Journal of Geoscience and Environment Protection, 2026, 14(5), 265-290
https://www.scirp.org/journal/gep
ISSN Online: 2327-4344
ISSN Print: 2327-4336

Mapping Community Vulnerability to Hurricane Hazards in Coastal North Carolina Using Machine Learning

Om Dahal[1], Satya Kalluri[2], Dambar Uprety[3], Donglian Sun[1*]

[1]Department of Geography and Geoinformation Science, George Mason University, Fairfax, VA, USA
[2]NOAA NESDIS, College Park, MD, USA
[3]Department of Information Systems and Business Analytics, Kent State University, Kent, OH, USA
Email: duprety@kent.edu, *dsun@gmu.edu

How to cite this paper: Dahal, O., Kalluri, S., Uprety, D., & Sun, D. L. (2026). Mapping Community Vulnerability to Hurricane Hazards in Coastal North Carolina Using Machine Learning. *Journal of Geoscience and Environment Protection, 14,* 265-290.
https://doi.org/10.4236/gep.2026.145016

Received: February 24, 2026
Accepted: May 26, 2026
Published: May 29, 2026

Copyright © 2026 by author(s) and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

Abstract

Extreme record-breaking hurricanes, followed by heavy rainfall and flooding, claim a lot of lives and billions of dollars' worth of property damage every year in the Atlantic coastal areas of the United States. The Atlantic coast areas are most vulnerable to hurricane hazards, but not all the communities are equally vulnerable due to their varying degrees of exposure and coping abilities. Thus, it is of vital importance to learn the extent of vulnerability of different communities for prevention, preparedness, response, and recovery efforts. Many physical, statistical, and data-driven methods have been employed to predict geophysical area-centered vulnerability to landslides and floods, primarily using geophysical explanatory variables, but not hurricane-induced hazards. This study makes three key contributions. First, it integrates geophysical, demographic, and social media data to assess community-level vulnerability to hurricane hazards. Second, it applies a Random Forest framework to model vulnerability at the census block level, capturing non-linear interactions among predictors. Third, it provides empirical evidence on the relative importance of explanatory variables, highlighting the role of real-time social media data in disaster vulnerability assessment. The results indicate strong predictive performance (R^2 = 0.93) and identify tweets, roads, elevation, NDVI, and water bodies as the most influential variables. The findings highlight the importance of integrating geophysical, demographic, and real-time social media data for accurate vulnerability assessment. This approach provides a scalable framework for disaster preparedness and risk management in coastal regions.

Keywords

Mapping Vulnerable Communities, Hurricane Hazards, Remote Sensing,

Social Media, Demographic Data, Machine Learning

1. Introduction

Disaster is an overall consequence of a hazard event (Klonner et al., 2016). Vulnerability is a function of exposure and coping ability, or people cannot deal with the hazards due to the physical and social backgrounds of the place of their residence (Wu et al., 2002). The vulnerability of communities varies with their coping ability, which is a combination of resistance and resilience (Rygel et al., 2006). Levels of risk depend on the hazard intensity and levels of vulnerability. Therefore, the same hazard may have different impacts on different communities or places depending on their exposure and coping ability (Klonner et al., 2016). Vulnerability has been conceptualized as pre-existing conditions that potentially expose humans to hazards, e.g., humans settled in hazardous areas. Loss of life and property is likely in the hazardous areas when there is a natural event. This is a vulnerability caused by biophysical settings of the area of residence (Rygel et al., 2006). The second way of conceptualizing vulnerability is social vulnerability that stems from social marginalization due to age, race, disability, or income (Rygel et al., 2006). Assessment of social or community vulnerability needs inclusion of selecting demographic data, essentially disability, vulnerable age groups (children, aged population), and poverty (economic factor) (Aubrecht et al., 2013). The third approach is the vulnerability of places that combine biophysical as well as social risk within a specific geographic area to assess vulnerability (Rygel et al., 2006). There are multiple frameworks to explain the root cause of vulnerability to natural disasters, from social conditions inherent in the community to the biophysical environment around the community, or a combination of both. It is crucial to consider a coupled human-environment system, associating it with the proximity to hazards to identify vulnerable communities (Cutter et al., 2008).

Identification and mapping of coastal communities at risk of hurricane flood hazards is crucial for every stage of disaster management, consisting of prevention, preparedness, response, and recovery. The use of remote sensing data analysis methods has been increasingly used for risk assessment from hurricane disasters (Hoque et al., 2017a). This is particularly promising given the increasing availability and high spatial resolution of remotely sensed data for hurricane risk assessment (Zhou et al., 2019).

The geotagged information from Twitter, Facebook, or Flickr has also been proven to be highly applicable in hurricane impact studies because they provide valuable information regarding geometries, attributes, and semantic information. The social media-generated data can have spatial patterns, and the social media posts during disaster strikes are closer to the affected areas, and they are likely to relate to the disaster effects. Therefore, social media-generated geographic information is a promising alternative to geospatial data for natural hazard analysis. For the natural disaster hazard model, the crowdsourced locations, messages, or

images can be used as valuable complementary data (Klonner et al., 2016).

Statistical, physical, and data-driven (e.g., machine learning) models are typically used in the prediction of natural hazard events by modelling to assess the risks. Even though the physical models have great capabilities of predicting natural hazards and risks, they require datasets collected from the ground, intensive computation, and a high level of expertise. The most notable drawback of this modeling is that the prediction cannot be carried out in a short time frame because data collection efforts take a long time. Similarly, numerical prediction models have systematic errors (Mosavi et al., 2018). To overcome the shortcomings of these models, data-driven prediction modeling has been widely used. The strengths of the machine learning models are that they do not require knowledge of underlying physical processes, are quicker to develop, and allow fast training, validation, testing, and evaluation. Moreover, this approach has outperformed the conventional approaches with higher prediction accuracy, and data-driven algorithms can predict beyond the range of training datasets spatially and temporally (Mosavi et al., 2018). Artificial Neural Networks (ANNs), Multilayer Perceptron (MLP), Adaptive Neuro-Fuzzy Inference System (ANFIS), Wavelet Neural Network (WNN), Support Vector Machine (SVM), Decision Tree (DT), and Ensemble Prediction Systems (EPSs) are the algorithms that have the highest favorability among the natural hazards modeling community (Mosavi et al., 2018).

In the hurricane hazard risk analyses literature, socio-economic variables have been preferred less than geophysical variables to analyze the vulnerability of coastal communities from hurricane hazards, although it is a multivariate nonlinear problem. Thus, this work has utilized both types of variables in the analysis.

High wind and storm surge, coupled with flooding, are the number one cause of infrastructure damage and loss of life and property in the coastal United States (Helderop & Grubesic, 2019). Extreme weather events (e.g., hurricanes, floods, fires) are in an increasing trend. Consequently, the effects on life and property are expected to increase in the future (Bouwer, 2018; Hoque et al., 2017b), which will make coastal human communities more vulnerable (Hoque et al., 2017b).

Consideration of the demographic scenario of coastal areas is vital in hurricane vulnerability analysis. About 94.7 million (29.1 percent of the total U.S. population) live in coastline regions, of which about 44.4 million people live in the Atlantic coastline. The Atlantic coastline had a 13.2 percent population growth between 2000 and 2017. The percentage population of 85 and older is higher in coastline counties compared to that of the United States (Cohen, 2019). The population in Atlantic coastal areas is most vulnerable to hurricanes because the frequency of devastating hurricanes is high in this region. It is evident from the fact that eight hurricanes made landfall in Atlantic coastal areas between 2000 and 2017, each of which caused more than 10 billion worth of damage (Cohen, 2019).

Hurricane Florence is another disaster event of the most devastating hurricane event that occurred in the 2018 hurricane season. It lasted until September 18 since it made landfall on September 14 with a forward motion of about 3 - 4 miles per hour with a zone of tropical storm force winds nearly 400 miles wide (Feaster

et al., 2018). This hurricane was at the intensity of category 1 along the southeastern coast of North Carolina. It caused a total of 52 fatalities and estimated damage of approximately $24 billion, of which a significant portion of the loss was in North Carolina. About one million households lost power only in North Carolina. Numerous trees were uprooted due to the force of hurricane winds, but most of the damages to homes and commercial buildings were caused by freshwater flooding, with approximately 74,563 structures being flooded (Stewart & Berg, 2019). The loss of agricultural farm products and livestock alone due to Hurricane Florence accounted for at least $1.1 billion (Feaster et al., 2018).

Florence produced 10 to more than 30 inches of rainfall in New Hanover County and surrounding areas due to slow movements and persistent rain bands before and after the hurricane made landfall by setting a new record of rainfall in two decades. This extreme rain resulted in record-breaking river floods across New Hanover County. Eighteen record-breaking peaks of streamflow were recorded in North Carolina, with some of them having the highest records since 1940 (Stewart & Berg, 2019).

New Hanover County is one of the coastline counties located in the tidewater area in the state of North Carolina. This county has a total area of 328 square miles (850 km^2), of which 192 square miles (500 km^2, 42%) is water (https://en.wikipedia.org/wiki/New_Hanover_County,_North_Carolina). This is one of the most densely populated counties in North Carolina, with a population of 227,198, and 95,097 households according to Census Bureau estimates of 2017 (https://www.census.gov/acs/www/data/data-tables-and-tools/supplemental-tables/). Wilmington is one of the largest cities in North Carolina, located in New Hanover County (**Figure 1** left).

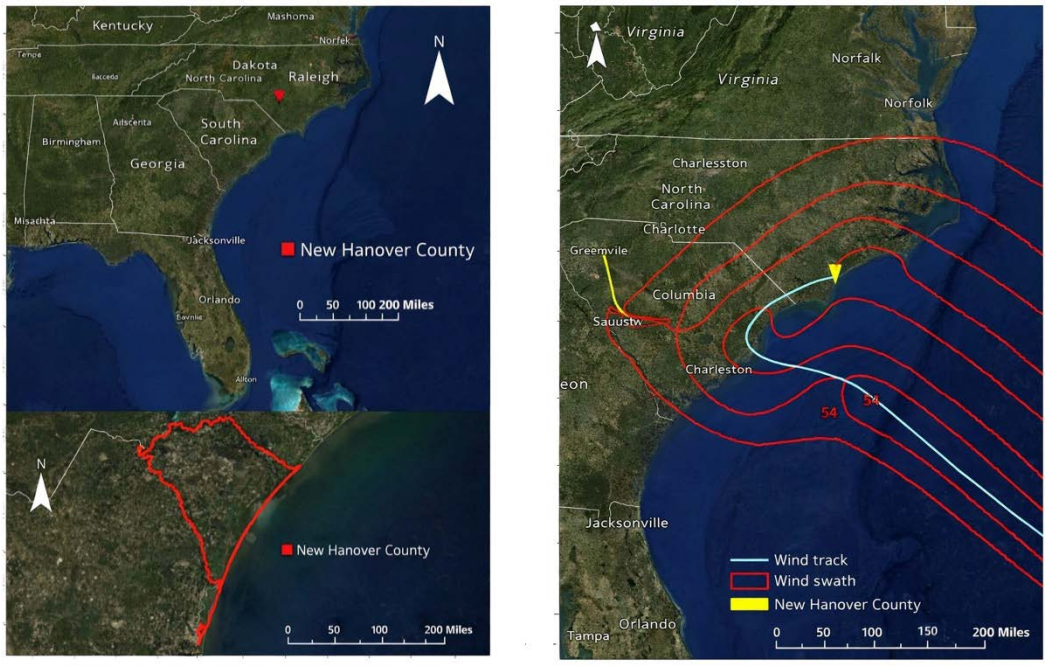

Figure 1. New Hanover County, North Carolina (left), and hurricane Florence wind track and swath (right).

New Hanover County is one of the hardest hit areas by Hurricane Florence on the North Carolina coast, where the worst flash floods were experienced in local history. Florence made landfall near Wrightsville Beach, which is in the coastal area of New Hanover County (**Figure 1**, right). In New Hanover County, up to 3 feet of flash flood water inundated Northchase, Writsboro, and Ogden neighborhoods. Downtown Wilmington was inundated by 2 feet of floodwater from the Cape Fear River. As a result, the entire county was generally isolated from the outside world due to access road closures for several days (Stewart & Berg, 2019).

Given the devastating hurricane, rainfall, and flooding events that occur frequently in the coastal United States, it is vital to learn the levels of vulnerability of different communities for damage prevention, preparedness, and rescue and recovery efforts. Moreover, it is necessary to know what variables should be given higher priority for the study. Similarly, whether the extensively used machine learning algorithm in prediction (mainly used to predict areas of potential landslides and flooding natural disasters), Random Forest (RF), can be useful to predict vulnerable communities from hurricane hazards. Thus, the objectives set for this study are as follows:

1) To identify the level of vulnerability of different communities in coastal New Hanover County, North Carolina, from Hurricane Florence using geophysical, socio-economic, and social media-generated variables.

2) To examine the usefulness and applicability of the Random Forests algorithm to make categorical predictions of vulnerability in coastal communities from hurricane hazards.

The remainder of this paper is organized as follows. Section 2 describes the datasets used in the study. Section 3 outlines the methodology, including the Random Forest modeling framework and variable construction. Section 4 presents the results and analysis of vulnerability prediction. Section 5 concludes with key findings and directions for future research.

2. Data Used

The datasets used in this study are listed here:

- Sentinel-2 multispectral data at a high resolution (10 m) were used for land-use/land-cover (LULC) classification and to calculate Normalized Difference Vegetation Index (NDVI).
- The digital elevation model (DEM) obtained from the U.S. Geological Survey (USGS) was used as an elevation dataset for the model.
- Road features were obtained from the North Carolina (NC) Department of Transportation (NCDOT).
- Major rivers distributed by the NC Center for Geographic Information and Analysis.
- American Community Survey (ACS) data at the block group level of age, disability, and poverty were used as demographic explanatory variables.

- Real-time Twitter stream during the Hurricane Florence period from September 14 to September 19, 2018.

3. Methodology

3.1. Random Forest Classification and Regression

Natural hazard risk prediction is a multivariate and non-linear task (Wang et al., 2015) due to the combined role of many disaster-inducing factors. Several methods and machine learning algorithms have been employed to solve predictive analysis, such as support vector machine (SVM), artificial neural networks (ANN), and decision trees (DT). The major weakness of these algorithms is their inability to estimate each conditioning factor's contribution to the total risk (Wang et al. 2015). CART decision trees are greedy. Even with bagging, the trees can have structural similarities that will result in high correlation in predictions. However, in Random Forest (RF), the trees are uncorrelated or at least correlated because learning algorithms just select a random sample of features from a random sample of variables consistent with standard Random Forest methodology (Breiman, 2001). RFs are modifications of classification and regression trees (CART) algorithms (Pourghasemi & Kerle, 2016). It is a supervised classification and regression method of modelling that allows growing an ensemble of trees and letting them vote for the most occurring class as the predicted class (Breiman, 2001). The RF is the algorithm that is capable of estimating the contribution of each factor to the total effect (Wang et al., 2015). The RF has high forecast accuracy, acceptable tolerance to outliers and noise, and easy avoidance of overfitting (Wang et al., 2015). RF algorithm generates numerous binary trees, which are collectively called forests (Park & Kim, 2019). In the RF, trees grow based on a bootstrap sample. For each node, random subsets of samples are selected. The "out-of-bag" error rate is calculated using samples out of the bootstrap sample (Park & Kim, 2019). Mean decreases in accuracy and mean in the Gini are calculated in the process, which is then used to calculate the variable importance score (Park & Kim, 2019).

Given an observation, for each tree in the model, RF predicts the outcome using a tree applied to an observation and stores the outcome as a list. If the model is a classifier, it returns the maximum count. If the model is a regression, it returns an average as depicted in **Figure 2**, consistent with standard ensemble learning frameworks

The RF algorithm uses a parallel ensemble method called "bagging" or bootstrap aggregation to generate classifiers. This is a method that averages multiple estimates that are measured from random subsamples of variables. A subset of observations is selected at random to form a subsample and used to train the model. The process is repeated to select the subset of samples from the original observation until the specified number of trees is reached. This process is known as bootstrapping, consistent with standard Random Forest methodology (Breiman, 2001). Random Forests are built by: specifying the number of trees,

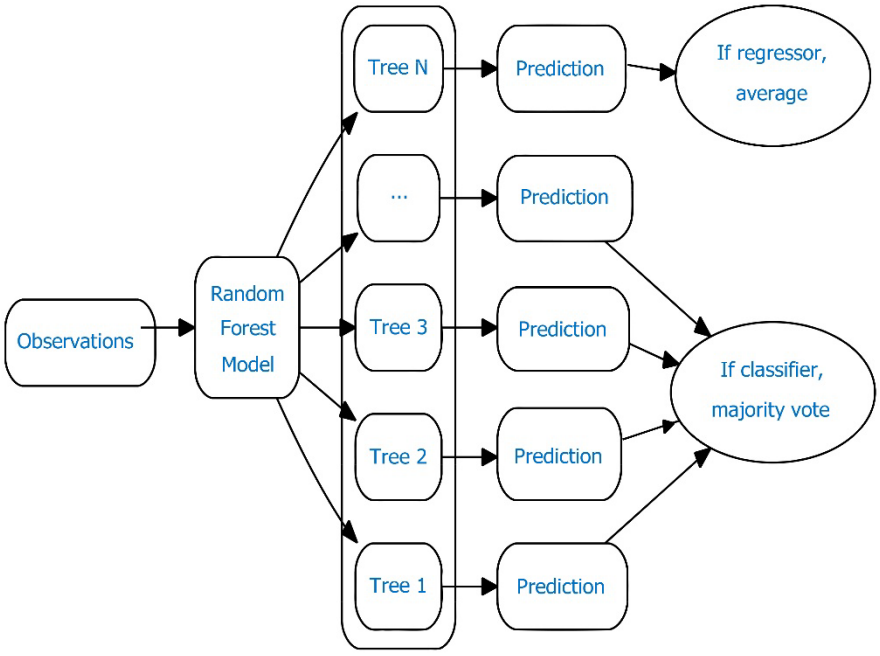

Figure 2. Random Forest process flow.

specifying the number of variables, specifying the number of features (columns) to be used in each tree. Then, for each tree: select some samples with replacement from all observations (rows), select given number of features randomly, train a decision tree with selected samples and features (Breiman, 2001). Select a specified number of samples from the original dataset, which is known as bootstrap samples. Randomly select variables from the sample for each node split. An unpruned classification tree is grown for each bootstrap sample. All the trees are aggregated to predict the new label by majority votes (Ai et al. 2014).

Variable importance

Mean decrease accuracy and mean decrease Gini are widely used for measuring, ranking, and selecting variable importance (Park & Kim 2019). Often in regression problems, the drop in the sum of squared errors, and classification problems, the Gini impurity is commonly used to evaluate node purity in tree-based models and is widely applied in Random Forest algorithms (Breiman, 2001; Hastie et al., 2009). The greater the impurity, the greater the importance of the variable, as commonly established in tree-based learning methods (Breiman, 2001; Hastie et al., 2009). The Gini impurity is computed by summing the probability of each item chosen multiplied by the probability of an error to classify that item into the correct class (Ai et al., 2014). The Gini impurity is obtained by the following algorithm (Equation (1)).

$$G(k) = \sum_{i=1}^{n} P(i) \times (1 - P(i)) \quad (1)$$

The Gini impurity of the parent node is higher than that of the child node (Wang et al., 2015). The Gini decrease of each explanatory variable is combined to estimate the total contribution of it in the prediction of vulnerability (Wang et

al., 2015). The variable importance is calculated by the given formula (Equation (2)).

$$P_k = \frac{\sum_{i=1}^{n}\sum_{j=1}^{t} D_{Gkij}}{\sum_{k=1}^{m}\sum_{i=1}^{n} D_{Gkij}} \quad (2)$$

where P_k is the variable importance, m = the total number of explanatory variables, n = the total number of classification trees, t = the total number of nodes, D_{Gkij} = Gini decrease value of the j^{th} node in the i^{th} tree that belongs to the k^{th} variable. Mean-squared error (MSE) is obtained by the given equation (Equation (3)).

$$\varepsilon = \left(V_{\text{observed}} - V_{\text{response}}\right)^2 \quad (3)$$

where ε is the mean squared error, V_{observed} is the variable from observed data, and V_{response} is the variable from the result (Lee et al. 2017).

Out-of-Bag (OOB) error

Each tree in the random forest is constructed from a random sample of observations, usually called bootstrap samples. The observations that are left out from constructing a tree during the classification process are called "out-of-bag" (OOB) observations, i.e., unseen data in classification (or out of bootstrap samples). Therefore, each tree is constructed from different samples from the whole dataset. The prediction for an observation made from the trees for which the observation was not used. The error rate that is estimated from these predictions is known as OOB error (Ai et al. 2014; Janitza & Hornung, 2018).

3.2. Training Features

Ground observations from social media were used for training. NAPSG Foundation, GISCorps, and CEDR Digital maintained a Story Map displaying 2018 hurricane crowdsourced photos collected from Instagram, Twitter, Facebook, and online news media (https://napsg.maps.arcgis.com/apps/StoryMapCrowdsource/index.html?appid=69b95886cf8e49a3a349c9d550174a91). This is the collection of photos with a brief description of events by social media users illustrating the incidences (e.g., hurricane impact, hurricane intensity, damage, storm surge, flooding, rescue efforts, etc.) during and after the hurricane event. After careful observation of photos and their descriptions, they were classified into four different levels of severity as class vulnerability levels (categories) and assigned the numbers from 1 to 4 - 1 being the most at risk (highest vulnerability) location and 4 being the least at risk (lowest or no vulnerability) location. Within the study area, 99 locations were identified from the story map that were appropriate for training input. Similarly, emergency shelters, shelter locations designated by New Hanover County to evacuate county residents during natural disaster emergencies, including hurricanes and floods. These were considered as no-risk or least risk locations, which were assigned to 5 and 6 in the vulnerability category. More locations were identified by observing satellite imagery and flood maps during Hurricane Florence and as-

signed numbers from 1 to 6 vulnerability categories depending on the severity of the impact observed. A total of 273 locations were identified for input as training features. These training features with vulnerability labels are summarized in Table 1 and displayed on a map in Figure 3.

Table 1. Vulnerability categories and the number of locations for model training.

Vulnerability Category		Number of Locations
1	Very High	59
2	High	77
3	Moderate	51
4	Low	31
5	Relatively Low	17
6	Very Low	38

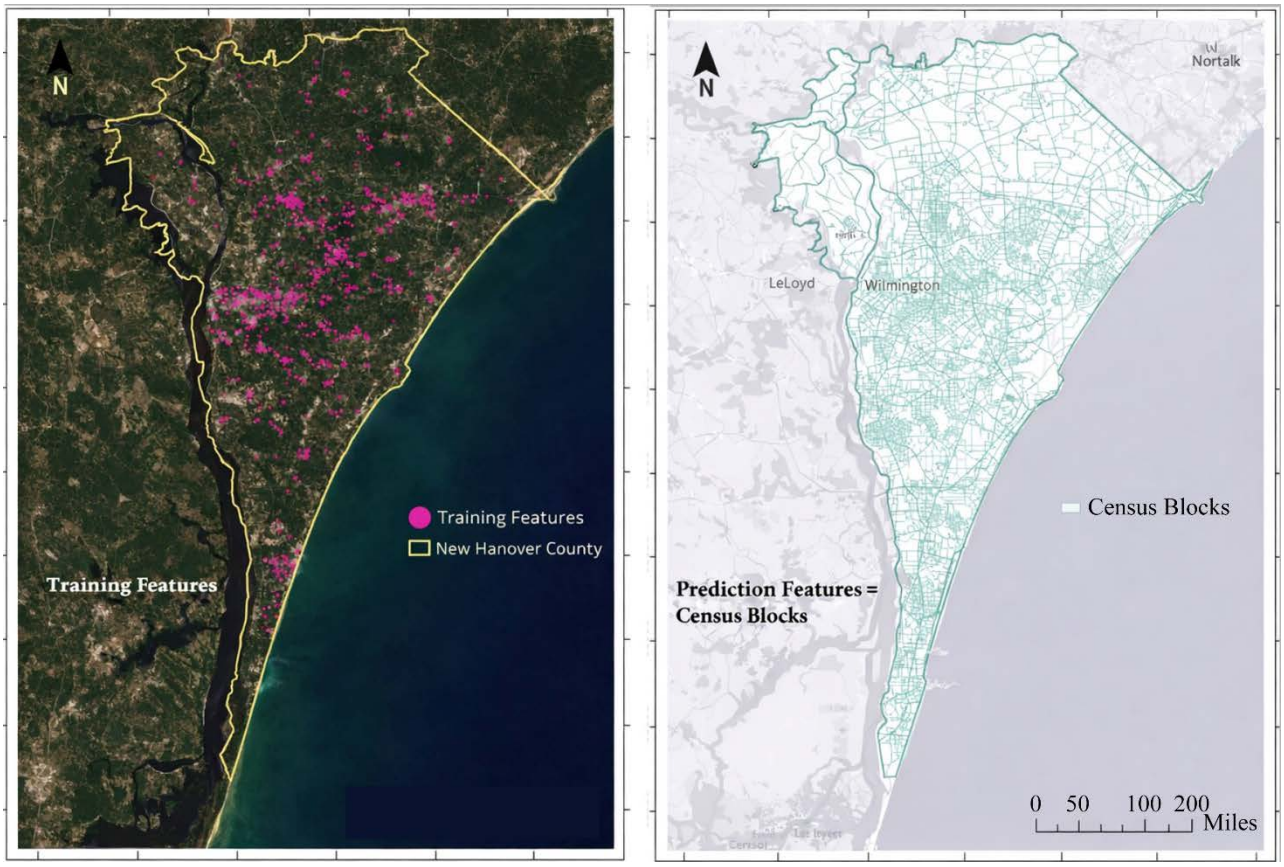

Figure 3. Training point features corresponding to Table 1 (left), and prediction polygon features (census blocks, right).

The distribution of training samples across vulnerability categories (Table 1) indicates moderate class imbalance, with fewer observations in extreme categories. This imbalance may influence model training by biasing predictions toward more frequent classes. While Random Forest is relatively robust to class imbalance

due to its ensemble structure, imbalanced class distributions can still affect predictive performance (Chen et al., 2004). Techniques such as stratified sampling, class weighting, or balanced subsampling could further improve model reliability.

Future work will explicitly address class imbalance and evaluate performance differences across categories, particularly for extreme vulnerability levels.

3.3. Labeling Criteria and Quality Control

Vulnerability categories (1 - 6) were assigned based on observable damage severity and contextual information from crowdsourced images, descriptions, and satellite imagery. The classification followed these general rules:

- **Category 1 (Very High):** severe structural damage, deep flooding, life-threatening conditions;
- **Category 2 (High):** significant flooding or infrastructure damage;
- **Category 3 (Moderate):** localized flooding or moderate disruption;
- **Category 4 (Low):** minimal visible damage;
- **Category 5 - 6 (Very Low):** safe zones such as shelters or unaffected areas.

To reduce subjectivity, ambiguous cases were reviewed iteratively and assigned based on consensus interpretation of image evidence and contextual metadata. A subset of locations was re-evaluated to ensure consistency in labeling decisions.

These photos are powered by NAPSG Foundation, GISCorps, and CEDR Digital, a Story Map (upper left), a general map with a cluster of locations with impacted locations; the map showing the location of the damaged gas station in Wilmington, NC on 9/14/2018 (upper right), map showing location of a downed tree on a house on 9/14/2018 (lower left); location on map and an abandoned car in Wilmington, NC on 9/15/2018 (lower right).

3.4. Prediction Polygon Features

This is a feature that represents polygons to receive the results of the predictions made by the models. Since the goal of this work is to make predictions for the vulnerability of communities, the census blocks would be ideal polygon features to predict on because census blocks are the areas that encompass small communities with distinctive geophysical and demographic similarities. New Hanover County consists of 5069 census blocks as delineated by the US Census Bureau in the 2010 census (**Figure 3**, right).

3.5. Explanatory Variables

The vulnerability is due to one or a combination of multiple geophysical, demographic, or socio-economic conditions of people or places. These conditions, information generated regarding these conditions, and information regarding the hurricane itself can be defined as explanatory variables for hurricane disaster analysis. There is no consensus as to which factors should be given higher priority when deciding the level of vulnerability of communities from hurricanes in coastal areas (Bathi & Das, 2016). This work used a combination of geophysical, demo-

graphic, and social media-generated information as explanatory variables, as discussed in forthcoming sections.

Geophysical variables

1) Land use/land cover

Sentinel-2, a sensor developed by the European Space Agency (ESA), provides high-resolution (10 m) multispectral imagery for surface reflectance, which was used for land-use/land-cover (LULC) classification. The imagery was classified using a Semi-automatic Classification Plugin for QGIS version 2.18. Out of 12 Sentinel-2 spectral bands, bands 1 (coastal aerosol), 9 (water vapor), and 10 (cirrus) were excluded from the classification dataset. The imagery was classified into nine different land-use and land-cover classes: a) forest, b) ocean, c) river, d) lake/pond, e) road, f) residential, g) agricultural, h) commercial, and i) marsh. The maximum likelihood algorithm was used to classify the imagery, which calculates the probability distribution for the classes, based on the Bayesian theorem, to determine which pixel belongs to the land cover class in training (Richards & Jia, 2006). The classified output raster was then resampled to 30 m to reduce the number of pixels to synchronize with the processing ability of ArcGIS Pro, Forest-based Classification and Regression tool.

2) Elevation

Digital elevation model (DEM) dataset at 1/9 arc seconds (approximately 1 m) resolution obtained from the 3D Elevation Program (3DEP) of the USGS National Map Services (https://www.usgs.gov/core-science-systems/ngp/3dep) was used as an elevation dataset for the model. The DEM was resampled to 30 m to overcome the computational limitation of the tool. The elevation of New Hanover County ranges from 0 m to 30 m.

3) Slope

The slope tells the steepness of a raster surface. The slope was calculated in degrees using the DEM data discussed in the previous section. The planar method parameter was used, where the slope is measured as the maximum rate of change in value from a cell to its immediate neighbors. The following slope algorithm was used (Equation (4)).

$$\text{Slope degrees} = \text{ATAN}\left(\left[\mathrm{d}z/\mathrm{d}x\right]^2 + \left[\mathrm{d}z/\mathrm{d}y\right]^2\right)^{1/2} * 180/\pi \tag{4}$$

where $\frac{\mathrm{d}z}{\mathrm{d}x}$ is the rate of change in the *x*-direction, and $\frac{\mathrm{d}z}{\mathrm{d}y}$ is the rate of change in the *y*-direction. The slope indicates the topographic change and variability of the surface. A lower slope means a flatter surface, which has a higher risk of flooding (Wang et al., 2015). The slope raster used as input is shown in the map.

4) Stream Power Index (SPI)

Stream Power Index (SPI) is a measure of the power of flowing water on the terrain surface. The higher the stream power index, the more erosion it can cause downstream. Stream power is a hydrological factor that can condition or explain how damaging the flood would be (Wang et al., 2015; Lee et al., 2017). The SPI

was calculated from slope and flow accumulation raster datasets obtained from terrain analysis of digital elevation models. The percent rise slope was used to calculate SPI by the formula in the ArcGIS raster calculator (Equation (5)).

$$\text{SPI} = \ln\left(\text{Flow accumulation raster} + 0.001\right) * \left(\left(\text{Slope raster}/100\right) + 0.001\right) \quad (5)$$

5) Normalized Difference Vegetation Index (NDVI)

Normalized Difference Vegetation Index (NDVI) measures the difference between near-infrared (NIR) and red values of wavelengths. NDVI values range from −1 to 1. Healthy vegetation has the highest NDVI value, i.e., inclined towards 1, and water is inclined towards −1. Other land cover values fall between these two extremes depending on the type, growth, soil moisture, and presence or absence of vegetation, snow, and soil roughness (Wang et al., 2015). NDVI of the area of interest was computed from the Sentinel-2 imagery bands, Band 4 (Red) and Band 8 (NIR), as given by the formula in Equation (6) below. NDVI input is as shown in the map.

$$\text{NDVI} = \frac{\text{NIR}\left(\text{Band 8}\right) - \text{Red}\left(\text{Band 4}\right)}{\text{NIR}\left(\text{Band 8}\right) + \text{Red}\left(\text{Band 4}\right)} \quad (6)$$

6) Major roads

Major roads play a critical role before, during, and after natural disasters from the perspective of evacuations, rescue, and recovery needs. The more roads and the wider roads are closer to a settlement, the easier it becomes to evacuate and provide post-event assistance. As a result, the communities could become safer from the impacts of hurricanes and floods. Thus, road features can be considered marked variables to explain vulnerability prediction. Road features were obtained from the North Carolina Department of Transportation (NCDOT).

7) Water features

The NC Center for Geographic Information and Analysis distributes the major hydrography data that includes major rivers and water bodies (lakes, ponds, dams, etc.) and floodwaters. Rivers and other water features are the areas where floods surge during hurricanes and heavy rainfall. People near these water features could be in danger of being affected by floods. For this reason, this is an important addition to the list of explanatory variables of vulnerability prediction.

Demographic variables

As shown in **Figure 4**, poverty, gender, race, ethnicity, age, and disability are demographic indicators of social vulnerability. Poor, women, children, people with disabilities, and aged people are vulnerable because of their inability to have access to resources that they need to protect themselves when disaster strikes and recover in the aftermath of disaster (Rygel et al., 2006). American Community Survey (ACS) 2017 data at block group levels of age, disability, and poverty were used as demographic explanatory variables. Age groups 0 - 14 and 65 plus, the population with disability, and the population with poverty are considered more vulnerable than the rest of the population in the event of natural disasters such as hurricanes.

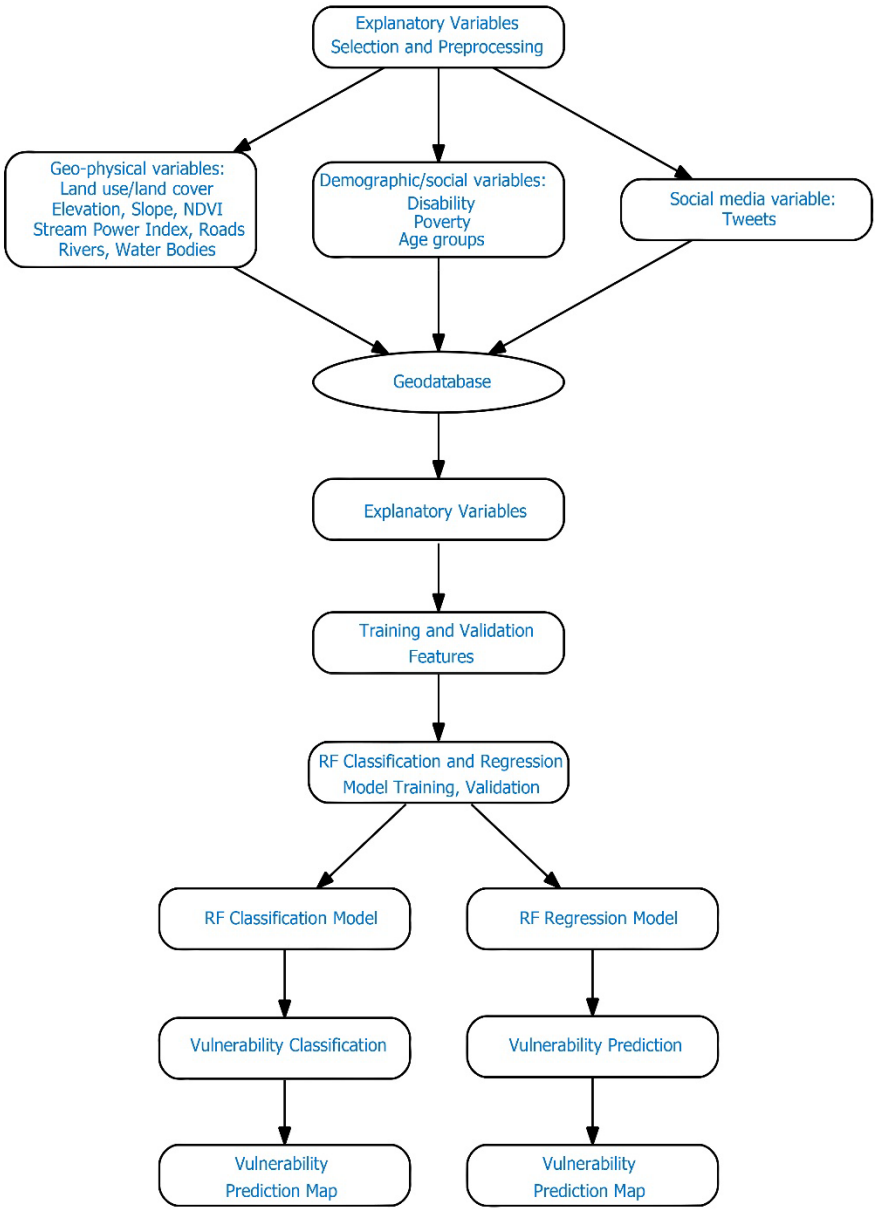

Figure 4. Workflow for hurricane vulnerability mapping and prediction.

Social media variables

Social media is a fundamental tool for individuals to access and disseminate real-time information regarding storm intensity, damage, safety, evacuation, and recovery during disaster events (Goodchild & Glennon, 2010; Klonner et al., 2016). The "tweets" variable represents the spatial presence of geotagged Twitter posts during Hurricane Florence. Specifically, the predictor was constructed as the count of geotagged tweets within proximity to each location, serving as a proxy for real-time human-reported impact intensity. Tweets were filtered using the keyword "#Florence" and restricted to geolocated posts within New Hanover County during September 14-19, 2018. A total of 65 geotagged tweets were retained after filtering.

Given the limited number of observations, this variable may be subject to spatial and sampling bias. Therefore, its influence on model performance should be interpreted cautiously. Future work will explore alternative representations such as kernel density estimation, spatial smoothing, or population-normalized tweet intensity to improve robustness.

4. Results

The Random Forests model was created and applied. The explanatory distance variables and explanatory raster variables described in the previous section were used to predict hurricane flood vulnerability by the RF regression model. The model was constructed based on "vulnerability levels," which are variables to predict. Variables to predict were represented as six ordered vulnerability levels from 1 to 6 (1 indicating the most vulnerable and 6 indicating the least vulnerable to hurricane hazards) as an attribute in training features. A combination of thirteen vector and raster geospatial datasets that conceivably would explain the vulnerability of communities in New Hanover County, North Carolina, was used as input variables in the analysis.

Explanatory variables were calculated by first finding distances from the nearest input distance features to each of the input training features. Explanatory variables are extracted from the raster input dataset for each point location. The distance attributes were calculated from the training feature to the closest segments of the polygons or lines of explanatory variables. The explanatory variables were then used to construct a model and predict the vulnerability of communities using census blocks as prediction areas. The input variables for training include LULC, elevation, NDVI, SPI (**Figure 5**), slope, major roads, major rivers, water bodies (**Figure 6**), poverty, disability, children, and aged population (**Figure 7**), and tweets (**Figure 8**), as shown below:

Regression Results and Analysis

The two-thousand decision trees parameter was found to be optimal during model construction. The predictions from regression models were made to census blocks to produce predicted vulnerability output corresponding to vulnerability levels in input training features. Explanatory variables were calculated from the distance feature and raster datasets.

Thirty percent of the training data was excluded from training the model for validation. After the model is trained, the validation data are used to predict the values of the test data; the predicted values are then compared to the observed values to provide a measure of prediction accuracy based on data that were not included in the training process.

The validation strategy employed a random 70/30 split, which may lead to optimistic performance estimates in spatial datasets due to spatial autocorrelation. In geographically structured data, nearby observations tend to share similar characteristics, potentially inflating predictive accuracy (Roberts et al., 2017).

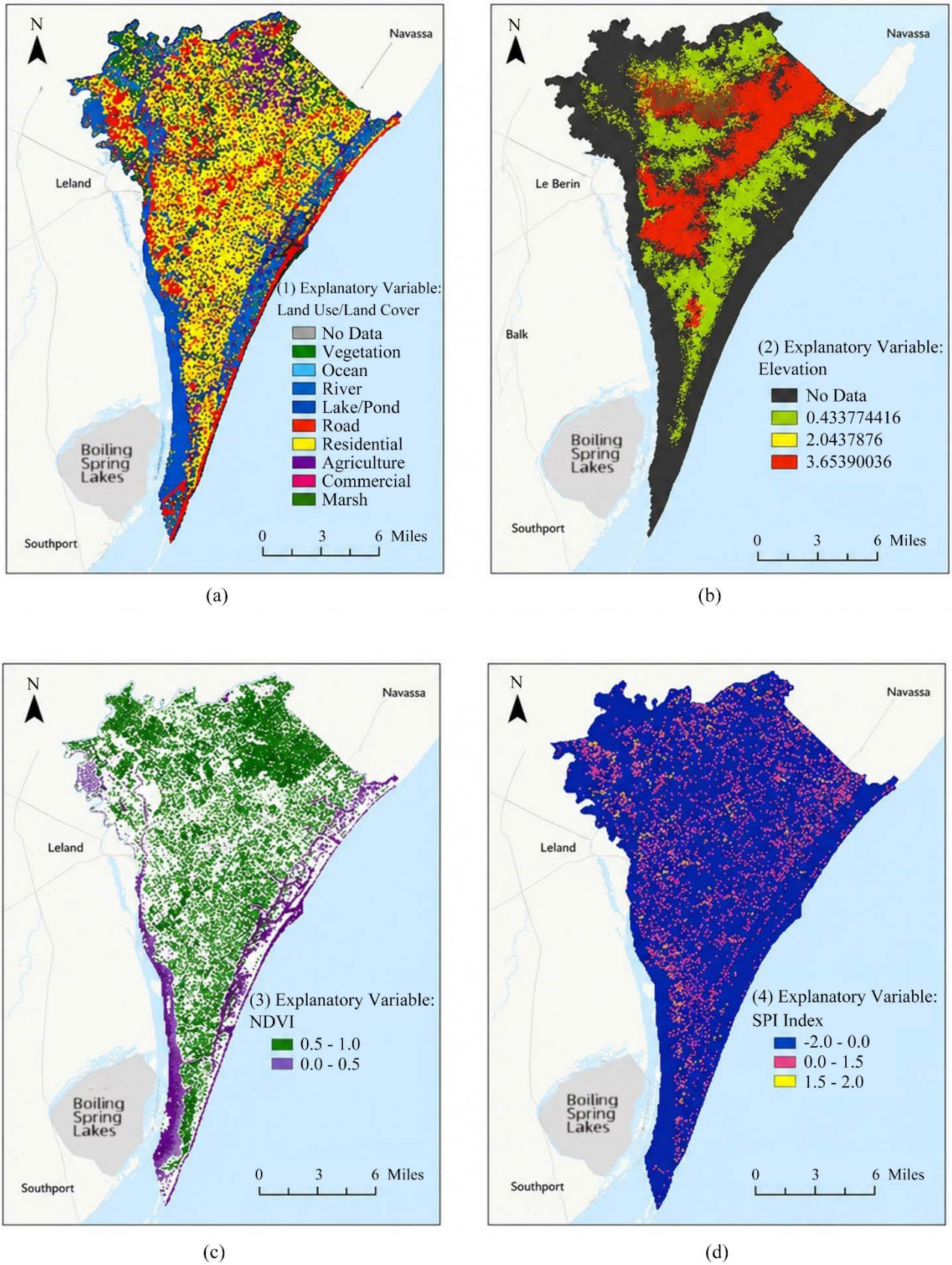

Figure 5. Explanatory geophysical variables: (a) land use/land cover, (b) elevation, (c) NDVI, and (d) SPI.

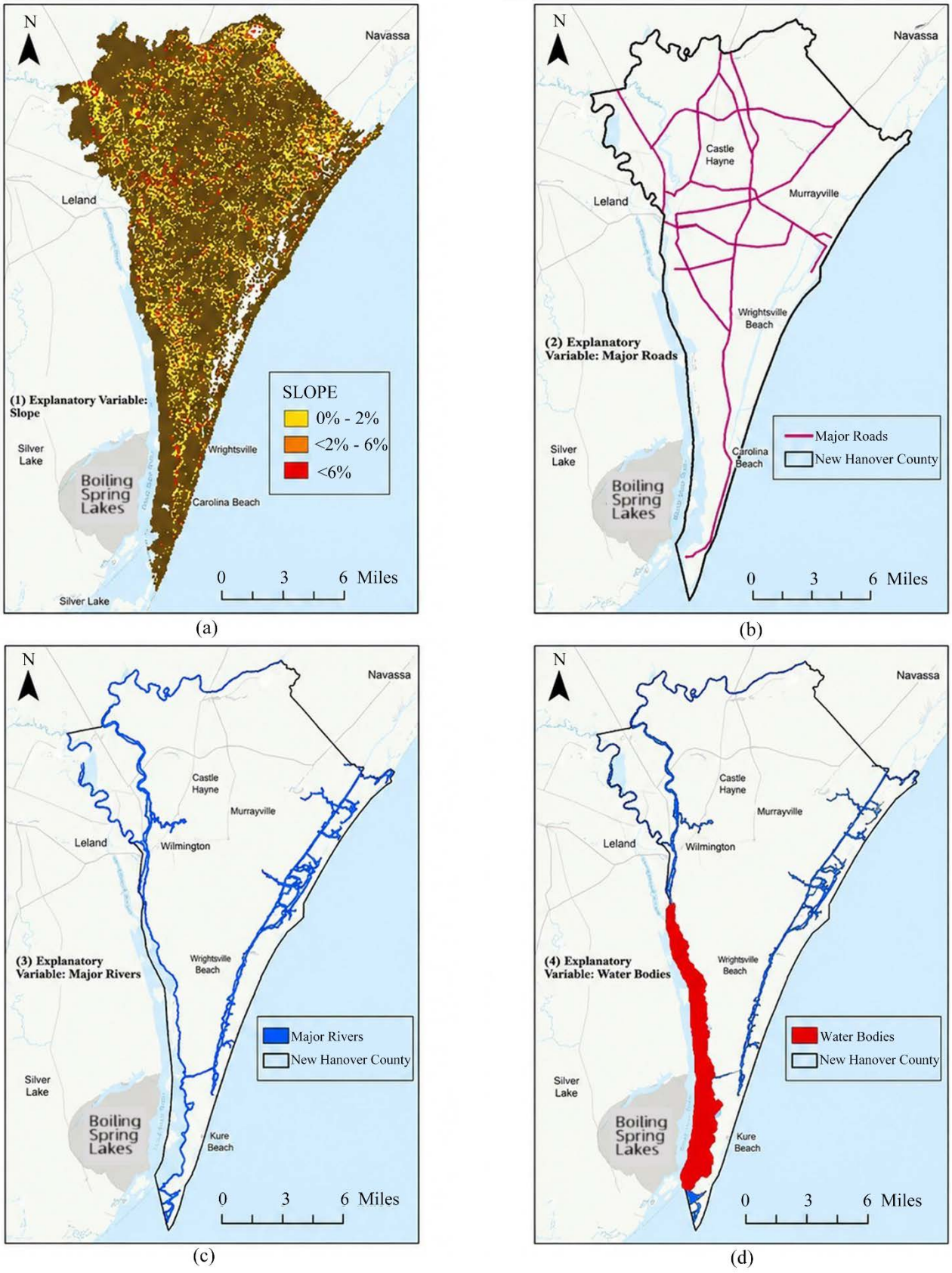

Figure 6. Explanatory variables: (a) slope, (b) major roads, (c) major rivers, and (d) water bodies.

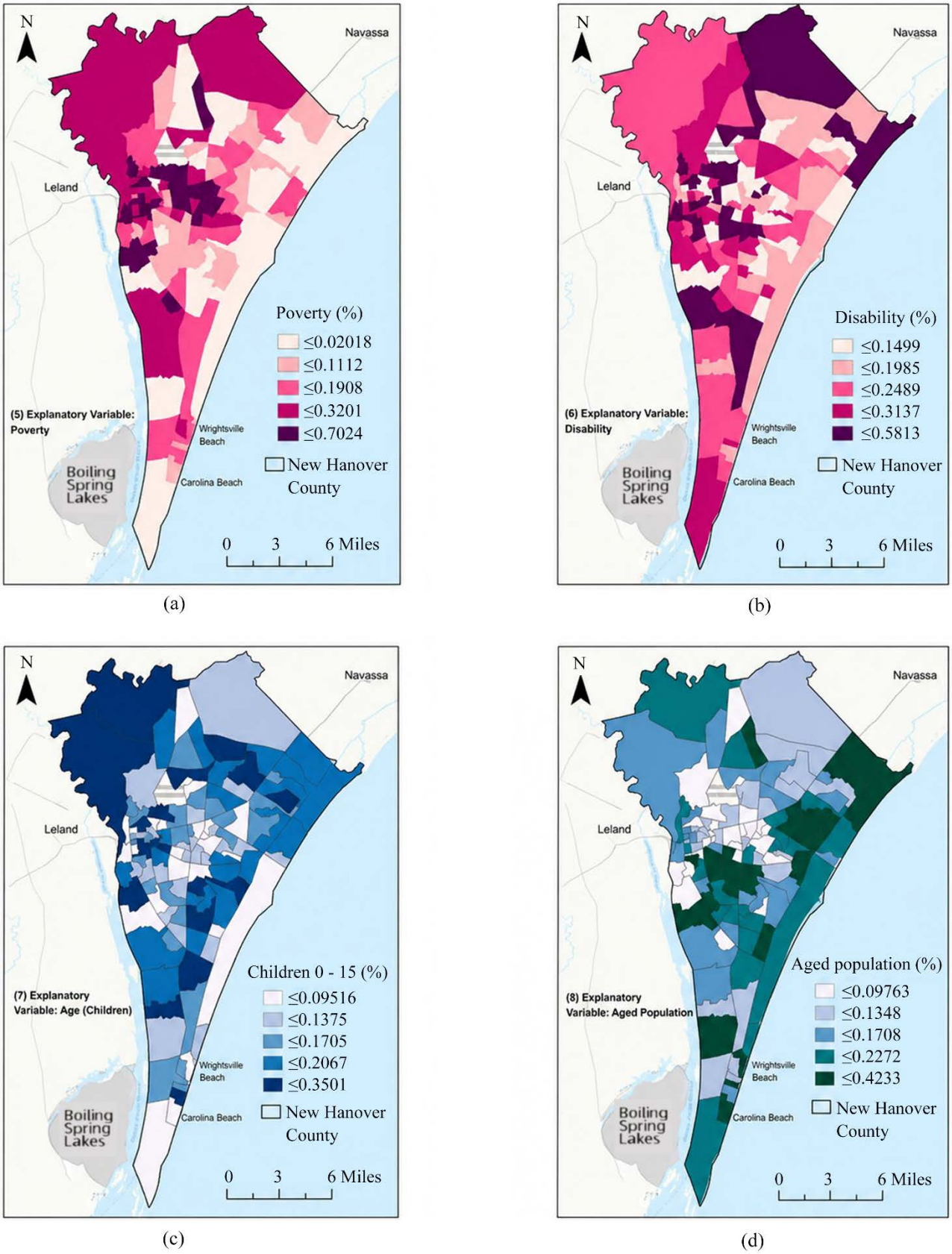

Figure 7. Explanatory demographic variables: (a) poverty, (b) disability, (c) children, and (d) aged population.

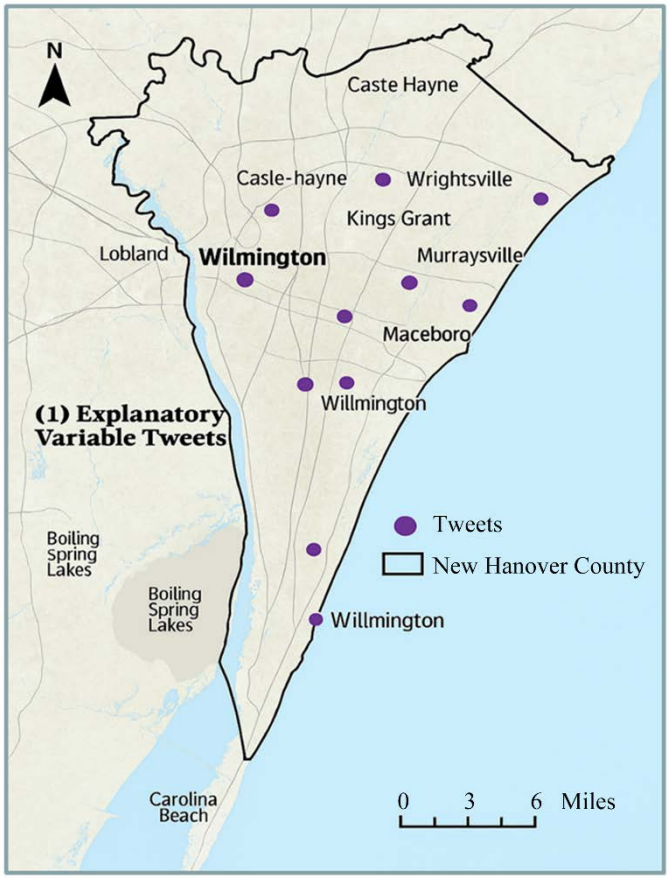

Figure 8. Explanatory variables: tweets.

A more robust approach would involve spatial cross-validation techniques, such as spatial blocking or leave-area-out validation, to better assess generalization across geographically distinct regions (Valavi et al., 2018). Due to data and computational constraints, this was not implemented in the current study; however, future work will incorporate spatial validation frameworks to provide more conservative and realistic performance estimates.

The leaf size parameter is the number of observations required to keep a terminal node without further splitting. The minimum leaf size parameter set for this regression model was 5, i.e., the tree stopped growing after it had a minimum observation of 5 at its terminal node. Tree depth means the depth of each tree in the tree. The tree depths in the forest range between 0 - 18 (this is data-driven), with having mean tree depth of five. The number of data points available to form per tree was set to 100%, and the number of randomly sampled variables for each tree was 3 (the square root of the total number of variables). The percentage of data excluded for validation for the regression model was 30.

Variable importance is a measure of how important a variable is in prediction. Complex interactions among the variables determine Random Forests. It is determined by looking at how much the prediction error increases when data for that variable is permuted while all others are left unchanged. The calculations are car-

ried out for each tree.

Mean decrease in accuracy measures to which a variable contributes to the mean decrease in accuracy of prediction during the OOB error calculation. The variables with a large mean decrease in accuracy are more important for classification. The more the accuracy of the random forest decreases due to the exclusion (permutation) of a single variable, the more important that variable is considered. The mean decrease in Gini measures how each variable contributes to the homogeneity of the nodes. Each time a particular variable is used to split a node, the Gini for the child nodes is calculated and compared to that of the original node. Variables that result in nodes with higher purity have a higher decrease in Gini.

Mean squared error (MSE) is the average squared difference between the predicted values and the actual values. This is a measure of the quality of the model. The values closer to zero are better. In this model, the MSE for the number of trees 1000 and 2000 are 1.728 and 1.729, respectively. While doubling the number of trees, the error decreased, but not significantly for the regression model.

The percent of variation explained is the determination of the degree of relationship in the patterns of variation, or how well the variation of one variable explains the variation of the other variable. The coefficient of determination, R^2, is a measure of the variation explained. The higher the value of R^2, the higher the predictive value of the regression. In this study, R^2 is 0.931 from the training and only 0.610 from the validation. The percent of variation explained may vary as the number of trees parameter is changed. In this analysis, the percent of variation explained is 1.728 and 1.719 for 1000 and 2000 trees respectively, a slight increase when the number of trees is doubled, indicating that predictive ability of the model increased as the number of tree parameter has increased from 1000 to 2000, but also it appears that it did not make a remarkable difference in the ability of model to predict. The difference between training ($R^2 = 0.931$) and validation performance ($R^2 = 0.610$) suggests moderate overfitting, which is common in spatial machine learning models with complex interactions. Despite this, the model retains acceptable predictive performance on unseen data, indicating reasonable generalization ability.

Since the outcome variable represents ordered categories, additional evaluation metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) would provide more interpretable measures of prediction error compared to R^2 alone (Hastie et al., 2009). Furthermore, if formulated as a classification problem, performance could be evaluated using per-class accuracy and confusion matrices to distinguish between minor and major misclassification errors (Powers, 2020).

In this study, R^2 and MSE are reported to maintain consistency with the regression framework; however, future extensions will incorporate these additional metrics to better align evaluation with the ordinal nature of vulnerability. *P*-value in regression analysis measures the relationship between the change in the predictor and response variables. Higher *P*-values mean the response variable is insignifi-

cant for prediction, and a value as low as (or lower than) <0.05 indicates a significant relationship with the predicted outcome. *P*-value in this analysis is zero (0), which means that the variables used are statistically significant, having a decent relationship with the predicted outcome.

Variable importance ranked in Table 2 and Figure 9 shows the contribution of each explanatory variable to predict the vulnerability situation in the study area from Hurricane Florence using a regression model. Tweets, roads, elevation, and NDVI have the highest contribution for predicting the vulnerable communities, whereas water body, land use/land cover, slope, demographic variables, and Stream Power Index (SPI) have a moderate contribution.

Table 2. Variable importance output from the RF regression.

Variable	Importance	%
TWEETS	92.30	18
ROADS	58.34	11
ELEV	53.66	10
NDVI	51.53	10
WTRBODY	44.64	9
RIVERS	43.56	9
LULC	34.23	7
SLOPE30	31.51	6
AGE	28.74	6
DISABILITY	27.26	5
POVERTY	23.72	5
SPI_INDX	21.71	4

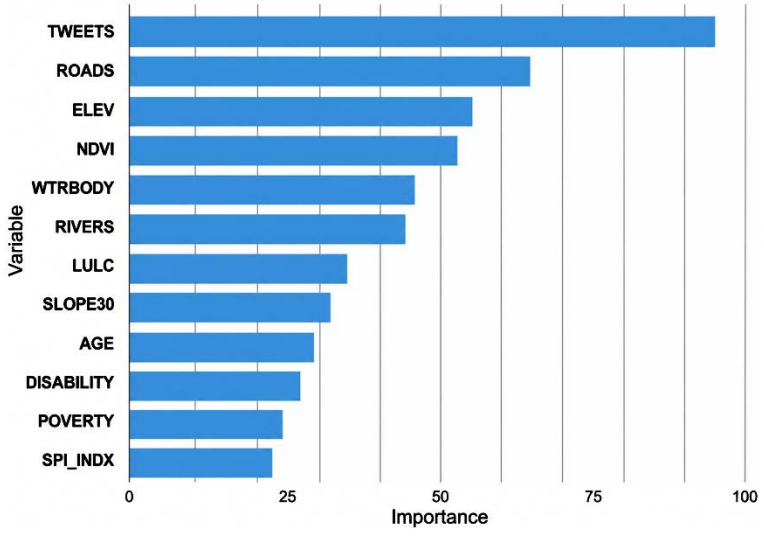

Figure 9. Summary of variable importance from the regression model.

The regression analysis predicted that approximately 10 percent of census blocks (482) would be classified as vulnerability level 1, approximately 47 percent (2311) as vulnerability level 2, and 31 percent (1538) as vulnerability level 3. The rest would be categorized as four or five (Figure 10).

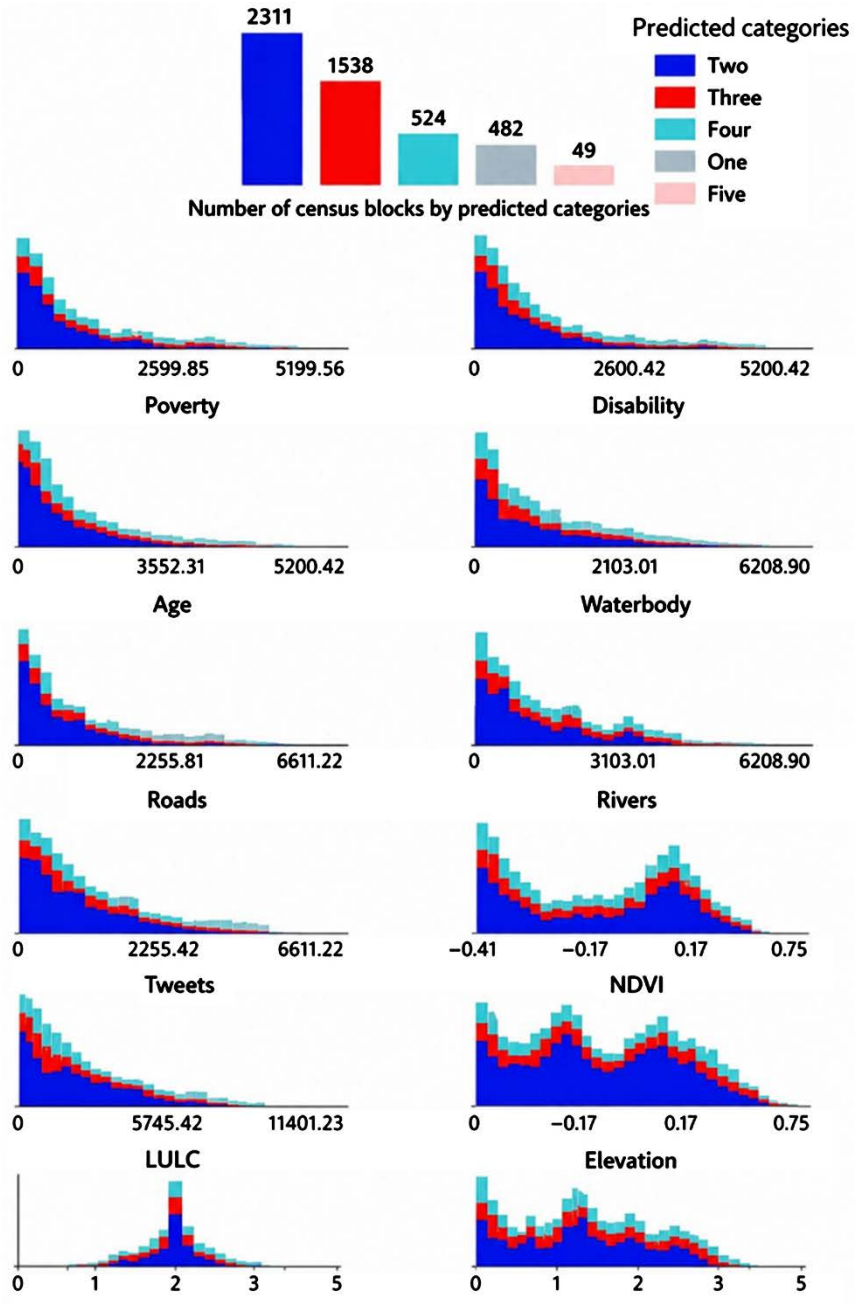

Figure 10. Predicted categories for different explanatory variables on census blocks from the RF regression model.

Figure 10 shows vulnerability categories by explanatory variables. This shows that about 10% of the communities had the highest level of vulnerability, nearly 47 percent of the communities corresponding to the census blocks had a high level

of vulnerability, about 31 percent of the communities were moderately vulnerable, and the rest, nearly 12 percent, had a low level of vulnerability to the risk associated with Hurricane Florence. It is found that the predicted map indicates that generally the areas around water bodies, including coast, rivers, and floodwaters, and lowland areas appear to have higher vulnerability compared to the areas away from them (Figure 11).

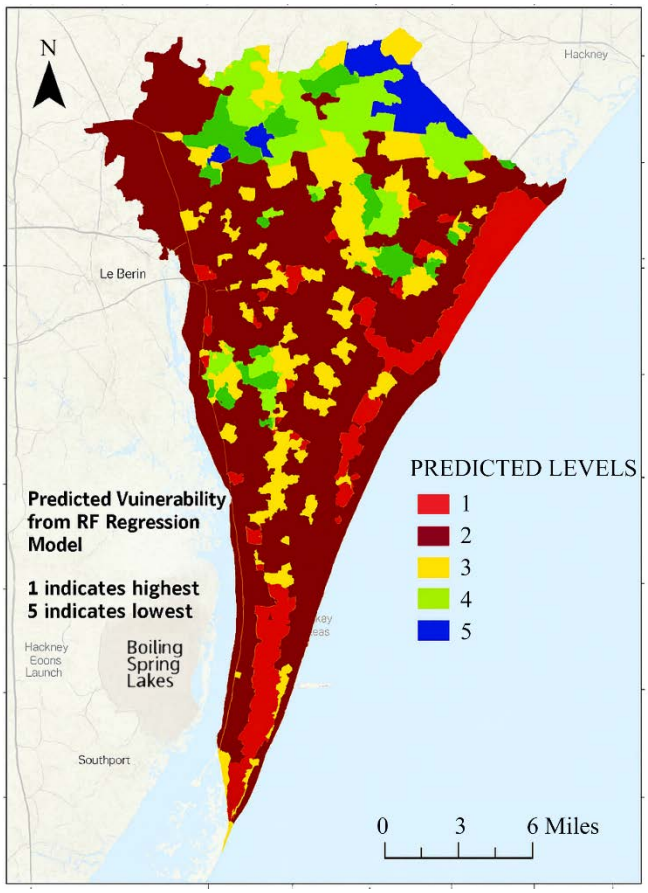

Figure 11. Predicted categories on census blocks from the RF regression model.

5. Limitations and Future Work

This study has several limitations. First, the training data were derived from crowdsourced and interpreted imagery, which introduces potential subjectivity in labeling. Second, the relatively small number of geotagged social media observations may limit the robustness of the "tweets" variable. Third, the use of a random validation split may overestimate predictive performance due to spatial autocorrelation.

Additionally, the regression framework approximates an ordinal outcome, which may not fully capture discrete category boundaries.

Future work will address these limitations by incorporating larger datasets, spatial cross-validation techniques, and alternative modeling approaches such as ordinal classification and deep learning frameworks.

6. Summary

Extreme hurricane events and the associated floods are in an increasing trend in the Atlantic Coast areas, making coastal communities more vulnerable. The growing population in the United States' coastal areas increases the risk of more loss of life and property damage. Hurricane Florence made landfall in New Hanover County, North Carolina, as a Category 1 storm, causing at least 24 billion worth of property damage and claiming dozens of human lives. The damage to property and loss of life were primarily caused by record-breaking heavy rainfall and flooding. The area of study for this work is a coastal county that comprises approximately 42% water area, which is one of the reasons why New Hanover County witnessed the most dangerous inundation flood, being isolated from the rest of the world for several days.

Random Forest regression modeling for geospatial predictive analysis of vulnerability to hurricane hazards was performed. Geophysical attributes were preferred over socio-demographic variables to carry out hurricane vulnerability modeling from machine learning. Given the lack of research on the use of a combination of variables that potentially could explain the vulnerability, this study demonstrates the value of integrating demographic and social media-generated variables alongside geophysical data to improve hurricane vulnerability prediction. The objectives are to make categorical predictions and map vulnerable communities using the Random Forests (RF) machine learning algorithm.

Disasters and thus vulnerability levels of communities demonstrate differences with the different demographic, socio-economic, and physical-environmental conditions of the place, i.e., exposure and coping ability. It is necessary to consider the coupled human-environment system when mapping vulnerability from natural hazards.

Among statistical, physical, and data-driven models used to predict natural hazards, data-driven methods are proven to be the most useful. A combination of geo-physical, demographic, and social media-generated data was used as explanatory variables for predicting vulnerability at the level of census blocks. Land use/land cover, water bodies, elevation, NDVI, Stream Power Index (SPI), slope, major roads, and major rivers are geo-physical variables; poverty, disability, and age are demographic variables; and tweets during hurricane events were social media-generated variables used to feed into the random forest regression model. Training data were collected from three different sources: 1) crowdsourced location features with photos from Instagram, Twitter, Facebook, and online news media during Hurricane Florence, 2) county-designated safe emergency shelter locations, and 3) imagery captured during the hurricane event. A total of 273 point locations were used as labelled feature data for model training. Census blocks were used as prediction polygon features since they represented areas with geophysical and demographic similarities.

The RF is an extensively used data modeling algorithm in natural hazard risk prediction, such as landslides and floods. However, the use of this modeling tech-

nique is very infrequent in hurricane vulnerability prediction. The RF is a supervised regression method of modeling by growing an ensemble of trees and letting them vote for the most occurring class as the predicted class. Trees grow based on bootstrap samples, and the "out-of-bag" error rate is calculated using samples out of the bootstrap samples for checking errors. Variable importance is an important output from the RF, and can be used to judge which variables are more useful than others to describe the impact of the disaster event.

For prediction by the RF regression, two thousand decision trees, three as a number of randomly sampled variables for constructing each tree, and 30 percent data were excluded for model validation. The MSEs for the numbers of trees 1000 and 2000 are 1.728 and 1.729, respectively. The regression model shows that while doubling the number of trees, the error decreased, but not significantly. Therefore, 2000 trees can be considered as an optimal number. However, the predictive ability does not appear to have increased remarkably by increasing the number of trees. Having an R^2 value of 0.931, *P*-value 0.000, and standard error 0.014 shows that the variables used are statistically significant, having good relationships with the predicted outcome. Even so, the R^2 value (0.610) appears lower than expected, and the standard error (0.048) appears higher for the predictions for the test data (excluded from model training) compared to the predictions for the data used to train the model. The variables, including tweets, roads, elevation, and NDVI, appear to have high importance for vulnerability prediction from hurricanes using a random forest regression model.

The RF model results show that geophysical and social media-generated variables have higher weight in terms of importance than demographic variables. The communities in the majority of census blocks have the highest level of vulnerability, whereas just around one-tenth of the communities are the least vulnerable in New Hanover County, North Carolina, from Hurricane Florence.

Conducting prediction analysis for vulnerability from hurricane risks using the RF algorithms, or other data-driven methods, for predicting the location of vulnerable communities is highly encouraged for future work. The prediction of community vulnerability could likely be performed at the building level for future work.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

Ai, F. F., Bin, J., Zhang, Z. M., Huang, J. H. et al. (2014). Application of Random Forests to Select Premium Quality Vegetable Oils by Their Fatty Acid Composition. *Food Chemistry, 143,* 472-478. https://doi.org/10.1016/j.foodchem.2013.08.013

Aubrecht, C., Özceylan, D., Steinnocher, K., & Freire, S. (2013). Multi-Level Geospatial Modeling of Human Exposure Patterns and Vulnerability Indicators. *Natural Hazards, 68,* 147-163. https://doi.org/10.1007/s11069-012-0389-9

Bathi, J. R., & Das, H. S. (2016). Vulnerability of Coastal Communities from Storm Surge

and Flood Disasters. *International Journal of Environmental Research and Public Health, 13,* Article 239. https://doi.org/10.3390/ijerph13020239

Bouwer, L. M. (2018). Observed and Projected Impacts from Extreme Weather Events: Implications for Loss and Damage. In *Loss and Damage from Climate Change: Concepts, Methods and Policy Options* (pp. 63-82). Springer.

Breiman, L. (2001). Random Forests. *Machine Learning, 45,* 5-32. https://doi.org/10.1023/a:1010933404324

Chen, C., Liaw, A., & Breiman, L. (2004). *Using Random Forest to Learn Imbalanced Data*. University of California.

Cohen, D. (2019). *About 60.2 Million Live in Areas Most Vulnerable to Hurricanes*. U.S. Census Bureau.

Cutter, S. L., Barnes, L., Berry, M., Burton, C., Evans, E., Tate, E. et al. (2008). A Place-Based Model for Understanding Community Resilience to Natural Disasters. *Global Environmental Change, 18,* 598-606. https://doi.org/10.1016/j.gloenvcha.2008.07.013

Feaster, T. D., Weaver, J. C., Gotvald, A. J., & Kolb, K. R. (2018). *Preliminary Peak Stage and Streamflow Data at Selected Streamgaging Stations in North Carolina and South Carolina for Flooding Following Hurricane Florence, September 2018. US Geological Survey Open-File Report*.

Goodchild, M. F., & Glennon, J. A. (2010). Crowdsourcing Geographic Information for Disaster Response: A Research Frontier. *International Journal of Digital Earth, 3,* 231-241. https://doi.org/10.1080/17538941003759255

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*. Springer.

Helderop, E., & Grubesic, T. H. (2019). Hurricane Storm Surge in Volusia County, Florida: Evidence of a Tipping Point for Infrastructure Damage. *Disasters, 43,* 157-180. https://doi.org/10.1111/disa.12296

Hoque, M. A. A., Phinn, S., Roelfsema, C., & Childs, I. (2017a). Tropical Cyclone Disaster Management Using Remote Sensing and Spatial Analysis: A Review. *International Journal of Disaster Risk Reduction, 22,* 345-354. https://doi.org/10.1016/j.ijdrr.2017.02.008

Hoque, M. A. A., Phinn, S., & Roelfsema, C. (2017b). A Systematic Review of Tropical Cyclone Disaster Management Research Using Remote Sensing and Spatial Analysis. *Ocean & Coastal Management, 146,* 109-120. https://doi.org/10.1016/j.ocecoaman.2017.07.001

Janitza, S., & Hornung, R. (2018). On the Overestimation of Random Forest's Out-of-Bag Error. *PLOS ONE, 13,* e0201904. https://doi.org/10.1371/journal.pone.0201904

Klonner, C., Marx, S., Usón, T., Porto de Albuquerque, J., & Höfle, B. (2016). Volunteered Geographic Information in Natural Hazard Analysis: A Systematic Literature Review of Current Approaches with a Focus on Preparedness and Mitigation. *ISPRS International Journal of Geo-Information, 5,* Article 103. https://doi.org/10.3390/ijgi5070103

Lee, K., Lee, H., Lee, K., & Shin, J. (2017). Training Confidence-Calibrated Classifiers for Detecting Out-of-Distribution Samples. arXiv preprint arXiv:1711.09325

Mosavi, A., Ozturk, P., & Chau, K. (2018). Flood Prediction Using Machine Learning Models: Literature Review. *Water, 10,* Article 1536. https://doi.org/10.3390/w10111536

Park, S., & Kim, J. (2019). Landslide Susceptibility Mapping Based on Random Forest and Boosted Regression Tree Models, and a Comparison of Their Performance. *Applied Sciences, 9,* Article 942. https://doi.org/10.3390/app9050942

Pourghasemi, H. R., & Kerle, N. (2016). Random Forests and Evidential Belief Function-

based Landslide Susceptibility Assessment in Western Mazandaran Province, Iran. *Environmental Earth Sciences*, **75**, Article Number 185. https://doi.org/10.1007/s12665-015-4950-1

Powers, D. M. (2020). *Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation*.

Richards, J. A., & Jia, X. (2006). *Remote Sensing Digital Image Analysis: An Introduction*. Springer.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G. et al. (2017). Cross-validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure. *Ecography, 40,* 913-929. https://doi.org/10.1111/ecog.02881

Rygel, L., O'sullivan, D., & Yarnal, B. (2006). A Method for Constructing a Social Vulnerability Index: An Application to Hurricane Storm Surges in a Developed Country. *Mitigation and Adaptation Strategies for Global Change, 11,* 741-764. https://doi.org/10.1007/s11027-006-0265-6

Stewart, S. R., & Berg, R. (2019). *Hurricane Florence* (*AL062018*): *31 August-17 September 2018*. National Hurricane Center Tropical Cyclone Report, National Oceanic and Atmospheric Administration (NOAA).

Valavi, R., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2018). *BlockCV: An R Package for Generating Spatially or Environmentally Separated Folds for K-Fold Cross-Validation of Species Distribution Models*.

Wang, Z., Lai, C., Chen, X., Yang, B., Zhao, S., & Bai, X. (2015). Flood Hazard Risk Assessment Model Based on Random Forest. *Journal of Hydrology, 527,* 1130-1141. https://doi.org/10.1016/j.jhydrol.2015.06.008

Wu, S., Yarnal, B., & Fisher, A. (2002). Vulnerability of Coastal Communities to Sea-Level Rise: A Case Study of Cape May County, New Jersey, USA. *Climate Research, 22,* 255-270. https://doi.org/10.3354/cr022255

Zhou, Z., Gong, J., & Hu, X. (2019). Community-Scale Multi-Level Post-Hurricane Damage Assessment of Residential Buildings Using Multi-Temporal Airborne Lidar Data. *Automation in Construction, 98,* 30-45. https://doi.org/10.1016/j.autcon.2018.10.018