Community flood vulnerability and risk assessment: An empirical predictive modeling approach

Yi (Victor) Wang | Antonia Sebastian

Department of Geological Sciences, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

Correspondence
Yi (Victor) Wang, Department of Geological Sciences, University of North Carolina at Chapel Hill, 104 South Road, Chapel Hill, NC 27514, USA.
Email: y.v.wang@unc.edu

Abstract
Effective assessment of flood vulnerability and risk is essential for communities to manage flood hazards. This paper presents an empirical modeling methodology to predict flood vulnerability and risk, considering factors of hazard distribution, property exposure, built environment, and socio-demographic and economic characteristics of a community. Vulnerability is empirically modeled as the expected fraction of property loss that is uninsured within a community (i.e., census tract) given water depth. Risk is derived as the expected annual uninsured property loss and loss ratio. The proposed framework is applied to the state of North Carolina in the United States. For model calibration, modeled flood loss data from Hurricanes Matthew in 2016 and Florence in 2018 and insurance claims data from the Federal Insurance and Mitigation Administration’s National Flood Insurance Program are used. The Federal Emergency Management Agency’s National Flood Hazard Layer is adopted, along with empirical probability distribution of water depth given flood event, to characterize hazard distribution. Results demonstrate how the presented methodology can be used to predict annual loss in terms of currency and to highlight hotspots of flood vulnerability and risk. Future work is needed to reduce uncertainty associated with limited hazard information available to the public.

KEYWORDS
built environment, disaster risk reduction, National Flood Hazard Layer, National Flood Insurance Program, North Carolina, social vulnerability, United States, zero-inflated modeling

1 | INTRODUCTION

Flooding is the most common and costly natural hazard globally (Guha-Sapir, 2020). To manage flood hazards effectively, a growing body of academic research has emerged aimed at improving the assessment of flood vulnerability and risk for communities around the world (e.g., Aerts et al., 2018; Cabrera & Lee, 2020; Das et al., 2020; de Moel et al., 2015; Feloni et al., 2020; Forero-Ortiz et al., 2020; Geng et al., 2020; Glas et al., 2020; Greço et al., 2020; Jun et al., 2020; Lin et al., 2020; Luu et al., 2020; Najafi et al., 2021; Pham et al., 2021; Sarkar & Mondal, 2020; Su, 2020; Tanoue et al., 2016; Tsakiris, 2014; Ward et al., 2020). Hydrologically oriented studies usually
emphasize estimation of extent of inundation, return periods of flood events, and potential physical exposures (e.g., Gori et al., 2020; Lindenschmidt et al., 2016; Mojaddadi et al., 2017; Ward et al., 2015; Zaharia et al., 2017). Some flood risk modelers develop or apply engineering approaches by using structural or system vulnerability curves in terms of damage or loss ratio as a function of flood intensity measure (e.g., Budiyono et al., 2016; D’Ayala et al., 2020; Karagiorgos et al., 2016; Nofal & van de Lindt, 2020; Thapa et al., 2020; Tincu et al., 2020). Other scholars with interests in urban planning apply analytical methods to investigate the underlying physical drivers of loss, such as land use/land cover (e.g., Brody et al., 2014; Brody et al., 2018).

Given the progress of studies on flood vulnerability and risk assessment, hazard scientists have further proposed to integrate social vulnerability factors into the framework for flood risk assessment and management (Cutter et al., 2013). Following this proposal, more scholarly efforts have been dedicated to quantification of flood risk accounting for social vulnerability manifested by communities’ sociodemographic and economic conditions (Rufat et al., 2015; Terti et al., 2015). Most of these social vulnerability-based works have adopted a traditional indexing approach to quantifying flood risk by assigning weights to vulnerability factors and aggregating these weights based on existing literature and criteria designed by the modelers (e.g., Allen et al., 2016; Hamidi et al., 2020; Pathak et al., 2020; Salazar-Briones et al., 2020; Tate et al., 2021; Vojinovic et al., 2016). Some researchers have implemented advanced data aggregation techniques to derive risk metrics after receiving feedbacks from invited stakeholders on perceived weights of considered variables (e.g., Hazarika et al., 2018). However, there is a lack of literature on assessing flood vulnerability and risk through an empirical predictive modeling approach based on records of flood losses to integrate social vulnerability factors into risk assessment framework.

In this paper, the authors introduce an empirical predictive modeling approach to assessing flood vulnerability and risk for communities by integrating information on hazard distribution, property exposure, and environmental and social vulnerability factors. This research contributes to the field of flood risk management in four aspects. First, we adopt a zero-inflated modeling technique to consider the effects of data points with zero losses. Second, we use the newly released National Flood Insurance Program (NFIP) redacted claims dataset from the Federal Insurance and Mitigation Administration (FIMA) to derive records of uninsured property losses and loss ratios for model calibration (Federal Emergency Management Agency (FEMA), 2020a). Third, we implement an empirical modeling methodology to predict expected adverse impacts and impact ratios for census areas. Fourth, we integrate information on flood zone, exposure value, built environmental characteristics, and social vulnerability factors to assess flood risk with one coherent framework. The presented methodology allows us to make predictions of expected annual uninsured property losses in currency and loss ratios per exposure as metrics of flood risk.

To demonstrate its utility, we apply the empirical methodology to derive geographic distributions of flood vulnerability and risk within the state of North Carolina (NC) of the United States. NC is one of the most flood-prone states in the United States and is prone to all major types of flood hazards including fluvial, pluvial, and coastal floods (North Carolina Flood Insurance (NCFI), 2020). We use the hindcast data on flood inundation and estimated building damage at 213,394 properties due to Hurricanes Matthew in 2016 and Florence in 2018 along with the NFIP claims data at census tracts. The modeling products show hotspots of flood vulnerability and risk in NC and highlight both the mountain and coastal plain areas as more flood-prone than the region of Piedmont plateau. Results also suggest that exposure of properties in flood zones is more significant a factor than community vulnerability in contributing to flood risk distribution. These products and findings, providing a picture of flood risk across the state, have the potential to be useful for emergency managers, decision makers, government officials, engineers, planners, and developers to facilitate flood mitigation and increase community flood resilience. The presented methodology can also be adopted for flood vulnerability and risk assessment for other states and the entire United States.

In the following sections, we first provide rationale for presenting our empirical methodology by reviewing existing modeling works on flood risk assessment. Next, we present our proposed approach for flood risk assessment, including formulations of risk, hazard, and vulnerability as well as the model selection process for vulnerability model. We then showcase our study area of NC. After that, we introduce the source and processing of data required for model calibration. The following section displays the outcomes of model calibrations and animations and maps showing geographic distributions of flood hazard, exposure, vulnerability, and risk. In the penultimate section, we outline the innovations of the proposed approach over existing ones and the potential challenges and limitations associated with the application of presented methodology that practitioners should consider. Finally, we conclude the paper by suggesting future directions to extend the proposed work on flood risk assessment.

2 | RATIONALE

Flood risk assessment involves integrative modeling of hazard, exposure, and vulnerability (Allen et al., 2016; de Moel et al., 2015; Foudi et al., 2015). In this paper,
refers to the geographic and frequency distribution of intensity measure of a potential flood event experienced at locations, where intensity measure is the metric indicating the expected adverse impact of the hazard event given a unit exposure of an entity with an average vulnerability. Exposure is the amount of values of the entity within an area with or without considering hazard conditions. Vulnerability, modeled as an expected loss rate given a spectrum of intensity measure, is the susceptibility of the entity to loss due to a flood event. Accordingly, risk, indicating an expected future loss or loss ratio, is a function of hazard, exposure, and vulnerability.

2.1 | Hazard and exposure

Hazard distribution and exposure of population and properties are the most basic aspects of flood risk assessment. As information on exposure is usually available and accessible for most parts of the world, many flood modelers have been focused on physical processes of flood hazard to create maps showing hazard distributions and subsequently how communities are exposed to flood hazards. For example, Mojaddadi et al. (2017) implemented an ensemble supervised machine learning technique based on support vector machine, geospatial approaches, and remote sensing data to estimate flood probability for risk assessment for the Damansara River catchment in Malaysia. Zaharia et al. (2017) adopted an analytical hierarchy process-based methodology to develop flood potential indices to characterize flood hazard for risk management in the Prahova catchment in Romania. More recently, Couasnon et al. (2018) implemented a copula-based Bayesian network approach to derive hazard distribution for compound hazard situations considering both fluvial and coastal floods for the catchment of Houston Ship Channel in Texas.

Despite these successful efforts at the catchment level, flood hazard mapping for large areas involving multiple catchments is rather difficult (Ward et al., 2020). There is only a limited number of products of hazard modeling for exposure analysis and further vulnerability and risk assessment for large areas such as a state. For the United States, the National Flood Hazard Layer (NFHL) from FEMA (2020b) is the only product of flood hazard modeling covering states and the entire country available to grassroots researchers and the public.

2.2 | Vulnerability curve

With information on hazard and exposure, flood risk assessment can be achieved via an engineering approach through incorporating flood vulnerability curves indicating losses, loss ratios, or probabilities of losses, as a function of flood intensity measure, where water depth is usually adopted as the intensity measure for flood vulnerability and risk assessment (Budiyono et al., 2015; Custer & Nishijima, 2015; De Risi et al., 2013; Foudi et al., 2015; Karagiorgos et al., 2016). When a probability of loss or failure of a structural or infrastructural system is computed with respect to intensity measure or the exceedance value of intensity measure, the probability curve may also be called a fragility curve (e.g., Anarde et al., 2018; Custer & Nishijima, 2015). Vulnerability curves may be created to estimate economic loss due to future flood events for administrative areas (e.g., Feyen et al., 2012). However, such works for administrative areas usually do not model the vulnerability curves based on integration of specific local built environmental or social vulnerability factors.

More often, modelers derive vulnerability curves for individual buildings, infrastructures, or system elements to assess flood vulnerability and risk by only considering built environmental vulnerability (Budiyono et al., 2015; Lindenschmidt et al., 2016). For example, Arrighi et al. (2013) implemented an analytical modeling method to create vulnerability curves for estimating flood risk in terms of structural monetary damage in the St. Croce district in Florence, Italy. Custer and Nishijima (2015) adopted an engineering approach to model structural fragility, considering damage processes due to water infiltration, as a function of water depth for a two-story masonry building. Foudi et al. (2015) used ordinary least squares method to estimate loss as a function of water depth for urban elements in Zaragoza, Spain. On another occasion, Karagiorgos et al. (2016) adopted a Weibull distribution to model loss ratios of buildings in Greece with respect to a relative water depth. Despite their attempt to consider both structural and social vulnerabilities, Karagiorgos et al. (2016) used a separate approach for treating social vulnerability and thus did not integrate social vulnerability factors into their vulnerability curve model. Without jointly considering both built environmental and social vulnerability factors, the derived vulnerability curves for flood risk assessment may not be capable of fully revealing the actual flood vulnerability distributions.

2.3 | Vulnerability indicators

Flood vulnerability is usually modeled as associated with both built environment and social conditions of communities exposed to flood hazard (Brody et al., 2012; Cutter et al., 2013; Karagiorgos et al., 2016). On the built environment side, land cover and land use (Brody et al., 2014), physical coastal characteristics (Martínez-Graña et al., 2016), residential location (Brody et al., 2018), and settlement elevation (Hazarika et al., 2018) have been identified as important drivers of flood loss and risk.
On the social vulnerability side, through a meta-analysis of 67 flood disaster case studies from 1997 to 2013, Rufat et al. (2015) summarized demographic, socioeconomic, and health metrics as the main indicators of social vulnerability to flood events. Among the existing works on social vulnerability, most tend to apply an indexing approach. For example, Bathi and Das (2016) used an equal weighting scheme to incorporate 11 socioeconomic vulnerability indicators to map flood vulnerability for three coastal counties in Mississippi. Equal weights were also adopted by Allen et al. (2016) to estimate risk of glacial lake outburst flood in Himachal Pradesh, India. In addition to equal weighting, other variable aggregation methods such as principal component analysis (PCA) have also been proposed for identifying and weighting social indicators as proxies of social vulnerability factors (e.g., Cutter et al., 2013). Beyond built environmental and social vulnerability factors, some researchers have also attempted to consider all possible indicators for risk assessment. As an example, Vojinovic et al. (2016) implemented a holistic approach to include risk perception factors along with indicators of hazard condition, built environment, and social vulnerability for flood risk assessment in areas with cultural heritage in Ayutthaya, Thailand.

Despite the abundance of quantitative models to assess flood risk accounting for hazard, exposure, and vulnerability factors, there is still a lack of effort to take into consideration a large amount of variables with an empirical predictive modeling approach based on evidence from data on losses due to historical flood events. In this regard, our proposed methodology and its application in NC demonstrates an advancement in flood vulnerability and risk assessment across large spatial scales using open-source data.

## 3 | METHODOLOGY

### 3.1 | Risk function

In general, the empirical predictive modeling approach to quantifying vulnerability and risk for natural hazards describes risk for a community occupying an area as (Wang et al., 2021),

\[
R_i = f(H_i, Ex_i, V_i),
\]

where \(i\) refers to the \(i\)th community, \(f(\cdot)\) is the risk function, and \(R_i, H_i, Ex_i,\) and \(V_i\) are, respectively, risk, hazard potential, exposure, and vulnerability. Considering \(R_{i,k}\) as an expected annual loss due to future flood events at the \(k\)th location of the \(i\)th community, we model \(R_{i,k}\) for flood hazard as

\[
R_{i,k} = \int_{-\infty}^{\infty} H_{i,k}(im)V_{i,k}(im)dim,
\]

where \(im\) is the intensity measure for flood, or water depth in this paper, and both \(H_{i,k}\) and \(V_{i,k}\) are functions of \(im\). For a community occupying a set of locations, its risk can be aggregated as

\[
R_i = \sum_k R_{i,k},
\]

and

\[
R_{Ui} = \frac{\sum_k R_{i,k}}{\sum_k Ex_{i,k}},
\]

where \(R_{Ui}\) is the risk considering unit exposure. In this study, we use the expected annual uninsured property loss, or expected total loss of property values subtracted by insurance payout, as metric of \(R_i\).

### 3.2 | Hazard quantification

\(H_{i,k}\) indicates the annual frequency distribution of intensity measure at the location. In this paper, it is modeled as

\[
H_{i,k} = \frac{g(im|Flood)}{RP_{i,k}},
\]

where \(g(im|Flood)\) is the probability distribution of \(im\) given a flood event and \(RP_{i,k}\) is the return period of flood events for the location. Here, we assume \(g(\cdot)\) to be universal for all locations within the study area of interest. In this study, \(g(\cdot)\) is empirically derived based on historical records of water depth due to flood events. \(RP_{i,k}\) can be obtained from flood zone maps, where flood zones are usually demarcated according to flood return period levels such as 100 or 500 year.

### 3.3 | Vulnerability formulation

\(V_{i,k}\) is vulnerability, indicating the expected loss ratio with respect to an intensity measure. Because data are aggregated at the community level in this study, the \(V_{i,k}\)’s within the same \(i\)th community are identical. From the empirical data end, \(V_{i,k}\) associated with a flood event can be computed as

\[
V_{i,k} = \frac{PL_i - IP_i}{Ex_i},
\]

where \(PL_i\) and \(IP_i\) are property loss and insurance payout, respectively.
Because loss rate can be zero, we apply a zero-inflated technique to model \( V_i \) as (also see Wang et al., 2020),

\[
V_i = VB(X_i, \beta_B) \cdot VL(X_i, \beta_L, \sigma),
\]

where \( VB(\cdot) \) is a logistic binomial regression model predicting the expected probability of experiencing loss, \( VL(\cdot) \) is a logistically transformed linear regression model giving the expected loss rate given observation of loss, \( X_i \) is the row vector of data matrix \( X \) for the \( i \)th data point corresponding to the \( i \)th community, \( \beta_B \) and \( \beta_L \) are, respectively, the column vectors of coefficients of columns of \( X \) of the \( VB(\cdot) \) and \( VL(\cdot) \) models, and \( \sigma \) is the dispersion parameter of \( VL(\cdot) \) model. Data matrix \( X \) include intercept, intensity measure variable of flood water depth, and built environmental and social vulnerability variables.

\( VB(\cdot) \) and \( VL(\cdot) \) models can be, respectively, further written as

\[
VB(X_i, \beta_B) = \frac{\exp(X_i \beta_B)}{\exp(X_i \beta_B) + 1},
\]

and

\[
VL(X_i, \beta_L, \sigma) = \frac{\exp(X_i \beta_L + \sigma \varepsilon)}{\exp(X_i \beta_L + \sigma \varepsilon) + 1},
\]

where \( \varepsilon \) is an independent standard normal random variable.

### 3.4 Model selection

To find the most desired set of vulnerability variables with non-zero coefficients of \( \beta_B \) and \( \beta_L \) in the binomial and linear regression models, respectively, we conduct a two-stage model selection process (also see Wang et al., 2019). For both binomial and linear regression models, as shown in Figure 1, we select individual vulnerability variables in the first stage for the second-stage model selection, while the second stage selects the best model with the most desired set of variables as the final model.

During the first stage, we examine all models with one vulnerability variable. We rank these models based on their Akaike information criterion (AIC) scores. An AIC indicates the relative quality of considered model in terms of prediction performance and model parsimony (Akaike, 1974). The formula of AIC is

\[
AIC_h = 2m_h - 2l(\theta_h),
\]

where \( m_h \) is the number of parameters in the \( h \)th model and \( l(\theta_h) \) denotes the maximum log-likelihood evaluated at the maximum likelihood estimator of the considered model with parameters \( \theta_h \). We then select 18 individual vulnerability variables according to the ranking of AIC scores of their corresponding first-stage models, also considering the criterion that a newly selected variable needs to have a Kendall’s tau with any one of the already selected variables less than 0.6.

In the second stage, we calibrate all possible models with intensity measure plus at most eight vulnerability variables that have been selected during the first stage. We then rank these second-stage models based on their AIC scores and select the best model with the lowest AIC score. For model calibration, we use the maximum likelihood approach with Raphson (1697) algorithm. The uncertainty metrics including confidence interval (CI) and prediction interval (PI) for vulnerability model are derived analytically (see Wang et al., 2019, 2020). To apply our methodology for flood vulnerability and risk assessment, we use the open-source software language Python version 3.7.6 (Python Software Foundation (PSF), 2020).

### 4 STUDY AREA

To demonstrate the utility of the proposed methodology, we apply it to assess flood vulnerability and risk for census tracts of NC. NC is geographically located between latitudes of 33°N and 37°N and between longitudes of 75°W and 85°W (Figure 2). The state is traditionally divided based on geologic characteristics into three regions, that is, the Appalachian Mountains in the west, the Piedmont plateau in the center, and the Atlantic coastal plains in the east (Figure 2). NC is prone to various types of floods (NCFI, 2020), particularly floods caused by tropical cyclones and heavy precipitation. Recent events of record include Hurricanes Matthew (2016) and Florence (2018). Matthew caused catastrophic flooding...
that killed 28 people and led to a property damage of 1.6 billion USD in NC (Price, 2017; Centers for Disease Control and Prevention (CDC), 2021). Florence resulted in 42 fatalities and an estimated damage of 16.7 billion USD in the state (National Weather Service (NWS), 2021). Among these losses by Hurricane Florence, the insured losses were only between 2.8 and 5 billion USD (Reuters, 2018), indicating the importance of examining and modeling flood vulnerability in terms of uninsured property loss. Figure 2 shows the geographic distribution of census tracts with buildings affected by floods associated with Hurricanes Matthew and Florence, with data provided by North Carolina Division of Emergency Management (NCDEM, 2020).

5 | DATA

The data used for this study is based on information on geographic distributions of flood hazard potentials, property values, water inundations and damages due to Hurricanes Matthew and Florence, flood insurance claims and payouts associated with the two hurricanes, built environmental attributes, and pre-event socio-demographic and economic indicators in NC. To illustrate flood hazard potential in terms of return period as $R_{pk}$ in Equation (5), we downloaded geodatabase files of NFHL for NC from FEMA (2020b) website and converted the polygons into three types, that is, 100-year flood zone, 500-year flood zone outside 100-year flood zone, and area outside 500-year flood zone. Details of classification of flood zones are listed in Table 1. We collected tax-assessed data on property values and built environmental attributes at the building parcel level from the North Carolina Floodplain Mapping Program (NCFMP, 2020). Data on flood intensity measures and building damages due to the two hurricanes were provided by the NCDEM (2020) at the parcel level. Intensity measure data from NCDEM were initially derived with implementation of the Sea, Lake, and Overland Surges from Hurricanes model for storm surge (National Oceanic and Atmospheric Administration (NOAA), 2020) and the Rapid Inundation Flood Tool of the Pacific Northwest National Laboratory for fluvial and pluvial flood (GeoPlatform.gov, 2020). Building damages were computed with the depth-damage functions of the United States Army Corps of Engineers (USACE, 1996). We also used the intensity measure data to compute the empirical probability distribution of water depth for $g(\cdot)$ function in Equation (5). To derive uninsured property losses, we downloaded the newly released FIMA NFIP redacted claims dataset (see, e.g., Wing et al., 2020) from FEMA (2020a) website. We also collected census data for years of 2015 (pre-Matthew), 2017 (pre-Florence), and 2018 (for future prediction) from the United States Census Bureau (USCB, 2020) to construct social vulnerability indicators.
For processing data, the authors used Python 3.7.6 (PSF, 2020) and proprietary software ArcMap version 10.7.1 of Esri (2020). Because the NFIP claims data only contained information on census tracts as the highest resolution for identifying building parcels, we aggregated building parcel data to census tracts for calibrating vulnerability model. For built environmental and social vulnerability indicators, we selected or derived means, medians, ratios, and indices based on the initial variables to ensure that the meaning and scale of each indicator were consistent for census tracts of all sizes. Indicators with 30 or more missing entries for any year were discarded. We also removed ratio indicators with 20% or more entries being zero as well as ratio indicators based on the same criterion for ones to preserve the representativeness of the remaining indicators. Logarithmic transformation and standard normalization were then conducted to keep the indicators with mean zero, standard deviation of one, and a range of $(-\infty, \infty)$. The final dataset for calibrating vulnerability model contains 1958 data points, each representing a census tract that experienced a zero or nonzero loss due to Hurricane Matthew or Florence. Within the dataset, there are 35 built environmental variables and 365 social vulnerability variables, as listed in Table S1, Supporting Information.

6 | RESULTS

6.1 | Hazard and exposure

Geographic distribution of flood hazard potential is depicted in Figure 3. We assumed locations within the 100-year flood zone to have an annual frequency of 1% and locations outside the 100-year but within the 500-year flood zone to have an annual frequency of 0.2%. We also simplified the risk assessment by assuming areas outside the 500-year flood zone with zero probability of flood. In addition to flood zone, we assumed that each location shared the same probability distribution of flood water depth (Figure 4), given a flood event. The empirical probability distribution of flood water depth in Figure 4 was computed with records of water depth at the building parcel level caused by Hurricane Matthew or Florence based on data from NCDEM (2020). The return periods and empirical probabilities correspond, respectively, to $R_{i,k}$ and $g(\cdot)$ in Equation (5).

Figure 5 illustrates the geographic distribution of exposure in terms of building value at the census tract level without considering hazard potential. When jointly considering flood hazard potential, we can plot the geographic distributions of building values within the flood zone (Figure 6) and their ratios to the total building values within census tracts (Figure 7). These maps of hazard plus exposure show that the coastal plains in the east and the mountain counties in the western portion of the state tend to have high exposure of property values within flood zones than counties in the Piedmont plateau.

6.2 | Vulnerability

The result of model selection is a final vulnerability model containing nine independent variables with nonzero coefficients for logistic binomial regression model and six for linear regression model, as listed in Table 2.
All listed independent variables in the final model have their estimated coefficients with a $p$-value less than 0.05, indicating statistical significance. Here, the other types of income of variable $S205$ include all types of income other than earnings; wage or salary income; self-employment income; interests, dividends, or net rental income; social security income; supplemental security income; public assistance income; and retirement income. Table 3 lists statistics of parameters of the final vulnerability model, where $\beta_{BInt}, \beta_{B4}, \beta_{B9}, \beta_{B25}, \beta_{B30}, \beta_{S13}, \beta_{S205},$ and $\beta_{S280}$ are, respectively, coefficients of intercept, $I0$, $B4$, $B25$, $B28$, $B30$, $S13$, $S205$, and $S280$ of binomial model and $\beta_{LInt}, \beta_{L3}, \beta_{LBS}, \beta_{LB25}, \beta_{LS185}, \beta_{LS280},$ and $\sigma$ are, respectively, coefficients of intercept, $I0$, $B3$, $B5$, $B28$, $S185$, and $S329$ and the dispersion parameter of linear model.

Because the independent variables are standard normalized, the estimated coefficients correspond to the correlation size of the variables to vulnerability. For the binomial model that estimates the expected probability of experiencing uninsured property loss, the ratio of concrete buildings ($B25$) seems to be the most dominating factor with a negative correlation. Meanwhile, water depth ($I0$), mean age of building ($B4$), median of first floor elevation ($B9$), ratio of steel buildings ($B28$), and ratio of buildings in the A zone ($B30$) are positively correlated with the expected loss probability, where A zone is within a 100-year flood zone as shown in Table 1. The social vulnerability variables of ratio of black alone population ($S13$), ratio of households with other types of income ($S205$), and ratio of females employed in education, legal, community service, arts, and media occupations ($S280$) have a negative correlation with the expected loss probability.

For the linear model that predicts the expected uninsured property loss rate given observation of loss, the most dominant factor is water depth ($I0$). Apart from median of building value ($B3$), all other independent variables of linear model, including median of building age ($B5$), ratio of steel buildings ($B28$), ratio of households with medium-low income ($S185$), and ratio of renter-occupied housing units ($S329$), are positively correlated with the expected loss rate. Because no scientific methods such as randomized controlled trials have been or may be applied for this study, these statistical significances shall not be interpreted as revelations of causalities. However, the identified independent variables in the final vulnerability model are interesting. In particular, the directions of correlation of $B9$ and $S13$ with flood vulnerability are counterintuitive and should be further investigated in future work. In addition, many of the other variables

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure3.png}
\caption{Hazard distribution in terms of flood zone}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure4.png}
\caption{Empirical probability distribution of flood water depth given a flood event}
\end{figure}
examined during the first-stage model selection but not included in the final model may also convey important information regarding flood vulnerability. The result of first-stage model selection for logistic binomial and linear regressions are listed respectively in Tables S2 and S3.

The performance of calibrated vulnerability model was evaluated via leave-one-out cross-validations (Rafaeilzadeh et al., 2009), as shown in Figure 8. Ideally, data points would line up along the dashed one-to-one lines. Data points with zero observed loss are relocated along the vertical axes. The size of a data point corresponds to its expected probability of experiencing loss. Comparison between predicted uninsured property loss rates and observations (Figure 8a) shows that the calibrated vulnerability model captures variations of data points with vulnerability indicators. Also considering property values exposed to flood events, Figure 8b compares the predicted uninsured property losses with the observed losses.
With the estimated model parameters, we can plot the vulnerability curve showing the uninsured property loss rate with respect to a spectrum of flood water depth for any census tract. An example is shown in Figure 9 for census tract 9305.01 in Beaufort County, where the circles represent the actual observations of uninsured property losses due to Hurricanes Matthew and Florence, respectively. Although it may be depicted as one solid curve for each census tract, vulnerability is a function of multiple variables. Each plotted vulnerability curve with its CI and PI is a projection of one hyperplane with its uncertainties from a multidimensional hyperspace.

The estimates of vulnerability curves for census tracts can also be plotted as a series of choropleth maps. Video 1 shows the animation of geographic distribution of flood vulnerability of census tracts with respect to a water depth increased from $C_0$ to 10 m. Using the empirical probability distribution of water depth given a flood event shown in Figure 4, we may approximately integrate over each vulnerability curve to derive the expected uninsured property loss rate given a flood event for census tract. Figure 10 depicts the geographic distribution of this integrated vulnerability. It indicates that, beside a few counties in the northeast (in particular, Halifax County and Northampton County), the western NC may have a higher level of flood vulnerability than in other parts of NC.

### 6.3 | Risk

Applying Equations (2)–(4), we may incorporate information on hazard potential, exposure, and vulnerability to produce the geographic distribution of flood risk in terms of expected annual uninsured property loss (Figure 11, corresponding to Equation (3)) and expected annual uninsured property loss ratio per total property value (Figure 12, corresponding to Equation (4)). The similarities between Figures 6, 7, 11, and 12 indicate that the flood risk distribution is likely to be influenced more by distribution of property values exposed in hazardous flood zones than by distribution of flood vulnerability. This implies that high-quality local flood hazard mapping is essential for effective flood risk assessment to guide flood risk management. In the meantime, regulating and restricting properties in the hazardous flood zones may be one of the most pertinent policies toward flood disaster risk reduction.

| Regression | Variable | Description | Effect on loss |
|------------|----------|-------------|----------------|
| Binomial   | $I_0$    | Mean water depth (cm) | Increasing |
|            | $B_4$    | Mean age of building (years) | Increasing |
|            | $B_9$    | Median of first floor elevation (cm) | Increasing |
|            | $B_{25}$ | Ratio of concrete buildings to buildings with provided building type | Decreasing |
|            | $B_{28}$ | Ratio of steel buildings to buildings with provided building type | Increasing |
|            | $B_{30}$ | Ratio of buildings in the A zone to buildings with provided flood zone information | Increasing |
|            | $S_{13}$ | Ratio of black alone population to population | Decreasing |
|            | $S_{205}$ | Ratio of households with other types of income in past 12 months to households | Decreasing |
|            | $S_{280}$ | Ratio of female civilians aged 16 or over employed in education, legal, community service, arts, and media occupations to female employed civilians aged 16 or over | Decreasing |
| Linear     | $I_0$    | Mean water depth (cm) | Increasing |
|            | $B_3$    | Median of maximum between building value and building replacement value (US$) | Decreasing |
|            | $B_5$    | Median of age of building (years) | Increasing |
|            | $B_{28}$ | Ratio of steel buildings to buildings with provided building type | Increasing |
|            | $S_{185}$ | Ratio of households with income of $30,000 to $34,999 in past 12 months in year-specific inflation-adjusted dollars to households | Increasing |
|            | $S_{329}$ | Ratio of renter-occupied housing units built between 1980 and 1989 to occupied housing units | Increasing |
**TABLE 3** Statistics of parameters of the final vulnerability model

| Regression | Parameter | Estimate | Standard error | t score | p-value |
|------------|-----------|----------|----------------|---------|---------|
| Binomial   | $\beta_{B0}$ | 2.2183  | 0.1038 | 21.3753 | $3.07 \times 10^{-91}$ |
|            | $\beta_{B1}$ | 0.3575  | 0.0678 | 5.2740  | $1.48 \times 10^{-7}$  |
|            | $\beta_{B4}$ | 0.2476  | 0.0718 | 3.4508  | $5.71 \times 10^{-4}$  |
|            | $\beta_{B9}$ | 0.2606  | 0.0976 | 2.6702  | $7.64 \times 10^{-3}$  |
|            | $\beta_{B25}$ | -0.6043 | 0.0983 | 6.1465  | $9.58 \times 10^{-10}$ |
|            | $\beta_{B28}$ | 0.3123  | 0.0713 | 4.3772  | $1.27 \times 10^{-5}$  |
|            | $\beta_{B30}$ | 0.3866  | 0.0802 | 4.8198  | $1.55 \times 10^{-6}$  |
|            | $\beta_{B33}$ | -0.2346 | 0.0945 | 2.4837  | $1.31 \times 10^{-2}$  |
|            | $\beta_{B35}$ | -0.1999 | 0.0771 | 2.5924  | $9.60 \times 10^{-3}$  |
|            | $\beta_{B38}$ | -0.2784 | 0.0816 | 3.4115  | $6.59 \times 10^{-4}$  |
| Linear     | $\beta_{L0}$ | -2.8304 | 0.0394 | 71.9229 | $<10^{-323}$             |
|            | $\beta_{L3}$ | 0.7906  | 0.0401 | 19.7070 | $1.05 \times 10^{-77}$  |
|            | $\beta_{L5}$ | -0.1406 | 0.0440 | 3.1935  | $1.43 \times 10^{-3}$  |
|            | $\beta_{L8}$ | 0.1612  | 0.0434 | 3.7193  | $2.11 \times 10^{-4}$  |
|            | $\beta_{L28}$ | 0.2086  | 0.0415 | 5.0239  | $5.62 \times 10^{-7}$  |
|            | $\beta_{L30}$ | 0.0900  | 0.0458 | 1.9657  | $4.95 \times 10^{-2}$  |
|            | $\beta_{L32}$ | 0.1623  | 0.0403 | 4.0234  | $6.00 \times 10^{-5}$  |
|            | $\sigma$ | 1.5647  | 0.0274 | 57.2014 | $<10^{-323}$             |

**FIGURE 8** Results of leave-one-out cross-validations for vulnerability model in terms of (a) predicted uninsured property loss rate and (b) predicted uninsured property loss

**FIGURE 9** Example of vulnerability curve with uninsured property loss rate displayed (a) with the original scale and (b) with a logarithmic scale
**VIDEO 1** Animation of vulnerability distribution in terms of uninsured property loss rate with respect to an increasing flood water depth [Video content can be viewed at https://onlinelibrary.wiley.com/doi/10.1111/jfr3.12739]

**FIGURE 10** Vulnerability distribution in terms of expected uninsured property loss rate given a flood event

**FIGURE 11** Flood risk distribution in terms of expected annual uninsured property loss (US$)
7 | DISCUSSION

7.1 | Scientific advances

As illustrated by its application in the study area of NC, the proposed methodology for assessment of flood vulnerability and risk has two major benefits over the traditional approaches. First, we adopted a predictive modeling approach with a two-stage variable selection process. This approach allows for an unlimited number of input variables to be initially considered. These input variables may include any hazard intensity measures, built environmental indicators, and social vulnerability indicators. This presents an advance over conventional damage models based on vulnerability curves derived for specific natural or built environment characteristics without consideration of social vulnerability factors. As indicated by the results of this presented study, social, demographic, and economic factors are likely to contribute to flood vulnerability.

Second, existing flood models which consider social vulnerability indicators tend to adopt an indexing, instead of a data-driven predictive modeling, approach. Indexing approaches usually only provide categories of vulnerability or risk for relative comparisons and, as a result, they cannot be used for technical computations to estimate the expected losses in terms of, for example, human life or property values due to future flood events (e.g., Salazar-Briones et al., 2020; Tate et al., 2021). Moreover, without referencing historical data on flood damage, it is difficult to verify the validity and accuracy of selected social indicators and the appropriateness of the weights of these indicators in their flood vulnerability or risk indices (also see Wang et al., 2021). Unlike indexing approaches, we derived the geographic distribution of flood vulnerability, based on historical records of flood damage, for technical computations of expected annual uninsured property losses due to future flood events. The vulnerability indicators and their weights were selected and determined using an empirical statistical process of variable selection and model calibration. As a result, the geographic distribution of flood vulnerability shown in Video 1 and Figure 10 may be closer to the actual flood vulnerability distribution than those created by an indexing approach.

7.2 | Limitations

Despite its benefits over the existing approaches, the proposed predictive modeling methodology for assessment of flood vulnerability and risk is not rid of challenges and limitations. A common challenge to all risk assessment tools for societal impacts is that, once results are produced, it is difficult to uncover their systemic bias and true uncertainties without access to a statistically sufficiently large number of disasters. When applying the risk assessment methodology presented in this paper, practitioners should also be aware of its potential prediction bias and inherent uncertainties.

The potential bias of risk prediction derived from the presented methodology may be germinated from three sources. First, the flood hazard map used for this study may not be accurate. As pointed out in numerous previous studies (e.g., Blessing et al., 2017; Brody et al., 2013; Quinn et al., 2019; Wing et al., 2017, 2018), the NFHL (FEMA, 2020b) may not be fully representative of the actual expected flood damage of an average community given a flood event. For example, the risks are underestimated in communities where headwater streams or tributaries have not been mapped (Wing et al., 2017). As highlighted previously, exposure within flood zones is one of the most pertinent contributors to flood risk. More
accurate flood zone mapping is needed for improving accuracy of flood risk assessment. Meanwhile, treating areas within a flood zone as sharing the same probability of flood is also not supported by empirical evidence (Brody et al., 2018). To overcome this issue, future flood hazard mapping needs to provide geographic distributions of, for example, occurrence frequencies with respect to intensity measure or exceedance values of intensity measure with respect to return period. With these more advanced flood hazard maps, we may also avoid using the spatially and temporally identical empirical probability distribution of water depth given a flood event, as shown in Figure 4, to further reduce bias due to the assumption of such identicality.

The second cause of bias may be related to data on flood impacts. Because only two hurricane events are considered, the constructed data points may not appropriately reflect the potential economic damages due to flood events that were not induced by hurricanes. To reduce the potential bias due to consideration of flood damage and extent associated with hurricanes only, it may be beneficial to rely on flood inundation data derived from other types of flood extent modeling approaches that capture hazards that may not be represented by physical models (e.g., Mobley et al., 2019). In addition, the damage data were initially estimated with analytical modeling tools instead of being collected based on actual observations on the ground. To provide more accurate predictive models for flood vulnerability and risk assessment, it will also be necessary to acquire actual historical records of flood losses for model calibration.

Beside flood hazard and damage data, a lack of accurate prediction of future vulnerability factors may also result in bias in model predictions of flood risk. In the presented work, we used 2018 census data to derive social vulnerability indicators as proxies of social factors for flood risk assessment for future years. These proxy indicators may underestimate or overestimate the true values of the indicators in the future if the indicator values are being changed significantly. To address this problem, modeling with time series forecasting techniques based on limited amount of yearly records of these indicators may need to be explored. Moreover, estimates of future built environmental attributes and exposure of property values in this study may also have limitations. Because there exists a small number of building records with missing entries of built environmental attributes or building values, the aggregated measures at census tracts may underestimate the true mean, median, or sum of these attributes and property values for some census tracts. As the building inventory data from NCFMP (2020) only contains one record for one building parcel along the temporal dimension, such records of built environmental attributes and property values may not appropriately correspond to future realities.

In addition to bias-related problems, it is also challenging to quantify the aleatory and epistemic uncertainties in the prediction results of presented risk assessment model. The uncertainty quantification for vulnerability model in this study was achieved via analytical approach (Wang et al., 2019, 2020). Beyond this, a lack of consistent descriptions of uncertainty measures for data on hazard, exposure, and vulnerability renders it difficult to quantify uncertainties for the entire risk assessment model. For example, although census data usually contains information on standard errors of measures of variables, such standard errors may not be useful for predictive modeling because the probability distributions of measures of variables are not given. Rather than merely providing the point estimate of each measure of variable and its standard error, a description of the probability distribution of such a measure in terms of values of pertinent parameters can be highly useful for predictive modeling for risk assessment for flood and other hazards.

As the presented methodology is only a tool for flood risk assessment and the further risk management, policy-oriented users may also need to consider the temporal, measurable, and inter-relational context of the independent variables as well as other factors for better understanding the precursors, processes, and outcomes of vulnerability factors and integrating societal behavior and behavioral adaptation dynamics into their risk assessment framework (Aerts et al., 2018; Rufat et al., 2015). To facilitate such consideration, some of the dynamic subjective vulnerability factors such as risk perception and cognitive mapping (Terti et al., 2015) may be included in the presented predictive modeling process, when pertinent data becomes available.

8 | CONCLUSION

Based on data on recent flood losses in NC as an example, in this paper, the authors presented a data-driven predictive modeling methodology to assess flood vulnerability and risk for communities. Products of the proposed methodology are maps that can be used to highlight hotspots of flood vulnerability and risk in terms of expected uninsured property loss and loss ratio given a flood water depth, a flood event, or a future year. Results of the study indicate that the proposed approach can be implemented to integrate both built environmental and social vulnerability factors to predict flood vulnerability and risk. Future works may further apply the methodology for flood vulnerability and risk assessment for other states in the United States as well as the entire country. Meanwhile, results also suggest
that flood hazard mapping for exposure computation is one of the most pertinent aspects for flood risk assessment. To improve prediction results, therefore, future works need to compare effects of different flood hazard mapping methods on risk assessment, to explore probability distributions of intensity measures given a flood event, and to develop more detailed flood hazard maps to better present geographic distributions of flood frequency and intensity measures.

ACKNOWLEDGMENTS
This research was supported by the North Carolina General Assembly through Session Law 2019-224. The authors are also grateful to North Carolina Policy Collaboratory and NCDEM for access to data.

CONFLICT OF INTEREST
The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

DATA AVAILABILITY STATEMENT
The data and codes that support the findings of this study are openly available in the University of North Carolina Dataverse at https://doi.org/10.15139/S3/C4NO7E.

ORCID
Yi (Victor) Wang  https://orcid.org/0000-0003-2228-7009
Antonia Sebastian  https://orcid.org/0000-0002-4309-2561

REFERENCES
Aerts, J. C. J. H., Botzen, W. J., Clarke, K. C., Cutter, S. L., Hall, J. W., Merz, B., Michel-Kerjan, E., Mysiak, J., Surminski, S., & Kunreuther, H. (2018). Integrating human behaviour dynamics into flood disaster risk assessment. *Nature Climate Change*, 8, 193–199. https://doi.org/10.1038/s41558-018-0085-1

Akaïke, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716–723. https://doi.org/10.1109/TAC.1974.1100705

Allen, S. K., Linsbauer, A., Randhawa, S. S., Huggel, C., Rana, P., & Kumari, A. (2016). Glacial lake outburst flood risk in Himachal Pradesh, India: An integrative and anticipatory approach considering current and future threats. *Natural Hazards*, 84, 1741–1763. https://doi.org/10.1007/s11069-016-2511-x

Anarde, K. A., Kameshwar, S., Irza, J. N., Nittrouer, J. A., Lorenzo-Trueba, J., Padgett, J. E., Sebastian, A., & Bedient, P. B. (2018). Impacts of hurricane storm surge on infrastructure vulnerability for an evolving coastal landscape. *Natural Hazards Review*, 19, 04017020. https://doi.org/10.1061/(ASCE)NH.1527-6996.0000265

Arrighi, C., Brugioni, M., Castelli, F., Franceschini, S., & Mazzanti, B. (2013). Urban micro-scale flood risk estimation with parsimonious hydraulic modelling and census data. *Natural Hazards and Earth System Sciences*, 13, 1375–1391. https://doi.org/10.5194/nhess-13-1375-2013

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, 239. https://doi.org/10.3390/ijerph13020239

Blessing, R., Sebastian, A., & Brody, S. D. (2017). Flood risk delineation in the United States: How much loss are we capturing? *Natural Hazards Review*, 18, 04017002. https://doi.org/10.1061/(ASCE)NH.1527-6996.0000242

Brody, S., Blessing, R., Sebastian, A., & Bedient, P. (2014). Examining the impact of land use/land cover characteristics on flood losses. *Journal of Environmental Planning and Management*, 57, 1252–1265. https://doi.org/10.1080/09640568.2013.802228

Brody, S., Sebastian, A., Blessing, R., & Bedient, P. B. (2018). Case study results from southeast Houston, Texas: Identifying the impacts of residential location on flood risk and loss. *Journal of Flood Risk Management*, 11, S110–S120. https://doi.org/10.1111/jfr3.12184

Brody, S. D., Peacock, W. G., & Gunn, J. (2012). Ecological indicators of flood risk along the Gulf of Mexico. *Ecological Indicators*, 18, 493–500. https://doi.org/10.1016/j.ecolind.2012.01.004

Budiyono, Y., Aerts, J., Brinkman, J., Marfai, M. A., & Ward, P. (2015). Flood risk assessment for delta mega-cities: A case study of Jakarta. *Natural Hazards*, 75, 389–413. https://doi.org/10.1007/s11069-014-1327-9

Budiyono, Y., Aerts, J. C. J. H., Tollenaar, D., & Ward, P. J. (2016). River flood risk in Jakarta under scenarios of future change. *Natural Hazards and Earth System Sciences*, 16, 757–774. https://doi.org/10.5194/nhess-16-757-2016

Cabrera, J. S., & Lee, H. S. (2020). Flood risk assessment for Davao City, Philippines using geographic information system-based multi-criteria analysis and the maximum entropy model. *Journal of Flood Risk Management*, 13, e12607. https://doi.org/10.1111/jfr3.12607

CDC. (2021, February 19). *Flooding from hurricane Matthew in North Carolina*. Author. Retrieved from https://www.cdc.gov/cpr/readiness/stories/nc.htm

Coussas, N., Sebastian, A., & Morales-Nápoles, O. (2018). A copula-based Bayesian network for modeling compound flood hazard from riverine and coastal interactions at the catchment scale: An application to the Houston Ship Channel, Texas. *Water*, 10, 1190. https://doi.org/10.3390/w10091190

Custer, R., & Nishijima, K. (2015). Flood vulnerability assessment of residential buildings by explicit damage process modelling. *Natural Hazards*, 78, 461–496. https://doi.org/10.1007/s11069-015-1725-7

Cutter, S. L., Emrich, C. T., Morath, D. P., & Dunning, C. M. (2013). Integrating social vulnerability into federal flood risk management planning. *Journal of Flood Risk Management*, 6, 332–344. https://doi.org/10.1111/jfr3.12018

D’Ayala, D., Wang, K., Yan, Y., Smith, H., Massam, A., Filipova, V., & Pereira, J. J. (2020). Flood vulnerability and risk assessment of urban traditional buildings in a heritage district of Kuala Lumpur, Malaysia. *Natural Hazards and Earth System Sciences*, 18, 3095–3107. https://doi.org/10.5194/nhess-18-3095-2018

Ducic, H., & totally confidential (2018). Assessing the health impacts of flooding from Hurricane Matthew in North Carolina. *Environmental Research*, 166, 465–471. https://doi.org/10.1016/j.envres.2018.08.003
Kirschbaum, D., Kunz, M., Mohr, S., Muis, S., Riddell, G. A., Schäfer, A., Stanley, T., Veldkamp, T. I. E., & Winsemius, H. C. (2020). Review article: Natural hazard risk assessments at the global scale. *Natural Hazards and Earth System Sciences, 20*, 1069–1096. https://doi.org/10.5194/nhess-20-1069-2020

Ward, P. J., Jongman, B., Salamon, P., Simpson, A., Bates, P., De Groeve, T., Muis, S., Coughlan de Perez, E., Rudari, R., Trigg, M. A., & Winsemius, H. C. (2015). Usefulness and limitations of global flood risk models. *Nature Climate Change, 5*, 712–715. https://doi.org/10.1038/nclimate2742

Wing, O. E. J., Bates, P. D., Sampson, C. C., Smith, A. M., Johnson, K. A., & Erickson, T. A. (2017). Validation of a 30 m resolution flood hazard model of the conterminous United States. *Water Resources Research, 53*, 7968–7986. https://doi.org/10.1002/2017WR020917

Wing, O. E. J., Bates, P. D., Smith, A. M., Sampson, C. C., Johnson, K. A., Fargione, J., & Morefield, P. (2018). Estimates of present and future flood risk in the conterminous United States. *Environmental Research Letters, 13*, 034023. https://doi.org/10.1088/1748-9326/aaac65

Wing, O. E. J., Pinter, N., Bates, P. D., & Kousky, C. (2020). New insights into US flood vulnerability revealed from flood insurance big data. *Nature Communications, 11*, 1444. https://doi.org/10.1038/s41467-020-15264-2

Zaharia, L., Costache, R., Prăvălie, R., & Ioana-Toroiu, G. (2017). Mapping flood and flooding potential indices: A methodological approach to identifying areas susceptible to flood and flooding risk. Case study: The Prahova catchment (Romania). *Frontiers of Earth Science, 11*, 229–247. https://doi.org/10.1007/s11707-017-0636-1

**SUPPORTING INFORMATION**

Additional supporting information may be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Wang, Y. (V.), & Sebastian, A. (2021). Community flood vulnerability and risk assessment: An empirical predictive modeling approach. *Journal of Flood Risk Management, 14*, e12739. https://doi.org/10.1111/jfr3.12739