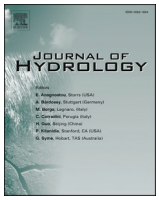
Journal of Hydrology 637 (2024) 131406

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol



Research papers

Predicting real-time roadway pluvial flood risk: A hybrid machine learning approach coupling a graph-based flood spreading model, historical vulnerabilities, and Waze data

Arefeh Safaei-Moghadam *, Azadeh Hosseinzadeh *, Barbara Minsker *

Department of Civil and Environmental Engineering, Southern Methodist University, Dallas, TX, USA

ARTICLE INFO

This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of Zhenxing Zhang, Associate Editor

Keywords:
Hybrid machine learning
Random forest
SVM: XGBoost
Baysian statistical model
Waze
GB-RFSM

ABSTRACT

Urban pluvial flash flooding (PFF), driven by extreme weather and urban expansion, introduces complex challenges that arise from the dynamic interaction of rainfall hazard, road vulnerability, and traffic exposure. These three critical components must be interconnected to provide a comprehensive prediction of roadway PFF risk. Our integrated approach combines historical data and real-time Waze flood alerts using a simplified physics-based PFF model and hybrid machine learning methods to predict flash flooding risk at the road segment scale. In a Dallas case study with four intersections, we trained multiple models with data from 15 storms and tested on 5 storms. The XGBoost method excels in test precision, while a Random Forest model offers better recall, and both methods outperform Support Vector Machines (SVM). The choice between models depends on factors such as negative class (prediction of unflooded areas) uncertainty and false positive cost (i.e., predicting no flooding incorrectly). For the case study, our approach could boost flood awareness, enhance safety, and improve urban flood management by correctly predicting 73% of risk observations during the test storm events.

1. Introduction

Urban flood risk is a complex phenomenon influenced by multiple factors. (Gouldby and Samuels, 2005) offer a concise and insightful definition of risk as a combination of three pivotal components: hazard, vulnerability, and exposure. According to this definition, the hazard is a phenomenon that has the potential to be destructive; vulnerability specifies how likely the system is to be damaged by the hazard; and exposure is the total number of receptors that the destruction may affect. In interpreting the risk of roadway flooding in particular, the hazard of the system is driven by precipitation; vulnerability is the potential of a road segment to accumulate water and form pluvial flash flooding (PFF) in the event of stormwater drainage system failure, which depends on the topographic and hydrologic characteristics of the roadway and surrounding catchments; and exposure is the traffic volume that confronts flooded road segments. In a highly urbanized area, flood formation is a highly complex and uncertain process due to the lack of information about underground and overland flow interaction (He et al., 2023; Santos et al., 2020). The impacts of flooding on roadway mobility also depend on various temporal and spatial variables. Moreover, the rapid onset and short lifetime of roadway PFF lead to a lack of observational data to quantify the historical PFF risk at a local scale. Hence, a comprehensive and integrated approach is needed that reflects the topographic and hydrologic specifications of road segments as well as temporal traffic levels to capture the heterogeneity of risk (Ren et al., 2022; Oneto and Canepa, 2023).

Flood prediction employs two primary modelling paradigms: physics-based models and data-driven models. Physics-based models play a crucial role in considering the geospatial and hydrologic relationships that drive PFF but are often limited by data scarcity and computational challenges of modelling the complexity of the urban environment. For example, 2D-1D coupled models can provide good accuracy but require significant data and face computational limitations (Noh et al., 2018; Bulti and Abebe, 2020; Hosseinzadeh et al., 2023). Researchers have sought alternatives to these data-intensive models through simplified physics-based approaches. Rapid Flood Spreading Models (RFSM), including Hierarchical Filling and Spilling Models (HFSM), offer efficient flood depth estimations (Lhomme et al., 2008a) by treating surface depressions in digital elevation models (DEMs) as hydrologic units. These models have gained traction in urban pluvial

* Corresponding authors.
E-mail addresses: asafaeimoghadam@mail.smu.edu (A. Safaei-Moghadam), ahosseinzadeh@mail.smu.edu (A. Hosseinzadeh), minsker@mail.smu.edu (B. Minsker).

https://doi.org/10.1016/j.jhydrol.2024.131406

Available online 24 May 2024

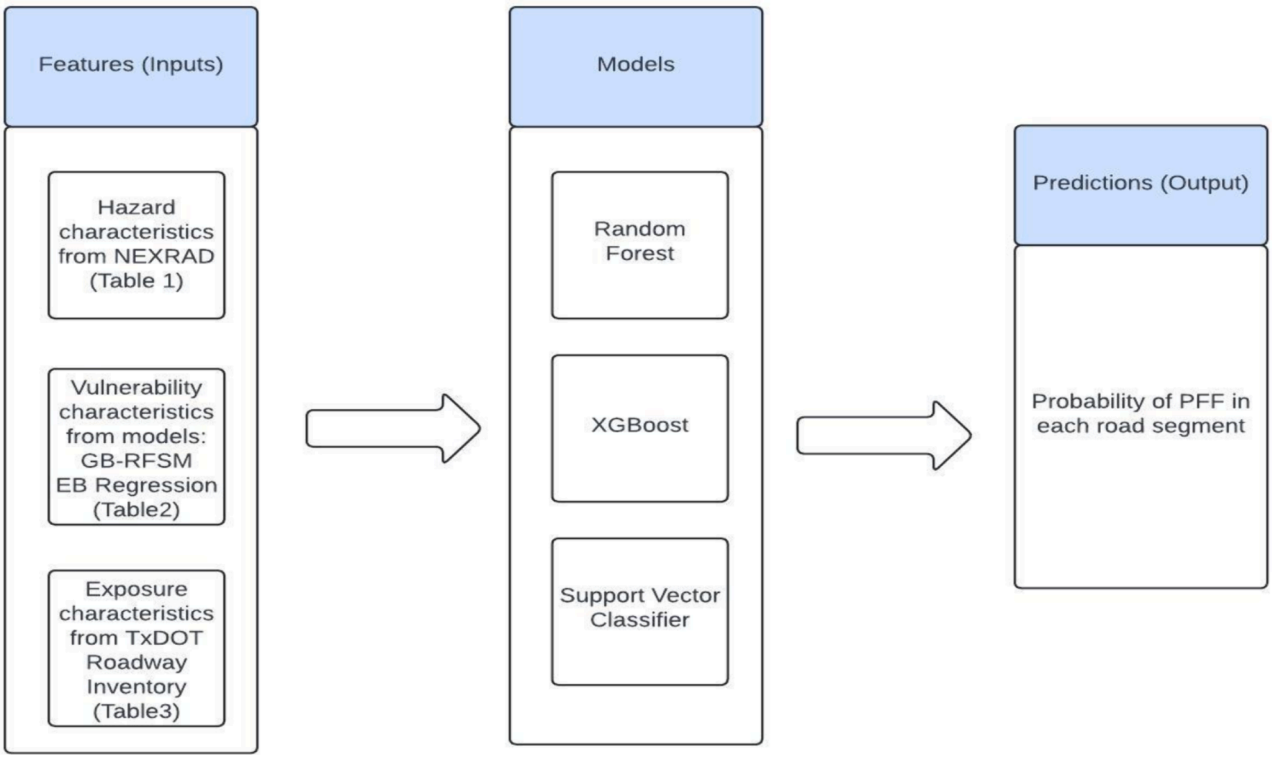


Fig. 1. Methodology flowchart.

flood (PFF) modeling due to their capacity to consider key retention features of urban runoff without significant computational effort.

Recent studies have emphasized the role of surface depressions in capturing complex urban catchment responses to rainfall (Guidolin et al., 2016; Lhomme et al., 2008b; Manfreda and Samela, 2019; Samela et al., 2020; Shen et al., 2016; Yong-He et al., 2009; Preisser et al., 2022; Zheng et al., 2018; Yao et al., 2016; Cristiano et al., 2017). These models delineate surface depressions, establish their nested hierarchy, and distribute flood volume among them using diverse approaches. Some models employ graph-based techniques, such as the level-set approach (Feng et al., 2022; Wu et al., 2019). While these methods efficiently delineate depressions, they don't always address the temporal flood accumulation process.

On the other hand, with recent technological advancements, the realm of data-driven modelling has significantly expanded, presenting a wealth of opportunities to improve flood forecasting. Crowdsourced data, which has become widely accessible and cost-effective, stands out as a valuable resource to augment sparse available datasets. However, by definition, crowdsourced data are collected by heterogeneous volunteer individuals of varying knowledge, experience, perceptions, and number (Bowler et al., 2022; Estellés-Arolas and González-Ladrón-de-Guevara, 2012). Social Media Geographic Information (SMGI) is a specific type of Volunteered Geographic Information (VGI) crowdsourced data that, in addition to geographic coordinates, contains time, user information, and multimedia content (Campagna, 2016). These and other geospatial data are used to train the hybrid model developed in this work. This builds on an increasing literature of hybrid flood prediction models (Berkhahn et al., 2019; Zahura et al., 2020; Farahmand et al., 2023; Moon et al., 2023; Li and Willems, 2020; Kim et al., 2019; and Fang et al., in review).

This study analyzes and assesses the quality of SMGI data for PFF prediction by coupling machine learning methods with a Graph-Based Rapid Filling and Spilling Model (Safaei-Moghadam et al., submitted) in a hybrid modelling approach.

We focus on SMGI from alerts posted to the Waze navigation App, which are geotagged points posted by Waze users, called Wazers, that express drivers' experience of road conditions in real time. Waze alerts include multimedia content describing road surface conditions (e.g., photographs and flood reports) and containing geographic coordinates, time, and user characteristics such as reputation and feedback from other users on each user's postings. From the context of roadway flooding, Waze alerts are not specifically designed for flood situational awareness and Wazers may not be aware that their shared experience will be interpreted as flooding data. Thus, Waze alerts are classified as implicitly SMGI crowdsourced data that was not shared for the specific purpose that the data are being employed (Craglia et al., 2012; Haworth and Bruce, 2015).

However, despite this extensive literature and a great variety of models, several key limitations in existing research are apparent that motivated this research. First, real-time flood risk prediction in highly urbanized areas with complex drainage systems requires models that can rapidly capture the dynamic nature of flood formation. While previous research has made substantial progress, it has often fallen short in addressing the temporal aspects of inundation and the rapid changes in flood conditions that characterize urban environments. Moreover, our research tackles the challenges posed by data limitations by incorporating crowdsourced data from platforms such as Waze.

Considering these limitations, our research seeks to extend and complement the existing body of work. Building upon the insights gained from our previous research (Safaei-Moghadam et al., 2023; Safaei-Moghadam et al., in review), which emphasized the value of crowdsourced data from the Waze navigation app in conjunction with detailed analysis of highly localized surface depressions and catchments for pluvial flash flooding (PFF) vulnerability at the road segment scale, this study extends the application of this valuable data source. In our present research, we leverage this knowledge to develop a hybrid machine learning model for real-time flood risk prediction, the primary focus of this paper. Moreover, we've incorporated the previously developed Graph-Based Rapid Flood Spreading Model (https://github.com/asmoghadam/GB-RFSM) into our approach.

Table 1
Features representing the hazard.

Feature	Equation	Definition
Precipitation	P_T	Precipitation at time T (mm)
Maximum preceding rain pulse	$Max(p_t)\ t\ in\ [0, T]$	Maximum precipitation per timestep since the beginning of storm
Total accumulation	$\sum_0^T p_t$	Total accumulated precipitation since the beginning of storm
n-hours preceding accumulation	$\sum_{t-n}^{t} p_t$	Total accumulated precipitation in last n timesteps
Count of rain pulses above h mm	$Count(p_t \geq h)$	Count of timesteps with precipitation higher than h millimeter
Count of rain pulses above mean	$Count(p_t \geq \mu)$	Count of timesteps with precipitation higher than average precipitation (μ) since the beginning of storm event
Time to the last maximum rain pulse	$t - t_{p=max(p)\_}$	Time interval between timestep of study and the last timestep with maximum precipitation
Time to the last minimum rain pulse	$t - t_{p=0}$	Time interval between timestep of study and the last timestep with 0 precipitation
Count of rainless intervals	$Count(p_t = 0)$	Count of rainless timesteps since the beginning of storm

Table 2
Features representing vulnerability.

Feature	Equation	Definition
Maximum depth	$Max(d_t)$	Maximum estimated depth on road surface
Area	$Count(d_t > 0) \times x^2$	Inundation area on road surface
Percentage of road segment inundated	$\frac{Count(d_t > 0)}{Count(d_t)}$	Volume of standing water on road surface
PFF likelihood-moderate storms	$p(i,j) = \frac{\hat{y}_{i,j}}{N_j}, j = moderate$	PFF likelihoods achieved from Empirical Bayes regression for moderate storms
PFF likelihood-severe storms	$p(i,j) = \frac{\hat{y}_{i,j}}{N_j}, j = severe$	PFF likelihoods achieved from Empirical Bayes regression for severe storms

The remainder of this paper is organized as follows: Section 2 details the methodology employed in for flood risk prediction. Section 3 summarizes the specific case study and data sources and values. In Section 4, we present the results obtained from our study, highlighting the key findings and insights. Finally, in Section 5, we offer conclusions drawn from this work and provide recommendations for future research directions.

2. HYPERLINK "SPS:id::Sec1" Methodology

For Waze flood alerts to be a reasonable proxy to the risk of roadway PFF, the alerts need to be associated with the hazard, vulnerability, and exposure. Safaei-Moghadam et al., 2023 demonstrated that the total number of flood alerts correlates with the rainfall's duration and maximum intensity (i.e., hazard data). Compared to moderate and light storm clusters, rainfall events that were part of the severe storm cluster had a higher incidence of alerts related to flooding, as expected. Regarding the vulnerability, Waze flood alerts are clustered around low-lying areas and depressions that can accumulate a large volume of excess runoff and are more prone to flooding in the future. Concerning exposure, Waze flood alerts are primarily posted from frequently travelled roads. Their findings showed a clear correlation between Waze flood alerts and exposure to road PFF, with increased traffic volume increasing the likelihood of observing a Waze roadway flooding alert under similar circumstances. Hence, it is reasonable to assume that the existence of Waze flood alerts is a function of hazard, vulnerability, and exposure and can be used as a proxy to the risk of roadway PFF. Fig. 1 shows the risk components and their relationship with Waze flood alerts and roadway PFF risk. More details on these three components are given in Sections 2.1 to 2.3, followed by descriptions of the modelling and model evaluation methods in Sections 2.4 and 2.5.

2.1. Hazard

High spatiotemporal precipitation data are required for good predictions due to the roadway PFF phenomenon's rapid onset, brief lifespan duration, and high spatial resolution (Karami et al., 2022). The precipitation time series from the start of the storm event until the study time interval is utilized to extract descriptive features that can characterize the storm event using the probability of PFF at each time interval. In addition to total rainfall, temporal patterns of storms play a significant role in PFF formation. When modeling PFF, it is important to consider the consistency, peaks, minimums, variations, zeros, and immediately preceding rain pulses of precipitation timeseries. As a result, features that may capture these qualities are calculated. Table 1 displays these features, where p_t is the precipitation timeseries from the storm start until time t, and μ is the average precipitation intensity since the storm started.

2.2. Vulnerability

The vulnerability of a road segment to roadway flooding is a function of its topographic and hydrologic characteristics that contribute to the excess runoff accumulation as well as the performance of the stormwater drainage system. While high-resolution hydrodynamic models of surface runoff and drainage systems may be used for this, their application for real-time prediction at this spatial scale can be computationally challenging and require data that may not be available. Therefore, in this work, we employ PFF likelihood features estimated from a previously-developed Empirical Bayes (EB) model (Safaei-Moghadam et al., 2023) and GB-RFSM (Safaei-Moghadam et al. in review, Safaei-Moghadam, 2023), a simplified DEM-based Hierarchical Filling and Spilling Algorithm (HFSA), to characterize the vulnerability of a road segment to PFF. The features are listed in Table 2 and described below.

The Empirical Bayes model predicts the historical frequency of PFF in each road segment with depressions, which serves as a proxy for flood susceptibility. The model estimates PFF frequency for three classes of historical storms – light, moderate, and severe – but only the predictions for moderate and severe categories are used as features in this work because the light storms do not produce significant flooding. For more details, please see Safaei-Moghadam et al., 2023.

GB-RFSM simulates the vulnerability of road surface depressions to water accumulation based on real-time rainfall and topography. The model dissects the terrain into small hydrologic units, considering averaged infiltration and drainage from each unit. Its use of graph-based calculations allows for accounting of the temporal evolution of inundation by translating the nested hierarchy of depressions and their catchments into a directed graph representing spilling and merging hierarchy. For this work, GB-RFSM estimates the potential inundation depth and inundation area in the worst-case event when the stormwater drainage system fails, while the EB-derived PFF likelihoods represent the contribution of site-specific unobserved variables, such as chronic debris generation in a location, in the roadway PFF vulnerability.

GB-RFSM requires a DEM-preprocessing phase to identify the nested hierarchy of depressions in the DEM of the watershed upstream of a road segment. We developed an ArcGIS Python toolkit for this purpose. The toolbox outputs a Network Common Data Form (NetCDF) and depression descriptor tables that contain the resulting information needed in the GB-RFSM step. The GB-RFSM step then converts the DEM to a directed graph dataset (DEM-graph) and applies rainfall to the DEM-graph by routing runoff through nodes and calculating the water mass balance until all rainfall is captured. Once the runoff is routed, the model computes the filled volume of the depressions.

Finally, water depth in each depression is retrieved from the depth-volume relationships extracted in the preprocessing stage and inunda-

Table 3
Features representing exposure.

Feature	Definition	Values
Average annual daily traffic (AADT)*	Commonly used measure showing the traffic load calculated by dividing the total annual volume of vehicle traffic by 365	Integer between 0 to 999,999
Functional system classification (F_system)*	Categories of roadways and highways based on the service they provide, such as volume of traffic and trip types	1:Interstate 2: Other freeway and expressway 3: Other principal arterial 4: Minor arterial 5: Major collector 6: Minor collector 7: Local road
Weekday	Whether the timestep is during weekends or weekdays	1: Weekends 2: Weekdays
Time of day (TOD)	The time of timestep	1: After midnight (00:00 to 4:00 AM) 2: Early morning (4:00 to 7:00 AM) 3: Morning (7:00 AM to 12:00 noon) 4: Afternoon (12:00 to 4:00 PM) 5: Evening (4:00 to 8:00 PM) 6: Night (8:00 PM to midnight)

*. Texas Department of Transportation (TxDOT) Roadway Inventory

tion maps are generated. In addition to the preprocessing outputs, a binary zone raster indicating road surface grid cells is used to extract statistics of the flooded road segment and features that represent the vulnerability. The list of attributes that reflect vulnerability is shown in Table 2, where d_t is the flood depth raster on road surface determined with the GB-RFSM, x is the resolution of the road surface grid cells, $p_{(i,j)}$ is the likelihood of flooding on depression i during storm type j, $\widehat{y}_{i,j}$ is the predicted number of floodings on depression i and storm type of j using the EB model and N_j is the number of storms in cluster j.

2.3. Exposure

Next, the exposure attributes are established, which define the number of drivers affected by a flooded road segment assuming typical prevailing traffic conditions and count of vehicles passing the road segment. Since access to historical traffic data is not publicly available at low cost, public data sources that are a proxy to the traffic volume on a road segment are implemented in this study. *Road classification*, average annual daily traffic (*AADT*), time of day (*TOD*), and *weekdays* are characteristics used in this work (Table 3).

2.4. Modeling

Three Machine Learning (ML) classification algorithms are employed to predict two classes of PFF risk (i.e., flood-related alerts exist or do not exist): Random Forest Classifier (RFC), Extreme Gradient Boosting Decision Tree (XGBoost), and Support Vector Classifier (SVC).

RFC and XGBoost are decision tree-based algorithms that employ if-else rules to maximize information gain and make final decisions (Chen and Guestrin, 2016; Tyralis et al., 2019). A decision tree consists of decision nodes and leaf nodes (end nodes). One advantage of tree-based algorithms is their ability to determine feature importance in model prediction power based on node impurities. Node impurity is a measure of homogeneity of the target values, PFF risk observations in this example, at each tree node. The normalized decrease in node impurity estimates the significance of a given feature when it is added to a tree.

Random Forest (RF) is a supervised ensemble machine learning algorithm that uses multiple decision tree learners to increase predictive performance (Pedregosa et al., 2011). The final prediction of RF is the average prediction of all decision trees; each tree is built from a bootstrap sample of observations and a subset of features. Bootstrap sampling involves random sampling with replacement, meaning that a particular observation may not be picked for sampling while it is allowed to appear once or more than once in training samples. RF has been widely used for data-driven modeling in the field of water resources (e.g., Sadler et al., 2018; Tyralis et al., 2019). This algorithm can handle large and imbalanced datasets by combining bootstrap sampling and ensemble learning to train each tree on a more balanced subsample as well as its capability of being cost-sensitive by assigning class weights in node impurity calculations. The aggregation of several trees and the binning process in decision trees makes RF resilient to bias and overfitting. In RFC, the overall feature importance is calculated as the average of the importance of a feature over all trees, weighted by the number of samples used in each split across all trees. Tuned RFC hyperparameters in this study are number of trees in forest, maximum depth of trees (maximum number of allowed splits), maximum number of features allowed to be used in each tree, and class weights used in favor of the minority class when calculating the impurity score of a split.

XGBoost is one of the most preferred ensemble tree boosting ML models because it has been shown to give state-of-the-art results in different fields (Chen and Guestrin, 2016), including water resources (Huang et al., 2021; Janizadeh et al., 2022; Sanders et al., 2022). In the Gradient Boosting approach, the model's loss function is minimized by adding weak learners trained on the remaining residuals of existing learners. The tuned hyperparameters include maximum depth of decision trees, number of trees, and class weights. In XGBoost, the contribution of each decision tree to the final prediction is calculated by minimizing the prediction error of the training dataset (called the training loss).

The SVM classifier is an optimization-based learning technique that divides classes by locating an optimum hyperplane with the greatest marginal distance between the two classes (Kecman, 2001). While SVM classifier was originally limited to linearly separable datasets, application of a kernel function transforms the data from a nonlinear input space into a linear representation to make the data separable. Standard kernel functions commonly used in SVM include linear, polynomial, radial basis function (rbf), and sigmoid function. The choice of kernel function is determined during the random search process of model selection. Details of SVM and kernel functions are provided by (Kecman, 2001). Application of SVM in water resources systems, flood prediction, and flood susceptibility mapping has been extensive (e.g., Han et al., 2007; Ke et al., 2020; Liu et al., 2022; Saha et al., 2021; Xiong et al., 2019).

In this study, the RF and SVM classifiers are executed using the Scikit-Learn library (Pedregosa et al., 2011) and XGBoost is implemented using the XGBoost package (Chen and Guestrin, 2016) in the Python environment. Recursive feature elimination is used to pick the most important attributes for modeling. This is a repetitive process in which the model is trained with all features and the least significant feature is dropped in every trial until performance declines significantly. A randomized grid search on a variety of model parameters is also utilized for hyperparameter tuning and 5-fold cross validation is performed to choose the optimal model and parameter combination.

2.5. Model evaluation

A total of 15 historical storm events are used to train the models and 5 storm events are used to test the models. The trained models predict PFF risks during test storms; the predicted risks are then compared with flood-related Waze alerts to evaluate the models' performance.

Given that flood risks and posting Waze flood alerts are rare binary events according to the Waze data, the classification model to predict the risk of roadway PFF deals with a highly imbalanced dataset (i.e.,

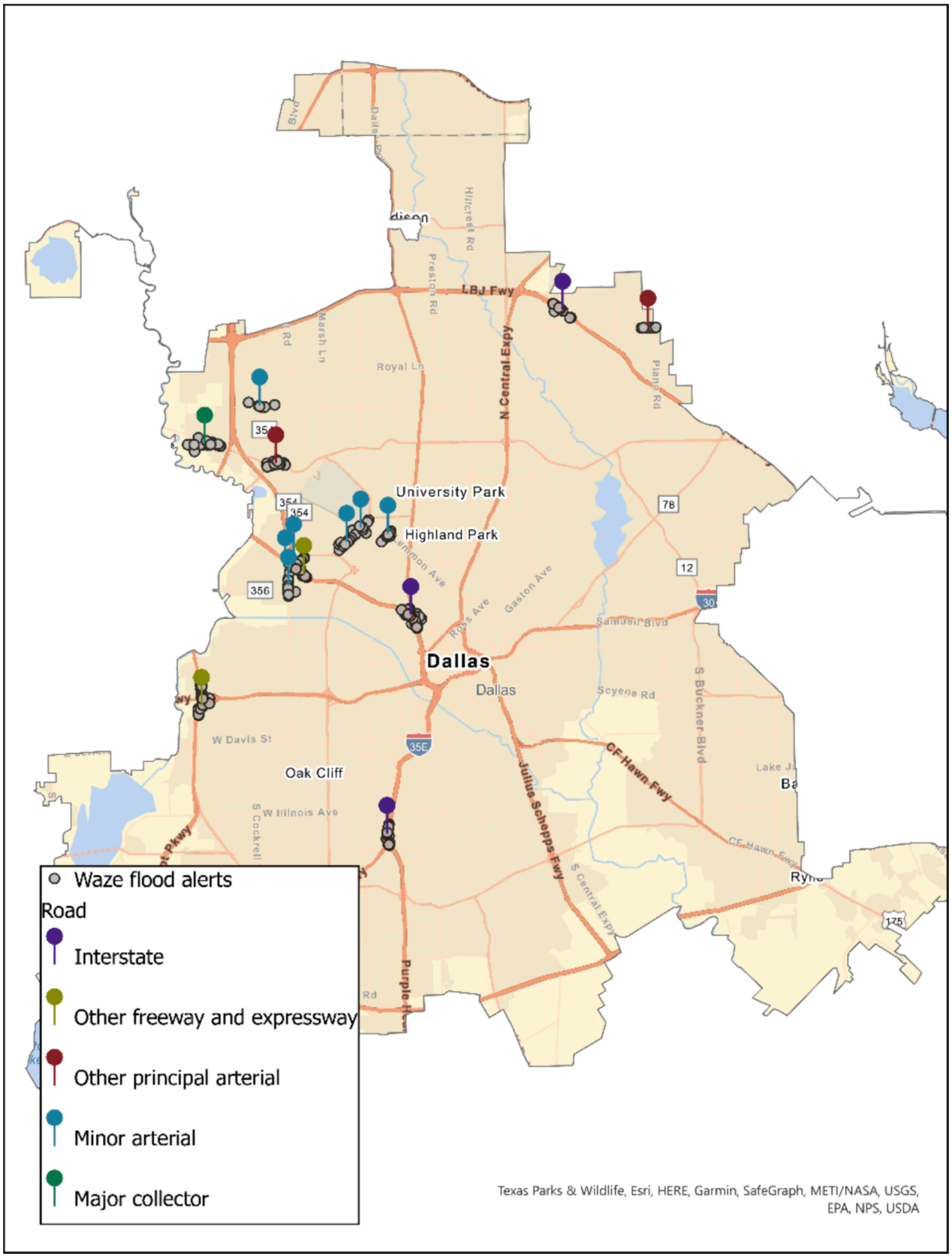


Fig. 2. Case study intersections and Waze flood alert.

Table 4
Storm information.

Storms	Max rainfall intensity (inch/hr)	Storm date	Duration (hrs)	Total rainfall depth (inch)	Number of Waze alerts
	1.14	2018-09-21 11:00:00	21	6.10	27
	1.62	2020-01-16 04:00:00	48	4.33	123
	0.95	2019-04-23 18:00:00	26	4.00	24
	1.85	2019-05-18 03:00:00	23	2.89	188
	0.65	2019-04-13 02:00:00	26	2.78	140
	0.99	2019-05-01 12:00:00	30	1.58	6
	2.01	2021-05-16 07:00:00	10	5.11	176
	2.26	2018-09-07 16:00:00	7	2.77	84
	1.41	2019-05-08 04:00:00	9	2.67	82
	0.74	2018-10-09 09:00:00	8	2.50	127
	1.67	2021-08-14 11:00:00	3	1.91	25
	0.33	2018-08-11 04:00:00	18	1.59	23
	0.63	2021-04-28 23:00:00	7	1.48	18
	1.09	2022-04-04 20:00:00	4	1.48	13
	1.1	2022-03-21 08:00:00	12	1.46	33
	0.68	2019-05-10 18:00:00	18	2.05	40
	1.35	2020-03-15 21:00:00	11	2.75	14
	1.31	2018-10-08 15:00:00	10	2.09	154
	0.65	2021-03-22 08:00:00	15	1.51	21

numerous negative examples of no flood alerts). To make models more sensitive to class imbalance and give higher importance to predicting instances when PFF risk has been observed (positive class), two steps are taken. First, the model is trained in a cost-sensitive approach by weighting loss values for minority (PFF risk observed) and majority (no PFF risk) classes. This method applies different weights to the loss computed for samples that are in the minority class when calculating the loss function. In this study, a higher weight is given to the misclassification of timesteps where PFF risk is observed than it is to the misclassification of timesteps where no flood alert was posted. In the model evaluation phase, frequently used performance metrics like accuracy might produce misleading assessments since they are insensitive to skewed data and give minority and majority classes the same priority. Therefore, metrics that are sensitive to class imbalance are used instead: Fβ score, area under the precision-recall curve (PRC-AUC), and area under the Receiver Operating Characteristic curve (ROC-AUC) (Buckland and Gey, 1994; Saito and Rehmsmeier, 2015).

Table 5
Configuration of tuned ML models.

Model	Six most important features	Model configuration
XGBoost	2-hours preceding accumulation Maximum inundation depth PFF likelihood in moderate storms Afternoon Time to the last maximum rain pulse Maximum preceding rain pulse	Maximum depth: 3 Number of trees: 1711 Positive class weight = 0.9
RFC	2-hours preceding accumulation Time to the last maximum rain pulse Maximum preceding rain pulse Maximum inundation depth PFF likelihood in moderate storms Count of rainless intervals	Maximum depth: 4 Number of features allowed: 6 Number of trees: 741 Positive class weight: 0.93
SVC	Maximum inundation depth Afternoon PFF likelihood in moderate storms Time to the last maximum rain pulse 2-hours preceding accumulation Maximum preceding rain pulse	Kernel = rbf gamma = 0.1 class weight = 0.90

Fβ score is an abstraction of F1 score in which recall's importance can be adjusted by a coefficient named β (Buckland and Gey, 1994). F1score, the harmonic means of recall (completion of retrievals, i.e., the probability of predicting PFF given its existence) and precision (purity of retrievals, i.e., the probability that PFF occurs given a predicted PFF), is a commonly used metric for imbalanced classification. However, utilizing the F1score assumes that recall and precision are equally significant, suggesting that stakeholders consider both false positive and false negative predictions similarly undesirable. Due to the ambiguity surrounding the crowdsourced proxy variable (Waze flood alerts), we cannot properly define the negative class. That is, the lack of a flood alert could indicate either no flooding or no driver posting an alert. As a result, prioritizing the positive class and reducing false negatives with the cost of increased false positive predictions can give us a higher confidence level overall. Consequently, higher importance should be assigned to the recall metric that focuses on the completion of minority class prediction.

To address this concern, precision, recall, and Fβ score are computed as shown in Equation (1) through (3).

$$\text{precision} = \frac{\text{tp}}{\text{tp} + \text{fp}} \tag{1}$$

$$\text{recall} = \frac{\text{tp}}{\text{tp} + \text{fn}} \tag{2}$$

where tp represent the number of observations in positive instances retrieved correctly, fp represents the number of negative observations retrieved incorrectly, and fn is the number of positive instances retrieved incorrectly. Fβ score is a metric that balances the recall and precision between 0 for the worst score and 1 for the perfect score (Equation (3).

$$\text{F}\beta\text{score} = \left(1 + \beta^2\right) \frac{\text{precision} \times \text{recall}}{\beta^2 \text{precision} + \text{recall}} \tag{3}$$

where β is a coefficient that controls the balance between precision and recall, with $\beta < 1$ when minimizing false positives and $\beta > 1$ when minimizing false negatives is the priority. In this study, $\beta = 2$ is to emphasize predicting observed risk incidents by making recall the priority. Given its benefits, Fβ score, with $\beta = 2$ (hereafter called F2score) is

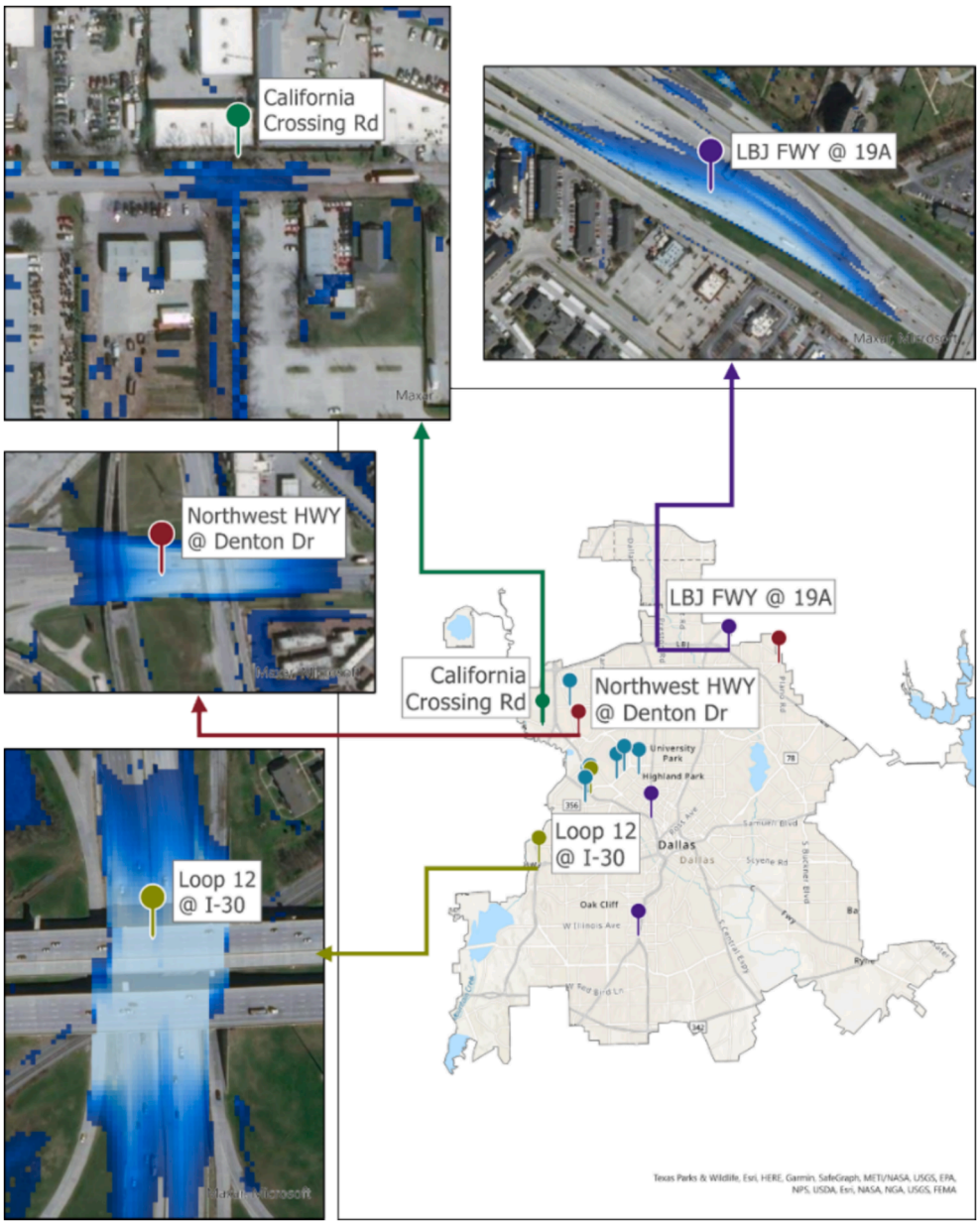


Fig. 3. Inundation extent on four road segments at the end of a severe storm on September 9, 2021.

utilized in cross-validated random search to rank models and find optimum hyper-parameters.

Precision, recall, and F2 score are single threshold measures and dependent on a cutoff threshold for the classifier to separate positive and negative classes, while the appropriate cutoff may vary depending on the application and dataset (Saito and Rehmsmeier, 2015). Model-wide threshold-free measurements are therefore needed to comprehensively assess the model's performance on an unknown test dataset. The Precision-Recall Curve (PRC) and Receiver Operating Characteristic (ROC) plots are model-wide measures that compare a model with a baseline no-skill performance benchmark. The ROC curve depicts the false positive rate versus the true positive rate over a range of thresholds. Although ROC is not biased by either the minority or the majority class, it might be deceptively optimistic when there are few positive samples focusing only on the positive class. The PRC curve is a more accurate measurement for unbalanced binary classification. The baseline performance of a model in the PRC curve can be calculated as Equation (4), where P and N represent the number of positive and negative samples, respectively. Since the baseline no-skill performance in the PRC curve is proportional to the number of the minority class, it is a better metric for imbalanced binary classification (Buckland and Gey, 1994; Saito and Rehmsmeier, 2015).

$$y = \frac{P}{P + N} \tag{4}$$

3. HYPERLINK "SPS:id::Sec2" Case study and datasets

The methods described above are evaluated at 15 intersections in the City of Dallas, Texas. The intersections are randomly selected from road segments prone to PFF. Fig. 2 shows the locations of the selected case study road segments and Table 4 provides a summary of the storm statistics used in the study (Table 5).

The Texas Department of Transportation's (TX-DOT) highway inventory and asset dataset contains features that describe traffic volume and road segment characteristics. The 15 road segments include arterials, major collectors, freeways, and interstates. (Fig. 2). The *AADT* of road segments ranges between 3,723 and 240,182 vehicles per day.

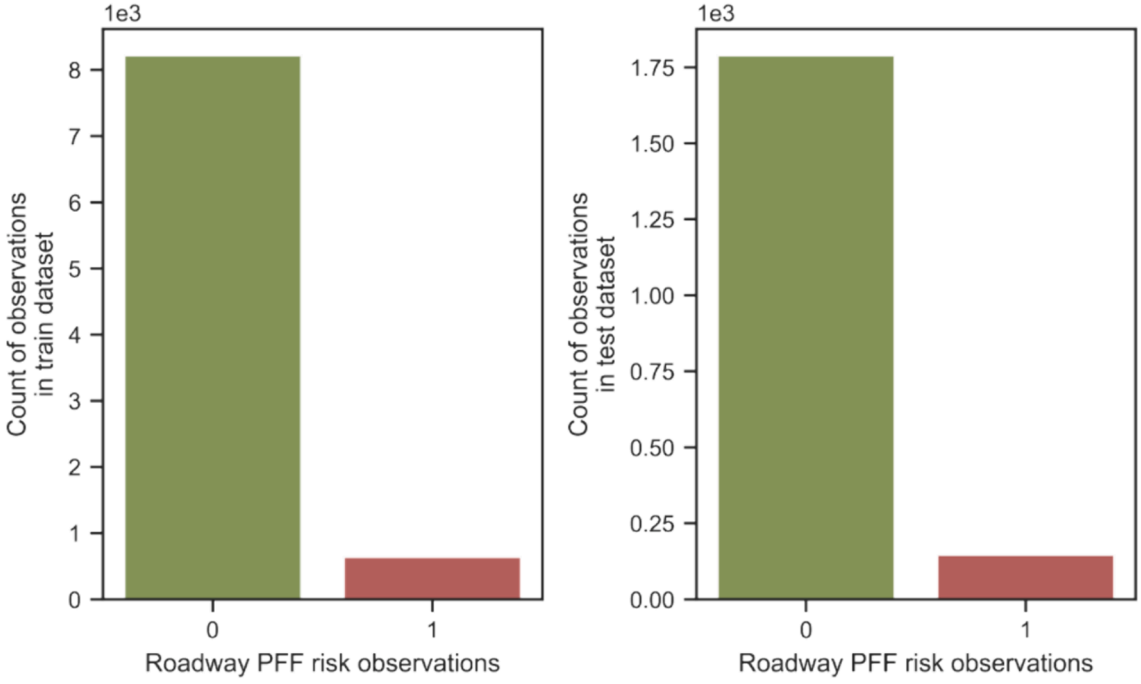


Fig. 4. Distribution of risk observations in training and testing datasets.

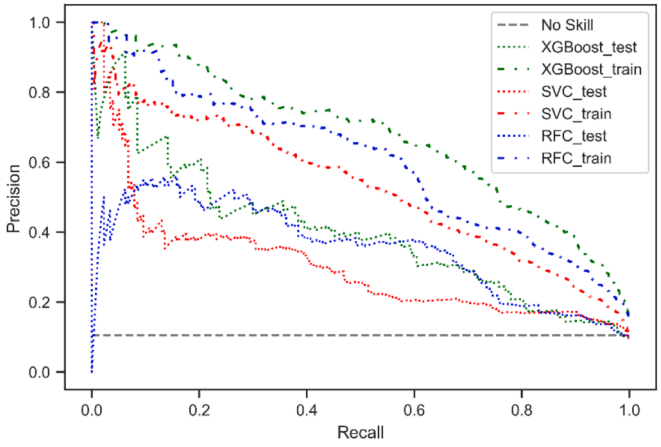


Fig. 5. Precision-recall curves (PRC) of ML models.

Merged depressions and delineated catchments extracted by the ArcGIS Python toolbox from 1-, 2-, and 3-meter resolution DEMs were identical; thus, to simplify the simulation, a DEM 3 m resolution is employed in the GB-RFSM model. Other data required in the GB-RFSM model include land cover, land use, and high-resolution catchments. NEXRAD precipitation data are obtained from National Center for Environmental Information (NCEI) for every 12-minute time interval and case study road segment with a spatial resolution of 3 m (TXDOT inventory). Twenty storm occurrences in total are selected from August 2018 to April 2022, although the Covid-19 shutdown period (March 1, 2020, to February 1, 2021) is not included in the research because of anomalies in traffic levels impacting crowdsourced data at that time. (Bureau of Transportation Statistic, 2021).

4. Results

In accordance with the methodology, the RFC, SVC, and XGBoost models are trained on the training dataset and their performance is evaluated on the unseen test dataset generated using storm events different from the training storms.

4.1. Data preparation

Fifteen of the storm events described in Table 4 are used to train the models and five storm events are held for testing the models. After retrieving precipitation data from NEXRAD, the optimum value of parameters n and h from Table 1 must be determined to generate hazard attributes. The correlation between the risk observation (target feature) and the accumulated rainfall during the previous n hours is used to determine the optimal *n*, testing several n values of 5, 4, 3, 2, and 1. The correlation is found to be greatest at n = 2 hours. In a similar process, the optimum h in Table 1 is found to be h = 4 mm with a correlation of 0.085.

Raw precipitation timeseries are used in the GB-RFSM model to estimate vulnerability features, including inundation area and the maximum depth on the accumulated road surface for 12-minute intervals (Table 2). Fig. 3 shows an example of the GB-RFSM inundation maps on four sample road segments at the end of a severe storm that happened on September 21, 2021.

4.2. Modeling

Distributions of observed roadway PFF risk in the training and testing datasets are depicted in Fig. 4. Both datasets have an uneven distribution of positive and negative classes, with negative classes being almost ten times more prevalent than positive classes, showing the imbalance in the datasets. Random grid search found the best model configuration for each algorithm, as shown in. The most important features are identified from models following the procedure described previously. XGBoost and SVC have at least one feature from hazard, vulnerability, and exposure, while no exposure attribute is significantly important for RFC. The important features also show that the most significant exposure attribute is whether the flood event occurred in the *afternoon TOD* or not. It appears that Waze users are more likely to post flood alerts between 12:00 and 4:00 PM.

Among the historical vulnerability attributes, PFF likelihood in a moderate storm and PFF likelihood in a severe storm, all three models

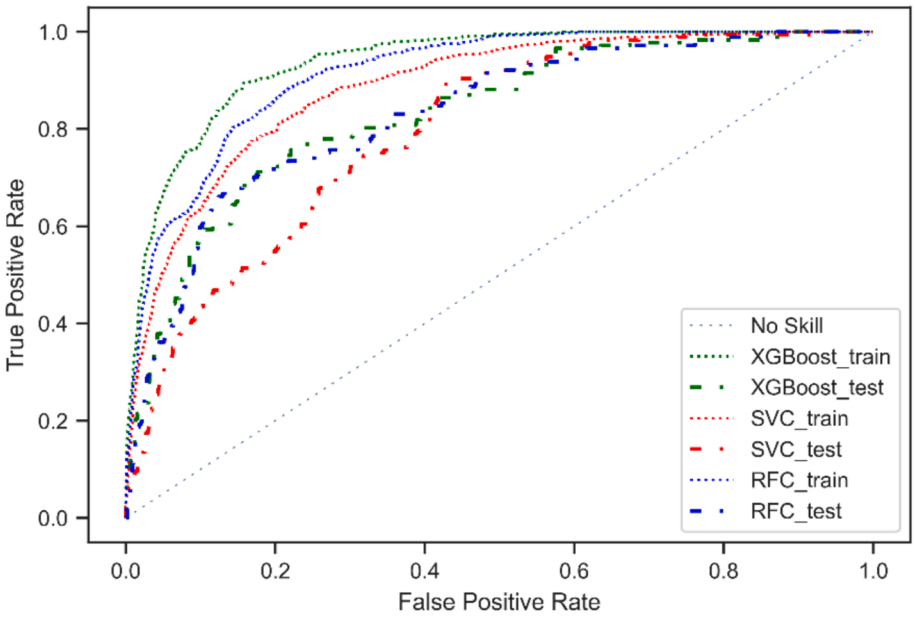


Fig. 6. ROC curve of ML models.

Table 6
ML model performances after tuning thresholds.

Model	PRC-AUC	ROC-AUC	Threshold	F2score	Precision	Recall
XGBoost	Train = 0.68 Test = 0.40	Train = 0.94 Test 0.83	0.48	Train: 0.70 Test: 0.54	Train: 0.37 Test: 0.29	Train:0.89 Test: 0.69
RFC	Train = 0.60 Test = 0.35	Train = 0.91 Test = 0.83	0.62	Train: 0.65 Test: 0.52	Train: 0.32 Test: 0.25	Train: 0.89 Test: 0.73
SVC	Train = 0.38 Test = 0.18	Train = 0.85 Test = 0.74	0.08	Train: 0.47 Test: 0.29	Train: 0.17 Test: 0.16	Train: 0.89 Test: 0.63

found EB estimations in a moderate storm class more significant than EB estimations in severe storms for predicting the risk. This may be due to the larger number of moderate storms used in calculating PFF likelihoods compared to severe storms. The larger sample size may cause the estimated likelihoods in the moderate cluster to be more accurate.

Figs. 5 and 6 show PRC and ROC curves, respectively, for three tuned ML models compared with a no-skill model. According to both the ROC and PRC curves, XGBoost performs better than RFC and SVC on both the training and testing datasets, as shown by the larger area under the curve (AUC) (Table 6). Further analysis of the PRC in Fig. 5 and the results in Table 6 reveal that RFC performs well in forecasting the positive class, driven by its higher recall. Conversely, XGBoost stands out in predicting the negative class, due to its higher precision and lower error rate in predicting the positive class. SVC shows limitations in predicting true positive instances (TP), indicating comparatively lower performance. In the ROC analysis (Fig. 6), XGBoost demonstrates a significantly greater distance from the no-skill method compared to RFC and SVC, suggesting higher true positive rates. This confirms XGBoost's superior performance in identifying negative instances, while SVC shows the lowest performance. Despite some differences, the trends in PRC and ROC curves generally align, indicating consistency in model evaluation across both metrics.

As mentioned in Section 0, the desired classification threshold (used to map predicted risk probabilities into risk classes) may vary depending on the data and application. Given the substantial uncertainty in the negative class due to the characteristics of crowdsourced data (i.e., smaller confidence in the negative class), forecasting the positive class is most important, as well as to reduce exposure to flooded areas. Therefore, to evaluate misclassification of each model when 90 % of risk instances are correctly predicted, the classification threshold is adjusted to map the projected probabilities to PFF risk and no-risk classes by setting the minimum required recall score to be 0.9. Using the adjusted threshold, the RFC, XGBoost and SVC predict 73 %, 69 % and 63 % of risk observations in the test storm events respectively, suggesting the superiority of the RFC model.

Table 6 shows the model performances and the adjusted threshold. XGBoost performs better on the testing and training dataset for the F2score. However, the recall score on the testing dataset is the highest for the RFC model. This conforms with the PRCs depicted in Fig. 5, which shows that the superiority of XGBoost to RFC in terms of PRC-AUC is due to higher performance on smaller recalls. However, in recalls higher than 0.2 their performance is almost identical on the test dataset.

The variability of important attributes representing hazard, vulnerability, and exposure (namely, last 2-hours precipitation, maximum inundation depth, and TOD) in true and false positive and negative predictions of the test dataset (tp, fp, tn, fn in Equation 2Equation 3) are shown in Figs. 7 to 9 for XGBoost, RFC, and SVC. The marginal histograms shown on these figures indicate the distributions of maximum depth estimated by GB-RFSM and last 2-hour precipitation for each set of predictions, respectively.

The distribution of predictions from XGBoost and RFC are fairly similar; for example, both predict minimal risks in instances where the accumulated precipitation in the last 2 h is less than 10 mm (Fig. 7-b, 7-d, 8-b, 8-d). Also, they cover a wide range of maximum inundation depths in the tn (true negative) category (Fig. 7-a, 8-a), which shows they successfully captured attributes other than maximum depth that impact PFF risk. These figures also suggest that XGBoost is not superior to RFC in predicting the positive class, given that it has fewer tp and more fn. Rather, its dominance over RFC comes from its ability to distinguish negative classes better, which are uncertain anyway (comparing Fig. 7-a, 8-a).

Unlike XGBoost and RFC, SVC does not show a wide range of maximum depths in the tn group (Fig. 9-a) suggesting that it has a wider range of maximum depth in the fp group (Fig. 9-b). Fig. 9-b and 9-

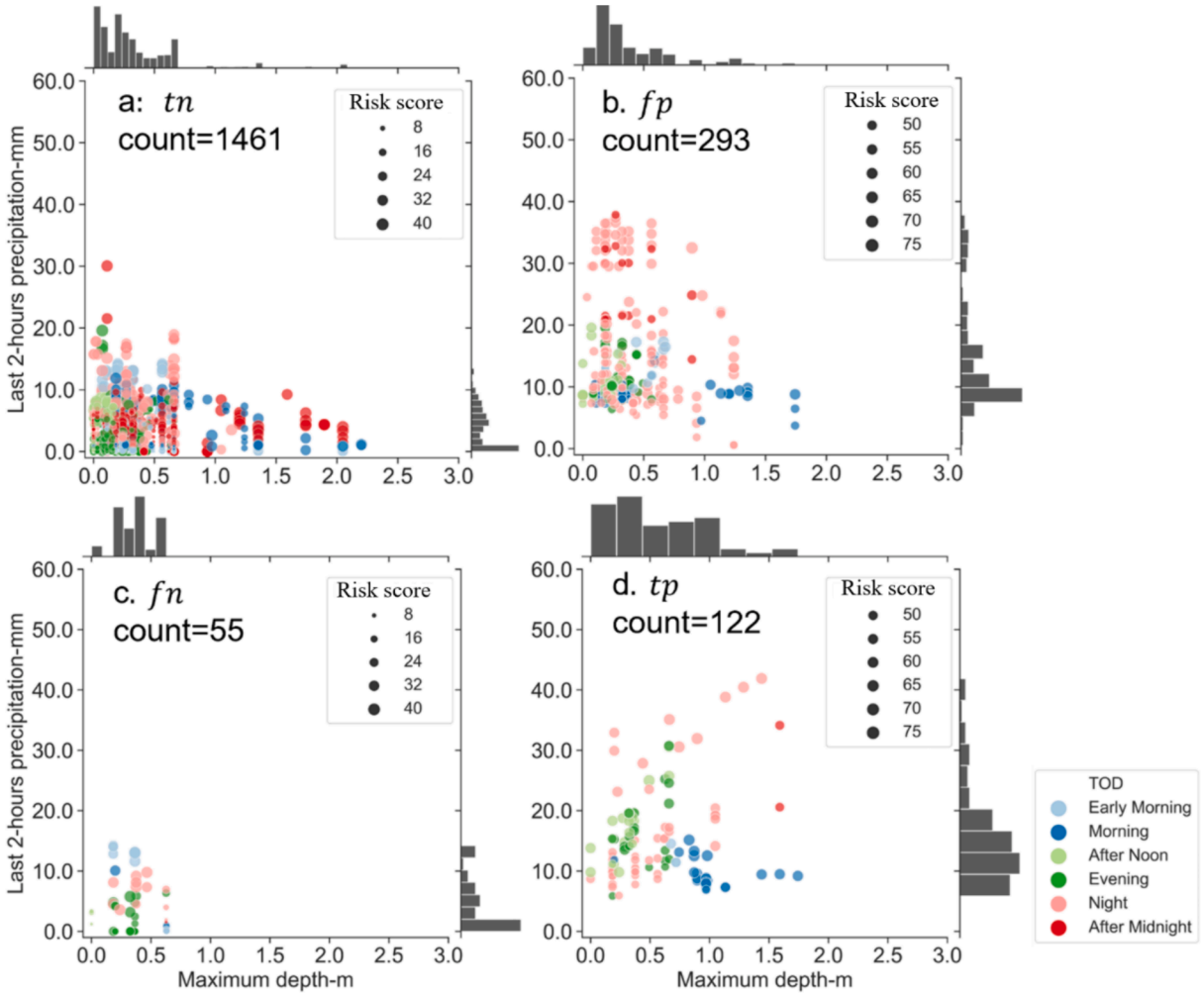


Fig. 7. Features distributions in confusion matrix of test predictions, a: True negative (tn) predictions, b: False positive (fp) predictions, c: False negative (fn) predictions, and d: True positive (tp) predictions – XGBoost.

d show that although SVC predicts low PFF risk in instances with the last 2-hours rainfall less than 10 mm, correctly retrieving 90 % of risk instances causes misclassification of low probabilities.

Fig. 7, Fig. 8, and Fig. 9 depict the confusion matrices corresponding to the predictions of the three machine learning models (XGBoost, RFC, and SVC). Each figure provides an overview of the model predictions categorized into true negative (tn), false positive (fp), false negative (fn), and true positive (tp) classifications. These visual representations offer insights into the performance and predictive capabilities of each model across various risk scenarios and conditions. The x-axis of each figure represents the precipitation in the last two hours, while the y-axis represents the maximum rainfall depth. In these figures, the color legend represents different times of the day, while the varying sizes of dots indicate the risk score (number of Waze flood reports).

Looking closely at Fig. 7-a and 8-a, it can be seen that instances with large maximum depths in the tn groups mostly occur after midnight and in the morning, when fewer vehicles are on the roads and preceding rainfall is not as high. This implies that these no-risk observations are due to low exposure and low hazard, while vulnerability is still high at locations that could cause roadway PFF.

In addition, the fn group of model predictions (Fig. 7-c, 8-c, 9-c) all have low maximum depth and recent 2 h of precipitation, which suggests low vulnerability and hazard, respectively. The range of recent precipitation between 30 and 40 mm is identical for the three models in the fp group (Fig. 7-b, 8-b, and 9-b). Even though all models indicate a high risk of PFF for these datapoints, there are no risk (Waze) observations available. All of these timesteps occur at night (8 pm to midnight) and after midnight, when there is less traffic and presumably fewer Waze users reporting risk observations. These findings underscore the importance of considering temporal dynamics and reporting biases in interpreting model predictions and assessing pluvial flash flood risk accurately.

5. Conclusions

This paper shows that crowdsourced traffic flood alerts, specifically Waze flood alerts, can be a valuable data source as a proxy to roadway PFF risk, which is a combination of hazard, vulnerability, and exposure. Other data sources used in this work include: (1) likelihoods of PFF in storm clusters (light, moderate, and severe), calculated by our previously-published EB regression model and maximum inundations from GB-RFSM, representing site-specific vulnerabilities of road segments to PFF; (2) hazard of a road segment to PFF, estimated based on NEXRAD precipitation and time; and (3) AADT and road classifications are used to estimate exposure.

Three ML models (XGBoost, RFC, and SVC) are trained using a curated dataset to predict the risk of roadway PFF. The three models identify the following key hazard and vulnerability features: maximum inundation depth, 2 h of preceding precipitation, PFF likelihoods during moderate storms, and time to the most recent maximum rain pulse. While AADT and road classification were not significant in either model, time of day is the most important factor among features that indicate

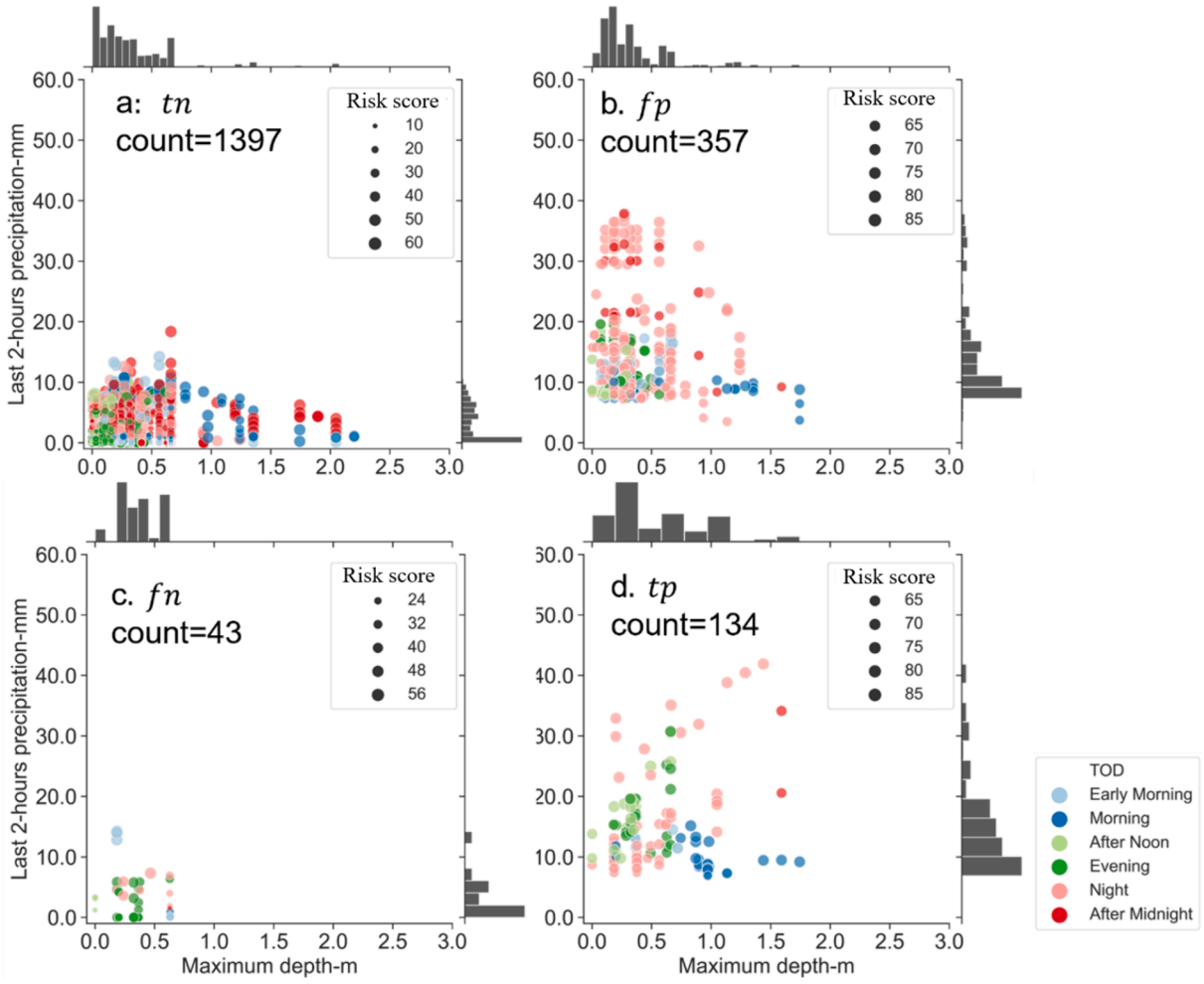


Fig. 8. Features distributions in confusion matrix of test predictions, a: True negative (tn) predictions, b: False positive (fp) predictions, c: False negative predictions (fn), and d: True positive (tp) predictions −RFC.

exposure.

Several strategies are employed to train the models in a way appropriate for imbalanced data and uncertainty in the negative class (no flood risk); First, cost-sensitive training is carried out, which assigns a higher weight to the minority (positive) class mispredictions. Second, in order to emphasize recall over precision when adjusting model hyperparameters, F2 score is utilized in the model selection and cross-validation process. Finally, the classification threshold is modified to extract a training recall score of 0.9 in order to compare models in the high recall area of the PRC and select the model that is more effective at predicting positive risk cases, which represent more risk to motorists. F2 score is then used to compare the models after mapping probabilities using the updated threshold.

The modeling results showed that XGBoost and RFC both performed superior to the SVC model by having higher PRC-AUC and ROC-AUC. Even though the XGBoost and RFC models both had greater AUCs, the threshold change showed that by setting the required training recall to 0.9, the XGBoost model outperforms the RFC model in terms of test precision while the RFC model outperforms the XGBoost model in terms of test recall. Therefore, deciding between the two models depends on the level of uncertainty in the negative class and the cost of false positive predictions. In vehicle routing applications, depending on changes in travel costs and times when rerouting vehicles, higher recall or higher precision may be preferred. Generally, if rerouting does not significantly increase travel costs and times, predicting false positives is justified to minimize the probability of exposure to a flooded road; hence a higher recall is preferred. However, if false positives significantly raise travel costs or turnaround times, it will be more reasonable to avoid predicting risk when there is no risk (fp), even if doing so results in missing certain risk instances with low probabilities. In such a case, the model would favor higher precision. The modeling approach taken in this work assumes that capturing true positives is more important than minimizing false positives to protect motorists from flood exposure.

Overall, the RFC model predicts 73 % of risk observations (i.e., Waze alerts) during the test storm events. The presented modeling approach is informative to roadway PFF awareness that could increase travel safety. Mispredictions with high hazard and vulnerability mostly occur at night when exposure is lower and fewer Waze alerts are available.

In terms of limitations, the vulnerability features (maximum inundation depth, and EB-derived PFF likelihoods) are simplified attributes. For instance, the GB-RFSM used to compute maximum inundation depth assumes that there is no subsurface drainage system for excess runoff (due to lack of data) and therefore accumulates all excess runoff into low-lying regions. Deploying this model in a hybrid approach allows such errors to be corrected by the machine learning model. Additionally, developing the site-specific PFF likelihoods involves manual pre-processing and judgment, adding additional uncertainty to the data. Like all machine learning methods, the models are limited by the training dataset and future extreme events beyond the current dataset may not be accurately predicted. Using reliable, high-fidelity records of historical flood depth measurements and storm drainage system configurations and performance in the models could improve the model

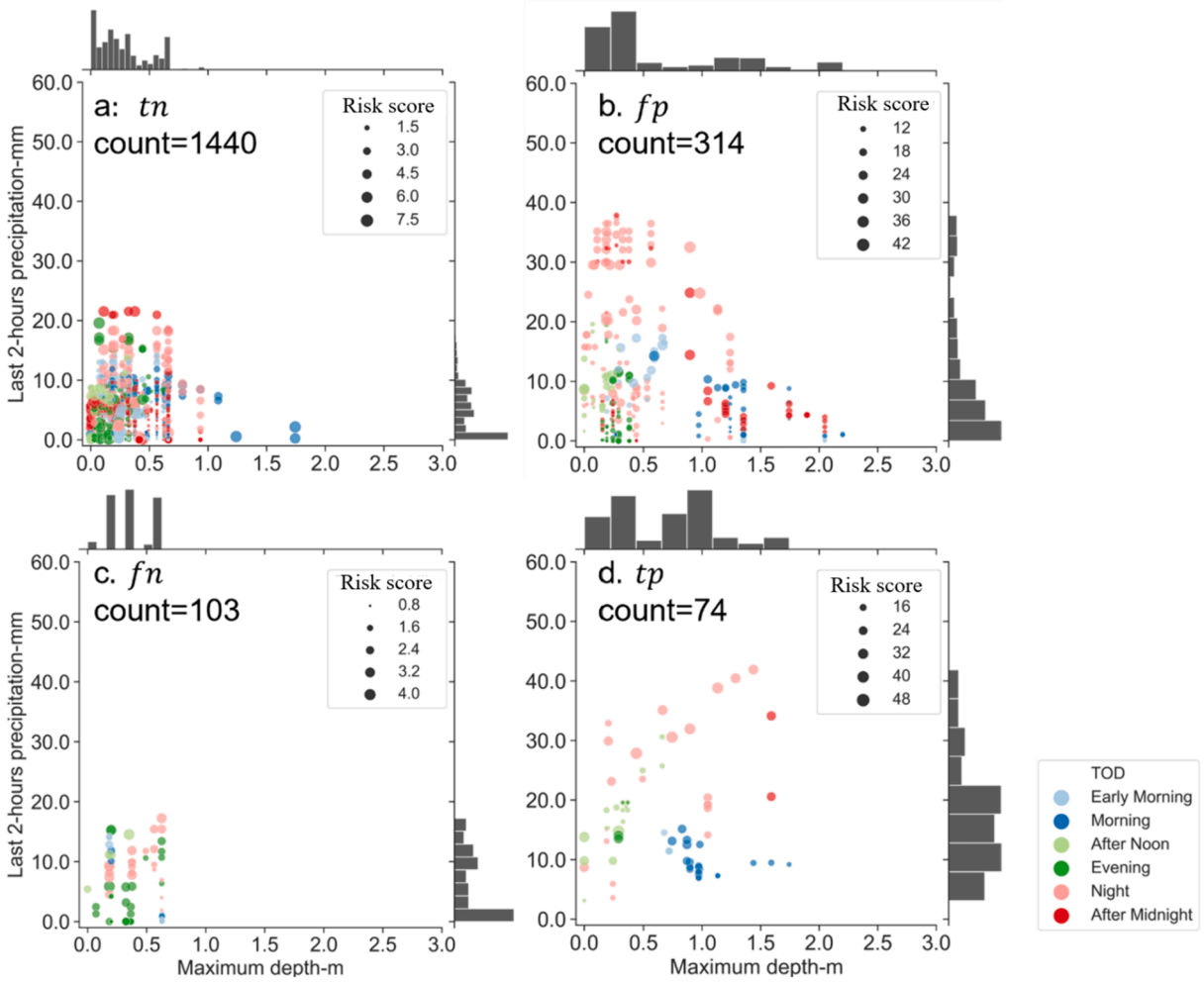


Fig. 9. Features distributions in confusion matrix of test predictions, a: True negative (tn) predictions, b: False positive (fp) predictions, c: False negative (fn) predictions, and d: True positive (tp) predictions – SVC.

performance. Finally, employing more complex calibrated hydraulic and hydrologic models that account for more PFF mechanisms may improve predictive performance, at the cost of significantly more detailed data requirements and higher computational effort.

CRediT authorship contribution statement

Arefeh Safaei-Moghadam: Software, Resources, Methodology, Formal analysis. **Azadeh Hosseinzadeh:** Writing – original draft, Visualization, Resources, Investigation. **Barbara Minsker:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This research was supported by the National Institute of Standards and Technology (grant no. 60NANB17D180). The North Central Texas Council of Governments provided access to the Waze flood alert data considered in the case study.

References

Berkhahn, S., Fuchs, L., Neuweiler, I., 2019. An ensemble neural network model for real-time prediction of urban floods. J. Hydrol. 575, 743–754.
Bowler, D.E., Bhandari, N., Repke, L., Beuthner, C., Callaghan, C.T., Eichenberg, D., Henle, K., Klenke, R., Richter, A., Jansen, F., 2022. Decision-making of citizen scientists when recording species observations. Sci. Rep. 12 (1), 11069.
Buckland, M., Gey, F., 1994. The relationship between recall and precision. J. Am. Soc. Inf. Sci. 45 (1), 12–19.
Bulti, D.T., Abebe, B.G., 2020. A review of flood modeling methods for urban pluvial flood application. Model. Earth Syst. Environ. 6, 1293–1302.
Bureau of Transportation Statistic. (2021). *Bureau of Transportation Statistic*.
Campagna, M., 2016. Social Media Geographic Information: Why social is special when it goes spatial. *European Handbook of Crowdsourced Geographic*. Information 45.
Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 785–794.
Craglia, M., Ostermann, F., Spinsanti, L., 2012. Digital Earth from vision to practice: Making sense of citizen-generated content. Int. J. Digital Earth 5 (5), 398–416.
Cristiano, E., ten Veldhuis, M., Van De Giesen, N., 2017. Spatial and temporal variability of rainfall and their effects on hydrological response in urban areas–a review. Hydrol. Earth Syst. Sci. 21 (7), 3859–3878.
Estellés-Arolas, E., González-Ladrón-de-Guevara, F., 2012. Towards an integrated crowdsourcing definition. J. Inf. Sci. 38 (2), 189–200.
Fang, X., Wu, J., Li, H., Jiang, P., Wang, X., Zhang, S., Wang, C., & Liu, K. (n.d.). A Hybrid Model Integrating Hydrodynamics Mechanism and Deep Learning for Real-Time Flood Routing Analysis. *Available at SSRN 4411807*.

Farahmand, H., Xu, Y., Mostafavi, A., 2023. A spatial–temporal graph deep learning model for urban flood nowcasting leveraging heterogeneous community features. Sci. Rep. 13 (1), 6768.
Feng, Y., Huang, X., Sester, M., 2022. Extraction and analysis of natural disaster-related VGI from social media: review, opportunities and challenges. Int. J. Geogr. Inf. Sci. 36 (7), 1275–1316.
Gouldby, B., & Samuels, P. (2005). *Language of risk, project definitions, FLOODsite project report T32-04-01*. EU GOCE-CT-2004-505420. http://www. floodsite. net/html/partner_area ….
Guidolin, M., Chen, A.S., Ghimire, B., Keedwell, E.C., Djordjević, S., Savić, D.A., 2016. A weighted cellular automata 2D inundation model for rapid flood analysis. Environ. Model. Softw. 84, 378–394.
Han, D., Chan, L., Zhu, N., 2007. Flood forecasting using support vector machines. J. Hydroinf. 9 (4), 267–276.
Haworth, B., Bruce, E., 2015. A review of volunteered geographic information for disaster management. Geogr. Compass 9 (5), 237–250.
He, H., Li, R., Pei, J., Bilodeau, J.-P., Huang, G., 2023. Current overview of impact analysis and risk assessment of urban pluvial flood on road traffic. Sustain. Cities Soc. 104993.
Hosseinzadeh, A., Behzadian, K., Rossi, P., Karami, M., Ardeshir, A., Torabi Haghighi, A., 2023. A new multi-criteria framework to identify optimal detention ponds in urban drainage systems. J. Flood Risk Manage. e12890.
Huang, R., Ma, C., Ma, J., Huangfu, X., He, Q., 2021. Machine learning in natural and engineered water systems. Water Res. 205, 117666.
Janizadeh, S., Vafakhah, M., Kapelan, Z., Mobarghaee Dinan, N., 2022. Hybrid XGboost model with various Bayesian hyperparameter optimization algorithms for flood hazard susceptibility modeling. Geocarto Int. 37 (25), 8273–8292.
Karami, M., Behzadian, K., Ardeshir, A., Hosseinzadeh, A., Kapelan, Z., 2022. A multi-criteria risk-based approach for optimal planning of SuDS solutions in urban flood management. Urban Water J. 19 (10), 1066–1079.
Ke, Q., Tian, X., Bricker, J., Tian, Z., Guan, G., Cai, H., Huang, X., Yang, H., Liu, J., 2020. Urban pluvial flooding prediction by machine learning approaches–a case study of Shenzhen city, China. Adv. Water Resour. 145, 103719.
Kecman, V., 2001. Learning and soft computing: Support vector machines, neural networks, and fuzzy logic models. MIT Press.
Kim, H.I., Keum, H.J., Han, K.Y., 2019. Real-time urban inundation prediction combining hydraulic and probabilistic methods. Water 11 (2), 293.
Lhomme, J., Sayers, P., Gouldby, B. P., Samuels, P. G., Wills, M., & Mulet-Marti, J. (2008a). Recent development and application of a rapid flood spreading method.
Li, X., Willems, P., 2020. A hybrid model for fast and probabilistic urban pluvial flood prediction. Water Resour. Res. 56 (6) e2019WR025128.
Liu, J., Wang, J., Xiong, J., Cheng, W., Li, Y., Cao, Y., He, Y., Duan, Y., He, W., Yang, G., 2022. Assessment of flood susceptibility mapping using support vector machine, logistic regression and their ensemble techniques in the Belt and Road region. Geocarto Int. 37 (25), 9817–9846.
Manfreda, S., Samela, C., 2019. A digital elevation model based method for a rapid estimation of flood inundation depth. J. Flood Risk Manage. 12, e12541.
Moon, H., Yoon, S., Moon, Y., 2023. Urban flood forecasting using a hybrid modeling approach based on a deep learning technique. J. Hydroinf. 25 (2), 593–610.
Noh, S.J., Lee, J.-H., Lee, S., Kawaike, K., Seo, D.-J., 2018. Hyper-resolution 1D–2D urban flood modelling using LiDAR data and hybrid parallelization. Environ. Model. Softw. 103, 131–145.
Oneto, G., Canepa, M., 2023. Addressing sustainable urban flood risk: Reviewing the role and scope of theoretical models and policies. Water Policy 25 (8), 797–814.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.
Preisser, M., Passalacqua, P., Bixler, R.P., Hofmann, J., 2022. Intersecting near-real time fluvial and pluvial inundation estimates with sociodemographic vulnerability to quantify a household flood impact index. Hydrol. Earth Syst. Sci. 26 (15), 3941–3964.
Ren, M., Zhang, Z., Zhang, J., Mora, L., 2022. Understanding the use of heterogenous data in tackling urban flooding: An integrative literature review. Water 14 (14), 2160.
Sadler, J.M., Goodall, J.L., Morsy, M.M., Spencer, K., 2018. Modeling urban coastal flood severity from crowd-sourced flood reports using Poisson regression and Random Forest. J. Hydrol. 559, 43–55.
Safaei-Moghadam, A., Tarboton, D., Minsker, B., 2023. Estimating the likelihood of roadway pluvial flood based on crowdsourced traffic data and depression-based DEM analysis. Nat. Hazards Earth Syst. Sci. 23 (1), 1–19.
Safaei-Moghadam, A., Tarboton, D., Heidari, B., Jaber, F., Minsker, B. Graph-Based Rapid Flood Spreading Model for Real-Time Estimation of Hyper-Local Roadway Flooding Vulnerability. [Manuscript in preparation].
Saha, A., Pal, S.C., Arabameri, A., Blaschke, T., Panahi, S., Chowdhuri, I., Chakrabortty, R., Costache, R., Arora, A., 2021. Flood susceptibility assessment using novel ensemble of hyperpipes and support vector regression algorithms. Water 13 (2), 241.
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.
Samela, C., Persiano, S., Bagli, S., Luzzi, V., Mazzoli, P., Humer, G., Reithofer, A., Essenfelder, A., Amadio, M., Mysiak, J., 2020. Safer_RAIN: A DEM-based hierarchical filling-&-Spilling algorithm for pluvial flood hazard assessment and mapping across large urban areas. Water 12 (6), 1514.
Sanders, W., Li, D., Li, W., Fang, Z.N., 2022. Data-driven flood alert system (FAS) using extreme gradient boosting (XGBoost) to forecast flood stages. Water 14 (5), 747.
Santos, P.P., Pereira, S., Zêzere, J.L., Tavares, A.O., Reis, E., Garcia, R.A., Oliveira, S.C., 2020. A comprehensive approach to understanding flood risk drivers at the municipal level. J. Environ. Manage. 260, 110127.
Shen, J., Tong, Z., Zhu, J., Liu, X., Yan, F., 2016. A new rapid simplified model for urban rainstorm inundation with low data requirements. Water 8 (11), 512.
TXDOT inventory. (n.d.). Https://gis-txdot.opendata.arcgis.com/datasets/TXDOT::txdot-roadway-inventory/.
Tyralis, H., Papacharalampous, G., Langousis, A., 2019. A brief review of random forests for water scientists and practitioners and their recent history in water resources. Water 11 (5), 910.
Wu, Q., Lane, C.R., Wang, L., Vanderhoof, M.K., Christensen, J.R., Liu, H., 2019. Efficient delineation of nested depression hierarchy in digital elevation models for hydrological analysis using level-set method. JAWRA J. Am. Water Resour. Assoc. 55 (2), 354–368.
Xiong, J., Li, J., Cheng, W., Wang, N., Guo, L., 2019. A GIS-based support vector machine model for flash flood vulnerability assessment and mapping in China. ISPRS Int. J. Geo Inf. 8 (7), 297.
Yao, L., Chen, L., Wei, W., 2016. Assessing the effectiveness of imperviousness on stormwater runoff in micro urban catchments by model simulation. Hydrol. Process. 30 (12), 1836–1848.
Yong-He, L., Wan-Chang, Z., Jing-Wen, X., 2009. Another fast and simple dem depression-filling algorithm based on priority queue structure. Atmos. Oceanic Sci. Lett. 2 (4), 214–219.
Zahura, F.T., Goodall, J.L., Sadler, J.M., Shen, Y., Morsy, M.M., Behl, M., 2020. Training machine learning surrogate models from a high-fidelity physics-based model: Application for real-time street-scale flood prediction in an urban coastal community. Water Resour. Res. 56 (10) e2019WR027038.
Zheng, X., Maidment, D.R., Tarboton, D.G., Liu, Y.Y., Passalacqua, P., 2018. GeoFlood: Large-scale flood inundation mapping based on high-resolution terrain analysis. Water Resour. Res. 54 (12), 10013–10033.