Predicting Flash Flood Economic Damage at the Community Scale: Empirical Zero-Inflated Model with Semicontinuous Data

Shi Chang[1]; Rohan Singh Wilkho[2]; Nasir Gharaibeh, M.ASCE[3]; Stacey Lyle[4]; and Lei Zou[5]

Abstract: Rainfall-induced flash floods are characterized by their rapid onset and small spatial scale. With little lead time for warning, floodwater can accumulate rapidly and its force can damage roads, swamp houses, destroy bridges, and scour out channels. Having data-driven estimates of potential economic losses from flash floods (before they occur) helps authorities make informed decisions about planning and prioritizing mitigation projects. This article provides a probabilistic predictive model to estimate flash flood economic damage at the census tract scale. To simplify model utilization and avoid strong assumptions about property value and replacement costs, the model predicts the total cost of property and infrastructure damages for individual census tracts (expressed in 2019 prices). The model was developed based on a flash flood data set for a 15-year period (2005–2019) in Texas. The data set was assembled by integrating disparate data from multiple platforms. The occurrence of economic damage was found to be a zero-inflated problem. Therefore, we developed a two-part mixed-effect model. The model first estimates the probability that economic damage will occur (zero-inflated part) and then predicts the dollar amount of the economic damage (continuous part). Utilization of the developed model was demonstrated in an application to Harris County, Texas.
DOI: 10.1061/NHREFO.NHENG-1729.

Author keywords: Flash flood; Economic loss; Two-part model; Zero inflation; Semicontinuous data.

Introduction

Over the past two decades, data reported by the National Oceanic and Atmospheric Administration's National Weather Service (NWS) have indicated that approximately 52% of flood-related economic losses in the US are attributed to flash flooding (NWS 2019). These economic losses include immediate and long-term damages to both private property and public infrastructure (Vinet 2008; Karagiorgos et al. 2016).

Flood economic-damage forecasting models have received increasing attention over the years to inform flood control and insurance policies (Karamouz et al. 2016; Mokhtari et al. 2017; Gutenson et al. 2017; Milanesi and Pilotti 2021). However, the spatial scales of these models remain either too large (e.g., county level) or too small (e.g., site-specific or building type–specific) for planning flash flood mitigation measures at the city or sub-city scale (Chang et al. 2023; Wilkho et al. 2023). For a detailed discussion of differences in the methodology and purpose of flood risk assessments at different spatial scales (supra-national, macro, meso, micro), we refer the reader to a review by De Moel et al. (2015). Furthermore, many flood damage estimation models are based on the depth of floodwater and associated damage at those depths using depth-damage functions. In flash floods, however, the concern is not just the depth of water but also the water flow velocity and accumulation time. In depth-damage functions, the main determinant of direct damage is the water depth. Pistrika et al. (2014) and Milanesi and Pilotti (2021) pointed out that many more factors— flow velocity, duration of flooding, sediment load, contamination, available flood warning systems, and effectiveness of emergency response during a flood event—affect the severity and extent of flood damages. Flood economic damage models rarely include all of these influencing factors (Pistrika et al. 2014; Baradaranshoraka et al. 2019), perhaps due to difficulties in obtaining reliable data for them. Other factors, such as population increase, urbanization, increasing vulnerabilities, and increasing extreme rainfalls, seem to have a substantial effect on the impacts of flash floods (Ahmadalipour and Moradkhani 2019). Gutenson et al. (2017) emphasized the need for considering the uncertainties associated with flood damage and economic loss estimations, and effectively communicating these uncertainties to end users. To address these limitations, we provide a new empirical model for forecasting flash flood economic damages at the census tract scale. Economic damage is defined here as the approximate cost of direct physical damage (expressed in 2019 US prices), including loss of property and costs of repairing damaged infrastructure (Downton and Pielke 2005).

We make provide two major advances to flash flood economic damage forecasting models in this article. First, the proposed model uses a spatial scale consistent with the scale of flash floods. We use the census tract as the analysis unit to capture the characteristics of the triggering storm and the built, natural, and social environments in which it occurs. Furthermore, data on the model inputs

[1]Graduate Research Assistant, Zachry Dept. of Civil and Environmental Engineering, Texas A&M Univ., College Station, TX 77843 (corresponding author). ORCID: https://orcid.org/0000-0002-5180-6296. Email: changs18@tamu.edu

[2]Graduate Research Assistant, Zachry Dept. of Civil and Environmental Engineering, Texas A&M Univ., College Station, TX 77843. Email: rohanswilkho_93@tamu.edu

[3]Professor, Zachry Dept. of Civil and Environmental Engineering, Texas A&M Univ., College Station, TX 77843. Email: ngharaibeh@civil.tamu.edu

[4]Associate Professor of Practice, Zachry Dept. of Civil and Environmental Engineering, Texas A&M Univ., College Station, TX 77843. Email: stacey.lyle@tamu.edu

[5]Assistant Professor, Dept. of Geography, Texas A&M Univ., College Station, TX 77843. Email: lzou@tamu.edu

Note. This manuscript was submitted on August 4, 2022; approved on March 24, 2023; published online on June 27, 2023. Discussion period open until November 27, 2023; separate discussions must be submitted for individual papers. This paper is part of the ***Natural Hazards Review***, © ASCE, ISSN 1527-6988.

are readily available at this spatial scale. The census tract (delineated by the US Census Bureau) is a relatively permanent subdivision of a county with a population size between 1,200 and 8,000 people (the optimum size is 4,000 people). Second, the model accounts for the inflated zeros in flash flood damages. As shown later in this paper, flash flood damage data are strongly skewed, containing large numbers of zeros (i.e., flash floods that have no or minimal damages), as well as extremely high damage values (outliers). This type of phenomenon is described as a zero-inflated problem in economics and epidemiology (Burger et al. 2009; Rose et al. 2006). It is difficult to obtain unbiased statistical inferences and predictions from such data using conventional statistical methods (Burger et al. 2009; Rose et al. 2006). Therefore, we employ a two-part mixed-effect model. The first part determines the probability that economic damage will occur, and the second part predicts the dollar amount of the damage.

Background

There exists a considerable body of research on flood damage assessment and forecasting models. However, the literature on forecasting flash flood damage at the community level remains scarce, despite the unique nature of flash floods. The conventional approach for estimating direct economic flood damage to structures is water depth-damage functions. In flash floods, however, damage to structures is often caused by swift floodwater (high velocity and high force) and it can involve the carrying away of valuable properties such as vehicles (Vuichard and Zimmermann 1986; Schroeder et al. 2016). Also, water depth-damage functions tend to ignore the harmful molds, producing allergens and irritants, which usually require specialized treatments performed by licensed technicians (Chew et al. 2006; Brandt et al. 2006). These treatments cost more than a thousand dollars each time and can be long term (Aalberts and Hoyt 2000). In addition to immediate and long-term damages to private property, flash floods result in damages to critical infrastructure. Large debris and floodwaters can cause structural damage to bridge and roadways. Power, telephone, and cable lines can be washed away or seriously damaged by flash floods as well. Interruptions to transportation systems, business closures, and logistical delays can lead to additional economic losses (Taguchi et al. 2022). A review by Merz et al. (2010) provides a comprehensive review of economic flood damage assessment models and future research directions.

The impacts of flash flooding have been analyzed at the county scale in the United States. Khajehei et al. (2020) evaluated the socioeconomic vulnerability to flash flooding at the county scale in accounting for flash flood characteristics, including duration, frequency, magnitude, and severity. Alipour et al. (2020) employed a random forest model to predict flash flood damages at the county scale in the Southeastern US. Such models indicate an association between flash flood–related economic losses and event precipitation level, event duration, unit peak discharge, surrounding natural and built environment, population density, and community social vulnerability.

In summary, the model provided in this paper has two key advantages. First, the model uses a spatial scale consistent with the spatial scale of flash flood events (i.e., sub-city scale, census tract). While the census tract scale may not be in perfect agreement with the flash flooding scale (e.g., low-populated areas where the scale might be smaller than the affected census tract), it provides closer agreement than currently available models (e.g., county scale). Second, the model does not rely on conventional water depth-damage functions, which may not be suitable for flash flooding where damage occurs because of not only floodwater depth but also floodwater velocity and accumulation time. Instead, the model estimates economic damage based on empirical data about the triggering storm and the built, natural, and social environments in which it occurs.

Finally, we recognize that previous research efforts addressed the safety risks (human fatalities and injuries) of flash flooding in the United States (Zahran et al. 2008; Ashley and Ashley 2008; Sharif et al. 2012, 2015; Ahmadalipour and Moradkhani 2019; Chang et al. 2023). However, these safety issues are beyond the scope of this article.

Study Area and Data

Flash Flood Event Data

Flash flood event data were obtained from NOAA's storm events database. A total of 2,866 flash flood events that occurred in Texas over a 15-year period (2005–2019) were included in this study. These events were recorded by law enforcement and emergency management officials. The storm events database contains spatial and temporal information about natural storm hazards (including flash flooding) that have sufficient intensity to cause loss of life, injuries, significant economic damage, and/or disruption to commerce.

Texas has the highest number of flash flood events in the US. These flash floods occur largely in the Hill Country and in Central Texas areas along the Balcones Escarpment, which lies between the Edwards Plateau and the coastal plain (Sharif et al. 2012, 2015). This region (often called Flash Flood Alley) is characterized by steep terrain, shallow soil, and unusually high rainfall rates (Baker 1975; Caran and Baker 1986). Additionally, hurricanes and tropical storms frequently visit Gulf of Mexico coastal areas in Texas and release enormous amounts of water in a very short time, which generates flash floods across communities and results in billions of dollars of damage or economic loss (Burnett 2008; Kousky et al. 2020).

Fig. 1 shows the number of flash flood events in each Texas county and the total dollar amounts (expressed in 2019 dollars) of economic damage from these events. Since the damage analysis in this study spans a 15-year period, a 2% annual average inflation rate, acquired from the US Bureau of Labor Statistics, was applied to convert the dollar amounts into 2019 equivalent values. It can be seen that areas with severe flash flood damage (>$10M) were found not only in areas where flash flooding happens recurrently (dark red areas) but also in in areas with occasional or infrequent flash flooding. This indicates that these damages occur due to a variety of influencing factors, such as storm characteristics, the built environment, site geographic and socioeconomic characteristics, and environmental stimulus.

Factors Influencing Economic Damage from Flash Flood Events

Studies have found that flash flood damage and human loss are related not only to behavior or poor judgment about potential dangers but also to the physical, natural, and social environments at the site of the event (Zahran et al. 2008; Sharif et al. 2015; Terti et al. 2019; Chang et al. 2021). Based on this research, we identified 14 candidate factors that represent these environments and influence the occurrence and magnitude of flash flood economic damage. These factors and their data sources are provided in Fig. 2.

The data on our candidate factors were acquired from different publicly available data sets and platforms with diverse formats and

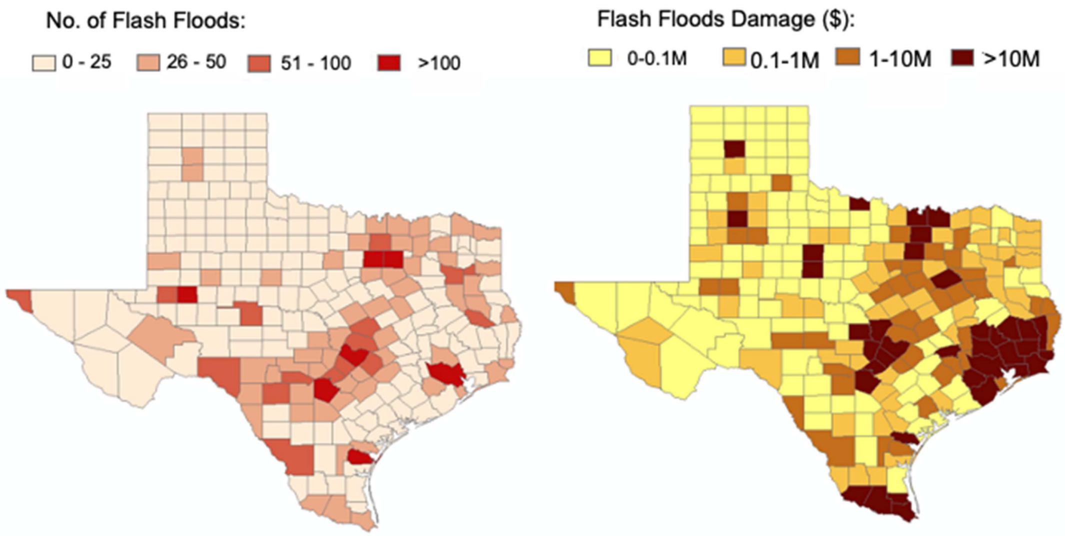

Fig. 1. Number of flash flood events and associated economic damages in Texas (2005–2019).

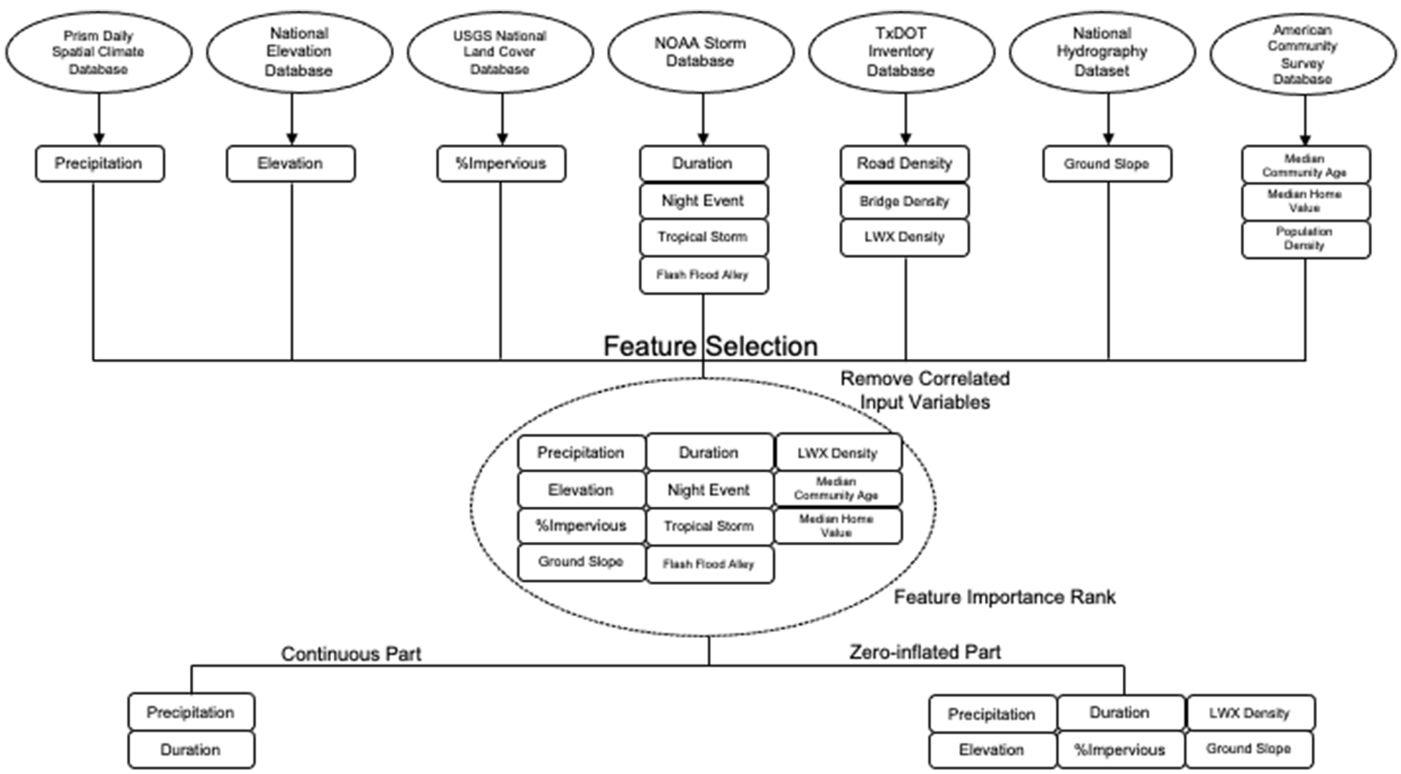

Fig. 2. Flash flood microdata and feature selection for input variables.

structures. Therefore, it was necessary to integrate these disparate data based on geographical coordinates (latitude and longitude), census tract, and year of event. The Google Earth Engine Python programming interface and Esri's ArcGIS software were used to implement data integration and geoprocessing at a fine spatial scale. The final data set used in this study included flash flood events and potential influencing factors at the census tract scale. The census tract was used as a surrogate for neighborhood/community and local environment to capture temporal and spatial complexities at the scene of the event.

Two-Part Model of Economic Damage

In this section, we describe the development of the flash flood economic damage predictive model. Fig. 3 shows that 70% of the flash flood events in the data set resulted in zero economic damage. Also, the nonzero part of the economic damages exhibits a log-normal distribution. Such skewed data with many zeros are typically referred to as semicontinuous data in economics, social science, and epidemiology research (Su et al. 2009; Hsu and Liu 2008; Boulton and Williford 2018).

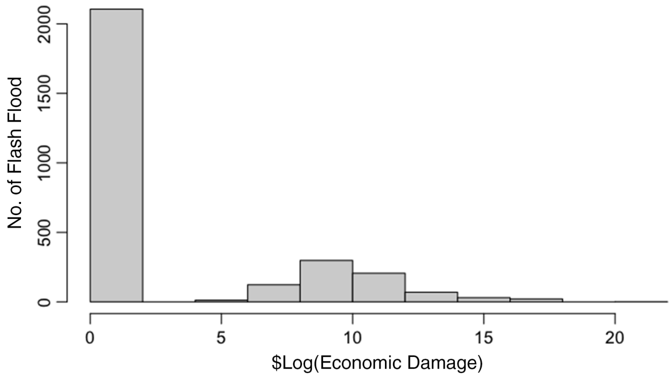

Fig. 3. Histogram of Texas flash flood economic damage in 2005–2019 (expressed in 2019 prices).

The skewed shape of distribution for the response variable, like that shown in Fig. 3, can be problematic for conventional statistical predictive models (i.e., linear regression, generalized linear regression), in which several assumptions must be satisfied, leading to a biased estimation and misleading inference for potential influencing factors (Olsen and Schafer 2001; Boulton and Williford 2018). On the other hand, simple discretization of the data into two categories (zero damages and nonzero damages) leads to loss of information and inflates relations between the response variables and the influencing factors. Therefore, semicontinuous variables are often thought to reflect observations from two distinct data-generating processes, one determining whether the outcome is zero and the other determining the actual value if the outcome is nonzero (Olsen and Schafer 2001; Lachenbruch 2002; Tooze et al. 2002; Su et al. 2009).

County-Level Random Intercept

In this study, the severity of flash flood damage was shown to vary significantly in different geographic areas, even under the same rainfall amount, onset time, and storm episode. This depends on local vulnerability and resilience to flash flooding, such as landscape terrains, soil types, land use types, housing structures, drainage system conditions, and household compositions (Zahran et al. 2008; Khajehei et al. 2020). To illustrate this point, Fig. 4 plots flash flood economic damage (in 2019 prices and log-transformed) versus event precipitation in four counties in Texas.

The fitted lines indicate that Travis County (the top line), located in central Texas, where the landscape is covered by steep terrain

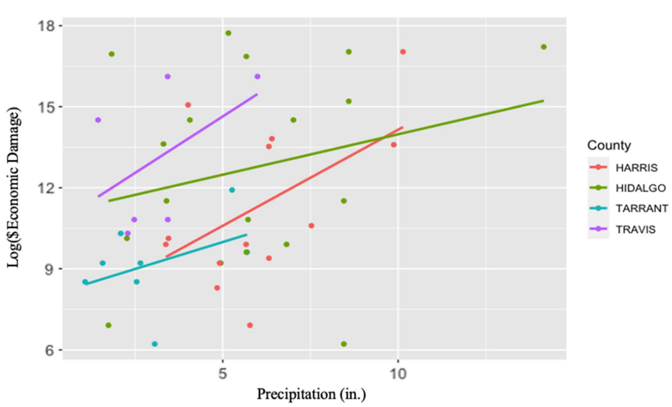

Fig. 4. Flash flood damages versus event precipitation in four Texas counties.

and shallow soil, experienced more severe economic damage than other counties in flash flood hazards. The economic damage in Hidalgo County was also high, possibly due to its poor infrastructure conditions and frequent tropical storms off the Gulf of Mexico. On the other hand, Harris County and Tarrant County, the two biggest metropolitan areas in Texas, exhibited better resistance to flash flood hazards in terms of economic loss. Therefore, this study treated county as a random intercept in the predictive model to represent a combination of unknown factors that are difficult to measure but that influence flash flood economic damage. Observation of the plot in Fig. 4 shows that for any given precipitation economic damage can vary greatly depending on county.

Mathematical Basis of Two-Part Mixed-Effect Model for Semicontinuous Data

Olsen and Schafer (2001) proposed a two-part mixed-effect model based on the generalized linear mixed model (GLMM) by introducing correlated random effects into both the model's binary and continuous parts. This model specifies a logistic regression for the dichotomous indicator that the outcome is zero or not, and a standard linear mixed-effect model for the logarithmic transformation of nonzero responses. In the literature, two-part mixed models has been proposed to analyze such continuous data with extra zeros (Su et al. 2009). Model notation and formulation are briefly described in this section.

Let Y_i be a semicontinuous variable for the ith $(i = 1, \ldots, N)$ event. This outcome variable can be represented by the occurrence variable Z_i, where

$$Z_i = \begin{cases} 0 & if\ Y_i = 0 \\ 1 & if\ Y_i > 0 \end{cases} \quad (1)$$

and the intensity variable $g(Y_i)$, given that $Y_i > 0$; $g(\cdot)$ = transformation function (log transform) that makes Y_i approximately normally distributed.

In the two-part mixed model, other than marginal distribution of Y_i, we focused on the distribution of the occurrence variable Z_i and the conditional distribution of the intensity variable $g(Y_i)$, given that $Y_i > 0$. Specifically, we assumed that Z_i follows a logistic regression model [Eq. (2)], so

$$\pi_i = \mathrm{logit}(Pr(Z_i = 0)) = X_i\theta \quad (2)$$

where π_i = probability of zero economic loss from event i; $X_i = 1 \times q$ explanatory variable vector; and $\theta = q \times 1$ regression coefficient vector. The intensity variable $g(Y_i)$, given $Y_i > 0$, follows a linear mixed model [Eq. (3)], so

$$g(Y_i)|Y_i > 0, \qquad g(Y_i) = X_i^*\beta + V_j + \epsilon_i \quad (3)$$

where $X_i^* = 1 \times p$ explanatory variable vector; $\beta = p \times 1$ regression coefficient vector; and V_j, again, = county-level random intercept. The error term ϵ_i is assumed to be distributed as $N(0, \sigma_e^2)$.

The coefficients in Eqs. (2) and (3) were estimated using the maximization likelihood function available in the R package GLMMAdpative. This function is based on a linear mixed-effects Laplacian approximation by adaptive Gaussian quadrature to approximate the likelihood developed by Pinheiro and Bates (1995).

The final model mixture response variable can be expressed as

$$Y_i = Exp(g(Y_i) \times (1 - \pi_i)) \quad (4)$$

where π_i = likelihood of zero economic damage from the model's zero-inflated part; $g(Y_i)$ = model estimation from the continuous part; and Y_i = estimated economic damage.

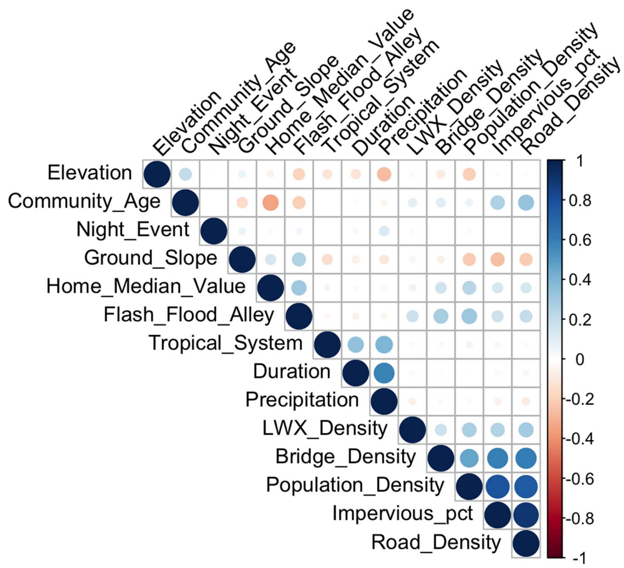

Fig. 5. Correlation matrix for flash flood influencing factors.

Feature Selection

Feature selection is an important process that removes redundant variables and reduces the number of input variables when developing a predictive model. It is desired to remove highly correlated variables to reduce the computational cost of modeling, improve the model's performance and robustness, avoid overfitting, and facilitate the model's utilization in practice.

We performed a correlation test to remove redundant features by analyzing the Pearson correlation matrix between the input variables. Fig. 5 shows the correlation matrix between all features in our flash flood data set. High correlation coefficients (> 0.5) were found between impervious surface percentage, road and bridge density, and population density (shown as large dots in the correlation matrix).

In addition to removing highly correlated features, we estimated the importance of features from the data by building a preliminary classification or regression model. We constructed a learning vector quantization (LVQ) model (classification) and a quantile regression forest (QRF) model (regression) to evaluate and rank the importance of input features for the zero-inflated part and the continuous part, separately. For the zero-inflated part, receiver operating characteristic (ROC) curve analysis was conducted for each predictor from the LVQ model. For the continuous part, the relationship between each predictor and the outcome was evaluated based on the R^2 statistic from the QRF model. The LVQ model is an artificial neural network algorithm developed for classification problems using a collection of codebook vectors (Kohonen 1995). The QRF model, a generalization of random forests, provides a nonparametric and accurate way of estimating conditional quantiles for high-dimensional predictor variables (Meinshausen and Ridgeway 2006). Details of these algorithms are beyond the scope of this work since we only used them for preliminary feature selection and estimating feature importance.

As shown in Fig. 6(a), ground slope is the most important feature among all input features to classify whether a flash flood event will cause economic damage. It is followed by road density, duration, impervious percentage, population density, and precipitation. On the other hand, according to Fig. 6(b), the magnitude of economic damage (in 2019 dollars) seems to be determined only by duration and precipitation.

Feature selection methods are believed to provide the most predictive features to be included in the two-part mixed model. The reduced amount of input variables will not only save computation time for fitting the two-part mixed model, but also generate a robust and nonoverfitted predictive model.

Model Results

Selection of Best Model

Table 1 lists four two-part mixed-effect models for four different subsets of input variables preselected using the feature selection method discussed earlier. As discussed earlier, the marginal distribution of the flash flood damage data is not normal due to the extra zeros. As a result, standard residual plots, when interpreted in the same way as linear models, are not suitable due to non-normality and heteroscedasticity, even if the model is correctly specified. Therefore, we used a simulation-based approach to create readily interpretable scaled (quantile) residuals to assess this GLMM model, with standardized residuals between 0 and 1. This approach is implemented in the R package DHARMa residual (DHARMa stands for diagnostics for hierarchical regression models) (Rizopoulos et al. 2010; Hartig and Hartig 2017). Its key advantage is that the defined residuals always have the same known distribution, independent of model fitting, if the model is correctly specified. Additionally, the following goodness-of-fit tests were performed on the scaled residuals and support visual inspection of the simulated residuals and model predictive power:

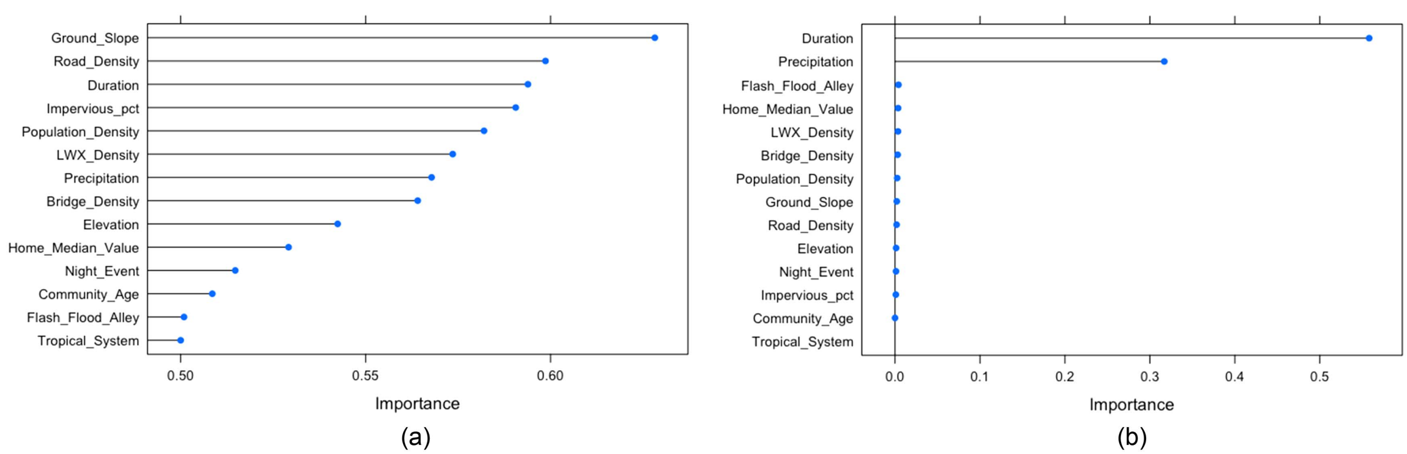

Fig. 6. (a) Feature importance ranking for model's zero-inflated part; and (b) feature importance ranking for model's continuous part.

Table 1. Candidate models and their performance metrics

Model	Input variables	AIC	BIC	KS test	Dispersion test
1	Precipitation + impervious percentage + ground slope + community age + duration + LWX density + elevation	6,454	6,502	Deviation not significant	Deviation significant
2	Precipitation + impervious percentage + ground slope + community age + duration + elevation	6,453	6,501	Deviation not significant	Deviation significant
3	Precipitation + impervious percent + ground slope + duration + community age	6,453	6,498	Deviation not significant	Deviation significant
4	Precipitation + impervious percent + ground slope + community age	6,451	6,492	Deviation not significant	Deviation not significant

- Kolmogorov–Smirnov (KS): tests if the overall distribution conforms to expectations.
- Dispersion: tests if the simulated dispersion is equal to the observed dispersion.

In this model, overdispersion or underdispersion deviation occurs when the discrepancies between the observed responses and their predicted values are larger or smaller than what the GLMM model predicts, possibly due to omitted covariates or nonsignificant variables involved. Based on Table 1, only Model 4 satisfies both goodness-of-fit tests on the scaled residuals, which indicates that it conforms to the expected distribution and shows no deviation for overdispersion or underdispersion problems. Therefore, Model 4 was selected as the best model.

In addition, we used Akaike's (1974) information criterion (AIC) and Schwartz's (1978) Bayesian information criterion (BIC) to describe the model's balance between goodness of fit and ability to avoid overfitting or underfitting problems. As shown in Table 1, Model 4 outperforms the other models with a lower value of AIC and BIC. While models with lower AIC and BIC are generally preferred, AIC and BIC may not be directly applicable to a model's significance for complex models, such as GLMM (Waagepetersen 2006; Pan and Lin 2005).

Model Coefficients

Table 2 provides coefficient estimates for the final model. For the zero-inflated part of the model, precipitation, duration, community age, and ground slope proved to have a significant correlation with the occurrence of economic damage during flash flood events. These inputs variables exhibit small p-values (<0.05). Further, the negative coefficient of precipitation and impervious percentage indicates that flash flood events with higher precipitation or occurring in areas with a greater number of impervious surfaces tend to have a greater chance of resulting in economic damage. In contrast, a positive coefficient of ground slope indicates that areas with a flatter ground slope tend to have higher levels of economic damage. We attributed this to accumulation of floodwater in flat downstream areas. Also, older neighborhoods exhibited less risk of economic damage during flash flood events than newer neighborhoods. This finding might be an indicator that older neighborhoods (in terms of median age of structures) in Texas tend to be located at higher elevations and at greater distances from floodways compared with newer neighborhoods.

For the continuous part of the model, a linear correlation was found between the amount of economic damage (expressed in 2019 prices) and the square-transformed level of precipitation and duration of the storm; both showed very small p-values ($<1 \times 10^{-4}$). The positive coefficients for both variables imply that flash flood events with heavier rainfall and longer duration generate greater economic damage. Further, greater flash flood economic damage tends to occur in communities where buildings are worth more in dollar value and are located in low-elevation areas.

Table 2. Model coefficients

Part	Parameter	Coefficient	Standard error	p-value
Zero-inflated part (probability of zero damage)	(Intercept)	−0.2436	0.1924	0.2055
	Precipitation[a]	−0.0064	0.0012	$<1 \times 10^{-4}$
	Impervious percentage	−0.0117	0.0031	0.0001
	Ground slope	0.0246	0.0026	$<1 \times 10^{-4}$
	Community age	0.0085	0.0038	0.0236
Continuous part (dollar amount)	(Intercept)	—	0.9578	$<1 \times 10^{-4}$
	Precipitation[a]	0.0144	0.0021	$<1 \times 10^{-4}$
	Duration[a]	0.0068	0.0015	$<1 \times 10^{-4}$
	Elevation[b]	−0.1686	0.0843	0.0454
	Home median value[b]	0.354	0.1771	0.0457

Note: See Fig. 7 and Appendix for county-specific random intercept values.
[a]Squared-transformation of variable.
[b]Log-transformation of variable.

The intercept value of the model's continuous part varies by county, as shown in Fig. 7 and listed in the Appendix. In this case, the county encompasses a random effect, which may be attributed to a combination of factors that are difficult to identify and quantify from the data set. Counties with darker color (and therefore greater intercept) exhibit more vulnerability to flash flood economic damages.

Final Model Evaluation

A simulation-based DHARMa residual analysis was performed to assess whether the final two-part mixed model has a visually homogenous residual. Fig. 8 shows the qq-plot and residual plot generated from the DHARMa residual simulation. The qq-plot on the left shows that the expected and observed residuals are lined up at the 45-degree line, without a significant deviation problem found. In the residuals against the predicted value plot, a scaled residual value of 0.5 means that half of the simulated data are higher than the observed value and half are lower. Several outliers at the top of the residual plot, shown as stars in Fig. 8, indicate high economic damage. These outliers are largely due to flash floods induced by hurricanes and tropical storms in the Texas coastal region. The overall uniform distribution of the scaled residuals in this plot indicates that the final model is correctly specified.

Finally, we performed a posterior predictive check on the final model, as shown in Fig. 9. We compared the empirical distribution

Nat. Hazards Rev., 2023, 24(4): 04023030

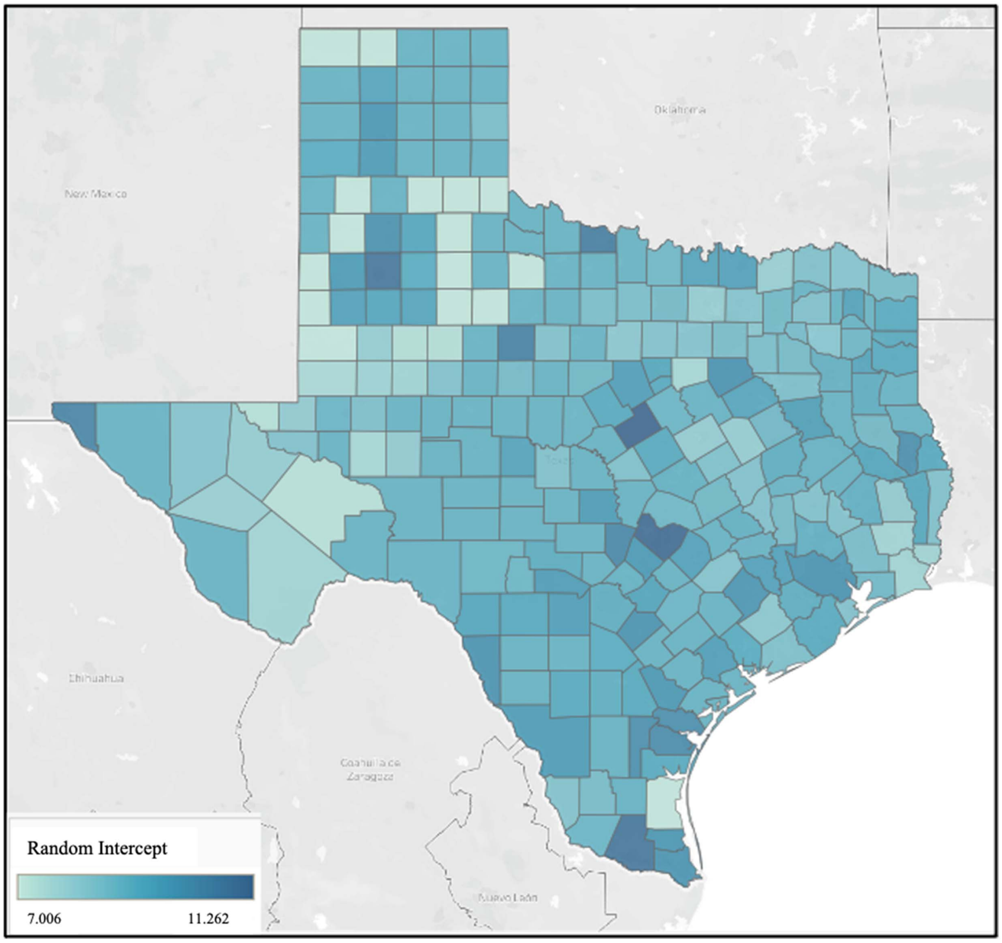

Fig. 7. Random intercept value based on the county variant. (© OpenStreetMap contributors.)

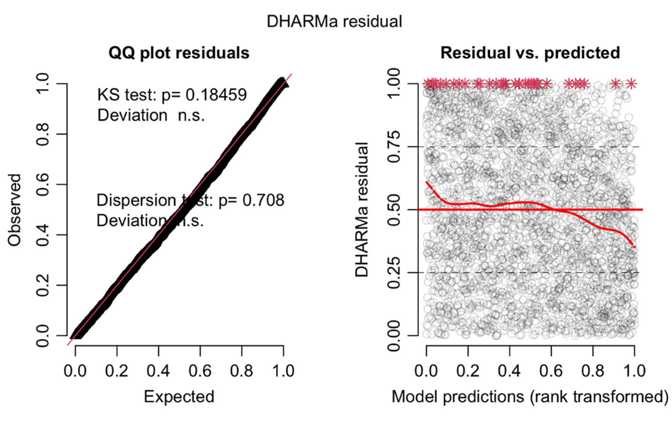

Fig. 8. qq-plot and DHARMa residuals (n.s. = not significant).

function of the observed data with estimates of the empirical distribution function of simulated/replicated data from the model. The figure indicates that the model-estimated economic damage (gray lines) has good agreement with observed economic damage (black line) at 30 iterations of simulation.

Model Application

To demonstrate use of the developed model for informing flash flood mitigation planning and economic risk assessment at the community scale (i.e., census tract), we applied it to estimate economic damage from the highest 1% of rainfall events in Harris County, Texas. According to the George Bush Intercontinental/Houston Airport (IAH) weather station, precipitation for these events at 6-, 12-, and 24-h duration are 27.7 mm (10.9 in.), 34.5 mm (13.6 in.), 41.6 mm (16.4 in.), respectively.

Fig. 10 shows the probability of economic damage if flash flooding occurs in each census tract in Harris County. For a 6-h extreme rainfall event, a 40% to 60% chance of economic damage is predicted for the majority of the county's communities. The chance of economic damage increases substantially if the rainfall events lasts for 12 and 24 h, as indicated by the darker areas. A few communities have a more than 80% chance of economic damages if such storm events occur. These communities tend to have a high percentage of impervious land cover and/or are built in low-lying areas where floodwater can easily accumulate.

Fig. 11 shows the dollar amount of economic damage (in 2019 prices) estimated by coupling the economic damage probability and the expected economic damage amount from model's continuous part. For a 6-h extreme rainfall event, the economic damage remains at $10,000 or less. However, it rises dramatically across communities when rainfall and duration increase, shown as darker for 12- and 24-h events. A few communities may expect millions of dollars of economic damages for such long and intense storms. These communities are generally located in the central Houston area, which has high home values and is highly urbanized.

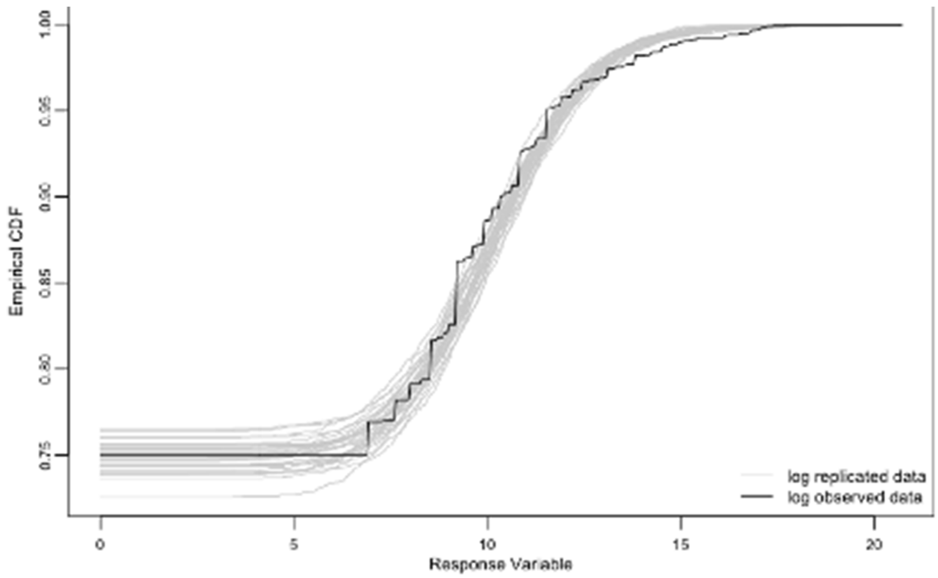

Fig. 9. Posterior predictive check for the final model.

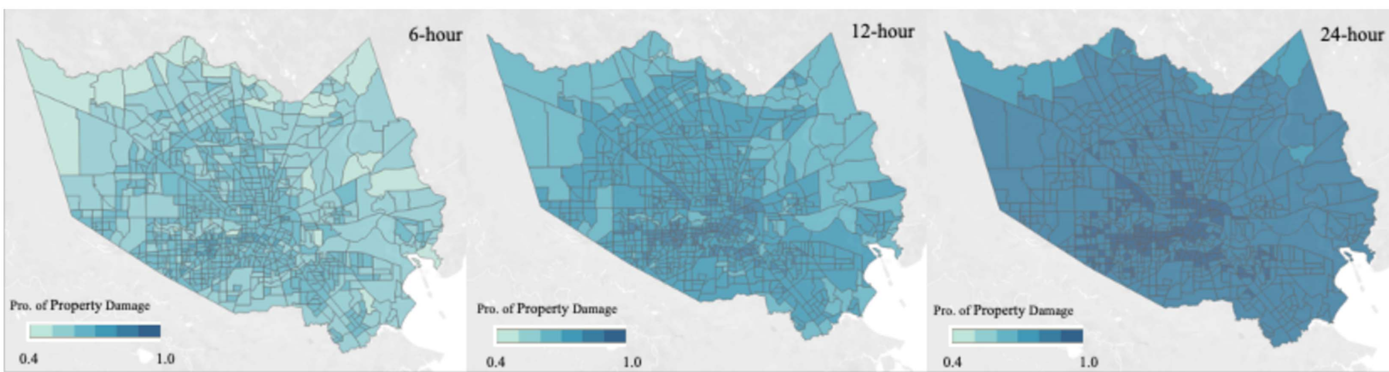

Fig. 10. Probability of economic damage in extreme rainfall events. (© OpenStreetMap contributors.)

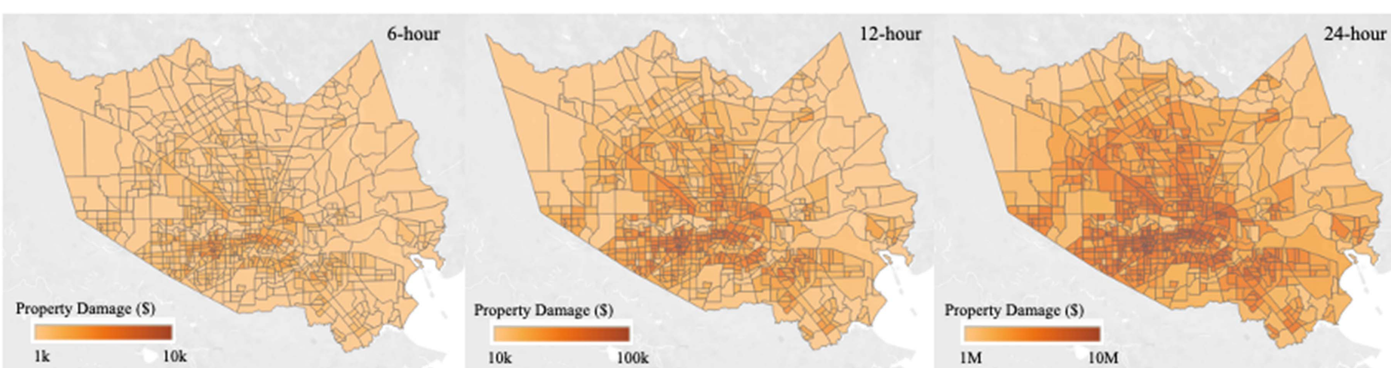

Fig. 11. Dollar amount of economic damage in extreme rainfall events. (© OpenStreetMap contributors.)

Summary, Conclusions, and Future Work

The occurrence of economic damage was found to be a zero-inflated problem: 70% of flash flood events result in zero economic damage based on observation of a 15-year period (2005–2019) in Texas, whereas the nonzero part conforms to a log-normal distribution. A two-part mixed-effect model was developed in this study to predict zero-inflated economic damage at the census tract scale. The model indicates that flash flood events with higher levels of precipitation or occurring in areas with higher percentages of impervious surfaces have a greater chance of resulting in economic damages. It also indicates that the effect of ground slope on increasing flash flood damage at the sub-city scale may be reversed given that floodwaters tend to accumulate in flatter downstream areas. These results suggest that flash flood damage cannot be explained by a single factor. Instead, it is explained by the combination of three fundamental issues: how soon runoff collection takes place, floodwater flow velocity, and floodwater depth.

Use of the developed model was demonstrated in an application to estimate the probability of occurrence and dollar amount of economic damage from extreme rainfall–induced flash flood events in Harris County, Texas. The developed model can provide insights for regional and state authorities in planning and prioritizing of flash flood mitigation projects.

Model limitations and future work to address them are as follows. One, since this was an empirical study, one must be very careful in extrapolating its results to areas beyond Texas. Future studies may develop other models for other regions in the United States with local influencing factors considered. Two, the occurrence of damage in flash flooding is a complex process with potentially many influencing variables. Future work might identify a more detailed variant than the county to account for hidden factors in the model's random effect component. Third, although the model was evaluated and its applicability demonstrated, further validation using separate data would be beneficial to confirm its predictive accuracy. Finally, the model can be integrated into simulation decision support systems to inform the planning of flash flood mitigation and safety projects.

Appendix. Random Effect Intercept

County	Intercept
Anderson	9.380956
Andrews	7.32277
Angelina	9.2835
Aransas	8.763674
Archer	8.763674
Armstrong	8.763674
Atascosa	8.90862
Austin	9.198015
Bailey	8.763674
Bandera	9.510173
Bastrop	9.38471
Baylor	8.763674
Bee	9.642205
Bell	9.043998
Bexar	8.937398
Blanco	10.22549
Borden	7.070281
Bosque	8.796952
Bowie	8.592224
Brazoria	8.727681
Brazos	8.115521
Brewster	7.529803
Brooks	8.864337
Brown	8.972698
Burleson	8.840071
Burnet	8.763674
Caldwell	8.763674
Calhoun	8.763674
Callahan	8.763674
Cameron	9.754217
Camp	8.763674
Carson	8.763674
Cass	9.021529
Chambers	7.544593
Cherokee	8.763674
Clay	8.763674
Coke	8.763674
Coleman	8.763674
Collin	8.005075
Collingsworth	8.763674
Colorado	9.647803
Comal	8.749377
Comanche	9.003894
Concho	9.365725
Cooke	9.378267
Coryell	8.990429
Cottle	8.763674
Crane	8.763674
Crockett	8.763674
Crosby	9.276274
Culberson	8.08386
Dallas	8.600488
Dawson	7.962285
De Witt	8.763674
Deaf Smith	9.257984
Delta	8.763674
Denton	7.66515
Dimmit	8.763674
Donley	8.763674
Duval	8.763674
Eastland	8.735558
Ector	9.006931
Edwards	8.766531
El Paso	10.371396
Ellis	9.853416
Erath	9.3996
Falls	8.442595
Fannin	7.89387
Fayette	8.102733
Fisher	8.763674
Floyd	9.341472
Foard	8.763674
Fort Bend	8.737717
Franklin	8.763674
Freestone	8.699667
Frio	8.763674
Gaines	7.12738
Galveston	7.921484
Garza	9.45351
Gillespie	8.763674
Glasscock	8.763674
Goliad	8.763674
Gonzales	8.561258
Gray	8.763674
Grayson	9.494468
Gregg	9.354148
Grimes	8.351476
Guadalupe	9.3983
Hale	9.928009
Hamilton	11.261738
Hansford	8.763674
Hardeman	8.763674
Hardin	7.517398
Harris	9.069818
Harrison	9.126006
Hartley	8.763674
Haskell	8.763674
Hays	9.587365
Hemphill	8.763674
Henderson	8.399621
Hidalgo	10.782907
Hill	8.666045

Appendix. (*Continued.*)

County	Intercept
Hockley	9.596809
Hood	9.073847
Hopkins	8.296038
Houston	8.87229
Howard	7.891936
Hudspeth	8.763674
Hunt	8.579403
Hutchinson	8.763674
Irion	8.763674
Jack	8.39682
Jackson	8.489897
Jasper	7.686141
Jeff Davis	8.763674
Jefferson	7.746858
Jim Hogg	8.447321
Jim Wells	9.817957
Johnson	7.618314
Jones	8.763674
Karnes	8.763674
Kaufman	8.407157
Kendall	8.763674
Kerr	9.051575
Kimble	8.763674
King	8.763674
Kinney	9.269886
Kleberg	9.112436
La Salle	9.101595
Lamar	8.294216
Lampasas	8.383396
Lavaca	8.763674
Lee	8.763674
Leon	8.921879
Liberty	8.145533
Limestone	7.664471
Lipscomb	8.763674
Live Oak	9.046656
Llano	9.389816
Loving	7.005927
Lubbock	10.86478
Lynn	9.216361
Madison	8.673184
Marion	9.085162
Martin	7.472931
Mason	8.763674
Matagorda	8.098542
Maverick	10.002265
Mcculloch	8.471851
Mclennan	7.803653
Mcmullen	8.92191
Medina	9.464084
Menard	8.763674
Midland	8.413078
Milam	8.680758
Mills	8.763674
Mitchell	8.194534
Montague	8.987995
Montague	8.125462
Montgomery	8.92191
Moore	9.304132
Morris	8.653933
Nacogdoches	8.654602
Navarro	9.003734
Newton	8.185618
Nolan	8.763674
Nueces	9.809476
Ochiltree	8.763674
Oldham	8.763674

Appendix. (*Continued.*)

County	Intercept
Orange	8.233246
Palo Pinto	8.005903
Panola	8.763674
Parker	8.207132
Parmer	8.763674
Pecos	7.228842
Polk	8.918924
Potter	9.839518
Presidio	8.763674
Rains	8.468775
Randall	9.52662
Reagan	7.787388
Real	8.709842
Red River	8.100367
Reeves	7.90682
Refugio	8.763674
Roberts	8.763674
Robertson	8.235995
Rockwall	7.978919
Runnels	8.763674
Rusk	8.942039
Sabine	8.859321
San Augustine	9.779242
San Jacinto	8.291942
San Patricio	9.620295
San Saba	8.763674
Schleicher	8.763674
Scurry	7.307483
Shackelford	8.215698
Shelby	8.763674
Smith	9.044134
Somervell	8.991642
Starr	8.876768
Stephens	8.763674
Sterling	8.763674
Sutton	8.763674
Swisher	8.763674
Tarrant	8.255634
Taylor	8.56278
Terrell	8.763674
Terry	9.48456
Throckmorton	8.763674
Titus	9.344153
Tom Green	8.763674
Travis	10.924512
Trinity	8.763674
Tyler	7.824816
Upshur	8.725422
Upton	7.512817
Uvalde	9.155334
Val Verde	8.826047
Van Zandt	8.520482
Victoria	9.573463
Walker	8.137177
Waller	9.33142
Ward	8.195995
Washington	8.132443
Webb	9.386166
Wharton	7.909092
Wheeler	8.444792
Wichita	10.669928
Wilbarger	8.763674
Willacy	9.821527
Williamson	8.550515
Wilson	9.779524
Winkler	7.802466
Wise	8.718033

Appendix. (*Continued.*)

County	Intercept
Wood	8.173361
Young	8.551549
Zapata	8.20151
Zavala	8.763674

Data Availability Statement

The data that support the findings of this study are available on request from the corresponding author, Shi Chang. The data will be publicly available at a future date with the completion and assessment of the NSF Project.

Acknowledgments

This material is based on work supported by the National Science Foundation under Grant No. 1931301. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Aalberts, B., and R. Hoyt. 2000. "Appraisers and toxic mold: Legal and valuation issues." *J. Real Estate Portfolio Manage.* 6 (2): 203–215. https://doi.org/10.1080/10835547.2000.12089605.

Ahmadalipour, A., and H. Moradkhani. 2019. "A data-driven analysis of flash flood hazard, fatalities, and damages over the CONUS during 1996–2017." *J. Hydrol.* 578 (Nov): 124106. https://doi.org/10.1016/j.jhydrol.2019.124106.

Akaike, H. 1974. "A new look at the statistical model identification." *IEEE Trans. Autom. Control* 19 (6): 716–723. https://doi.org/10.1109/TAC.1974.1100705.

Alipour, A., A. Ahmadalipour, P. Abbaszadeh, and H. Moradkhani. 2020. "Leveraging machine learning for predicting flash flood damage in the Southeast US." *Environ. Res. Lett.* 15 (2): 024011. https://doi.org/10.1088/1748-9326/ab6edd.

Ashley, S. T., and W. S. Ashley. 2008. "Flood fatalities in the United States." *J. Appl. Meteorol. Climatol.* 47 (3): 805–818. https://doi.org/10.1175/2007JAMC1611.1.

Baker, V. R. 1975. "Flood hazards along the Balcones escarpment in central Texas: Alternative approaches to their recognition, mapping, and management." In *Virtual landscapes of Texas*. Austin, TX: Univ. of Texas.

Baradaranshoraka, M., J.-P. Pinelli, K. Gurley, M. Zhao, X. Peng, and A. Paleo-Torres. 2019. "Characterization of coastal flood damage states for residential buildings." *ASCE-ASME J. Risk Uncertainty Eng. Syst., Part A: Civ. Eng.* 5 (1): 04019001. https://doi.org/10.1061/AJRUA6.0001006.

Boulton, A. J., and A. Williford. 2018. "Analyzing skewed continuous outcomes with many zeros: A tutorial for social work and youth prevention science researchers." *J. Soc. Social Work Res.* 9 (4): 721–740. https://doi.org/10.1086/701235.

Brandt, M., C. Brown, J. Burkhart, N. Burton, J. Cox-Ganser, S. Damon, H. Falk, S. Fridkin, P. Garbe, and M. McGeehin. 2006. "Mold prevention strategies and possible health effects in the aftermath of hurricanes and major floods." *Morbidity Mortality Weekly Rep.: Recommendations Rep.* 55 (8): 1883–1889. https://doi.org/10.1289/ehp.9258.

Burger, M., F. Van Oort, and G. J. Linders. 2009. "On the specification of the gravity model of trade: Zeros, excess zeros and zero-inflated estimation." *Spatial Econ. Anal.* 4 (2): 167–190. https://doi.org/10.1080/17421770902834327.

Burnett, J. 2008. *Flash floods in Texas*. College Station, TX: Texas A&M University Press.

Caran, S. C., and V. R. Baker. 1986. "Flooding along the Balcones escarpment, central Texas." In *The Balcones escarpment—geology, hydrology, ecology and social development in central Texas*, 1–14. Boulder, CO: Geological Society of America.

Chang, H., et al. 2021. "Assessment of urban flood vulnerability using the social-ecological-technological systems framework in six US cities." *Sustainable Cities Soc.* 68 (May): 102786. https://doi.org/10.1016/j.scs.2021.102786.

Chang, S., R. S. Wilkho, N. Gharaibeh, G. Sansom, M. Meyer, F. Olivera, and L. Zou. 2023. "Environmental, climatic, and situational factors influencing the probability of fatality or injury occurrence in flash flooding: A rare event logistic regression predictive model." *Nat. Hazards* 116 (3): 3957–3978. https://doi.org/10.1007/s11069-023-05845-x.

Chew, G. L., J. Wilson, F. A. Rabito, F. Grimsley, S. Iqbal, T. Reponen, M. L. Muilenberg, P. S. Thorne, D. G. Dearborn, and R. L. Morley. 2006. "Mold and endotoxin levels in the aftermath of Hurricane Katrina: A pilot project of homes in New Orleans undergoing renovation." *Environ. Health Perspect.* 114 (12): 1883–1889. https://doi.org/10.1289/ehp.9258.

De Moel, H., B. Jongman, H. Kreibich, B. Merz, E. Penning-Rowsell, and P. J. Ward. 2015. "Flood risk assessments at different spatial scales." *Mitigation Adapt. Strategies Global Change* 20 (Aug): 865–890. https://doi.org/10.1007/s11027-015-9654-z.

Downton, M. W., and R. A. Pielke. 2005. "How accurate are disaster loss data? The case of US flood damage." *Nat. Hazards* 35 (2): 211–228. https://doi.org/10.1007/s11069-004-4808-4.

Gutenson, J. L., A. A. Oubeidillah, A. N. S. Ernest, L. Zhu, X. Zhang, and S. T. Sadeghi. 2017. "Investigating uncertainty in developing regional building inventories for flood damage prediction." *Nat. Hazards Rev.* 18 (3): 04016013. https://doi.org/10.1061/(ASCE)NH.1527-6996.0000240.

Hartig, F., and M. F. Hartig. 2017. "DHARMa: Residual diagnostics for hierarchical (multi-level/mixed regression models." Accessed September 8, 2022. https://mran.microsoft.com/snapshot/2018-04-06/web/packages/DHARMa/vignettes/DHARMa.html.

Hsu, A. C., and S. C. Liu. 2008. "The hurdle models choice between truncated normal and lognormal." *Appl. Econ.* 40 (2): 201–207. https://doi.org/10.1080/00036840600749581.

Karagiorgos, K., T. Thaler, J. Hübl, F. Maris, and S. Fuchs. 2016. "Multi-vulnerability analysis for flash flood risk management." Supplement, *Nat. Hazards* 82 (S1): 63–87. https://doi.org/10.1007/s11069-016-2296-y.

Karamouz, M., M. Fereshtehpour, F. Ahmadvand, and Z. Zahmatkesh. 2016. "Coastal flood damage estimator: An alternative to FEMA's HAZUS platform." *J. Irrig. Drain. Eng.* 142 (6): 04016016. https://doi.org/10.1061/(ASCE)IR.1943-4774.0001017.

Khajehei, S., A. Ahmadalipour, W. Shao, and H. Moradkhani. 2020. "A place-based assessment of flash flood hazard and vulnerability in the contiguous United States." *Sci. Rep.* 10 (1): 1–12. https://doi.org/10.1038/s41598-019-57349-z.

Kohonen, T. 1995. "Learning vector quantization." In *Self-organizing maps*, 175–189. Berlin: Springer.

Kousky, C., M. Palim, and Y. Pan. 2020. "Flood damage and mortgage credit risk: A case study of Hurricane Harvey." Supplement, *J. Hous. Res.* 29 (S1): 86–120. https://doi.org/10.1080/10527001.2020.1840131.

Lachenbruch, P. A. 2002. "Analysis of data with excess zeros." *Stat. Methods Med. Res.* 11 (4): 297–302. https://doi.org/10.1191/0962280202sm289ra.

Meinshausen, N., and G. Ridgeway. 2006. "Quantile regression forests." *J. Mach. Learn. Res.* 7 (6): 983–999.

Merz, B., H. Kreibich, R. Schwarze, and A. Thieken. 2010. "Review article 'Assessment of economic flood damage'." *Nat. Hazards Earth Syst. Sci.* 10 (8): 1697–1724. https://doi.org/10.5194/nhess-10-1697-2010.

Milanesi, L., and M. Pilotti. 2021. "Coupling flood propagation modeling and building collapse in flash flood studies." *J. Hydraul. Eng.* 147 (12): 04021047. https://doi.org/10.1061/(ASCE)HY.1943-7900.0001941.

Mokhtari, F., S. Soltani, and S. A. Mousavi. 2017. "Assessment of flood damage on humans, infrastructure, and agriculture in the Ghamsar Watershed using HEC-FIA software." *Nat. Hazards Rev.* 18 (3): 04017006. https://doi.org/10.1061/(ASCE)NH.1527-6996.0000248.

NWS (National Weather Service). 2019. "NWS preliminary US flood fatality statistics." Accessed September 1, 2020. https://www.weather.gov/arx/usflood.

Olsen, M. K., and J. L. Schafer. 2001. “A two-part random-effects model for semicontinuous longitudinal data.” *J. Am. Stat. Assoc.* 96 (454): 730–745. https://doi.org/10.1198/016214501753168389.

Pan, Z., and D. Y. Lin. 2005. “Goodness-of-fit methods for generalized linear mixed models.” *Biometrics* 61 (4): 1000–1009. https://doi.org/10.1111/j.1541-0420.2005.00365.x.

Pinheiro, J. C., and D. M. Bates. 1995. “Approximations to the log-likelihood function in the nonlinear mixed-effects model.” *J. Comput. Graphical Stat.* 4 (1): 12–35.

Pistrika, A., G. Tsakiris, and I. Nalbantis. 2014. “Flood depth-damage functions for built environment.” *Environ. Processes* 1 (4): 553–572. https://doi.org/10.1007/s40710-014-0038-2.

Rizopoulos, D., G. Verbeke, and G. Molenberghs. 2010. “Multiple-imputation-based residuals and diagnostic plots for joint models of longitudinal and survival outcomes.” *Biometrics* 66 (1): 20–29. https://doi.org/10.1111/j.1541-0420.2009.01273.x.

Rose, C. E., S. W. Martin, K. A. Wannemuehler, and B. D. Plikaytis. 2006. “On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data.” *J. Biopharm. Stat.* 16 (4): 463–481. https://doi.org/10.1080/10543400600719384.

Schroeder, A. J., et al. 2016. “The development of a flash flood severity index.” *J. Hydrol.* 541 (Oct): 523–532. https://doi.org/10.1016/j.jhydrol.2016.04.005.

Schwarz, G. 1978. “Estimating the dimension of a model.” *Ann. Stat.* 6 (2): 461–464. https://doi.org/10.1214/aos/1176344136.

Sharif, H. O., M. M. Hossain, T. Jackson, and S. Bin-Shafique. 2012. “Person-place-time analysis of vehicle fatalities caused by flash floods in Texas.” *Geomatics Nat. Hazards Risk* 3 (4): 311–323. https://doi.org/10.1080/19475705.2011.615343.

Sharif, H. O., T. L. Jackson, M. M. Hossain, and D. Zane. 2015. “Analysis of flood fatalities in Texas.” *Nat. Hazards Rev.* 16 (1): 04014016. https://doi.org/10.1061/(ASCE)NH.1527-6996.0000145.

Su, L., B. D. Tom, and V. T. Farewell. 2009. “Bias in 2-part mixed models for longitudinal semicontinuous data.” *Biostatistics* 10 (2): 374–389. https://doi.org/10.1093/biostatistics/kxn044.

Taguchi, R., M. Tanoue, D. Yamazaki, and Y. Hirabayashi. 2022. “Global-scale assessment of economic losses caused by flood-related business interruption.” *Water* 14 (6): 967. https://doi.org/10.3390/w14060967.

Terti, G., I. Ruin, J. J. Gourley, P. Kirstetter, Z. Flamig, J. Blanchet, A. Arthur, and S. Anquetin. 2019. “Toward probabilistic prediction of flash flood human impacts.” *Risk Anal.* 39 (1): 140–161. https://doi.org/10.1111/risa.12921.

Tooze, J. A., G. K. Grunwald, and R. H. Jones. 2002. “Analysis of repeated measures data with clumping at zero.” *Stat. Methods Med. Res.* 11 (4): 341–355. https://doi.org/10.1191/0962280202sm291ra.

Vinet, F. 2008. “Geographical analysis of damage due to flash floods in southern France: The cases of 12–13 November 1999 and 8–9 September 2002.” *Appl. Geogr.* 28 (4): 323–336. https://doi.org/10.1016/j.apgeog.2008.02.007.

Vuichard, D., and M. Zimmermann. 1986. “The Langmoche flash-flood, Khumbu Himal, Nepal.” *Mt. Res. Dev.* 6 (1): 90–94. https://doi.org/10.2307/3673345.

Waagepetersen, R. 2006. “A simulation-based goodness-of-fit test for random effects in generalized linear mixed models.” *Scand. J. Stat.* 33 (4): 721–731. https://doi.org/10.1111/j.1467-9469.2006.00504.x.

Wilkho, R. S., N. G. Gharaibeh, S. Chang, and L. Zou. 2023. “FF-IR: An information retrieval system for flash flood events developed by integrating public-domain data and machine learning.” *Environ. Modell. Software* 107: 105734. https://doi.org/10.1016/j.envsoft.2023.105734.

Zahran, S., S. D. Brody, W. G. Peacock, A. Vedlitz, and H. Grover. 2008. “Social vulnerability and the natural and built environment: A model of flood casualties in Texas.” *Disasters* 32 (4): 537–560. https://doi.org/10.1111/j.1467-7717.2008.01054.x.