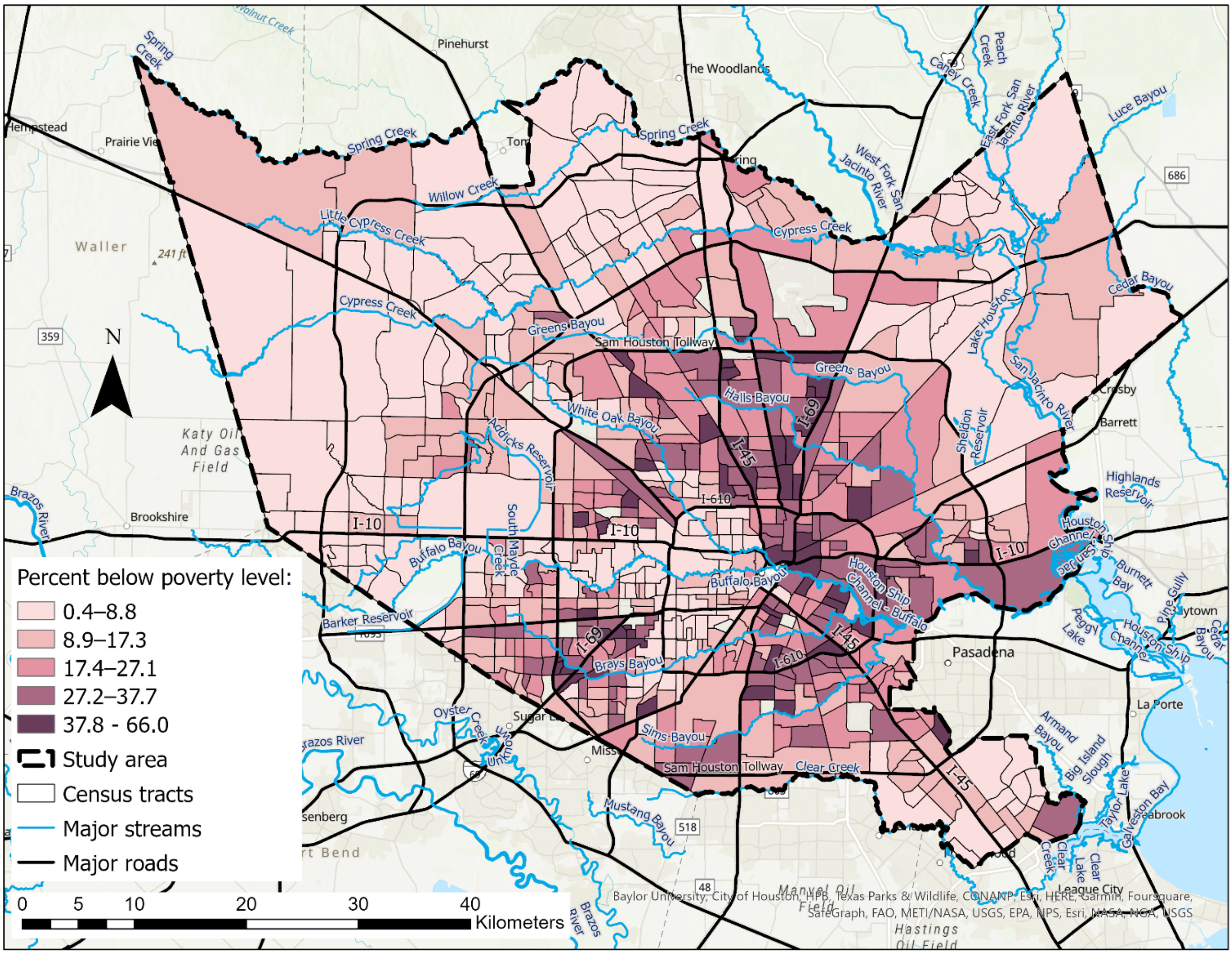


Figure 8: Percentages of the population living below the poverty level in census tracts of the study area based on the 2017 American Community Survey (ACS) by the U.S. Census Bureau (2017).

As analyzed in the study, the unemployment rate among the population aged 16 years or over ranged from zero to 22.3 percent in the census tracts. As seen in Fig. 9, the majority of the tracts with the highest rates were located in the eastern part of the city. According to the results, the ratio of African-American individuals varied from zero to 94.8 percent across the studied census tracts (Fig. 10). Among these tracts, 72 showed a percentage exceeding 50% indicating a significant presence of this racial group while 14 tracts showed no presence of African-Americans.

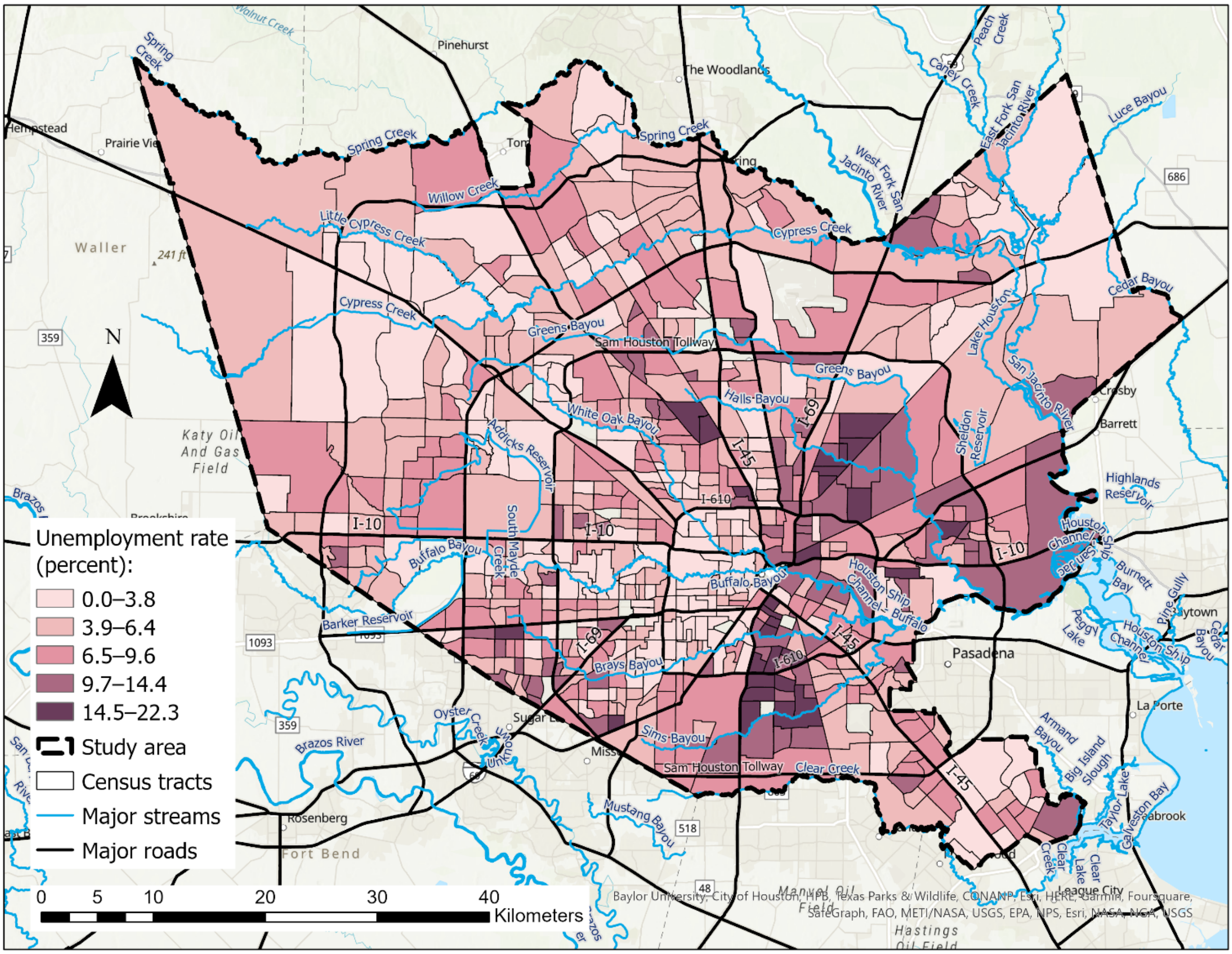


Figure 9: Distribution of the unemployment rate in census tracts of the study area based on the 2017 American Community Survey (ACS) by the U.S. Census Bureau (2017).

The percentage of Hispanic or Latino residents in the census tracts exhibited a range of 2.53 to 97.2 percent. In 229 tracts which accounts for approximately 33% of the total tracts, this ratio exceeded 50% indicating a significant presence of this ethnic group in those areas. The minimum and maximum percentages of the individuals who were not a U.S. citizen were 0.6% and 62.2%, respectively. The median age range was from 21.3 to 57.8 years. Additionally, the percentage of individuals aged 25 or older with no high school diploma ranged from zero to 64.4 percent. In 39 census tracts, this ratio was over 50 percent.

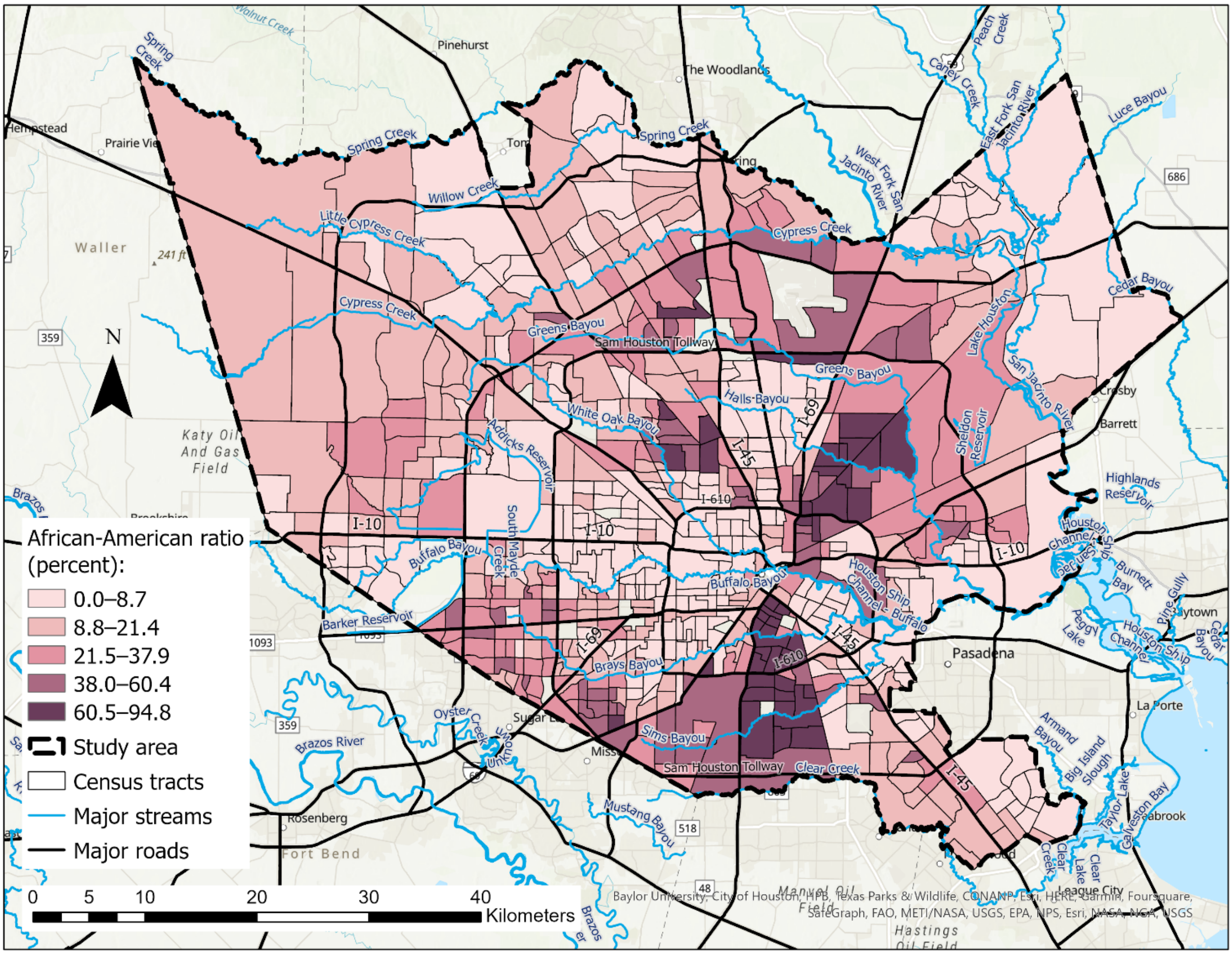


Figure 10: African-American ratios in the census tracts across the study area based on the 2017 American Community Survey (ACS) by the U.S. Census Bureau (2017).

350 Regarding housing characteristics, the study observed a considerable variation in the median housing values across the census tracts, ranging from $20,000 to $2 million. Specifically, ten tracts had median housing values exceeding 1 million dollars while a substantial number of tracts, 192 in total, had median housing values below 100,000 dollars. In addition, the ratio of housing units occupied by renters showed a wide variation ranging from 0.7 to 90.9 percent across the study area. Notably, in 209 tracts, the proportion of renter-occupied units exceeded 50%.

355 3.4 Correlation and importance analyses

The random forest model was trained using a subset of 474 randomly selected entries, accounting for 70% of the dataset. All the predictor variables were included in the training process. Following the training phase, the model's performance was evaluated on the remaining 30% of the dataset which was held out for validation. By examining the root of mean square error (RMSE) and mean absolute error (MAE) values, we assessed the level of accuracy the model achieved in the prediction of the unseen data. The results indicated that the model achieved an RMSE of 12.8 percent measuring the average magnitude of the errors between the predicted and actual damage ratios. Additionally, the MAE was found to be 10.7 percent representing the average absolute difference between the predicted and actual values.

To further investigate the contribution of each variable to the damage ratios, a conditional permutation importance (CPI) analysis was conducted with a threshold of 0.65. To better understand the variables' importance in a more comparable way, we employed a min-max scaling method to normalize the positive CPI scores. This involved assigning weights between 0 and 1 to the scores. We determined the maximum and minimum scores, excluding negative values, and calculated the range. Subsequently, we subtracted the minimum CPI from each score and divided it by the range, Eq. (1):

$$Normalized\ CPI\ weight = \frac{CPI - minimum\ CPI}{maximum\ CPI - minimum\ CPI}, \tag{1}$$

We assumed zero CPI weights for all negative scores. The produced scores and normalized weights for the predictor variables are presented in Table 2. The variables are listed in descending order based on their respective CPI scores, highlighting their relative importance in influencing the average damage ratios.

As seen in the table, it is evident that the percentage of poorly drained soils held the highest importance in the model concerning the average damage losses. Moreover, population density and the proportion of medium-intensity developed areas were identified as two other variables with relatively high influence in the model. Conversely, vulnerability-related variables made lesser contributions to the model. Among these variables, the unemployment rate and African-American ratio exhibited relatively higher CPI scores. To provide a visual summary of the variable importance results, Fig. 11 was created.

Table 2: Predictor variables' generated conditional permutation importance (CPI) scores and normalized weights.

Predictor variable	Conditional permutation importance (CPI) score	Normalized weight
Percentage of poorly drained soils	0.3734	1.00
Population density	0.1389	0.37
Percentage of medium-intensity developed	0.1356	0.36
Median curvature	0.0652	0.17
Median cost distance	0.0248	0.06
Unemployment rate	0.0228	0.06
African-American ratio	0.0225	0.06
Percentage of low-intensity developed	0.0169	0.04
Percent below poverty level	0.0160	0.04
Renter-occupied ratio	0.0096	0.02
Non-citizen ratio	0.0042	0.01
Median housing value	0.0027	0.00
Median elevation	0.0014	0.00
Hispanic or Latino ratio	0.0011	0.00
Median age	-0.0023	0.00
No high school diploma ratio	-0.0024	0.00
Percentage of high-intensity developed	-0.0050	0.00

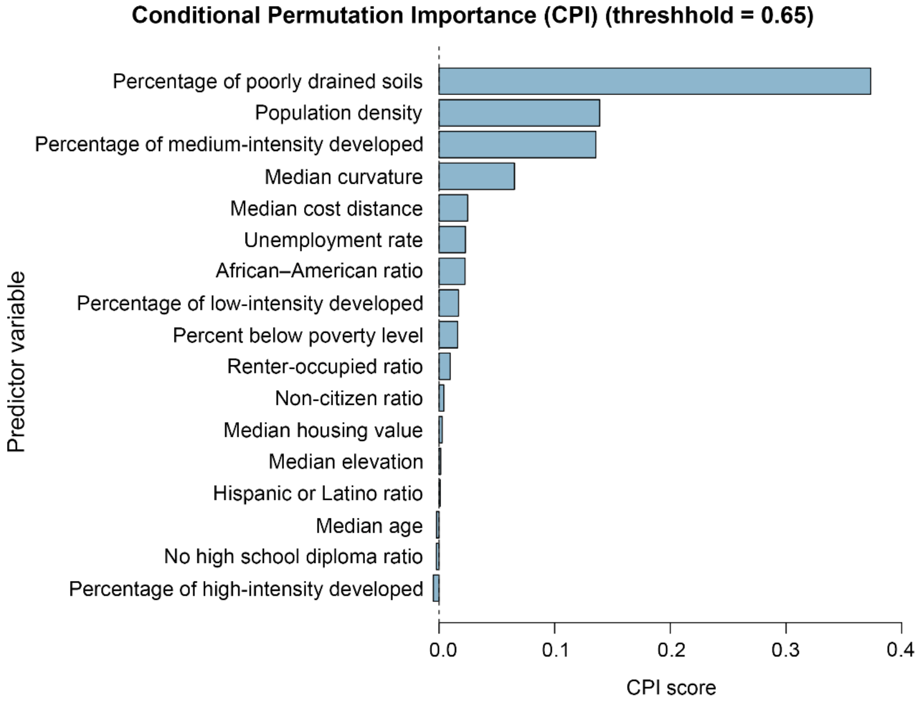


Figure 11: Bar chart of the predictor variables' conditional permutation importance (CPI) scores.

4 Discussion

Based on the findings of this study, the variable associated with the percentage of poorly drained soils in the census tracts emerged as significantly more important in predicting average damage ratios. Following this, exposure-related variables of population density and the proportion of medium-intensity developed areas exhibited similar contribution levels. Conversely, the vulnerability-related variables derived from socioeconomic data demonstrated considerably lower levels of importance, as indicated by the model results. These outcomes can be discussed and interpreted from various perspectives.

The utilization of the National Flood Insurance Program (NFIP) claims data can provide valuable insights into the analysis of flood loss. However, it is important to acknowledge that these data may introduce biases that can affect the predictive models due to their limitations. The NFIP operates based on flood maps that designate zones with a one percent chance of flooding in a given year (100-year) commonly referred to as the Special Flood Hazard Area (SFHA). FEMA is responsible for administering and providing flood insurance coverage for eligible communities in the United States. Within the SFHA, there exists a mandatory obligation to purchase flood insurance for properties that are being financed by federally regulated lending institutions. To address instances of noncompliance, lenders and loan servicers have been more rigorously enforcing this requirement. They are now required to assess and document whether a structure falls in the SFHA and ensure that the borrower maintains flood insurance throughout the loan (Highfield et al., 2013). One major limitation stems from the SFHA maps themselves. These maps have been subject to criticism due to their inherent inaccuracies, incomplete geographic coverage, lack of consideration for climate change impacts, exclusion of pluvial (overland) floods, reliance on a binary classification of properties as “inside” or “outside” flood zones, and long (5-year) update periods that fail to account for rapid land alterations (Brody et al., 2018; Highfield et al., 2013; Pralle, 2019).

Investigations into the flooding caused by Hurricane Harvey in 2017 revealed that nearly 75 percent of the affected residential structures in the City of Houston and surrounding areas in Harris County were located outside the above designated 100-year flood zone subject to the regulatory flood insurance measures (Pralle, 2019). The owners whose buildings are not located in the SFHA are less likely to buy flood insurance. In cases where properties outside the SFHA are not insured, the owners can still file flood claims to seek assistance through the Individual and Households Program (IHP). But the support provided to uninsured properties through the IHP is typically minimal compared to the assistance offered to insured properties. The lack of insurance coverage and limited availability of federal support in areas outside the SFHA may lead to a potential underreporting of flood damages (Knighton et al., 2020). This issue poses another limitation with the NFIP claims data which could potentially influence the results and findings of this study.

Due to data availability considerations, the experiment was conducted at the scale of census tracts. The NFIP claims records were redacted by FEMA to ensure individual privacy protection. In addition, the used detailed socioeconomic data were publicly available at the tract level, as provided by the Census Bureau. Therefore, the study was preferred to be conducted at this scale. Moreover, the environmental variables were extracted for the building centroids and then aggregated at the tract level as median values or percentages. This approach offered advantages by focusing the analysis on locations of insurable buildings rather than undeveloped areas or open spaces. However, it is important to acknowledge that uncertainties could arise from summarizing the damage ratios and predictor variables within the census tracts as average, median, or percentage values. The tracts that were predominantly non-residential such as those containing airports and universities were not included to ensure a comprehensive and accurate analysis of residential demographics.

The analyses conducted in this study are subject to a potential limitation, namely the spatial resolution. In the development of hazard-related variables, we used a 10 m DEM, while for development of the physical exposure indicators, we relied on the 30 m NLCD land cover dataset. The discrepancy between these resolutions could pose a challenge for the research; however, we opted not to resample any of these datasets to avoid introducing additional uncertainty. It is worth emphasizing that the cell size of the datasets employed had an impact on the study results. Access to raster grids at a higher resolution would likely result in more accurate outcomes. For future studies, an alternative approach could involve extracting such variables through remote sensing techniques utilizing high-resolution imagery.

5 Conclusions

The goal of this study was to investigate the relationships between direct flood loss to buildings, as the response variable, and a combination of socioeconomic attributes and environmental characteristics, as the conditioning factors, in Houston, TX. The findings partially supported the research hypothesis in the study area. It was observed that flood damage ratios in the census tracts were correlated with the presence of poorly drained soils which emerged as the most significant factor. This correlation can be attributed to the predominance of impermeable clay soils in the region reducing hydraulic conductivity. According to the model's analysis, this particular factor exerted a greater influence on the loss ratios even compared to the cost distance to the streams. This emphasizes the essential role of geotechnical considerations in land use planning and zoning regulations to consider this problem. Moreover, increasing the implementation of green infrastructure and natural water retention features such as wetlands, bioswales, and green roofs can help absorb and retain rainwater, mitigating the impact of impermeable surfaces. Integrating these nature-based solutions into urban development plans can improve water management and reduce flood loss. In addition, updating and enforcing building codes and floodplain management regulations can enhance the resilience of structures in the flood-prone areas of the city. Measures such as requiring elevated foundations, flood-resistant construction materials, and proper drainage systems can minimize flood damage and improve community resilience.

The next important variables, with similar degrees of contribution, were population density and the proportion of medium-intensity developed lands. Specifically, in the southwestern part of the urban area, census tracts with higher population densities and a dominance of medium-intensity developed lots could potentially contribute to the correlation with flood loss. In this type of development, impervious surfaces make up 50% to 79% of the total land cover, and the majority of residential structures consist of one- or two-story single-family homes. Since the entirety of a one-story building is typically exposed to floodwaters, the ratio of flood damage to the property value is often higher for such structures. Thus, the combination of a high population, significant imperviousness, and a low number of stories in buildings could contribute to higher ratios of flood loss. Given the above situation, promoting smart growth principles (United States Environmental Protection Agency (EPA), 2011) and compact development strategies in such urban regions can be a solution. Encouraging higher-density developments in appropriate areas, as recommended by such principles, can minimize impervious surfaces and mitigate flood risk. This approach can include mixed-use developments, higher building densities, and the integration of green spaces within communities.

The results of the importance analysis did not align with the expectations of the environmental justice concept. Variables related to social vulnerability were found to have low significance in predicting the damage ratios. In addition to the limitations of the NFIP claims data discussed earlier, there are different perspectives that can explain these outcomes. For example, one reason could be the cost of mandatory flood insurance which may discourage economically disadvantaged individuals from choosing to live in flood-prone areas or prompt them to relocate if they are already residing in such zones (Cutter et al., 2018). Knighton et al. (2020) mention two other potential reasons. Firstly, minorities with limited access to financial resources for recovery are more likely to completely relocate from hazardous zones rather than claim for assistance after a disaster. Secondly, historical racial segregation efforts in the region may have resulted in minorities being clustered in other urban districts that do not intersect with the floodplains. Future studies could be developed to geospatially analyze the impact of such historical efforts in relation to flood exposure and loss.

Hale et al. (2018) offer a different perspective considering a location-based approach. They attribute the presence of less socially vulnerable individuals in flood-prone locations to the added value of aesthetic advantages near waterbodies or water-based amenities in urban areas making such regions less affordable for the minorities and disadvantaged groups. This proposition may hold true for census tracts with higher median housing values or those containing new green spaces and recreational developments along the bayous in Houston. As another potential direction for future research, it would be worth considering the affordability of such areas for individuals with varying levels of wealth and income.

This study, summarizing and analyzing millions of records and spatial features, can be an example of handling both spatial and aspatial big data through the integration of geospatial technologies and AI. It built upon previous studies such as those conducted by Alipour et al. (2020), Kalaycıoğlu et al. (2023), and Knighton et al. (2020) to incorporate the relevant methods in a comprehensively to consider the effects of environmental and socioeconomic factors on flood loss. This article can also contribute to the evolving field of applying machine learning techniques to comprehensive disaster risk assessment processes. Future research can enhance the current findings by considering additional factors associated with alterations of precipitation patterns and hydrological cycles resulting from climate change. It would be beneficial to include more detailed data on rainfall and stream gauges in such studies. This would enable a more thorough analysis of these variables and their potential influence on the study outcomes.

Data availability

The raw data can be provided by the corresponding author upon request.

Author contributions

BB, AEM and MPS: Conceptualization; BB: Data Curation; BB: Formal analysis; BB, AEM and MPS: Investigation; BB, AEM and MPS: Methodology; BB: Project administration; BB and AEM: Resources; AEM and MPS: Supervision; AEM and MPS: Validation; BB: Visualization; BB: Writing - Original Draft; AEM and MPS: Writing - Review & Editing

Competing interests

The authors declare that they have no conflict of interest.

References

Alipour, A., Ahmadalipour, A., Abbaszadeh, P., and Moradkhani, H.: Leveraging machine learning for predicting flash flood damage in the Southeast US, Environ. Res. Lett., 15, 024011, https://doi.org/10.1088/1748-9326/ab6edd, 2020.

Bedient, P., Blackburn, J., Gori, A., and Juan, A.: Tropical storm Harvey summary report - No. 1, Tech. rep., Severe Storm Prediction, Education, & Evacuation from Disasters (SSPEED) Center, Rice University, Houston, USA, 13 pp., https://www.sspeed.rice.edu/harvey-reports, 2017.

Birkmann, J. and Welle, T.: Assessing the risk of loss and damage: Exposure, vulnerability and risk to climate-related hazards for different country classifications, Int. J. Global Warm., 8, 191-212, https://doi.org/10.1504/IJGW.2015.071963, 2015.

Brivio, P. A., Colombo, R., Maggi, M., and Tomasoni, R.: Integration of remote sensing data and GIS for accurate mapping of flooded areas, Int. J. Remote Sens., 23, 429-441, https://doi.org/10.1080/01431160010014729, 2002.

Brody, S. D., Sebastian, A., Blessing, R., and Bedient, P. B.: Case study results from southeast Houston, Texas: Identifying the impacts of residential location on flood risk and loss, J. Flood Risk Manag., 11, S110-S120, https://doi.org/10.1111/jfr3.12184, 2018.

City of Houston: Annexation, https://www.houstontx.gov/planning/Annexation/ (last access: 21 February 2022).

Collins, T. W., Grineski, S. E., and Chakraborty, J.: Environmental injustice and flood risk: A conceptual model and case comparison of metropolitan Miami and Houston, USA, Reg. Environ. Change, 18, 311-323, https://doi.org/10.1007/s10113-017-1121-9, 2018.

Crichton, D.: UK and global insurance responses to flood hazard, Water Int., 27, 119-131, https://doi.org/10.1080/02508060208686984, 2002.

Cutter, S. L. and Finch, C.: Temporal and spatial changes in social vulnerability to natural hazards, P. Natl. Acad. Sci. USA, 105, 2301-2306, https://doi.org/10.1073/pnas.0710375105, 2008.

Cutter, S. L., Boruff, B. J., and Shirley, W. L.: Social vulnerability to environmental hazards, Soc. Sci. Quart., 84, 242-261, https://doi.org/10.1111/1540-6237.8402002, 2003.

Cutter, S. L., Emrich, C. T., Gall, M., and Reeves, R.: Flash flood risk and the paradox of urban development, Nat. Hazards Rev., 19, 05017005, https://doi.org/10.1061/(ASCE)NH.1527-6996.0000268, 2018.

Debeer, D. and Strobl, C.: Conditional permutation importance revisited, BMC Bioinformatics, 21, 1-30, https://doi.org/10.1186/s12859-020-03622-2, 2020.

Debeer, D., Hothorn, T., and Strobl, C.: permimp: Conditional Permutation Importance Version: 1.0-2 [code], https://CRAN.R-project.org/package=permimp, 2021.

Dewan, A. M.: Floods in a megacity: Geospatial techniques in assessing hazards, risk and vulnerability, Springer, Dordrecht, the Netherlands, 199 pp., ISBN 978-94-007-5874-2, 2013.

Environment Systems Research Institute (ESRI) and United States Department of Agriculture (USDA), Natural Resources Conservation Service (NRCS): SSURGO Downloader [data set], https://www.arcgis.com/home/item.html?id=cdc49bd63ea54dd2977f3f2853e07fff (last access: 10 June 2023), 2022.

Environment Systems Research Institute (ESRI): ArcGIS Pro software version 3.1[code], 2023.

Federal Emergency Management Agency (FEMA): OpenFEMA Dataset, Disaster Declarations Summaries - v2 [data set], https://www.fema.gov/openfema-data-page/disaster-declarations-summaries-v2 (last access: 4 June 2023), 2023. “This product uses the FEMA OpenFEMA API, but is not endorsed by FEMA. The Federal Government or FEMA cannot vouch for the data or analyses derived from these data after the data have been retrieved from the Agency’s website(s)”.

Ferguson, A. P. and Ashley, W. S.: Spatiotemporal analysis of residential flood exposure in the Atlanta, Georgia metropolitan area, Nat. Hazards, 87, 989-1016, https://doi.org/10.1007/s11069-017-2806-6, 2017.

Figueiredo, R. and Martina, M.: Using open building data in the development of exposure data sets for catastrophe risk modelling, Nat. Hazard Earth Sys., 16, 417, https://doi.org/10.5194/nhess-16-417-2016, 2016.

Flanagan, B. E., Gregory, E. W., Hallisey, E. J., Heitgerd, J. L., and Lewis, B.: A social vulnerability index for disaster management, J. Homel. Secur. Emerg., 8, https://doi.org/10.2202/1547-7355.1792, 2011.

Flores, A. B., Collins, T. W., Grineski, S. E., and Chakraborty, J.: Social vulnerability to Hurricane Harvey: Unmet needs and adverse event experiences in Greater Houston, Texas, Int. J. Disast. Risk Re., 46, 101521. https://doi.org/10.1016/j.ijdrr.2020.101521, 2020.

Hale, R. L., Flint, C. G., Jackson-Smith, D., and Endter-Wada, J.: Social dimensions of urban flood experience, exposure, and concern, JAWRA J. Am. Water Resour. As., 54, 1137-1150, https://doi.org/10.1111/1752-1688.12676, 2018.

Hallegatte, S.: Natural disasters and climate change: An economic perspective, Springer, Washington, DC, USA, 194 pp., ISBN 978-3-319-08932-4, 2014.

Hammond, M. J., Chen, A. S., Djordjević, S., Butler, D., and Mark, O.: Urban flood impact assessment: A state-of-the-art review, Urban Water J., 12, 14-29, https://doi.org/10.1080/1573062X.2013.857421, 2015.

Highfield, W. E., Norman, S. A., and Brody, S. D.: Examining the 100-year floodplain as a metric of risk, loss, and household adjustment, Risk Analysis: An International Journal, 33, 186-191, https://doi.org/10.1111/j.1539-6924.2012.01840.x, 2013.

Houston–Galveston Area Council (H-GAC): GIS Datasets, USGS DEM 10m [data set], https://www.h-gac.com/rds/gis-data/gis-datasets.aspx (last access: 11 January 2019), 2019.

Jha, A. K., Bloch, R., and Lamond, J.: Cities and flooding: A guide to integrated urban flood risk management for the 21st century, The World Bank, Washington, DC, USA, 631 pp., ISBN: 978-0-8213-9477-9, 2012.

Joy, J., Kanga, S., and Singh, S. K.: Kerala flood 2018: Flood mapping by participatory GIS approach, Meloor Panchayat, International Journal on Emerging Technologies, 10, 197-205, https://sdma.kerala.gov.in/wp-content/uploads/2020/08/FloodMapping-Joy-2020.pdf, 2019.

Kalaycıoğlu, O., Akhanlı, S. E., Menteşe, E. Y., Kalaycıoğlu, M., and Kalaycıoğlu, S.: Using machine learning algorithms to identify predictors of social vulnerability in the event of a hazard: Istanbul case study, Nat. Hazard Earth Sys., 23, 2133-2156, https://doi.org/10.5194/nhess-23-2133-2023, 2023.

Kaźmierczak, A. and Cavan, G.: Surface water flooding risk to urban communities: Analysis of vulnerability, hazard and exposure, Landscape Urban Plan., 103, 185-197, https://doi.org/10.1016/j.landurbplan.2011.07.008, 2011.

Knighton, J., Buchanan, B., Guzman, C., Elliott, R., White, E., and Rahm, B.: Predicting flood insurance claims with hydrologic and socioeconomic demographics via machine learning: Exploring the roles of topography, minority populations, and political dissimilarity, J. Environ. Manage., 272, 111051, https://doi.org/10.1016/j.jenvman.2020.111051, 2020.

Kubal, C., Haase, D., Meyer, V., and Scheuer, S.: Integrated urban flood risk assessment-Adapting a multicriteria approach to a city, Nat. Hazard Earth Sys., 9, 1881-1895, https://doi.org/10.5194/nhess-9-1881-2009, 2009.

Lin, J. M. and Billa, L.: Spatial prediction of flood-prone areas using geographically weighted regression, Environmental Advances, 6, 100118, https://doi.org/10.1016/j.envadv.2021.100118, 2021.

Maldonado, A., Collins, T. W., Grineski, S. E., and Chakraborty, J.: Exposure to flood hazards in Miami and Houston: Are Hispanic immigrants at greater risk than other social groups, Int. J. Env. Res. Pub. He., 13, 775, https://doi.org/10.3390/ijerph13080775, 2016.

Maxwell, A. E., Sharma, M., Kite, J. S., Donaldson, K. A., Thompson, J. A., Bell, M. L., and Maynard, S. M.: Slope failure prediction using random forest machine learning and LiDAR in an eroded folded mountain belt, Remote Sens., 12, 486, https://doi.org/10.3390/rs12030486, 2020.

Maxwell, A. E., Warner, T. A., and Fang, F.: Implementation of machine-learning classification in remote sensing: An applied review, Int. J. Remote Sens., 39, 2784-2817, https://doi.org/10.1080/01431161.2018.1433343, 2018.

Merz, B., Kreibich, H., and Lall, U.: Multi-variate flood damage assessment: A tree-based data-mining approach, Nat. Hazard Earth Sys., 13, 53-64, https://doi.org/10.5194/nhess-13-53-2013, 2013.

Merz, B., Kreibich, H., Schwarze, R., and Thieken, A.: Review article "Assessment of economic flood damage", Nat. Hazard Earth Sys., 10, 1697-1724, https://doi.org/10.5194/nhess-10-1697-2010, 2010.

Merz, B., Kreibich, H., Thieken, A., and Schmidtke, R.: Estimation uncertainty of direct monetary flood damage to buildings, Nat. Hazard Earth Sys., 4, 153-163, https://doi.org/10.5194/nhess-4-153-2004, 2004.

Microsoft Corporation: Microsoft Building Footprints v2.0, United States GitHub [data set], https://www.microsoft.com/en-us/maps/building-footprints (last access: 6 September 2021), 2018.

Mohanty, M. P. and Simonovic, S. P.: Understanding dynamics of population flood exposure in Canada with multiple high-resolution population datasets, Sci. Total Environ., 759, 143559, https://doi.org/10.1016/j.scitotenv.2020.143559, 2021.

Mosavi, A., Ozturk, P., and Chau, K. W.: Flood prediction using machine learning models: Literature review, Water, 10, 1536, https://doi.org/10.3390/w10111536, 2018.

Multi-Resolution Land Characteristics Consortium (MRLC): National Land Cover Database class legend and description, https://www.mrlc.gov/data/legends/national-land-cover-database-class-legend-and-description (last access: 10 June 2023).

Multi-Resolution Land Characteristics Consortium (MRLC): NLCD 2016 Land Cover (CONUS) [data set], https://www.mrlc.gov/data/nlcd-2016-land-cover-conus (last access: 10 June 2023), 2016.

Murphy, K. P.: Machine learning: A probabilistic perspective, The MIT Press, Cambridge, Massachusetts, USA, 1067 pp., ISBN 978-0-262-01802-9, 2012.

Poole, D. L. and Mackworth, A. K.: Artificial Intelligence: Foundations of computational agents (2nd edition), Cambridge University Press, New York, USA, 820 pp., ISBN 978-1-107-19539-4, 2017.

Pralle, S.: Drawing lines: FEMA and the politics of mapping flood zones, Climatic Change, 152, 227-237, https://doi.org/10.1007/s10584-018-2287-y, 2019.

R Core Team: R: A language and environment for statistical computing, R Foundation for Statistical Computing (R-4.2.3) [code], https://www.R-project.org/, 2023.

Rose, A.: Economic principles, issues, and research priorities in hazard loss estimation, in: Modeling spatial and economic impacts of disasters, edited by: Okuyama, Y., and Chang, S. E., Springer, New York, USA, 13-36, 2004.

Spearman, C.: The proof and measurement of association between two things, Int. J. Epidemiol., 39, 1137–1150, https://doi.org/10.1093/ije/dyq191, 2010.

Taramelli, A., Righini, M., Valentini, E., Alfieri, L., Gatti, I., and Gabellani, S.: Building-scale flood loss estimation through vulnerability pattern characterization: Application to an urban flood in Milan, Italy, Nat. Hazard Earth Sys., 22, 3543-3569, https://doi.org/10.5194/nhess-22-3543-2022, 2022.

Tate, E., Rahman, M. A., Emrich, C. T., and Sampson, C. C.: Flood exposure and social vulnerability in the United States, Nat. Hazards, 106, 435-457, https://doi.org/10.1007/s11069-020-04470-2, 2021.

Tehrany, M. S., Pradhan, B., and Jebur, M. N.: Spatial prediction of flood susceptible areas using rule based decision tree (DT) and a novel ensemble bivariate and multivariate statistical models in GIS, J. Hydrol., 504, 69-79, https://doi.org/10.1016/j.jhydrol.2013.09.034, 2013.

Thieken, A. H., Olschewski, A., Kreibich, H., Kobsch, S., and Merz, B.: Development and evaluation of FLEMOps–a new Flood Loss Estimation MOdel for the private sector, WIT Trans. Ecol. Envir., 118, 315-324, https://www.witpress.com/elibrary/wit-transactions-on-ecology-and-the-environment/118/19311, 2008.

United States Census Bureau, Geography Division: 2017 TIGER/Line® Shapefiles: Census Tracts [data set], https://www.census.gov/cgi-bin/geo/shapefiles/index.php?year=2017&layergroup=Census+Tracts (last access: 19 June 2023), 2017.

United States Census Bureau: DP1 | Profile of general population and housing characteristics, https://data.census.gov/table?g=160XX00US4835000&d=DEC+Demographic+Profile&tid=DECENNIALDP2020.DP1 (last access: 23 June 2023), 2020.

United States Census Bureau: Explore Census Data, Advanced Search [data set], https://data.census.gov/advanced (last access: 1 June 2023), 2017.

United States Environmental Protection Agency (EPA): Smart growth: A guide to developing and implementing greenhouse gas reductions programs, Tech. rep., EPA, Washington, DC, USA, 40 pp., https://www.epa.gov/sites/default/files/2017-06/documents/sm_growth_guide.pdf, 2011.

United States Geological Survey (USGS): The National Map Downloader, National Hydrography Dataset (NHD) [data set], https://apps.nationalmap.gov/downloader/#/ (last access: 10 June 2023), 2023.

Wagenaar, D., Curran, A., Balbi, M., Bhardwaj, A., Soden, R., Hartato, E., et al.: Invited perspectives: How machine learning will change flood risk and impact assessment, Nat. Hazard Earth Sys., 20, 1149-1161, https://doi.org/10.5194/nhess-20-1149-2020, 2020.

Wagenaar, D., Jong, J. D., and Bouwer, L. M.: Multi-variable flood damage modelling with limited data using supervised learning approaches, Nat. Hazard Earth Sys., 17, 1683-1696, https://doi.org/10.5194/nhess-17-1683-2017, 2017.

Wang, Z., Lai, C., Chen, X., Yang, B., Zhao, S., and Bai, X.: Flood hazard risk assessment model based on random forest, J. Hydrol., 527, 1130-1141, https://doi.org/10.1016/j.jhydrol.2015.06.008, 2015.

Watson, K. M., Harwell, G. R., Wallace, D. S., Welborn, T. L., Stengel, V. G., and McDowell, J. S.: Characterization of Peak Streamflows and Flood Inundation of Selected Areas in Southeastern Texas and Southwestern Louisiana from the August and September 2017 Flood Resulting from Hurricane Harvey, Tech. rep., United States Geological Survey (USGS) in cooperation with Federal Emergency Management Agency (FEMA), Austin, USA, 44 pp., https://doi.org/10.3133/sir20185070, 2018.