Natural Hazards (2023) 116:3957–3978
https://doi.org/10.1007/s11069-023-05845-x

ORIGINAL PAPER

Environmental, climatic, and situational factors influencing the probability of fatality or injury occurrence in flash flooding: a rare event logistic regression predictive model

Shi Chang[1] · Rohan Singh Wilkho[1] · Nasir Gharaibeh[1] · Garett Sansom[2] · Michelle Meyer[3] · Francisco Olivera[1] · Lei Zou[4]

Received: 5 July 2022 / Accepted: 30 January 2023 / Published online: 20 February 2023

Abstract
Flash flooding is considered one of the most lethal natural hazards in the USA as measured by the ratio of fatalities to people affected. However, the occurrence of injuries and fatalities during flash flooding was found to be rare (about 2% occurrence rate) based on our analysis of 6,065 flash flood events that occurred in Texas over a 15-year period (2005 to 2019). This article identifies climatic, environmental, and situational factors that affect the occurrence of fatalities and injuries in flash flood events and provides a predictive model to estimate the likelihood of these occurrences. Due to the highly imbalanced dataset, three forms of logit models were investigated to achieve unbiased estimations of the model coefficients. The rare event logistic regression (Relogit) model was found to be the most suitable model. The model considers ten independent situational, climatic, and environmental variables that could affect human safety in flash flood events. Vehicle-related activities during flash flooding exhibited the greatest effect on the probability of human harm occurrence, followed by the event's time (daytime vs. nighttime), precipitation amount, location with respect to the flash flood alley, median age of structures in the community, low water crossing density, and event duration. The application of the developed model as a simulation tool for informing flash flood mitigation planning was demonstrated in two study cases in Texas.

Keywords Flash flooding · Fatality and injury · Rare event · Logistic regression · Hazard impact

✉ Shi Chang
changs18@tamu.edu

[1] Zachry Department of Civil & Environmental Engineering, Texas A&M University, 3136 TAMU, College Station, TX 77843-3136, USA

[2] Department of Environmental and Occupational Health, Texas A&M University, College Station, TX, USA

[3] Department of Landscape Architecture and Urban Planning, Texas A&M University, College Station, TX, USA

[4] Department of Geography, Texas A&M University, College Station, TX, USA

1 Introduction

1.1 Background

Flash flooding, usually triggered by heavy rainfall associated with severe thunderstorms, hurricanes or tropical storms, is considered one of the most lethal natural hazards in USA (Ahmadalipour and Moradkhani 2019; Terti et al. 2017; Kelsch et al. 2001) define flash flooding as a phenomenon in which the rainfall-runoff hydrologic processes occur on the same temporal and spatial scales as the triggering intense precipitation. Rainfall-induced flash floods are characterized by their rapid onset (usually under 6 h) and small spatial scale (Terti et al. 2015; Ahmadalipour and Moradkhani 2019) found that the average duration of flash floods between 1996 and 2017 was about 3.5 h, but in rare cases, they lasted for two days. The fact that they hit with little lead time for warning and are of such velocity and force makes flash floods one of the most unsafe types of natural hazard when measured by the ratio of fatalities to people affected. Over the past two decades, data reported by the National Weather Service (NWS) indicate that approximately 72% of all flood-related fatalities, 72% of all flood-related injuries, and 52% of flood-related economic losses in the USA are attributed to flash flooding (NWS 2019). Figure 1 indicates that the number of fatalities from flash floods largely exceeded that from other flood types in the USA during the period 2000–2019.

Texas has the highest number of flash flood fatalities among all states. From 1959 to 2008, there were 840 flood fatalities reported in Texas, 442 of which occurred in flash flood events (52.6%) (Shah et al. 2017). These flash flood fatalities occurred largely in Central Texas "Hill Country" along the Balcones Escarpment, which lies between the Edwards Plateau and the coastal plain (Sharif et al. 2012, 2015). This region (known as "the Flash Flood Alley) contains 44 counties and is characterized by steep terrain, shallow soil, and high rainfall rates (Baker 1975; Caran and Baker 1986).

Although rainfall patterns are complex and difficult to predict, climate models suggest that precipitation will likely occur more frequently and with greater intensity in the future in many regions around the globe (Wang and Zhang 2008; Sharif et al. 2012; Wobus et al. 2014). Since 1991, extreme rainfall events in the USA, defined as the heaviest 1% of daily

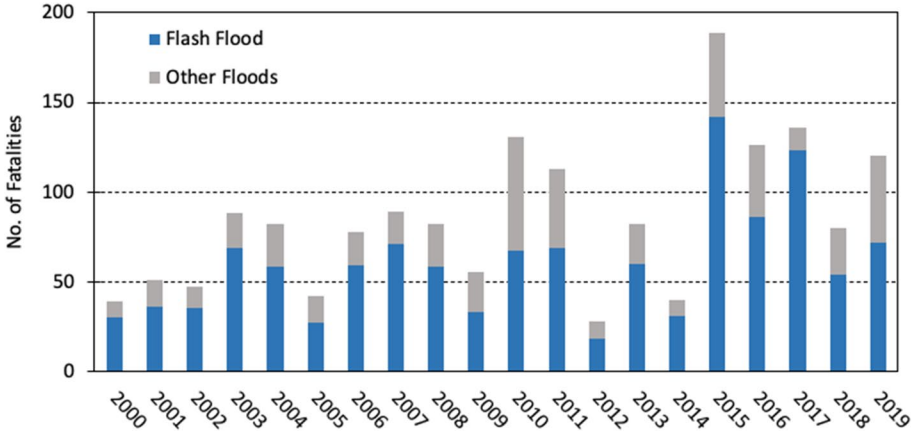

Fig. 1 Number of flood fatalities in the USA (2000–2019)

events, have increased in both frequency and magnitude (Walsh et al. 2014). Urbanization, which causes changes to land use, removal of vegetation and soil, grading of land surface, and construction of drainage networks, affect flood peak discharge and volume, and consequently human safety (Anderson 1970; Bailey 1989; Konrad 2002). In this study, we analyze the role that rainfall patterns and built and social environment characteristics play in predicting the occurrence of fatalities and injuries during flash flood events. This work informs hazard mitigation planning to reduce impacts from flash flooding (Masterson et al. 2014).

1.2 Prior work and knowledge gaps

Prior studies suggest that flash flood human fatalities and injuries are influenced by behavioral factors, the surrounding natural and built environments, and storm characteristics. Next, we discuss these groups of factors and identify the knowledge gaps that this paper addresses.

Behavioral factors refer to actions willingly taken by individuals, leading to injury or death in flash flood events. These actions are often rooted in beliefs that flash flooding would pose no or little risk to human life (Benight et al. 2007; Hamilton et al. 2016). Almost half of people driving vehicles on roadways who enter floodwater reported that they did not think it was unsafe to do so (Ruin et al. 2007). A study conducted by Sharif et al. (2015) revealed that of 616 flood fatalities in Texas, 471 (76%) were vehicle-related, such as people driving over what may have seemed like a low-water crossing. About 16.5% (102 individuals) died after walking into flood waters to cross an area that was flooded. In a separate study, Terti et al. (2017) found that more than 60% of the reported flash flood fatalities were related to vehicles and involved men. Ashley and Ashley (2008) found that the age of those who die in flash flooding is either between 10 and 29 years or above 60 years. Similar findings about human behavior around floodwaters have been reported in various parts of the world, such as Greece (Diakakis 2020) and Australia (Hamilton et al. 2016). The identification of these social and behavioral factors provides guidance for non-structural interventions (such as public awareness campaigns) to curb risky behaviors during flood situations (Lindell and Perry 1992).

Environmental and situational factor—(the focus of this study)—refer to the characteristics of the flood area and community, human interaction with vehicles, and the triggering natural hazard. Understanding these factors is important for planning flash flood mitigation and safety strategies.

Zahran et al. (2008) and Terti et al. (2019) developed flood casualty predictive models at the county scale. Zahran et al. (2008) developed a zero-inflated negative binomial model to predict the odds of a flood casualty (considering all flood types combined) from the impacts of hurricanes, tropical storms, and tornados. Terti et al. (2019) trained a random forest model to predict the likelihood of vehicle-related fatal incidents in flash flood events at the county scale using data from Texas and Oklahoma. Both of these models indicate an association between the odds of flood casualty occurrence and the event precipitation amount, event duration, unit peak discharge, population density, and social vulnerability. Zahran et al. (2008) model considers all flood types combined, its applicability to flash flooding may be limited. Terti et al. (2019) county-based model has the advantage of being specific to flash flooding and therefore could be used in decision-making around flash flood threats at the county level or larger scales (e.g., NWS flash flood warnings).

To address these limitations, we provide a new model for identifying climatic, environmental, and situational factors that affect the occurrence of human fatalities and injuries in flash flood events at a finer spatial resolution. The model uses these influencing factors to predict the probability that a flash flood event will lead to at least one fatality or injury at the census tract scale. Key advantages of the proposed model compared to existing models include:

- The model uses a spatial scale consistent with the local nature of flash floods. We use the census tract as the analysis unit to capture temporal and spatial complexities at the scene of the incident. The census tract (delineated by the US Census Bureau) is a relatively permanent subdivision of a county with a population size between 1200 and 8000 people (the optimum size is 4000 people) (U.S. Census 2022). In urban areas, census tracts are relatively small geographic areas, although in rural areas they can be larger. Census tracts are commonly used as a proxy for neighborhood in many studies. Since much hazard planning is completed at the county scale, intracounty geographic scales like census tracts allow practitioners to address vulnerabilities within their jurisdiction (Lindell et al. 2006).
- The model accounts for the rare occurrence of flash flood fatalities and injuries. As it will be shown later in this paper, flash flood events that resulted in human fatalities or injuries are rare compared to the number of events that did not result in human harm (i.e., the data are highly imbalanced). It is difficult to obtain unbiased statistical inferences from these data using conventional statistical methods. Therefore, we use a "rare-event" modeling technique, commonly used in economics, social anthropology, and natural hazard and earth sciences (Sanders et al. 2002; Clauset and Woodard 2013; King and Zeng 2001a, b; Guns and Vanacker 2012).

1.3 Research objective

The objective of this study is twofold: (a) identify the climatic, environmental and situational factors that affect the occurrence of human harm (fatalities and injuries) in flash flood events, and (b) develop a predictive model to estimate the likelihood of human harm occurrence in flash flood events. The developed model has the potential to enhance public safety by informing the planning of structural and non-structural flood mitigation and risk reduction projects at the local community scale (e.g., neighborhood, town, sub-city).

1.4 Paper organization

The remainder of this paper is organized into six sections. Section 2 describes the data used in the study and the process of assembling these data from disparate sources. Section 3 describes the modeling methodology and philosophy of a binary logistic model applied to rare events. The logit model evaluation methods and the results of different logistic models are discussed in Sect. 4. The interpretation of the final model coefficient and model performance evaluation are summarized in Sect. 5. Section 6 provides two case studies where the model is used to predict the likelihood of human harm occurrence in hypothetical flash flood events. Section 7 discusses the key advantages, contributions, and the limitations of the proposed model and provides explanations for the factors that influence human safety during flash flooding. Section 8 summarizes the study conclusions and recommendations.

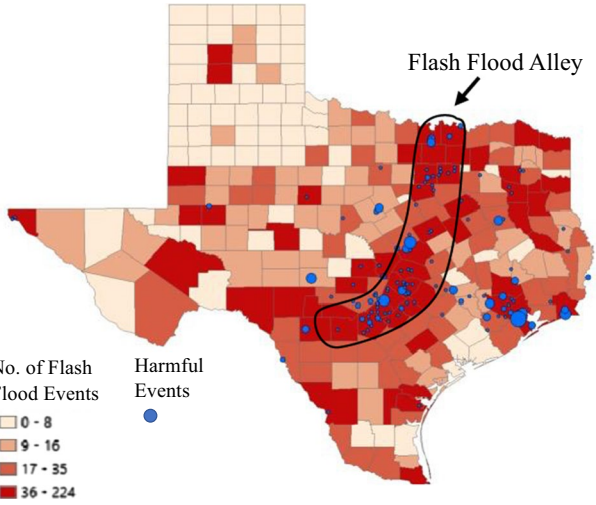

Fig. 2 Number of flash flood events and associated fatalities and injuries in Texas (2005–2019). (Dot size is proportional to the number of fatalities and injuries in the flash flood event. Approximate boundaries of the flash flood alley are sketched in black.)

2 Data

2.1 Flash flood event data

A total of 6065 flash flood events that occurred in Texas over a 15-year period (2005–2019) are included in this study. These events were obtained from the National Oceanic and Atmospheric Administration (NOAA) Storm Events database (NWS 2022). 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 property damage, and/or disruption to commerce.

Human injury or death, called "human harm" in this study, occurred in 128 out of 6065 flash flood events. Therefore, the occurrence rate of human harm is only about 2%. Figure 2 shows the number of flash flood events in each Texas county and the relative number of fatalities and injuries. The blue dots represent events that resulted in human harm. The size of dot represents the relative number of human fatalities and injuries. The larger the size of the blue dot, the greater the number of human fatalities and injuries during that flash flood event. It can be seen that human harm was not only found in areas where flash flooding is more common (dark red counties), but also could happen in areas with occasional or infrequent flash flooding.

2.2 Factors influencing human safety during flash flood events

Based on previous studies of environmental and situational factors associated with human fatalities and injuries in flash flood events, we identified 15 candidate factors as model predictors. These factors, their data sources, and the rationale for considering them in this study are provided in Table 1. Collectively, these factors represent the external stimuli that could influence the likelihood of human harm occurrence during flash flood events.

The data on these factors were acquired from different publicly available datasets and platforms, which have diverse formats and structures. Therefore, it was necessary to

Table 1 Potential factors influencing human safety during flash flood events

Potential influencing factor	Data source	Rationale
Flash flood event duration, (hour)	NOAA storm events database	Rapid onset of flash flooding generates surprising and swift moving water. Longer event duration increases the amount of runoff and risk exposure.
Precipitation during Event, (inch)	Prism daily spatial climate dataset	Greater precipitation increases runoff amount and velocity.
Census tract average ground elevation *, (feet)	National elevation dataset	Flash floods are usually characterized by raging torrents after heavy rains that rapidly flood into lower-elevation areas.
Median age of built structures in the census tract in the event year, (year)	American community survey	The physical and natural environments of communities change over time. Older neighborhoods might have older infrastructure, but newer neighborhoods might have been built in high flood risk areas.
Census tract population density *, (people/sqmi)	American community survey	Densely populated areas, usually urban areas, tend to have highly developed land, more impervious surfaces, and higher runoff. Also, higher population density leads to greater number of people being exposed to the event.
Time of flood event (Binary variable: Night Event = 1 and Day Event = 0)	NOAA storm events database	Visibility affects people's ability to react to flood warnings, making nighttime flash flood events more dangerous than daytime events.
Type of triggering storm (Binary variable: Tropical Storm or Hurricane = 1, Other Storm Types = 0)	NOAA storm events database	Tropical storms and hurricanes tend to produce greater rainfall and stronger wind than thunderstorms and affect wider areas, they also offer warning time and potential evacuation affecting regular population movement patterns.
Vehicle-related incident (Binary variable: Vehicle-Related incident = 1, Non-Vehicle-Related Incident = 0)	NOAA storm events database	Flash flood fatalities and injuries tend to be related to vehicles, such as attempting to drive through floodwater, low water crossing, and flooded bridges.
Flash flood alley (Binary variable: Event located in Flash Flood Alley = 1, Event located outside Flash Flood Alley = 0)	NOAA storm events database	Past research indicates that communities in the Flash Flood Alley suffered a high number of fatalities and injuries in flash flood events.
Census tract developed impervious surface *, (%)	USGS national land cover database	During heavy rainfall, impervious surfaces reduce the amount of water that infiltrates into the ground and therefore increase the severity of flash floods.

Table 1 (continued)

Potential influencing factor	Data source	Rationale
Census tract bridge density *, (count/sq.mi)	TxDOT bridge inventory dataset	Flooded bridges increase the risk of human harm to pedestrians and vehicles.
Census tract low water crossing density *, (count/sq.mi)	Texas natural resources information system inventory of low water crossings	Drivers may underestimate the risk of driving through low water crossings during storms. As in bridges, flooded water crossings increase the risk of human harm to both pedestrians and vehicles.
Census tract road density *, (mi/sq.mi)	TxDOT roadway inventory dataset	Greater road density increases impervious surface area and could lead to more vehicle-related flash flood injuries and fatalities.
Census tract average ground slope *, (degree)	National hydrography dataset	Areas with steep ground slope could increase the runoff velocity, increasing the risk of water moving swiftly across communities.
Soil type (runoff potential: A– low; B– moderately low; C– moderately high; D- high)	National hydrography dataset	Different soil types have different abilities to absorb water and correspond to different levels of runoff potential during flooding.

*The year of the influencing factor is the closest or most recent available data from the sources

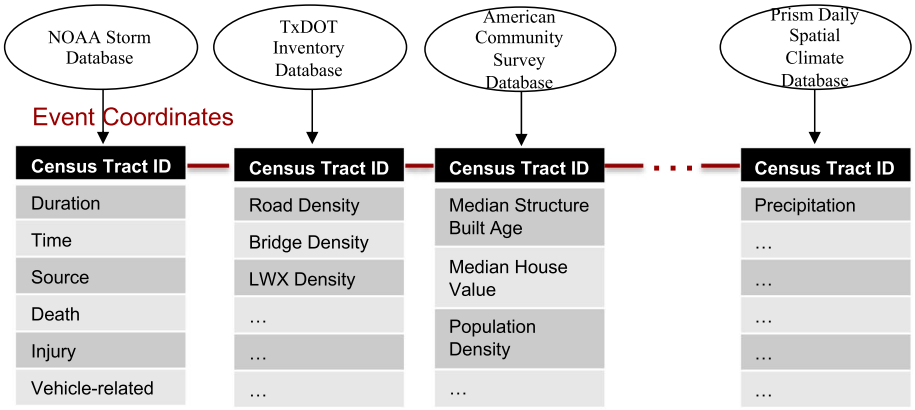

Fig. 3 Assembling of the study dataset

integrate these disparate data based on geographical coordinates (latitude and longitude), census tract, and year of the flash flood event. Google Earth Engine Python Application Programming Interface (API) and ESRI's ArcGIS software were used to implement data integration and geoprocessing. The final dataset used in this study included flash flood events and potential influencing factors at the census tract scale. The census tract is used to capture the characteristics of the natural, built, and social environments of the area affected by the flash flood event. The assembling of the study dataset is illustrated in Fig. 3.

3 Modeling approach

In this section, we present the theoretical basis of our rare-event modeling approach for predicting and explaining the occurrence of human harm in flash flood events. We adopted a logistic modeling approach over other more sophisticated modeling techniques (e.g., machine learning models) due to their enhanced transparency and interpretability. Logistic models allow the model users (including both practitioners and scientists) to contextualize the model's inputs and outputs and understand the mechanism of flash flooding safety through relatively simple mathematical formulas.

Logistic regression model (logit model) is a commonly used statistical method for predicting the probability of a binary outcome. However, conventional logistic regression can grossly underestimate the probability of rare events (Imbens 1992; Cosslett 1981; Lancaster and Imbens 1996; King and Zeng 2001a, b). Thus, appropriate statistical corrections must be applied carefully. This is a challenging problem because conventional logistic models do not always lead to a robust inference of controlling factors, as the results can be strongly sample dependent (Guns and Vanacker 2012).

3.1 Conventional logistic regression

In a logit model, a single outcome variable $Y_i (i = 1 \ldots n)$ is used to represent harmful event ($Y_i = 1$) and non-harmful event ($Y_i = 0$). It follows a Bernoulli probability function that takes on the value 1 with probability π_i and 0 with probability $1 - \pi_i$.

$$Y_i \sim Bernoulli\left(Y_i|\pi_i\right) \tag{1}$$

π_i varies over the observations as an inverse logistic function of a vector of influencing factors (X_i) and their coefficients (β):

$$\pi_i = \frac{1}{1+e^{-X_i\beta}} \tag{2}$$

The Bernoulli probability function is as follows:

$$P\left(\mathrm{Y_i}|\pi_\mathrm{i}\right) = \pi_\mathrm{i}^{\mathrm{Y_i}}\left(1-\pi_\mathrm{i}\right)^{1-\mathrm{Y_i}} \tag{3}$$

Assuming independence over the observations (i.e., the occurrence of flash flood event A has no effect on the occurrence of event B), it is common to use the maximum likelihood to estimate the parameters of the likelihood function:

$$L(\beta|y) = -\prod_{i=1}^{n} \pi_\mathrm{i}^{\mathrm{Y_i}}\left(1-\pi_\mathrm{i}\right)^{1-\mathrm{Y_i}} \tag{4}$$

The log-likelihood function is denoted as follows:

$$\begin{aligned} lnL(\beta|\mathrm{y}) &= \sum_{\{Y_i=1\}} \ln\left(\pi_i\right) + \sum_{\{Y_i=0\}} \ln\left(1-\pi_i\right) \\ &= -\sum_{\mathrm{i=1}}^{\mathrm{n}} \ln\left(1+\mathrm{e}^{(1-2\mathrm{Y_i})\mathrm{X_i}\beta}\right) \end{aligned} \tag{5}$$

Greene (1993) suggested using maximum likelihood estimation (MLE) analysis to find the value of β that gives the maximum value of this function, which is $\widehat{\beta}$. The estimated $\widehat{\beta}$ is considered consistent and asymptotically efficient when observations are randomly selected from the population. However, it is well known that MLE is only asymptotically unbiased and its estimators may be heavily biased when many covariates exist or highly correlate (Gao and Shen 2007). The bias could be exacerbated with rare events parameter estimation when a very small number of ones (Y = 1) exist in the observations (Leitgöb 2020). Moreover, conventional logistic regression strongly underestimates the $\pi_i = Pr\left(Y_i = 1|x_i\right)$ in rare events data in which the "ones" are more statistically informative than the "zeros" (Imbens 1992; Cosslett 1981; Lancaster and Imbens 1996; King and Zeng 2001a, b). This can be seen by analyzing the variance of the estimated $\widehat{\beta}$, as follows:

$$Var\left(\widehat{\beta}\right) = \left[\sum_{i=1}^{n} \pi_i\left(1-\pi_i\right)\boldsymbol{x_i'x_i}\right]^{-1} \tag{6}$$

where x_i' = the inverse of the influencing factors vector ($\boldsymbol{x_i}$).

In models that exhibit sufficient explanatory power, the term π_i is larger (and closer to 0.5) for $Y_i =$ lthan for $Y_i = 0$. Thus, additional "ones" yield a larger $\pi_i\left(1-\pi_i\right)$ and a smaller variance than additional zeros, making the model more informative.

3.2 Bias correction

While logistic regression is a powerful tool to predict a binary output, it could lead to strong biases in the coefficient estimates with heavily imbalanced data, like rare events

data, and result in widely varying predictions. In order to remedy the underestimates of probability for ones in rare events, King and Zeng (2001a, b) proposed a method of bias correction in logistic regression with finite sample for rare events such as war, vetoes, and epidemiological infections. The likelihood of those rare events is usually underestimated by conventional predictive models, like MLE logit model.

King and Zeng (2001a, b) developed a prior correction for the logit model for the intercept term β_0, which is statistically consistent:

$$\beta_0 = \widehat{\beta_0} - ln\left[\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)\right] \tag{7}$$

where τ = the true population fraction of events. $\bar{y}$ = the fraction of events in the sample. $\widehat{\beta_0}$ = uncorrected intercept term.

They also noted that the bias term can be simply subtracted from the estimated parameter, denoted as $\tilde{\beta}$,

$$\tilde{\beta} = \widehat{\beta} - bias\left(\widehat{\beta}\right) \tag{8}$$

Based on the analytical approximations from McCullagh and Nelder (1989), bias term in rare events can be written as,

$$bias\left(\widehat{\beta}\right) = \left(\boldsymbol{X'WX}\right)^{-1}\boldsymbol{X'}W\xi \tag{9}$$

where

$$\xi_i = 0.5Q_{ii}\left[\left(1+\omega_1\right)\widehat{\pi}_i - \omega_1\right]$$

$$\mathbf{Q} = \mathbf{X}\left(\boldsymbol{X'WX}\right)^{-1}\boldsymbol{X'}$$

$$\boldsymbol{W} = diag\left\{\widehat{\pi}_i\left(1-\widehat{\pi}_i\right)\omega_i\right\}$$

where $\omega_i$is the weight, computed from weighted log-likelihood function as follows:

$$\omega_0 = (1-\tau)/\left(1-\bar{y}\right)$$

$$\omega_1 = \tau/\bar{y}$$

$$\omega_i = \omega_1 Y_i + \omega_0\left(1-Y_i\right)$$

Computing this bias term involves solving a weighted least-square regression with X as the explanatory variable and ξ as the dependent variable with weight $\boldsymbol{W}$. Then, the computed $\tilde{\beta}$, as the bias corrected coefficient estimator, is used to calculate the probability function π_i, as follows:

$$\pi_i = \frac{1}{1+e^{-X_i\tilde{\beta}}} \tag{10}$$

This adjusted logit model with bias correction and weighting has been tested through multiple simulations to exhibit improvement over conventional MLE, especially when limited number of observations collected and event occurrence is less than 5%. This method of logistic regression with bias adjustment is referred as rare event logistic regression (Relogit).

The large sample size (a total of 6065 flash flood events) and low occurrence rate of harmful events (approximate 2%) make predicting human harm from flash flooding suitable for the Relogit method. In this study, Relogit is performed through the Zelig R package (Imai et al. 2008; Choirat et al. 2020).

3.3 Penalized maximum likelihood estimation

The King and Zeng (2001a, b) method is not the only statistical approach to adjusting for rare events. The bias of the MLE method in estimating parameter β can be expanded as:

$$Bias(\beta) = E\left(\widehat{\beta}\right) - \beta = \frac{B_1(\beta)}{n} + \frac{B_2(\beta)}{n^2} + \ldots \tag{11}$$

The common approaches to correct the bias in MLE are to remove the term $B_1(\beta)/n$ from the asymptotic bias (Cox and Hinkley 1979; Quenouille 1949, 1956). However, these approaches rely on the existence of the MLE estimators from the sample and then correct it afterwards. Firth (1993) noted that it is not uncommon that the MLE estimator is infinite in some samples, especially with small to medium sample size. This scenario is more likely to happen with linear logistic models for a binary response (Albert and Anderson 1984; Clogg et al. 1991).

In order to solve this problem, Firth's (1993) introduced a penalization parameter to the likelihood function that equal to the square root of the determinant of the information matrix $|I(\beta)|^{\frac{1}{2}}$. This correction scheme is equivalent to the Jeffery's invariant prior when the parameter is the canonical parameter of an exponential family (Jeffreys 1946). The penalized likelihood function can be written as:

$$L(\beta)^{*} = L(\beta) \cdot |\boldsymbol{I}(\beta)|^{\frac{1}{2}} \tag{12}$$

By taking the natural logarithm:

$$\ell^{*}(\beta) = \ell(\beta) + 0.5log|\boldsymbol{I}(\beta)| \tag{13}$$

where $I(\beta)$ denotes Fisher's information of the sample $\boldsymbol{I}(\beta) = \boldsymbol{X^T W X}$, where $\boldsymbol{W} = diag\left[\pi_i\left(1-\pi_i\right)\right]$. The information matrix also defined as the negative expected value of the first derivative of score function $U(\beta) = \sum_{i=1}^{n}\left(y_i - \pi_i\right)x_{ir} = 0$. Firth (1993) proposed a modification to the score function based on simple triangular geometry shown in Fig. 4. The first-order bias of $B_1(\beta)/n$ can be removed by shifting the score function by $I(\beta)B_1(\beta)/n$, where the gradient of $U(\beta)$ is given by $\partial U(\beta)/\partial\beta = -I(\beta)$.

The modified score function is:

$$U^{*}(\beta) = U(\beta) - I(\beta)B_1(\beta)/n \tag{14}$$

This bias-preventive approach offers a systematic corrective procedure applied to the score function instead of correcting it after it is estimated. When applied to a binary logit model, this approach is known as Firth's logistic regression or penalized

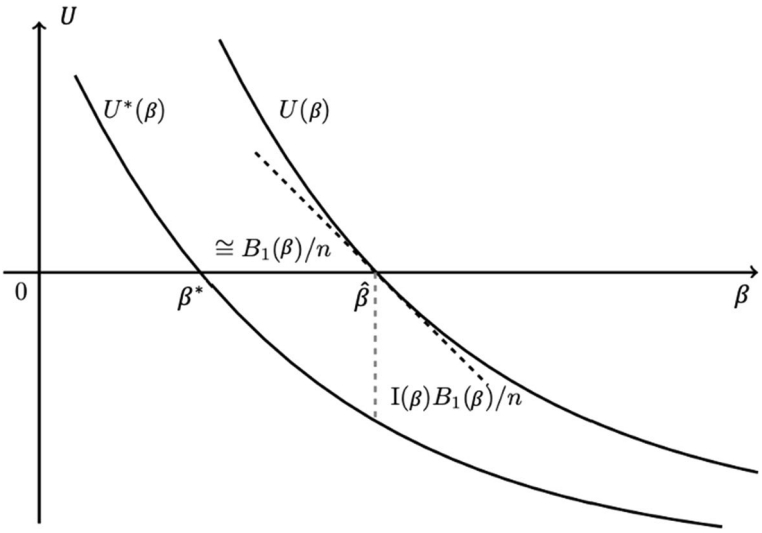

Fig. 4 Modification of the score function (Firth 1993)

maximum likelihood estimation (PMLE). Firth's PMLE approach has been proven to be consistently superior to conventional MLE for datasets with separation or small to medium sample size (Heinze and Schemper 2002; Bull et al. 2002). In this study, Firth's PMLE method was implemented using the logistf R package (Heinze et al. 2013). In the next section, we compare results of Firth's PMLE model to that of the conventional logistic regression model and King and Zeng's Relogit model to understand how the bias adjustment effects the parameter estimators (i.e., the model coefficients).

4 Training and evaluation of alternative models

In this section, we evaluate the conventional MLE, Firth's PMLE, and King and Zeng's Relogit models to determine the most robust model for predicting the likelihood of fatality or injury occurrence in flash flood events and for explaining the factors that can potentially affect human safety during flash flooding.

Table 2 Data splitting for training and testing

Attribute	Total dataset	Training dataset	Testing dataset
Total number of flash flood events	6065	4549	1516
Total number of harmful events	128	99	29
Harmful event occurrence rate (%)	2.11	2.18	1.91

4.1 Data splitting

The dataset in this study (containing 6065 flash flood events in Texas) was randomly split into training and testing datasets at a 75:25 ratio. Table 2 shows the distribution of harmful events for training and testing datasets. Logit models with MLE, Firth's PLME, and King and Zeng's Relogit were trained on the same training dataset and evaluated on the same validation dataset.

4.2 Performance of alternative models

The performance of the three alternative models was evaluated using the receiver operating characteristics (ROC) curve and Precision-Recall (PR) curve.

The ROC curve describes the trade-off between the model's sensitivity and false-positive rate (FPR) at varying probability cutoff thresholds that delineate harmful flash flood events from non-harmful ones. The area under the ROC curve (AUROC) is suitable for evaluating the alternative models because it is insensitive to class distribution and is threshold invariant (Guns and Vanacker 2012). The PR curve describes the trade-off between the model's precision and recall at varying threshold values. In the context of public safety, the positive class (i.e., flash flood events that resulted in at least one injury or fatality) is of greater interest than the negative class (i.e., non-harmful events). In this context, the area

Table 3 Performance of evaluated models using the testing dataset

Model	AUROC	AUPRC
MLE	0.90	0.24
PMLE	0.94	0.33
Relogit	0.94	0.41

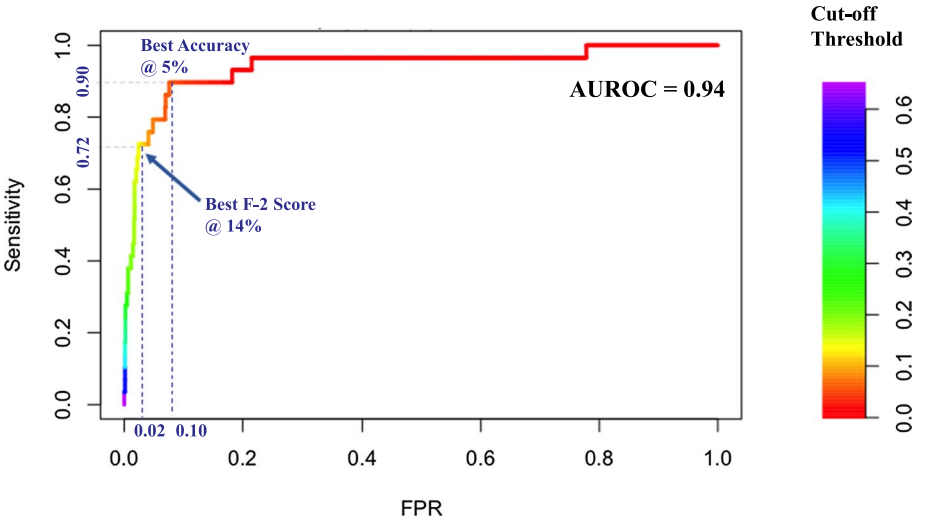

Fig. 5 ROC curve for the Relogit model

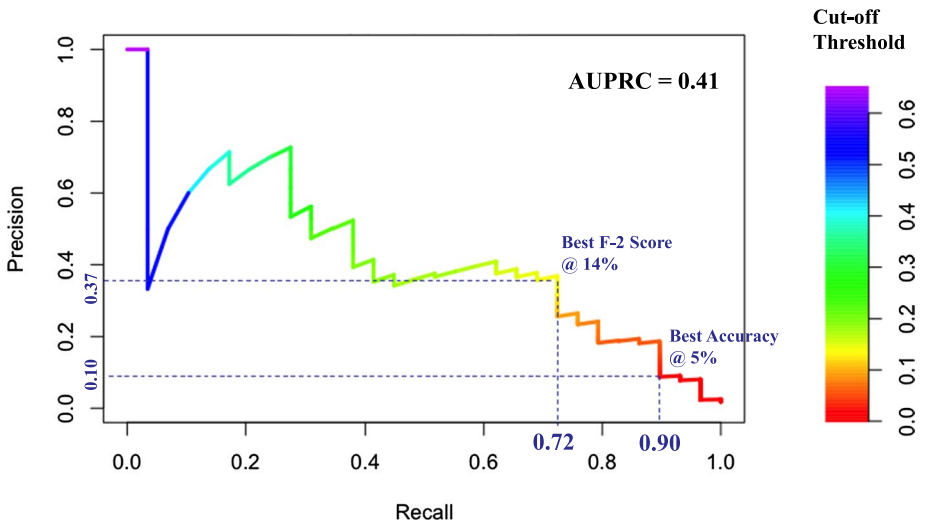

Fig. 6 PR curve for the Relogit model

under the PR curve (AUPRC) is also a suitable performance metric because it will not be swamped by the large proportion of true negatives in the data (Saito 2015; Sofaer et al. 2019; Pinker 2018). As shown in Table 3, all three models display a good performance in terms of AUROC (greater than 0.90). However, the Relogit model outperforms both the MLE and PMLE models in terms of AUPRC. The ROC and PR curves for the Relogit model are shown in Figs. 5 and 6, respectively.

To evaluate the precision of the Relogit model, a cutoff point needs to be selected to determine whether a flash flood event is considered harmful or not. Flash flood events with a probability higher than the cutoff point would be classified as harmful, whereas events with a probability lower than the cutoff point would be classified as non-harmful. For rare-event data, however, the selection of cutoff point (called optimal threshold tuning) is a trade-off between precision and recall. Precision is the fraction of true positives (harmful events) among all predicted positives, whereas recall is the fraction of true positives among all observed positives. In this study, we determined the optimal cutoff point in two ways: (1) maximize the model's accuracy (which is a function of the sum of the true-positive rate and true-negative rate), and (2) maximize the F-2 score (which combines precision and recall as a harmonic mean with additional weight on precision). These two methods yielded a precision of 10% (at 5% optimum cutoff point) and 37% (at 14% optimum cut-off point) (see Figs. 5 and 6). These low-precision values suggest that the Relogit model may not be suitable as a binary classifier of flash flood events as harmful and non-harmful. Instead, it is best suited for predicting the probability of human harm occurrence.

4.3 Monte carlo simulation for coefficient estimates

A Monte Carlo (MC) simulation was designed in this study to evaluate bias in the estimation of the logit model coefficients with different sample sizes. Since the harmful events have an occurrence rate of 2% in the study dataset, the MC simulation was designed at a

Table 4 MC simulation results of mean slope (β)

Model	Sample size						
	50	100	200	500	1000	2000	5000
MLE	15.412*	3.228*	2.141*	2.06*	2.024	2.015	2.004
PMLE	1.925	1.979	1.976	1.992	1.998	2.006	2.001
Relogit	NA	NA	2.038	1.994	2.005	2.003	1.999

*The 95% confidence interval does not contain the true slope of $\beta_0 = 2$

2% prevalence rate for the positive class with the sample size varying from 50 to 5,000. The setup of the MC simulation is summarized below:

$$\eta_{0.02} = \alpha + \beta * x \tag{15}$$

where $\eta_{0.02}$ = A logit transformed linear predictor with 2% occurrence rate; α = intercept of linear predictor defined as $\alpha = -log((1-p)/p)$; β = slope of linear predictor; x = simulated variable that follows a normal distribution between [0,1]; Event probability $p = 0.02$; Sample size = 50, 100, 200, 500, 1000, 2000, 5000; True slope (coefficient) $\beta_0 = 2$; Number of replications = 1000.

The results of the MC simulation for the coefficient estimates for the conventional MLE, Firth's PMLE, and King and Zeng's Relogit models are summarized in Table 4. Conventional MLE, without bias correction, fails to include the true slope ($\beta_0 = 2$) within 95% confidence interval of the estimates until the sample size exceeds 1000. In contrast, the coefficient estimations from the PMLE and Relogit models are close to the true slope ($\beta_0 = 2$) even with a small to medium sample size. Additionally, for 5000 observations, the PMLE and Relogit models still outperform the conventional MLE with mean value of estimation closer to the true slope. Therefore, bias adjustment is necessary for logit models to achieve correct statistical inferences about the occurrence of fatalities or injuries in flash flood events.

5 Selected model

5.1 Model variables and coefficients

As discussed earlier, the Relogit model was selected as the final model due to its superior performance over the MLE and PMLE models and unbiased estimation of model coefficients. The values of the final Relogit model coefficients are provided in Table 5. The interpretation of coefficient value is different for binary and numeric independent variables. For example, the coefficient for the binary variable Nighttime (0.6437) indicates that a flash flood that occurs at night has $e^{0.6437} = 1.90$ times the odds of being harmful compared to those that occur during the daytime. On the other hand, the coefficient for the numeric variable median age of structures (− 0.03423) implies that an increase of one year in the median age of structures for a census tract multiplies the odds of flash flood human harm by $e^{-0.03423} = 0.97$. Some potential influencing factors (such as road density and soil type) are removed from final Relogit model due to either the lack of statistical significance or not contributing to the model's predictive power.

Table 5 Final Relogit model independent variables and coefficients

Variable	Estimate	Std. error	Odds ratio (per unit)	*P* value	Signif. Codes[a]
(Intercept)	− 6.42E + 00	5.79E-01	–	< 2.00E-16	***
Duration (hour)	1.80E-02	5.11E-03	1.0182	4.24E-04	***
Precipitation (inch)	1.28E-01	2.39E-02	1.1366	9.14E-08	***
Median age of structures (year)	− 3.42E-02	8.45E-03	0.9663	5.14E-05	***
Population density (people/sq.mi)	7.79E-05	7.09E-05	1.0001	2.72E-01	+
Nighttime event	6.44E-01	2.55E-01	1.9035	1.17E-08	*
Flash flood alley	1.12E + 00	2.32E-01	3.0557	1.55E-06	***
Vehicle involvement	2.95E + 00	2.77E-01	19.0678	< 2.00E-16	***
Bridge density (count/sq.mi)	5.33E-02	2.46E-02	1.0547	0.030355	*
LWX density (count/sq.mi)[b]	2.36E-01	6.40E-02	1.2659	0.000231	***
Ground slope (degree)	1.16E-02	7.29E-03	1.0117	0.111786	+

[a]Signif. codes: *** < 0.001; ** < 0.01; * < 0.05., + > 0.05 but improve model's performance

[b]LWX = Low water crossing

Independent variables with lower *P* values have greater influence on the model's predictions. Thus, the independent variables can be ranked based on their *P* value or odds ratio to assess their effect on the likelihood of human injury or fatality in a flash flood event. Vehicle-related activities during flash flooding exhibited the smallest *P* value and greatest odds ratio (i.e., has the greatest effect on the model's predictive power) among all independent variables, followed by the event's precipitation amount, location with respect to the Flash Flood Alley (inside or outside), median age of structures, low water crossing density, and event duration. While independent variables with a *P* value slightly greater than 0.05, such as population density and ground slope, may be considered statistically insignificant, they do enhance the model's predictive power.

The variability of the logit model coefficient is indicated by the standard error. The smaller the standard error, the more precise the estimate of the coefficient value. No multicollinearity was found in the model (variance inflation factor less than 2 for all predictors).

5.2 Predicted likelihood of human harm occurrence

The likelihood of human harm occurrence is predicted using the final Relogit model as follows:

$$P(\mathrm{Y}=1)=\frac{1}{1+e^{-\mathrm{X}\beta}} \tag{16}$$

where $P(Y=1)$ is the probability that a flash flood event will lead to at least one fatality or injury; $\tilde{\beta}$ is bias adjusted coefficient estimates $\tilde{\beta_0}, \tilde{\beta_1}, \ldots, \tilde{\beta_{10}}$ for the independent variables coefficients and intercept from the final Relogit model and X represents input values of Matrix $x_1, x_2, \ldots, x_{10}$ from the independent variables. Values for the coefficients in the final Relogit model are listed earlier in Table 5.

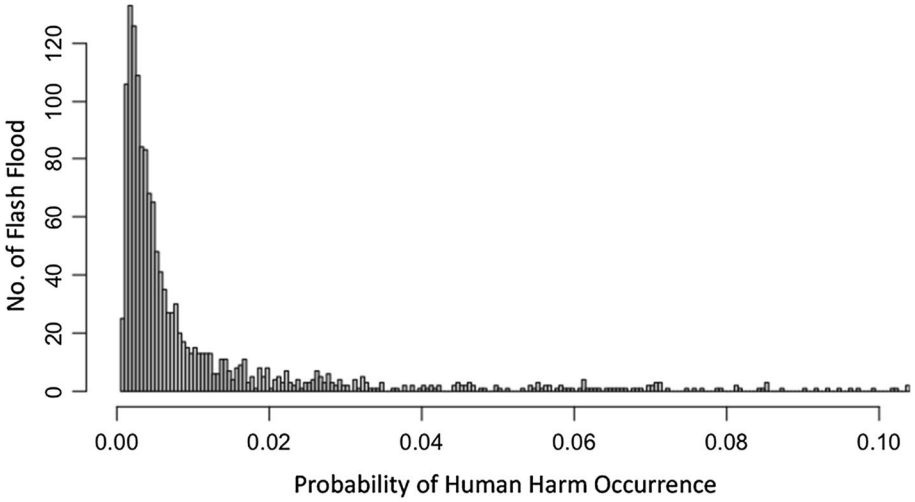

Fig. 7 Predicted probabilities of human harm occurrence for the testing dataset

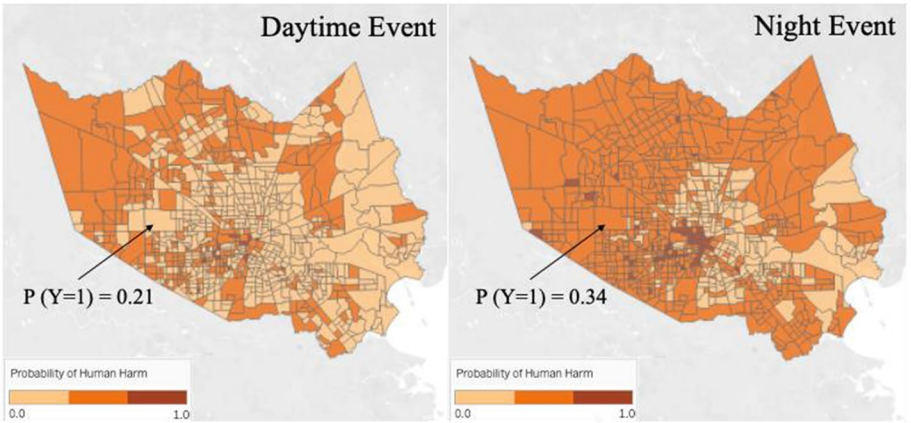

Fig. 8 Probability of non-vehicle-related human harm during flash flooding in harris county (hypothetical event duration = 72 h, hypothetical precipitation = 35 inches)

Figure 7 displays the histogram of probabilities of harmful flash flood events estimated using the final Relogit model for the testing dataset. For the majority of cases, the probability of human harm occurrence is under 2%, which aligns with the actual occurrence rate.

6 Study cases

To demonstrate the utilization of the developed model as a tool for informing better mitigation planning for flash flooding, we applied the model to two study cases in Texas. The hydrometeorological data used in these hypothetical cases are based on actual historical storms in the study regions.

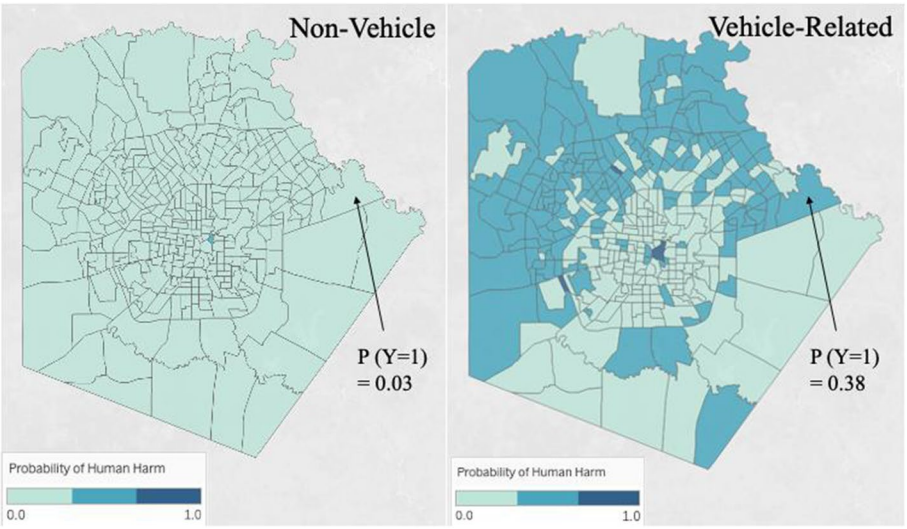

Fig. 9 Probability of human harm during flash flooding in bexar county (daytime event, hypothetical event duration = 14 h, hypothetical precipitation = 6 inches)

6.1 Study case 1: nighttime effect on flash flood safety in harris county

This study case shows the probability of human harm occurrence in a flash flood event generated by a storm similar to Hurricane Harvey (August 2017) in Harris County, Texas, where the city of Houston is located. This hurricane-induced flash flood event yielded 35 inches of rainfall across the county for a total of three days and resulted in numerous fatalities and injuries. Figure 8 maps the probability of the occurrence of human harm during this storm due to flash flooding in each census tract in Harris County, during nighttime and daytime. The estimated probabilities of human harm are presented in three equally spaced categories: low (<0.33), medium (0.33–0.66), and high (>0.67). In this simulation, the Houston downtown area has the greatest risk of human harm due to higher population density and greater bridge density. The probability of human harm increases for the entire county when the simulated flash flood event occurs at night (right side of Fig. 8). For example, the probability of human harm in census tract FIPS:48,201,543,200 (highlighted in Fig. 8 increases from 21% during daytime to 34% during nighttime.

6.2 Study case 2: effect of vehicle activities on flash flood safety in bexar county

This study case shows the probability of human harm occurrence from a flash flood event brought by a tropical storm similar to storm Erin (August 2007) in Bexar County, Texas, where the city of San Antonio is located. This storm had a precipitation of 6 inches and lasted for 14 h. The probability of the occurrence of human harm during this simulated flash flooding in Bexar County is plotted in Fig. 9. The majority of communities exhibit low risk of human harm as the rainfall amount and duration remain

low in this simulated event. However, the probability of human harm increases substantially for people exposed to vehicles during the event (both inside and outside the vehicle). It can be observed that nearly one half of the communities in Bexar County turn to medium risk, and some of them become high risk (i.e., high probability of fatality or injury occurrence). For example, census tract FIPS: 48,029,131,601 (highlighted in Fig. 9) has a 3% probability of non-vehicle human harm, but that probability jumps to 38% for vehicle-related incidents.

7 Discussion

The developed Relogit model has important advantages that improve the prediction of flash flooding safety risks. First, the data used in this model have finer spatial resolution than that used in prior studies. Prior interdisciplinary models for predicting human harm in flood events (Terti et al. 2019; Zahran et al. 2008) have been at the county scale. We use a finer spatial resolution (i.e., census tract) that is more consistent with the spatial scale of flash flooding. Second, our modeling technique accounts for the rareness of flash flood events that resulted in fatalities or injuries (i.e., imbalanced data). Despite these advantages, the model has limitations that stem from imperfections in the dataset, including the potential for inaccuracies in the event location and the possibility of missing (unrecorded) events.

Our model shows that vehicle involvement has the largest impact on human safety during flash flood events, which agrees with previous studies (e.g., Sharif et al. 2012). Quantitatively, our model shows that vehicles increase the odds of human harm by 19 times, compared to human harm that does not involve vehicles (e.g., drowning without vehicle involvement). The odds of human harm almost doubles when flash flooding occurs at night, perhaps due to the victims limited vision affecting their assessment of the depth and speed of floodwater. The odds of human harm are worse (three times more) if the affected community is located in the Flash Flood Alley.

The risk of human harm increases for newer and more densely populated neighborhoods. Neighborhoods with older structures exhibited less risk of human harm during flash flood events than newer neighborhoods. This finding could be an indicator that older neighborhoods (measured in terms of median age of structures) in Texas tend to be located in higher elevation areas and further away from floodways compared to newer neighborhoods. Densely populated communities tend to be more vulnerable to human harm during flash flooding perhaps because of greater urbanization and human movements in these communities.

Lastly, for all situations and geographic locations, the risk of human harm increases with rainfall duration and intensity, steep topography, higher density of bridges, and higher density of low water crossings.

8 Conclusions and recommendations

The occurrence of injuries and fatalities (termed “human harm” in this study) during flash flooding was found to be a rare event (about 2% occurrence rate) based on 6,065 flash flood events that occurred in Texas over a 15-year period (2005–2019). We found that the bias adjusted Relogit model is the most suitable logistic model for predicting the likelihood

of human harm occurrence (fatalities or injuries) in flash flooding. The model considers ten independent variables that could affect human safety in flash flood events at the census tract scale. These variables represent situational factors, storm characteristics, and the built, natural, and social environments in which the storm occurs. The Relogit model has a better precision-recall performance (measured as AUPRC) and similar sensitivity-FPR performance (measured as AUROC) compared to other logistic models, invariant of the cutoff threshold.

The utilization of the developed model as a simulation tool for informing flash flooding mitigation and safety planning was demonstrated in two study cases: flash flooding triggered by a hurricane in Harris County and flash flooding triggered by a tropical storm in Bexar County. Future work could include: (1) further assessment of the probabilistic nature of the trained model considering future flash flood events; (2) development of similar models for other regions in the USA using the process described here; (3) analysis of digital elevation models (DEM) to determine if older neighborhoods tend to be located in higher elevation areas and further away from floodways compared to newer neighborhoods; and (4) building a simulation platform that use the Relogit model for informing the planning of flash flood mitigation and safety strategies.

Author contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by SC and RSW. The first draft of the manuscript was written by SC and all authors commented on previous versions of the manuscript. Dr. Nasir Gharaibeh performed review, editing, and project administration.

Funding This material is based on work supported by the National Science Foundation (NSF) under Grant # 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.

Data availability The data and code are available at the following DesignSafe-ci.org DOI: https://doi.org/10.17603/ds2-e91y-cv92.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

Ahmadalipour A, Moradkhani H (2019) A data-driven analysis of flash flood hazard, fatalities, and damages over the CONUS during 1996–2017. J Hydrol 578:124106

Albert A, Anderson JA (1984) On the existence of maximum likelihood estimates in logistic regression models. Biometrika 71(1):1–10

Anderson DG (1970) Effects of urban development on floods in northern Virginia. US Government Printing Office, p 22

Ashley ST, Ashley WS (2008) Flood fatalities in the United States. J Appl Meteorol Climatol 47(3):805–818

Bailey JF (1989) Estimation of flood-frequency characteristics and the effects of urbanization for streams in the Philadelphia, Pennsylvania area. Department of the Interior, US Geological Survey, pp 87–4194

Baker VR (1975) Flood hazards along the Balcones escarpment in central Texas; alternative approaches to their recognition, mapping, and management. Virtual Landscapes of Texas

Benight CC, Gruntfest EC, Hayden M, Barnes L (2007) Trauma and short-fuse weather warning perceptions. Environ Hazards 7(3):220–226

Bull SB, Mak C, Greenwood CM (2002) A modified score function estimator for multinomial logistic regression in small samples. Comput Stat Data Anal 39(1):57–74

Caran SC, Baker VR (1986) Flooding along the balcones escarpment, central Texas. KIP Articles. 2088
Choirat C, Honaker J, Imai K, King G, Lau O (2020) Zelig: everyone's statistical software. Version 5.1.7, https://zeligproject.org/
Clauset A, Woodard R (2013) Estimating the historical and future probabilities of large terrorist events. Annal Appl Stat 7(4):1838–1865
Clogg CC, Rubin DB, Schenker N, Schultz B, Weidman L (1991) Multiple imputation of industry and occupation codes in census public-use samples using bayesian logistic regression. J Am Stat Assoc 86(413):68–78
Cosslett SR (1981) Maximum likelihood estimator for choice-based samples. Econom J Econom Soc 49:1289–1316
Cox DR, Hinkley DV (1979) Theoretical statistics. CRC Press, Florida
Diakakis M (2020) Types of behavior of flood victims around floodwaters. Correlation with situational and demographic factors. Sustainability 12(11):4409
Firth D (1993) Bias reduction of maximum likelihood estimates. Biometrika 80(1):27–38
Gao S, Shen J (2007) Asymptotic properties of a double penalized maximum likelihood estimator in logistic regression. Stat Probab Lett 77(9):925–930
Greene W (1993) Econometric analysis, 2nd edn. Macmillan, New York
Guns M, Vanacker V (2012) Logistic regression applied to natural hazards: rare event logistic regression with replications. Nat Hazards Earth Syst Sci 12(6):1937–1947
Hamilton K, Peden AE, Pearson M, Hagger MS (2016) Stop there's water on the road! Identifying key beliefs guiding people's willingness to drive through flooded waterways. Saf Sci 89:308–314. https://doi.org/10.1016/j.ssci.2016.07.004
Heinze G, Schemper M (2002) A solution to the problem of separation in logistic regression. Stat Med 21(16):2409–2419
Heinze G, Ploner M, Dunkler D, Southworth H (2013) logistf: firth's bias reduced logistic regression. R package version 1.20. Available at: http://cran.r-project.org/web/packages/logistf/index.html
Imai K, King G, Lau O (2008) Toward a common framework for statistical analysis and development. J Comput Graphical Stat 17(4):892–913
Imbens GW (1992) An efficient method of moments estimator for discrete choice models with choice-based sampling. Econom J Econom Soc 60:1187–1214
Jeffreys H (1946) An invariant form for the prior probability in estimation problems. Proc R Soc London Ser A Math Phys Sci 186(1007):453–461
Kelsch M, Carporali E, Lanza LG (2001) Hydrometeorology of flash floods. In: Gruntfest E, Handmer J (eds) Coping with flash floods. Kluwer Academic Publishers, Dordrecht, pp 19–35
King G, Zeng L (2001a) Explaining rare events in international relations. Int Org 55(3):693–715
King G, Zeng L (2001b) Logistic regression in rare events data. Political Anal 9(2):137–163
Konrad CP, Booth DB (2002) Hydrologic trends associated with urban development for selected streams in the Puget Sound Basin, Western Washington, vol 2. US Geological Survey, 4040
Lancaster T, Imbens G (1996) Case-control studies with contaminated controls. J Econ 71(1–2):145–160
Leitgöb H (2020) Analysis of rare events. SAGE Publications Limited
Lindell MK, Perry RW (1992) Behavioral foundations of community emergency planning. Hemisphere Publishing Corp
Lindell MK, Prater C, Perry RW (2006) Wiley pathways introduction to emergency management. Wiley
Masterson JH, Peacock WG, Van Zandt SS, Grover H, Schwarz LF, Cooper JT (2014) Planning for community resilience: a handbook for reducing vulnerability to disasters. Island Press
McCullagh P, Nelder JA (1989) Generalized linear models II
NWS (2019) NWS Preliminary US Flood Fatality Statistics (2019). https://www.weather.gov/arx/usflood
NWS (2022) Storm events database. Available online: https://www.ncdc.noaa.gov/stormevents/ftp.jsp. Accessed Novemb 11, 2022
Pinker E (2018) Reporting accuracy of rare event classifiers. NPJ Digit Med 1(1):1–2
Quenouille MH (1949) Problems in plane sampling. Ann Math Stat 20:355–375
Quenouille MH (1956) Notes on bias in estimation. Biometrika 43(3/4):353–360
Ruin I, Gaillard JC, Lutoff C (2007) How to get there? Assessing motorists' flash flood risk perception on daily itineraries. Environ Hazards 7(3):235–244
Saito T, Rehmsmeier M (2015) The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS one 10(3):e0118432
Sanders DEA, Brix A, Duffy P, Forster W, Hartington T, Jones G, Wilkinson M (2002) The management of losses arising from extreme events. Convention general insurance study group GIRO, London
Shah V, Kirsch KR, Cervantes D, Zane DF, Haywood T, Horney JA (2017) Flash flood swift water rescues, Texas, 2005–2014. Clim Risk Manage 17:11–20

Sharif HO, Hossain MM, Jackson T, Bin-Shafique S (2012) Person-place-time analysis of vehicle fatalities caused by flash floods in Texas. Geomatics Nat Hazards Risk 3(4):311–323
Sharif HO, Jackson TL, Hossain MM, Zane D (2015) Analysis of flood fatalities in Texas. Nat Hazards Rev 16(1):04014016
Sofaer HR, Hoeting JA, Jarnevich CS (2019) The area under the precision-recall curve as a performance metric for rare binary events. Methods Ecol Evol 10(4):565–577
Terti G, Ruin I, Anquetin S, Gourley JJ (2015) Dynamic vulnerability factors for impact-based flash flood prediction. Nat Hazards 79(3):1481–1497
Terti G, Ruin I, Anquetin S, Gourley JJ (2017) A situation-based analysis of flash flood fatalities in the United States. Bull Am Meteorol Soc 98(2):333–345. https://doi.org/10.1175/BAMS-D-15-00276.1
Terti G, Ruin I, Gourley JJ, Kirstetter P, Flamig Z, Blanchet J, Anquetin S (2019) Toward probabilistic prediction of flash flood human impacts. Risk Anal 39(1):140–161
U.S. Census Bureau (2022) Glossary. Available online: https://www.census.gov/programs-surveys/geography/about/glossary.html#par_textimage_13. Accessed Novemb 9, 2022
Walsh J, Wuebbles D, Hayhoe K, Kossin J, Kunkel K, Stephens G, Somerville R (2014) Our changing climate. Climate change impacts in the United States: The third national climate assessment, 19, 67
Wang J, Zhang X (2008) Downscaling and projection of winter extreme daily precipitation over North America. J Clim 21(5):923–937
Wobus C, Lawson M, Jones R, Smith J, Martinich J (2014) Estimating monetary damages from flooding in the United States under a changing climate. J Flood Risk Manag 7(3):217–229
Zahran S, Brody SD, Peacock WG, Vedlitz A, Grover H (2008) Social vulnerability and the natural and built environment: a model of flood casualties in Texas. Disasters 32(4):537–560