Quantification of continuous flood hazard using random forest classification and flood insurance claims at large spatial scales: a pilot study in southeast Texas

William Mobley[1], **Antonia Sebastian**[1,2], **Russell Blessing**[1], **Wesley E. Highfield**[1], **Laura Stearns**[1], and **Samuel D. Brody**[1]

[1]Department of Marine Sciences, Texas A&M University at Galveston, Galveston, Texas, USA
[2]Department of Geological Sciences, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

Correspondence: William Mobley (wmobley@tamu.edu)

Abstract. Pre-disaster planning and mitigation necessitate detailed spatial information about flood hazards and their associated risks. In the US, the Federal Emergency Management Agency (FEMA) Special Flood Hazard Area (SFHA) provides important information about areas subject to flooding during the 1 % riverine or coastal event. The binary nature of flood hazard maps obscures the distribution of property risk inside of the SFHA and the residual risk outside of the SFHA, which can undermine mitigation efforts. Machine learning techniques provide an alternative approach to estimating flood hazards across large spatial scales at low computational expense. This study presents a pilot study for the Texas Gulf Coast region using random forest classification to predict flood probability across a 30 523 km^2 area. Using a record of National Flood Insurance Program (NFIP) claims dating back to 1976 and high-resolution geospatial data, we generate a continuous flood hazard map for 12 US Geological Survey (USGS) eight-digit hydrologic unit code (HUC) watersheds. Results indicate that the random forest model predicts flooding with a high sensitivity (area under the curve, AUC: 0.895), especially compared to the existing FEMA regulatory floodplain. Our model identifies 649 000 structures with at least a 1 % annual chance of flooding, roughly 3 times more than are currently identified by FEMA as flood-prone.

1 Introduction

In the US, pluvial and fluvial flood events are some of the most damaging environmental hazards, averaging USD 3.7 billion annually, totaling over USD 1.5 trillion in total losses since 1980 (NOAA National Centers for Environmental Information, NCEI, 2021). This trend represents an increase of about 15 % in flood losses per year since 2002, despite large-scale efforts to mitigate losses over the same period (Kousky and Shabman, 2017). To offset the rising costs associated with extreme flood events, pre-disaster planning and mitigation necessitate detailed spatial information about flood hazards and their associated risks.

In the US, the Federal Emergency Management Agency (FEMA) Special Flood Hazard Area (SFHA) – the area of inundation associated with a 1 % annual exceedance probability – provides a basis for community and household planning and mitigation decisions (Blessing et al., 2017). These maps, intended to be used to set flood insurance rates, have become the de facto indicator of flood risk nationwide and are the primary reference point when making a vast array of decisions related to flood risk such as where it is safe to develop, household protective actions, and local mitigation policies. However, because the SFHA is binary, little information is provided about the distribution of risk to properties inside of the mapped flood hazard areas and, even more so, for residual risks to properties outside of the mapped flood hazard areas. Thus, the floodplain boundaries result in "dichotomous decisions", whereby a house is treated the same whether it is 1 m or 1 km outside of the floodplain boundary (Brody

et al., 2018; Morss et al., 2005). As a result, homes are being built and purchased in at-risk areas, and actions that would increase flood resilience, such as meeting National Flood Insurance Program (NFIP) minimum design and construction requirements, are not being adopted.

Further compounding the SFHA's inability to indicate at-risk areas is that many of the nation's flood hazard maps are out of date or, worse, non-existent. For example, a 2017 study found that over a quarter of high-risk counties have US flood hazard maps over 10 years old, failing to capture recent changes in climatology and land use and land cover that can heighten risk (Office, 2017). Although FEMA's Map Modernization (Map Mod) program updated many of the nation's flood maps from 2003 to 2008, it still struggles to keep them up to date (ASFPM, 2020) in part because the models used to produce the FEMA SFHAs are discontinuous across large spatial scales and often commissioned on a patchwork, community-by-community basis that is both slow to implement and resource-intensive (FEMA, 2015). A recent study found that 15 % of the NFIP communities have flood hazard maps that are over 15 years old and that only a third of that nation's streams have flood maps (ASFPM, 2020). Updating the nations floodplain maps and mapping previous unmapped areas constitute a costly solution at an estimated USD 3.2 billion to USD 11.8 billion, with an additional USD 107 million to USD 480 million per year to maintain them (ASFPM, 2020).

An alternative approach that is gaining popularity are models that can leverage both big data (e.g., large-scale remotely sensed imagery) and high-performance computing environments to efficiently estimate flood hazard at continental (Wing et al., 2018; Bates et al., 2020) and global scales (Ward et al., 2015; Ikeuchi et al., 2017). These models overcome some of the limitations of community-by-community watershed-scale modeling by providing a methodologically consistent approach using comprehensive spatial datasets that is less resource-intensive. Although capable of accurately identifying general patterns of flood hazard over large areas, these models often struggle to correctly estimate small-scale variability in stormwater dynamics within highly developed areas. That is, model performance at community and household scales is highly dependent on detailed information about local-scale hydrology and flood control infrastructure, with limited observational data to validate the results. Moreover, these models still suffer from the same sources of uncertainty found at the watershed scale including, specifically, natural uncertainty arising from physical processes driving runoff (Singh, 1997), uncertainty found within model parameterization (Moradkhani and Sorooshian, 2009), and uncertainty arising from the quality or availability of input data (Rajib et al., 2020). Accounting for uncertainty at locally relevant scales would likely require significant increases in observational hydrometeorological data and immense financial resources.

Machine learning (ML) methods provide a potential alternative to estimate flood hazards based on historical records of flood loss, especially in resource-limited areas. One such ML algorithm that has shown to be particularly effective within flood hazard mapping are random forests, which have recently been used to create a spatially complete floodplain map of the conterminous United States (Woznicki et al., 2019) and to predict flood insurance claims at US Census tract and parcel levels (Knighton et al., 2020). Although initial work has shown that random forests improve the prediction of flood hazard, there have been no studies that have used historic records of structural flood damage to estimate a probabilistic floodplain. Also, there has been little to no effort to compare random forest predictions against existing regulatory floodplains.

We address these gaps by introducing a novel method to map flood hazards continuously across large spatial scales using a random forest classification procedure trained on 40 years of historic flood damage records from the NFIP. Using the NFIP data and high-resolution geospatial data (e.g., topographic, land use and land cover, soil data), we generate flood hazard maps for a large coastal area in the Texas Gulf Coast region. We then compare our modeled outputs against the FEMA floodplains using multiple metrics at regional and community scales. The following sections provide further background information on machine learning algorithms and their application to hazards research (Sect. 2), describe the methods and data used for our analysis (Sect. 3), and present the model results (Sect. 4). This is followed by a discussion (Sect. 5) and conclusions (Sect. 6).

2 Background

Data-driven models – those that use statistical or machine learning algorithms for empirical estimations – are prevalent in water resource research (Solomatine and Ostfeld, 2008) and are rapidly gaining popularity for prediction and estimation of flooding in the hydrological sciences (Mosavi et al., 2018). For instance, models predict stream discharge rates (Albers and Déry, 2015) and estimate insured flood damage either for the household (Wagenaar et al., 2017) or aggregated (Brody et al., 2009). Data-driven approaches have also identified flood probability during a flood event (Mobley et al., 2019) (i.e., flood hazard). When data in an area are sparse, these models can help better describe the system (Rahmati and Pourghasemi, 2017). Often however, these approaches require large datasets for an accurate representation (Solomatine and Ostfeld, 2008). In answer to the data problem, many data-driven models rely on non-traditional data sources. For example, using video frames, flood waters can be identified for a given location (Moy de Vitry et al., 2019).

Flood hazard models use dichotomous variables and driver layers to predict the likelihood of flooding across the landscape. This dichotomous variable can come from a variety of

datasets, such as high water marks or flood losses (Knighton et al., 2020). Depending on data availability, models have predicted flood hazard across time (Darabi et al., 2019) or for specific events (Lee et al., 2017), producing one of two outputs: a probability of flooding (Darabi et al., 2019) or discrete classes of susceptibility (Darabi et al., 2019; Dodangeh et al., 2020), both of which give more information than the current SFHA dichotomy, in or out of the flood zone. By modeling flood hazard through this process, researchers are able to map large geographic areas (Hosseini et al., 2019) with potential to scale further given sufficient data availability. This method has been shown to quickly (Mobley et al., 2019) and accurately estimate flooding (Bui et al., 2019).

The number of data available and how they are structured affect the techniques available for predicting flood hazard. Data-driven flood hazard models require point locations where flooding has occurred. If non-flooded locations are unavailable, pseudo-absences can be randomly generated for modeling (Barbet-Massin et al., 2012). Common algorithms used for predicting flood hazard models include neural networks (Janizadeh et al., 2019; Bui et al., 2019); support vector machines (SVMs) (Tehrany et al., 2019a); and decision trees, such as random forests (Woznicki et al., 2019; Muñoz et al., 2018). While other algorithms have been used for predicting flood hazard (Bui et al., 2019), these three algorithms are often used in machine learning due to their maturity in research and their generalizability. SVMs are ideal in areas with small sample sizes, but computation times are quadratic as sample sizes increase (Li et al., 2009). The computational complexity of the model removes the scalability of the model, a primary benefit over physical models. Neural networks are often cited as highly accurate (Mosavi et al., 2018) but often come with a reproducibility problem (Hutson, 2018).

Finally, computational requirements for decision trees are lower than the other two algorithms. The random forest algorithm comes from the decision tree family of models. The random forest model is highly generalizable within the drivers' parameters. Decision trees are non-parametric and use logic to branch at different values within the independent variables to best fit the classification (Quinlan, 1986). By creating numerous trees and democratizing the decision, ensemble classifiers reduce overfitting of the final model while maintaining accuracy (Breiman, 2001). Each tree is given a random subsample of the independent variables to predict the dependent variable. These ensemble classifiers are computationally efficient and maintain a high degree of sensitivity (Belgiu and Drăgut, 2016). Random forests are capable of identifying interactions between independent variables and the dependent variable regardless of the effect size (Upstill-Goddard et al., 2013).

3 Methods and materials

3.1 Study area

We use the US Geological Survey (USGS) Watershed Boundary Dataset to delineate a region encompassing 13 eight-digit hydrologic unit code (HUC) watersheds that drain to Galveston Bay and the intercoastal waterway as well as the lower Trinity and Brazos rivers and the San Bernard River (Fig. 1). The resulting study area encompasses $28\,000\,\mathrm{km^2}$ and includes two economic population centers: the Houston–Woodlands–Sugar Land area, also known as Greater Houston, home to 4.7 million people and the Beaumont–Port Arthur–Orange Area, also known as the Golden Triangle, home to 0.39 million people (Bureau, 2019). Including rural areas, the total population of the study region is estimated to be around 8 million, accounting for around 25 % of the population of Texas (Bureau, 2019).

The region is prone to damaging flood events, resulting in USD 16.8 billion in insured loss between 1976 and 2017 across 184 826 insurance claims. Predominately clay soils and low topographic relief coupled with extreme precipitation result in wide and shallow floodplains. Regional flood events are driven by several dominant mechanisms, including mesoscale convective systems (MCSs) and tropical cyclones (Van Oldenborgh et al., 2017). Recent examples of MCS-driven events include the Memorial Day Flood (2015) and the Tax Day and Louisiana–Texas floods (2016). Several stalled tropical cyclones, including tropical storms Claudette (1979) and Allison (2001) and Hurricane Harvey (2017), have also resulted in record-setting precipitation. Frequency estimates of tropical cyclone landfall range from once every 9 years in the eastern part of the study region to once every 19 years in the southwest part of the region (NHC, 2015). Historical storm surge ranges as high as 6 m along the coast. Notable surge events include the Galveston hurricanes (1900, 1915), Hurricane Carla (1961), and Hurricane Ike (2008). Several previous studies have demonstrated that flood hazards are not well represented by the FEMA SFHA (Highfield et al., 2013; Brody et al., 2013; Blessing et al., 2017), suggesting a large proportion of the population along the Texas coast is not only vulnerable to flooding, but the decision makers lack the information to properly account for it.

3.2 Data collection

3.2.1 NFIP claims

NFIP flood claims were used as the predictive variable. NFIP claims between 1976 and 2017 were available for the study, totaling 184 826 claims, each of which was geocoded to the parcel centroid. The NFIP flood losses dataset provides the location, total payout, and structural characteristics. Claims provide a reliable indication of the presence of flooding but fail to identify locations that have not flooded, mak-
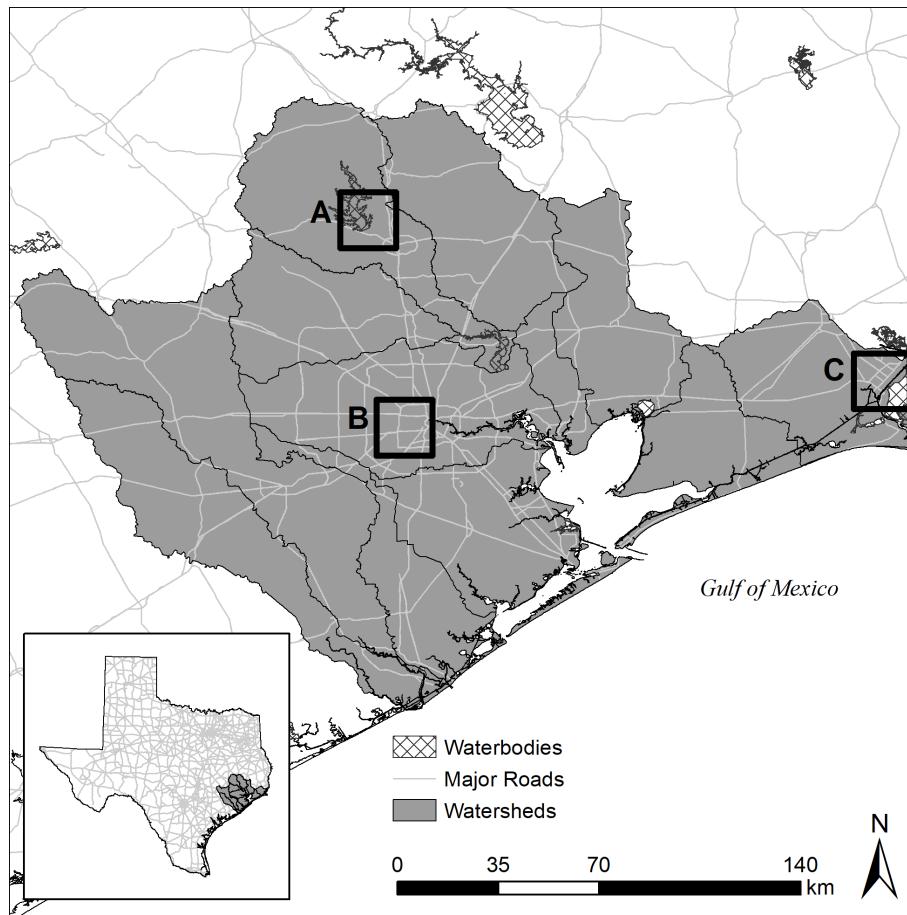
Figure 1. Map of the study area. Data originated from the US Census Bureau and the US Geological Survey.

ing the dataset presence only. Random forest algorithms require a binary dependent variable identifying presence and absence locations. A pseudo-random sample of background values can be used as a proxy for locations where flooding is absent (Barbet-Massin et al., 2012). When used in this way, the pseudo-absences essentially represent the population that could potentially flood. Therefore, the background sample locations were based on a random selection of all structures within the study area. The Microsoft building footprints dataset was used due to the open-source availability and high accuracy (https://github.com/Microsoft/USBuildingFootprints, last access: 4 March 2020). The study matched structures and claims using a one-to-one sample by watershed and year, meaning that for each claim, a structure is selected given the same year and is located within the same watershed. This one-to-one matching reduces potential bias from an unbalanced dataset (Chen et al., 2004). The final dataset removed any sample where an independent variable had null values, but the final ratio remained close to 50 % of the claims, with a sample size of 367 480.

3.2.2 Contextual variables

To parameterize flood hazard, several contextual variables were considered which represent the potential predictors of flooding across the study area (see Table 1). These variables can be divided into two main categories: (1) topographic (elevation and distance features which drive watershed response) and (2) hydrologic (overland and soil characteristics which govern infiltration and runoff). The variables were collected at different scales based on data availability. All variables were resampled to a 10 m raster and snapped to the Height Above Nearest Drainage (HAND) dataset. Below, we outline the variables, the reasoning behind their inclusion, and previous data-driven flood hazard studies that used them.

3.2.3 Hydrologic

Hydrologic variables explain how stormwater moves across the landscape and, therefore, can help differentiate between low- and high-flood-potential areas. The amount of stormwater that a given area can receive is a function of its flow accumulation potential, which is primarily mediated by three factors: the soil's ability to absorb water (i.e., saturated

Table 1. Concept measurements.

Name	Description	Initial resolution	Range
Accumulated flow	Hydrologic accumulation for contributing cells	30 m	$0–2.7 \times 10^6$ cells
Hydraulic conductivity (K_{sat})[1]	Average soil water transmission for all contributing cells	30 m	0–140 μm/s
Manning's roughness coefficient[2]	Average land cover roughness for all contributing cells	30 m	0.001–0.39
Elevation[3]	Digital elevation model (DEM) from USGS	30 m	0–155 m
Distance to coast[4]	Euclidean distance from the coast	10 m	0–155 km
Distance to stream[4]	Euclidean distance from the nearest ordered stream	10 m	0–15 km
Height Above Nearest Drainage (HAND)[5]	Relative vertical height compared with nearest stream, lake, or coastline	10 m	0–59 m
Imperviousness[6, 7]	Percent impervious (NLCD)	30 m	0 %–100 %
Topographic wetness index (TWI)	Topographic Wetness Index adjusted to have no 0 values	30 m	8.35–39.53

[1] USGS SSURGO database website: https://www.nrcs.usda.gov/ (last access: 4 March 2020).
[2] Years: 2001, 2004, 2006, 2008, 2011, 2013, 2016.
[3] LANDFIRE database website: https://www.landfire.gov (last access: 4 March 2020).
4 NHDPlus: https://www.usgs.gov/core-science-systems/ngp/national-hydrography/nhdplus-high-resolution (last access: 4 March 2020).
[5] Height Above Nearest Drainage: https://web.corral.tacc.utexas.edu/nfiedata/ (last access: 4 March 2020).
[6] NLCD website: https://www.mrlc.gov/data (last access: 4 March 2020).
[7] Years: 2001, 2006, 2011, 2016.

hydraulic conductivity), roughness (i.e., Manning's n), and imperviousness. Flow accumulation measures the total upstream area that flows into every raster cell based on a flow direction network as determined by the National Elevation Dataset (NED) (Jenson and Domingue, 1988). Areas with high flow accumulation are more susceptible to receiving larger amounts of stormwater during a given rainfall event. Soil infiltration influences the speed and amount at which stormwater can be absorbed into the ground. When stormwater cannot move into the ground easily, it may result in additional runoff, particularly in urbanized and downstream areas. Two measures of soil infiltration include lithology and saturated hydraulic conductivity (K_{sat}), both of which have been shown to be strong predictors of flood susceptibility (Brody et al., 2015; Janizadeh et al., 2019; Hosseini et al., 2019; Mobley et al., 2019). For this study, K_{sat} values were assigned to soil classes obtained from the Soil Survey Geographic Database (SSURGO) of the Natural Resources Conservation Service (NRCS) using the values presented in Rawls et al. (1983) and then averaged across the contributing upstream area for each cell.

Imperviousness is another strong indicator of an area's ability to infiltrate water. Previous studies have found that increasing impervious surfaces as a result of urbanization reduces infiltration and causes increased surface runoff and larger peak discharges, making it an important aspect in determining flood frequency and severity (Anderson, 1970;

Hall, 1984; Arnold Jr. and Gibbons, 1996; White and Greer, 2006). Imperviousness has been previously shown to be highly important in the Houston region, where urban sprawl has greatly increased imperviousness in the region and contributed to higher volumes of overland runoff (Brody et al., 2015; Gori et al., 2019; Sebastian et al., 2019). For this study, the percentage of impervious surface was measured using the Percent Impervious Surface raster from the National Land Cover Database (NLCD) for the years 2001, 2006, 2011, and 2016. Values range from 0 %–100 % and represent the proportion of urban impervious surface within each 30 m cell. Because comparable remote-sensing imagery only exists for these years, impervious surface was associated with the nearest date of each claim.

Roughness influences the speed at which stormwater can move across the landscape as well as the magnitude of peak flows in channels (Acrement and Schneider, 1984). Previous engineering studies have corroborated the relationship between roughness and overland water flow using Manning's roughness coefficient (Anderson et al., 2006; Thomas and Nisbet, 2007). This coefficient, called Manning's n, has become a critical input in many hydrological models and has also been shown to be a good predictor of event-based flood susceptibility (Mobley et al., 2019). For this study, roughness values were assigned to each NLCD land cover class using the values suggested by Kalyanapu et al. (2010) and, like K_{sat}, were averaged across the contributing upstream area for

each raster cell for the years 2001, 2004, 2006, 2008, 2011, 2013, and 2016.

3.2.4 Topographic

Elevation and slope are topographic variables frequently used to model flood hazard (Lee et al., 2017; Tehrany et al., 2019a; Rahmati and Pourghasemi, 2017; Bui et al., 2019; Hosseini et al., 2019; Mobley et al., 2019; Darabi et al., 2019). Low-lying areas tend to serve as natural drainage pathways, making them more susceptible to flooding and ponding during rainfall events. Elevation and slope were calculated using the National Elevation Dataset (NED), which was provided as a seamless raster product via the LAND-FIRE website at a 30 m resolution (LandFire, 2010).

Three continuous proximity rasters were used in this study: distance to stream, distance to coast, and Height Above Nearest Drainage (HAND). Proximity to streams and the coastline has been shown to be a significant indicator of flood damage (Brody et al., 2015) as these areas are typically much more prone to overbanking and surge, respectively. More recent flood hazard studies have used proximity to streams (Lee et al., 2017; Dodangeh et al., 2020; Janizadeh et al., 2019; Tehrany et al., 2019a, b; Rahmati and Pourghasemi, 2017; Bui et al., 2019; Hosseini et al., 2019; Mobley et al., 2019; Darabi et al., 2019), whereas proximity to coasts has been less common (Mobley et al., 2019). Distance to both stream and coast was calculated based on the National Hydrography Dataset (NHD) stream and coastline features.

HAND is calculated by defining the height of a location above the nearest stream to which the drainage from that land surface flows (Garousi-Nejad et al., 2019). Areas with high measures of HAND are more buffered from flooding because it requires increasingly more stormwater of short durations to create the peak flows that would reach those locations. This measure has been used to calculate flood depths, the probability of insured losses from floods (Rodda, 2005), soil water potential (Nobre et al., 2011), groundwater potential (Rahmati and Pourghasemi, 2017), and flood potential (Nobre et al., 2011). HAND was downloaded from the University of Texas' National Flood Interoperability Experiment (NFIE) continental flood inundation mapping system (Liu et al., 2016) at a 10 m resolution.

The Topographic Wetness Index (TWI) (Beven and Kirkby, 1979) is a popular measure of the spatial distribution of wetness conditions and is frequently used to identify wetlands. The TWI is used to quantify the effects of topography on hydrologic processes and is highly correlated with groundwater depth, and soil moisture (Sörensen et al., 2006). This measure has been found to be an influential and, in some cases, a significant predictor for estimating flood hazard (Lee et al., 2017; Tehrany et al., 2019a; Bui et al., 2019; Hosseini et al., 2019; Tehrany et al., 2019b). The TWI is calculated by the following equation:

$$\text{TWI} = \ln \frac{(A \cdot 900) + 1}{\tan\left(\frac{(S_0 \cdot \pi)}{180}\right)}, \tag{1}$$

where A is the contributing area (or flow accumulation), and S_0 is the average slope over the contributing area. High values of TWI are associated with areas that are concave, low-gradient areas where water often accumulates and pools, making them more vulnerable to flooding.

3.3 Random forest model

Random forest models categorize samples based on the highest predicted probability for each class. We developed a random forest algorithm at the parcel level (Fig. 2). The NFIP flood claims dataset was split between a training and test dataset. The test dataset was based on 30 % of the initial dataset. Within the training dataset, cross-validation is used to decide the final variables and parameters. A k-fold sampling method uses 90 % of the training dataset to predict the other 10 %. This sampling method helps to measure the robustness of the model as variables are pruned, and parameters are tweaked. The cross-validation assessment uses a 10-sample stratified k-fold. Structures were randomly sampled, reducing the potential for storm-event-based bias. In addition to the k-fold cross-validation, a year-by-year assessment was performed by removing 1 year as a validation sample for a model based on all other years. For example, 1976 through 2016 were used to predict flood hazard in 2017, then 1977 through 2017 were used to predict 1976, and so forth. This year-by-year assessment helps to identify the limitations of this flood hazard approach. All random forest computations were performed with the scikit-learn package (Pedregosa et al., 2011) in Python version 3.8.

Creating a properly calibrated random forest requires tuning two parameters to minimize error, and variable selection to improve generalizability. The two parameters tuned were the number of trees and the maximum tree depth. The number of trees controls the size of the forest used for the predictive model. Increasing the number of trees reduces error rates and increases the attributes used in the decision (Liaw and Wiener, 2002), but it comes at the cost of increasing computation time. Tree depth controls the maximum number of decisions that can be made, but too large of a tree will increase the chance of the model overfitting the data and reduce generalizability. The model used 200 trees and a maximum tree depth of 90 after optimizing error rates using the k-fold analysis. Variable reduction reduces the complexity of the model and decreases the likelihood of the model overfitting while speeding up the final training and raster predictions. An out-of-bag error score (OOB) isolates a subset of the training dataset which is used to measure error rates of each variable (Breiman, 1996) and generates feature importance. Initially, all variables were added to the model; those variables with
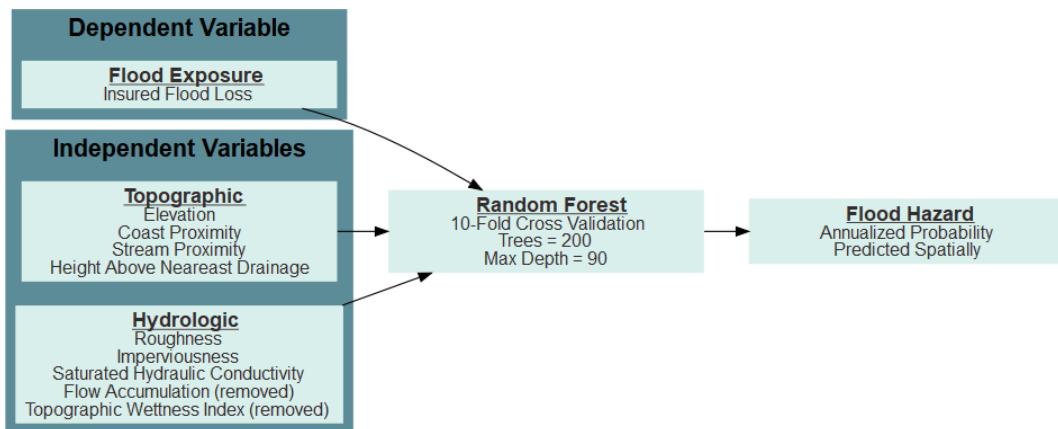
Figure 2. Conceptual framework to develop the annualized flood probability.
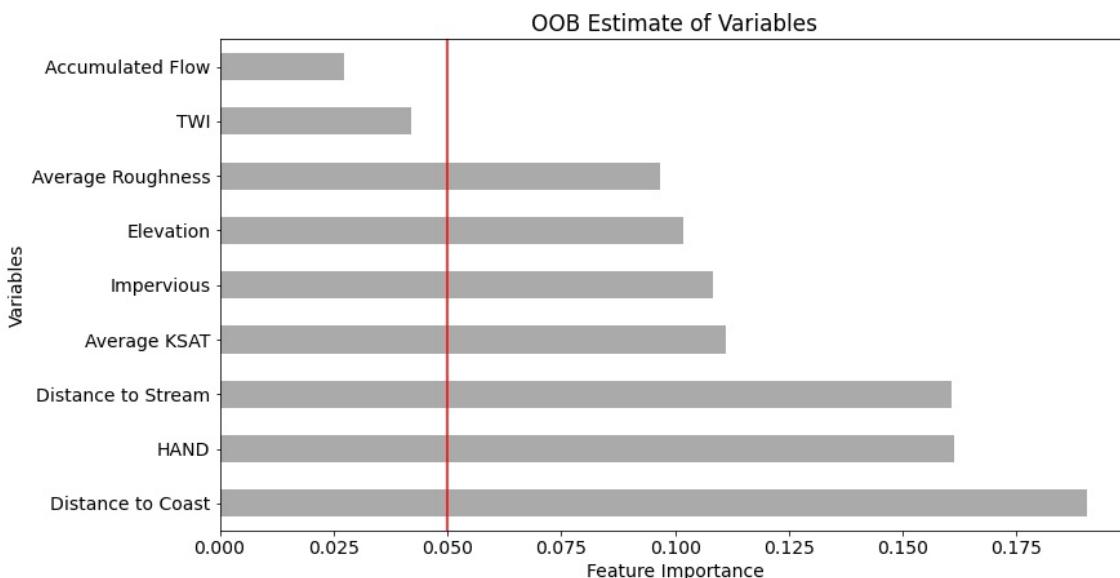
Figure 3. Feature importance chart. Note that variables below 0.05 were removed from the final model. The red line denotes the cutoff value for the final model.

the lowest contribution to feature importance were removed from the final model. Two variables were removed from the final model for not meeting the required threshold TWI and flow accumulation (Fig. 3).

A series of metrics were used to identify whether the model was properly calibrated. Average accuracy and sensitivity are measured for each iteration of the k-fold analysis. Accuracy measures the percentage of correctly identified flooded and non-flooded samples. Sensitivity estimates the probability that the flooded sample will be predicted, with a higher likelihood of flooding than a non-flooded sample (Metz, 1978) and based on the area under the curve (AUC) of the receiver operating characteristic (ROC). In practice AUCs are robust for continuous probabilities and are not limited by the number of features in a model or a threshold to use for classification. Both accuracy and sensitivity were com-

pared against the final model and the test sample. Ideally, the final model should be similar in accuracy and sensitivity to the average found in the k-fold analysis. A calibration plot shows how well the model predicts probabilities given the proportion of flooded points in each bin. A properly configured model will fall along a diagonal on the plot. Random forest outputs can either be a classification or represented as a probability. In this study, the final output was based on probability of flooding.

As an additional validity check we compared the random forest prediction against the FEMA SFHAs by examining the amount of area and number of structures exposed to specific annual exceedance probabilities. The random forest model predicts the probability that a location floods over the 42 years of the study. To convert the random forest probabilities to annual exceedance probabilities, we used the follow-
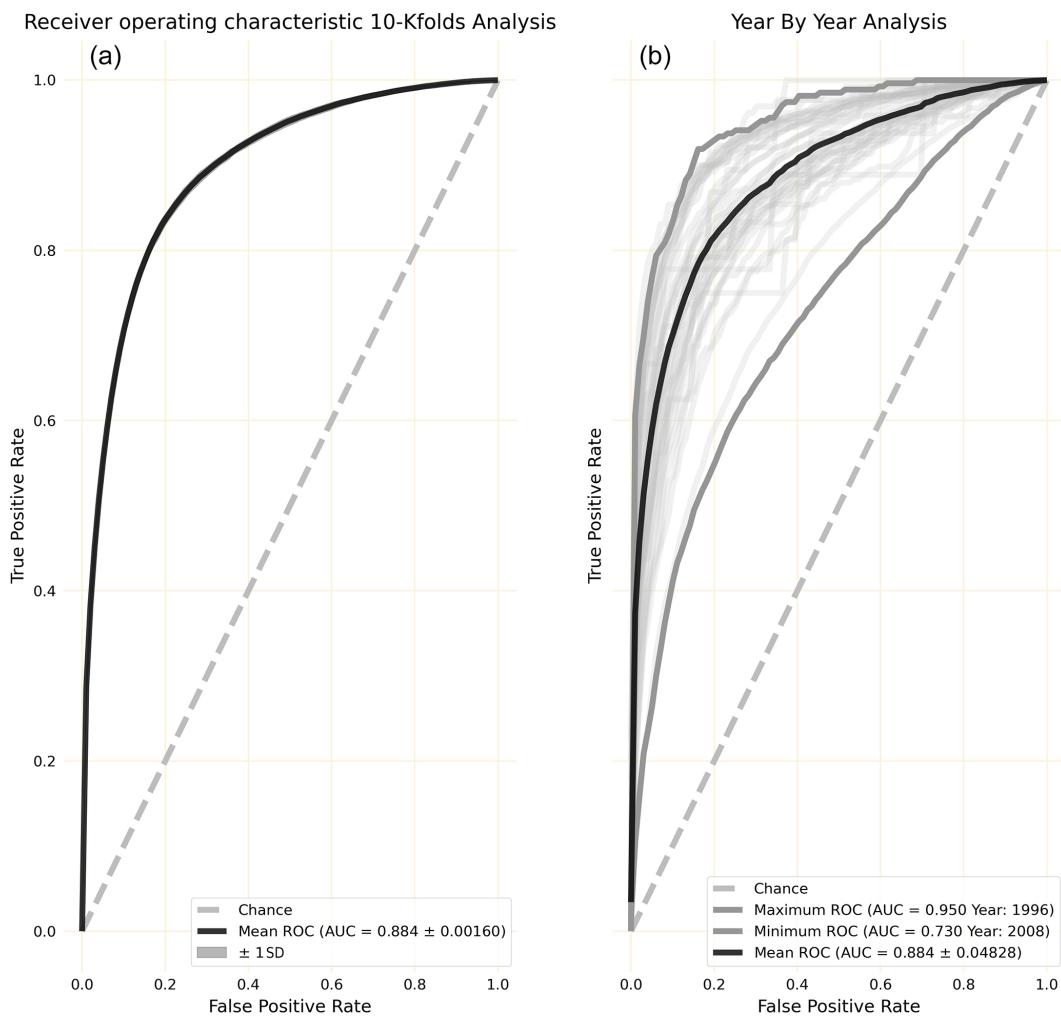
Figure 4. (a) The receiver operating characteristic (ROC) shown for 10-iteration k-fold cross-validation. **(b)** The ROC shown for the year-by-year analysis. The year-by-year analysis shows generalizability of the model by predicting how well stand-alone years with flooding will fit within the model. This approach also identifies years with storm outliers.

ing equation:

$$\text{exceedence} = 1 - (1 - p_{\text{flood}})^{(1/T)}. \tag{2}$$

This procedure allows for a more direct comparison between the FEMA SFHA and flood hazard.

4 Results

4.1 Model performance

Ten-iteration k-fold cross-validation was performed to predict flood hazard across 12 watersheds in the study area. From the k-fold analysis, the mean model accuracy was $81.9\% \pm 0.00198$, while the final test accuracy is 82.2%. The model had an average sensitivity of 0.893 ± 0.00184 (Fig. 4, left), while the final model produced a sensitivity of 0.895. In the year-by-year analysis, years predicted with a high sen-

sitivity fall within relatively normal events, while the extreme events such as Hurricane Harvey (1.5+ m of rain) and Hurricane Ike (high storm surge) perform poorer in relation. The year-by-year analysis showed more variation in sensitivity (Fig. 4, right); mean AUC is 0.884 ± 0.0482. The year with Hurricane Ike had the lowest sensitivity, at 0.730, and the year with Hurricane Harvey performed poorly as well, at 0.761. The highest sensitivity was predicted in 1997, at 0.929. This analysis sheds light on how the model performs for extreme events. For example, years such as 2017 perform worse as the conditions brought on by Hurricane Harvey are rarely seen.

The calibration plot (Fig. 5) suggests that the model slightly underpredicts flooded points at lower probabilities and slightly overpredicts flooded points at higher probabilities. The underprediction is explained by the histogram below the plot. Most non-flooded structures have a probability
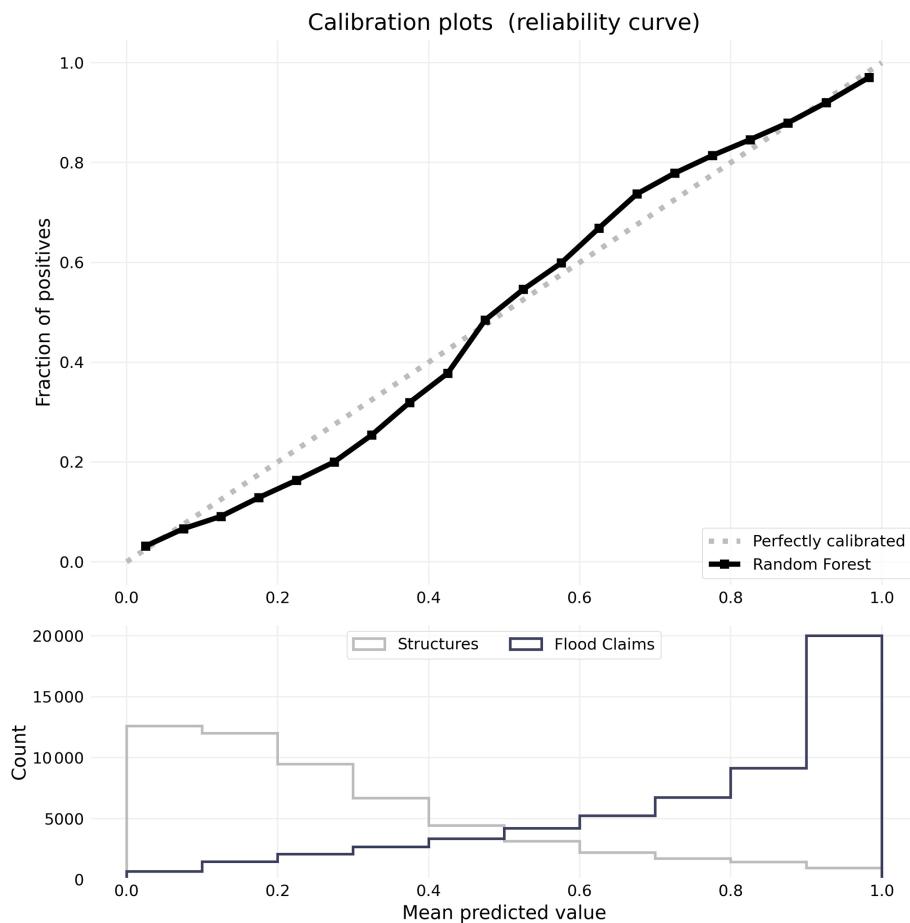
Figure 5. Calibration plot (top) with histogram of flooded and non-flooded samples and their predicted values.

Table 2. Represents the number of structures with at least a 1.0 % or 0.2 % probability of flooding within the modeled flood hazard and the SFHA.

Layer	Probability	Study area	Lake Conroe	Houston (Meyerland)	Port Arthur
Flood hazard	1.0 %	649 000	6040	87 300	26 000
	0.2 %	1 810 000	23 100	198 000	44 300
SFHA	1.0 %	207 000	1210	27 500	3080
	0.2 %	500 000	1540	58 000	32 900

below 30 %, while the largest proportion of flood loss points occur above 90 %.

4.2 Comparison with the FEMA SFHA

Flood hazard probabilities were converted to annual exceedance (Fig. 6), which allowed us to compare the amount of area and number of structures exposed to specific annual exceedance probabilities with the FEMA SFHAs (Table 2). Note as a reminder that the 100-year floodplain has an annual exceedance of 1 %, and the 500-year floodplain has an annual exceedance of 0.2 %. Based on the modeled flood hazard, 13 810 km^2 of land and 649 140 structures have

a 1 % chance of flood any given year, and 26 348 km^2 and 1.81 million structures have a 0.2 % chance of flooding any given year. In contrast, 8000 km^2 are classified within the 1 % SFHA, encompassing 207 000 structures, while the 0.2 % floodplain increases in size to 9 900 km^2 and encompasses 500 000 structures.

Focusing on three flood-prone areas, the model shows high flood hazards currently not identified by the FEMA SFHA (Fig. 7). In Conroe (Fig. 7a–c), for example, the 1 % areas appear visually similar for the SFHA and the flood hazard model, but when counted, the flood hazard model predicts 7.5 times more structures around Lake Conroe that have at
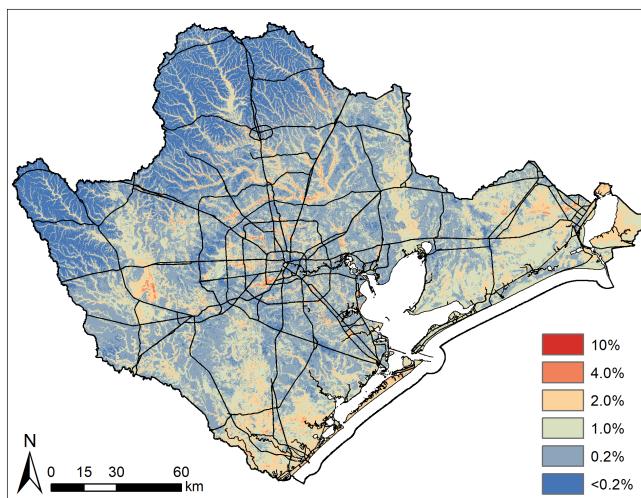
Figure 6. Continuous flood hazard map for the pilot study area based on model output. Road layers originated from the US Census Bureau.

least a 1 % chance annually of flooding compared with the 1 % SFHA. In the Meyerland area (Fig. 7d–f) both the FEMA SFHA and the flood hazard model clearly predict flooding along the rivers and identify areas where the floodplain expands. However, the flood hazard model identifies a much larger area with at least a 1 % chance of annualized flooding. Finally, within Port Arthur (Fig. 7g–i) the flood hazard model expects that the whole area has a flood hazard of at least 0.2 %, a significantly larger area than FEMA's 0.2 % floodplain.

5   Discussion

The results illustrate that flood hazards can be accurately estimated using a machine learning algorithm. The model is computationally efficient and scalable and can be used to predict flood hazards over relatively large regions. Training and image prediction were run in 1 h on a high-end desktop computer; in comparison physical models can run for 20 min or up to 10 h for an area 0.15 % of the size of the study area (Apel et al., 2009). Overall, the model demonstrates strong predictive power for estimating historic damages and accurately represents the spatial distribution of the flood hazard across those 40 years of flood damages. When compared against similar studies using machine learning approaches to predict flood hazards (Dodangeh et al., 2020; Janizadeh et al., 2019), our model demonstrates lower sensitivity. The differences are likely attributed to the high topographic relief of these regions where fluvial flood hazards predominate (Tehrany et al., 2019b), likely contributing to the predictive capacity of the models. In contrast, southeast Texas is characterized by little topographic relief where flooding may emanate from pluvial, fluvial, and marine sources, mak-

ing flood prediction more complex. In fact, when comparing model sensitivities across the study region, we find that the model performance increases in more steeply sloped inland areas than flat coastal areas.

Another aspect impacting model performance was sample selection. A systematic approach that identifies areas that did not flood (Darabi et al., 2019) can be important to increasing model performance (Barbet-Massin et al., 2012). While the model is based on a comprehensive record of observed flood claims in the study area from 1976 to 2017, there is no comparable record for structures that have not flooded. One option would be to randomly sample areas that have no claims; however this would not control for bias in the absence data and would come at the expense of model performance (Wisz and Guisan, 2009). To overcome this potential bias, we generated pseudo-absences by randomly selecting a sample of non-flooded structures by watershed and year to minimize this selection bias (Phillips et al., 2009). Based on the calibration plot (shown in Fig. 5), this is an appropriate assumption.

Another source of bias comes from how contextual variables were associated with the claims dataset. That is, the contextual variables that were attached to the claims came from the closest year for which the data were made available. However, this can cause some skewed behavior, particularly for the older claims, because some of the contextual variables only start in 2001, whereas the claims date back to 1976. Variables impacted the most by this are the result of large-scale changes in urbanization over time, which include imperviousness and roughness. More specifically, the imperviousness and roughness conditions attached to claims are going to be less representative of the actual conditions during the time of loss the further one goes back in time. However, a large majority of the claims (approximately 75 %) occurred after 2001, and most of the contextual variables do not change or change minimally over time (e.g., HAND, elevation, distance to coast, etc.), mitigating the influence of this bias.

Statistically generated flood hazard maps that can be compared to FEMA's regulatory floodplains are a novel outcome of this analysis. That is, we used the predicted model likelihoods to generate 1 % and 0.2 % annual exceedance probability thresholds or, equivalently, the 100- and 500-year flood hazard areas. The statistically generated flood hazard areas differed from the regulatory floodplains in that they are (1) nearly 3 times as large and (2) captured areas that are hydrologically disconnected from streams and waterbodies. The findings suggest that this approach can better capture small-scale variability in flood hazard by implicitly detecting underlying drivers that manifest themselves through subtle changes in historic damage patterns and trends. This corroborates previous research findings showing that the current 100-year floodplain underestimates and fails to accurately represent flood risk, particularly in urban areas (Blessing et al., 2017; Galloway et al., 2018; Highfield et al., 2013).
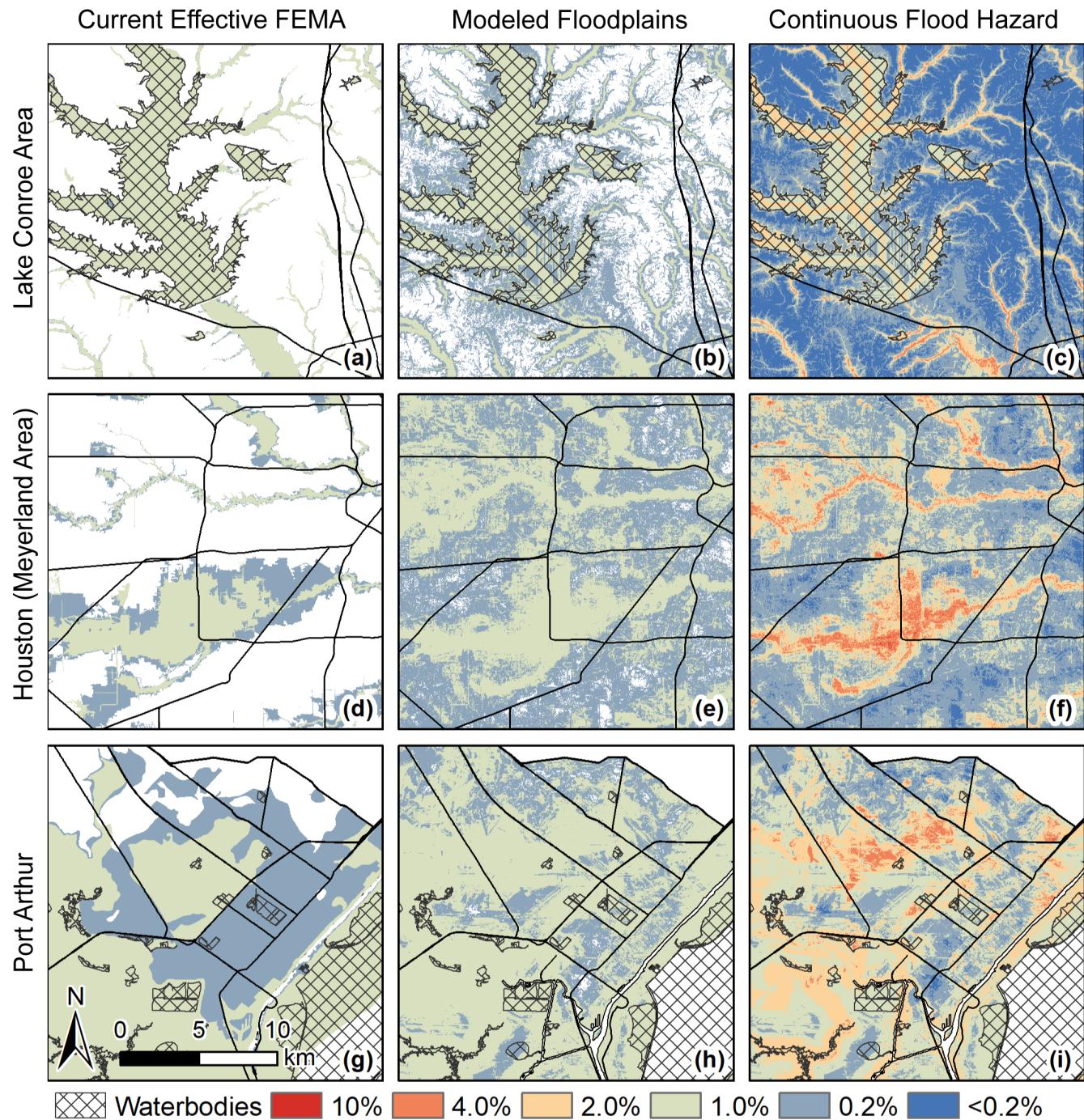
Figure 7. Side-by-side visual comparison for three areas affected by differing flood types. **(a–c)** Lake Conroe Area, **(d–f)** Houston (Meyerland Area), and **(g–i)** Port Arthur. Maps compare **(a, d, g)** current effective 100- and 500-year FEMA SFHA (ca. 2016), **(b, e, h)** areas within the 1 % and 0.2 % annual flood zones based on the random forest estimate, and **(c, f, i)** continuous flood hazard map. Road layers originated from the US Census Bureau, and floodplains originated from the Federal Emergency Management Agency.

It should be noted that the interpretation of the predicted random forest flood hazard areas differs from existing regulatory floodplains in that they are detecting the return period for structurally damaging inundation. The differences between the statistically generated flood hazard areas and regulatory floodplains are likely a result of multiple advantages of a data-driven approach in identifying the conditions in the underlying drivers (Elith et al., 2011), in other words, a data-driven model to better capture the reality of flood hazard by using actual historic impacts and simultaneously identifying

small-scale variations in flood exposure. By using historic losses, the random forest model accounts for extreme events that have occurred, but projecting future hazard is currently limited. The model does not currently incorporate precipitation patterns. Future work should examine the sensitivity of the model to precipitation as an input. With precipitation added to the model, there is still potential for error by underestimating the probability of future extreme events.

6 Conclusions

In this paper we demonstrate the efficacy of a random forest statistical model in spatially identifying flood hazards in southeast Texas, encompassing the Houston metropolitan area. In comparison with FEMA SFHA, we show that a statistical machine learning flood hazard model can (1) better capture the reality of flood hazard by using actual historic impacts; (2) better capture small-scale variability in flood hazard by implicitly detecting underlying drivers that manifest themselves through subtle changes in historic damage patterns and trends; (3) avoid the uncertainty associated with estimating rainfall return periods, stormwater infrastructure characteristics, and flood depths; (4) easily include alternative drivers of flood hazard such as HAND; and (5) be quickly updated using recent insurance claim payouts.

Appendix A
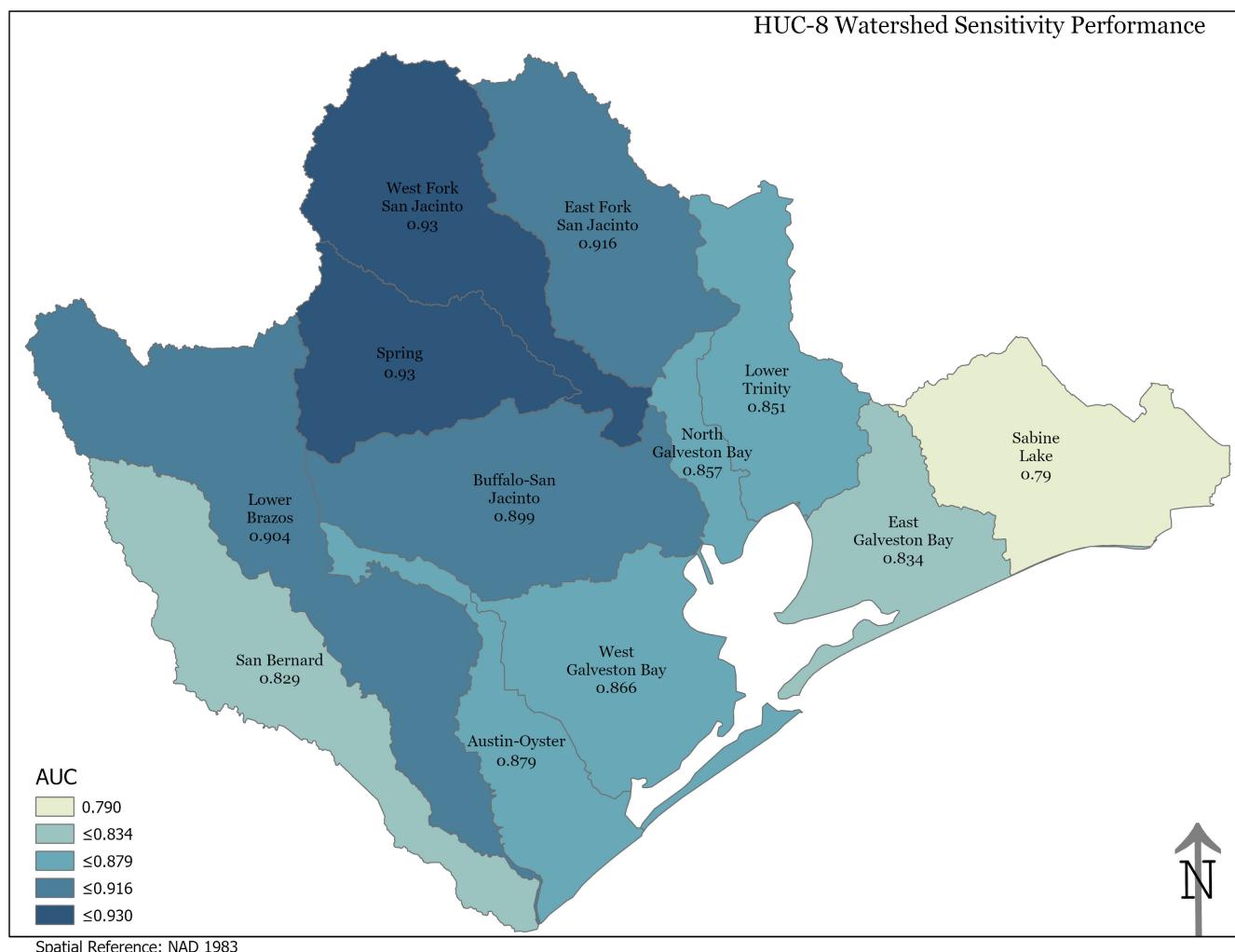
Figure A1. Sensitivity performance for each watershed. Note that the higher elevations occur in the northwest portion of the study area.

References

Albers, S. J. and Déry, S. J.: Flooding in the Nechako River Basin of Canada: A random forest modeling approach to flood analysis in a regulated reservoir system, Can. Water Resour. J., 41, 250–260, https://doi.org/10.1080/07011784.2015.1109480, 2015.

Anderson, B., Rutherfurd, I., and Western, A.: An analysis of the influence of riparian vegetation on the propagation of flood waves, Environ. Modell. Softw., 21, 1290–1296, 2006.

Anderson, D. G.: Effects of urban development on floods in northern Virginia, US Government Printing Office, Washington DC, 27 pp., 1970.

Apel, H., Aronica, G., Kreibich, H., and Thieken, A.: Flood risk analyses – how detailed do we need to be?, Nat. Hazards, 49, 79–98, 2009.

Arnold Jr., C. L. and Gibbons, C. J.: Impervious surface coverage: the emergence of a key environmental indicator, J. Am. Plann. Assoc., 62, 243–258, 1996.

ASFPM: Flood Mapping for the Nation: A Cost Analysis for Completing and Maintaining the Nation's NFIP Flood Map Inventory, Tech. Rep., Association of State Floodplain Managers, Madison, Wisconsin, USA, 2020.

Barbet-Massin, M., Jiguet, F., Albert, C. H., and Thuiller, W.: Selecting pseudo-absences for species distribution models: how, where and how many?, Methods Ecol. Evol., 3, 327–338, 2012.

Bates, P. D., Quinn, N., Sampson, C., et al.: Combined modelling of US fluvial, pluvial and coastal flood hazard under current and future climates, Water Resour. Res., e2020WR028673, 2020.

Belgiu, M. and Drăguţ, L.: Random forest in remote sensing: A review of applications and future directions, 114, 24–31, https://doi.org/10.1016/j.isprsjprs.2016.01.011, 2016.

Beven, K. J. and Kirkby, M. J.: A physically based, variable contributing area model of basin hydrology/Un modèle à base physique de zone d'appel variable de l'hydrologie du bassin versant, Hydrolog. Sci. J., 24, 43–69, 1979.

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

Breiman, L.: Out-of-bag estimation, Techinical report, Department of Statistics, University of California, 1996.

Breiman, L.: Random forests, Mach. Learn., 45, 5–32, 2001.

Brody, S., Blessing, R., Sebastian, A., and Bedient, P.: Examining the impact of land use/land cover characteristics on flood losses, J. Environ. Plann. Man., 57, 1252–1265, https://doi.org/10.1080/09640568.2013.802228, 2013.

Brody, S., Sebastian, A., Blessing, R., and Bedient, P.: Case study results from southeast Houston, Texas: identifying the impacts of residential location on flood risk and loss, J. Flood Risk Manag., 11, 110–120, 2018.

Brody, S. D., Zahran, S., Highfield, W. E., Bernhardt, S. P., and Vedlitz, A.: Policy Learning for Flood Mitigation: A Longitudinal Assessment of the Community Rating System in Florida, Risk Anal., 29, 912–929, https://doi.org/10.1111/j.1539-6924.2009.01210.x, 2009.

Brody, S. D., Highfield, W. E., and Blessing, R.: An Analysis of the Effects of Land Use and Land Cover on Flood Losses along the Gulf of Mexico Coast from 1999 to 2009, J. Am. Water Resour. As., 51, 1556–1567, 2015.

Bui, D. T., Ngo, P. T. T., Pham, T. D., Jaafari, A., Minh, N. Q., Hoa, P. V., and Samui, P.: A novel hybrid approach based on a swarm intelligence optimized extreme learning machine for flash flood susceptibility mapping, Catena, 179, 184–196, https://doi.org/10.1016/j.catena.2019.04.009, 2019.

Bureau, U. C.: American FactFinder – Results, available at: https://data.census.gov/cedsci/table?q=populationtotal&g= 3900000US480840,483362&tid=DECENNIALDPSF22000. DP1&hidePreview=false (last access: 24 February 2021), 2019.

Chen, C., Liaw, A., and Breiman, L.: Using random forest to learn imbalanced data, University of California, Berkeley, USA, 2004.

Center for Texas Beaches and Shores (CTBS): https://dataverse.tdl. org/dataverse/M3FR, last access: 31 March 2020.

Darabi, H., Choubin, B., Rahmati, O., Torabi Haghighi, A., Pradhan, B., and Kløve, B.: Urban flood risk mapping using the GARP and QUEST models: A comparative study of machine learning techniques, J. Hydrol., 569, 142–154, https://doi.org/10.1016/j.jhydrol.2018.12.002, 2019.

Dodangeh, E., Choubin, B., Eigdir, A. N., Nabipour, N., Panahi, M., Shamshirband, S., and Mosavi, A.: Integrated machine learning methods with resampling algorithms for flood

susceptibility prediction, Sci. Total Environ., 705, 135983, https://doi.org/10.1016/j.scitotenv.2019.135983, 2020.

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., and Yates, C. J.: A statistical explanation of MaxEnt for ecologists, Divers. Distrib., 17, 43–57, 2011.

Galloway, G. E., Reilly, A., Ryoo, S., Riley, A., Haslam, M., Brody, S., Highfield, W., Gunn, J., Rainey, J., and Parker, S.: The Growing Threat of Urban Flooding: A National Challenge, College Park and Galveston: University of Maryland and Texas A&M University, 44 pp., 2018.

Garousi-Nejad, I., Tarboton, D. G., Aboutalebi, M., and Torres-Rua, A. F.: Terrain analysis enhancements to the height above nearest drainage flood inundation mapping method, Water Resour. Res., 55, 7983–8009, 2019.

Gillies, S., Stewart, A. J., Snow, A. D., et al.: Rasterio: geospatial raster I/O for Python programmers, available at: https://github.com/mapbox/rasterio (4 March 2020), 2013.

Gori, A., Blessing, R., Juan, A., Brody, S., and Bedient, P.: Characterizing urbanization impacts on floodplain through integrated land use, hydrologic, and hydraulic modeling, J. Hydrol., 568, 82–95, 2019.

Hall, M. J.: Urban hydrology, Elsevier Applied Science Publishing, London, UK, 299, 1984.

Harris, C. R., Millman, K. J., van der Walt, S. J., et al: Array programming with NumPy, Nature, 585, 357–362, https://doi.org/10.1038/s41586-020-2649-2, 2020.

Highfield, W. E., Norman, S. A., and Brody, S. D.: Examining the 100-Year Floodplain as a Metric of Risk, Loss, and Household Adjustment, Risk Anal., 33, 186–191, https://doi.org/10.1111/j.1539-6924.2012.01840.x, 2013.

Hosseini, F. S., Choubin, B., Mosavi, A., Nabipour, N., Shamshirband, S., Darabi, H., and Haghighi, A. T.: Flash-flood hazard assessment using ensembles and Bayesian-based machine learning models: Application of the simulated annealing feature selection method, Sci. Total Environ., 711, 135161, https://doi.org/10.1016/j.scitotenv.2019.135161, 2019.

Hutson, M.: Artificial intelligence faces reproducibility crisis, American Association for the Advancement of Science, 359, 725–726, https://doi.org/10.1126/science.359.6377.725, 2018.

Ikeuchi, H., Hirabayashi, Y., Yamazaki, D., Muis, S., Ward, P. J., Winsemius, H. C., Verlaan, M., and Kanae, S.: Compound simulation of fluvial floods and storm surges in a global coupled river-coast flood model: Model development and its application to 2007 Cyclone Sidr in Bangladesh, J. Adv. Model. Earth Sy., 9, 1847–1862, 2017.

Janizadeh, S., Avand, M., Jaafari, A., Phong, T. V., Bayat, M., Ahmadisharaf, E., Prakash, I., Pham, B. T., and Lee, S.: Prediction Success of Machine Learning Methods for Flash Flood Susceptibility Mapping in the Tafresh Watershed, Iran, Sustainability-Basel, 11, 5426, https://doi.org/10.3390/su11195426, 2019a.

Jenson, S. K. and Domingue, J. O.: Extracting topographic structure from digital elevation data for geographic information system analysis, Photogramm. Eng. Rem. S., 54, 1593–1600, 1988.

Kalyanapu, A. J., Burian, S. J., and McPherson, T. N.: Effect of land use-based surface roughness on hydrologic model output, Journal of Spatial Hydrology, 9, 51–71, 2010.

Knighton, J., Buchanan, B., Guzman, C., Elliott, R., White, E., and Rahm, B.: Predicting flood insurance claims with hydrologic and socioeconomic demographics via machine learn-

ing: exploring the roles of topography, minority populations, and political dissimilarity, J. Environ. Manage., 272, 111051, https://doi.org/10.1016/j.jenvman.2020.111051, 2020.

Kousky, C. and Shabman, L.: Policy nook: "Federal funding for flood risk reduction in the US: Pre- or post-disaster?", Water Economics and Policy, 3, 1771001, https://doi.org/10.1142/S2382624X17710011, 2017.

LandFire: Homepage of the LANDFIRE Program, available at: https://www.landfire.gov/ (last access: 4 March 2020), 2010.

Lee, S., Kim, J.-C., Jung, H.-S., Lee, M. J., and Lee, S.: Spatial prediction of flood susceptibility using random-forest and boosted-tree models in Seoul metropolitan city, Korea, Geomat. Nat. Haz. Risk, 8, 1185–1203, https://doi.org/10.1080/19475705.2017.1308971, 2017.

Li, B., Wang, Q., and Hu, J.: A fast SVM training method for very large datasets, in: Neural Networks, 2009. IJCNN 2009. International Joint Conference on, 1784–1789, 2009.

Liaw, A. and Wiener, M.: Classification and regression by random-Forest, R news, 2, 18–22, 2002.

Liu, Y. Y., Maidment, D. R., Tarboton, D. G., Zheng, X., Yildirim, A., Sazib, N. S., and Wang, S.: A CyberGIS approach to generating high-resolution height above nearest drainage (HAND) raster for national flood mapping, CyberGIS, 16, 24–26, 2016.

Metz, C. E.: BASIC PRINCIPLES OF ROC ANALYSIS, Seminars in Nuclear Medicine, 8, 283–298, https://doi.org/10.1016/s0001-2998(78)80014-2, 1978.

Mobley, W.: Flood Hazard Modeling Output, V2, Texas Data Repository, https://doi.org/10.18738/T8/FVJFSW, 2020.

Mobley, W., Sebastian, A., Highfield, W., and Brody, S. D.: Estimating flood extent during Hurricane Harvey using maximum entropy to build a hazard distribution model, J. Flood Risk Manag., 12, e12549, https://doi.org/10.1111/jfr3.12549, 2019.

Moradkhani, H. and Sorooshian, S.: General review of rainfall-runoff modeling: model calibration, data assimilation, and uncertainty analysis, in: Hydrological modelling and the water cycle, edited by: Sorooshian, S., Hsu, K. L., Coppola, E., Tomassetti, B., Verdecchia, M., and Visconti, G., Springer, Berlin, Heidelberg, 1–24, https://doi.org/10.1007/978-3-540-77843-1_1, 2009.

Morss, R. E., Wilhelmi, O. V., Downton, M. W., and Gruntfest, E.: Flood risk, uncertainty, and scientific information for decision making: lessons from an interdisciplinary project, B. Am. Meteorol. Soc., 86, 1593, https://doi.org/10.1175/BAMS-86-11-1593, 2005.

Mosavi, A., Ozturk, P., and Chau, K.-w.: Flood prediction using machine learning models: Literature review, Water, 10, 1536, https://doi.org/10.3390/w10111536, 2018.

Moy de Vitry, M., Kramer, S., Wegner, J. D., and Leitão, J. P.: Scalable flood level trend monitoring with surveillance cameras using a deep convolutional neural network, Hydrol. Earth Syst. Sci., 23, 4621–4634, https://doi.org/10.5194/hess-23-4621-2019, 2019.

Muñoz, P., Orellana-Alvear, J., Willems, P., and Célleri, R.: Flash-Flood Forecasting in an Andean Mountain Catchment – Development of a Step-Wise Methodology Based on the Random Forest Algorithm, Water, 10, 1519, https://doi.org/10.3390/w10111519, 2018.

NHC: https://www.nhc.noaa.gov/climo/ (last access: 4 March 2020), Tropical Cyclone Climatology, 2015.

NOAA National Centers for Environmental Information (NCEI): US Billion-Dollar Weather and Climate Disasters, available at:

https://www.ncdc.noaa.gov/billions/ (last access: 4 March 2020), 2021.

Nobre, A. D., Cuartas, L. A., Hodnett, M., Rennó, C. D., Rodrigues, G., Silveira, A., and Saleska, S.: Height Above the Nearest Drainage – a hydrologically relevant new terrain model, J. Hydrol., 404, 13–29, 2011.

Office, C. B.: The National Flood Insurance Program: Financial Soundness and Affordability, Report, Congress of the United States congressional budget office, Washington D.C., 44 pp., 2017.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, J. Mach. Learn. Res., 12, 2825–2830, available at: https://dl.acm.org/doi/10.5555/1953048.2078195 (last access: 4 March 2020), 2011.

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., and Ferrier, S.: Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data, Ecol. Appl., 19, 181–197, https://doi.org/10.1890/07-2153.1, 2009.

Quinlan, J. R.: Induction of decision trees, Mach. Learn., 1, 81–106, 1986.

Rahmati, O. and Pourghasemi, H. R.: Identification of Critical Flood Prone Areas in Data-Scarce and Ungauged Regions: A Comparison of Three Data Mining Models, Water Res. Manag., 31, 1473–1487, https://doi.org/10.1007/s11269-017-1589-6, 2017.

Rajib, A., Liu, Z., Merwade, V., Tavakoly, A. A., and Follum, M. L.: Towards a large-scale locally relevant flood inundation modeling framework using SWAT and LISFLOOD-FP, J. Hydrol., 581, 124406, https://doi.org/10.1016/j.jhydrol.2019.124406, 2020.

Rawls, W. J., Brakensiek, D. L., and Miller, N.: Green-Ampt Infiltration Parameters from Soils Data, J. Hydraul. Eng., 109, 62–70, https://doi.org/10.1061/(ASCE)0733-9429(1983)109:1(62), 1983.

Rodda, H. J.: The development and application of a flood risk model for the Czech Republic, Nat. Hazards, 36, 207–220, 2005.

Sebastian, A., Gori, A., Blessing, R. B., van der Wiel, K., and Bass, B.: Disentangling the impacts of human and environmental change on catchment response during Hurricane Harvey, Environ. Res. Lett., 14, 124023, https://doi.org/10.1088/1748-9326/ab5234, 2019.

Singh, V.: Effect of spatial and temporal variability in rainfall and watershed characteristics on stream flow hydrograph, Hydrol. Process., 11, 1649–1669, 1997.

Solomatine, D. P. and Ostfeld, A.: Data-driven modelling: Some past experiences and new approaches, J. Hydroinform., 10, 3–22, https://doi.org/10.2166/hydro.2008.015, 2008.

Sørensen, R., Zinko, U., and Seibert, J.: On the calculation of the topographic wetness index: evaluation of different methods based on field observations, Hydrol. Earth Syst. Sci., 10, 101–112, https://doi.org/10.5194/hess-10-101-2006, 2006.

Tehrany, M. S., Jones, S., and Shabani, F.: Identifying the essential flood conditioning factors for flood prone area mapping using machine learning techniques, Catena, 175, 174–192, https://doi.org/10.1016/j.catena.2018.12.011, 2019a.

Tehrany, M. S., Kumar, L., and Shabani, F.: A novel GIS-based ensemble technique for flood susceptibility mapping using evidential belief function and support vector machine: Brisbane, Australia, PeerJ, 7, e7653, https://doi.org/10.7717/peerj.7653, 2019b.

Thomas, H. and Nisbet, T.: An assessment of the impact of floodplain woodland on flood flows, Water Environ. J., 21, 114–126, 2007.

Upstill-Goddard, R., Eccles, D., Fliege, J., and Collins, A.: Machine learning approaches for the discovery of gene-gene interactions in disease data, Brief. Bioinform., 14, 251–60, https://doi.org/10.1093/bib/bbs024, 2013.

van der Walt, S., Colbert, S. C., and Varoquaux, G.: The NumPy Array: A Structure for Efficient Numerical Computation, Comput. Sci. Eng., 13, 22–30, 2011.

Van Oldenborgh, G. J., Van Der Wiel, K., Sebastian, A., Singh, R., Arrighi, J., Otto, F., Haustein, K., Li, S., Vecchi, G., and Cullen, H.: Attribution of extreme rainfall from Hurricane Harvey, August 2017, Environ. Res. Lett., 12, 124009, https://doi.org/10.1088/1748-9326/aaa343, 2017.

Wagenaar, D., de Jong, J., and Bouwer, L. M.: Multi-variable flood damage modelling with limited data using supervised learning approaches, Nat. Hazards Earth Syst. Sci., 17, 1683–1696, https://doi.org/10.5194/nhess-17-1683-2017, 2017.

Ward, P. J., Jongman, B., Salamon, P., Simpson, A., Bates, P., De Groeve, T., Muis, S., De Perez, E. C., Rudari, R., Trigg, M. A., and Winsemius, H. C.: Usefulness and limitations of global flood risk models, Nat. Clim. Change, 5, 712–715, 2015.

McKinney, W.: Data Structures for Statistical Computing in Python, in: Proceedings of the 9th Python in Science Conference, 28 June–3 July, Austin Texas, USA, 56–61, https://doi.org/10.25080/Majora-92bf1922-00a, 2010.

White, M. D. and Greer, K. A.: The effects of watershed urbanization on the stream hydrology and riparian vegetation of Los Penasquitos Creek, California, Landscape Urban Plan., 74, 125–138, 2006.

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

Wisz, M. S. and Guisan, A.: Do pseudo-absence selection strategies influence species distribution models and their predictions? An information-theoretic approach based on simulated data, BMC Ecology, 9, 8, https://doi.org/10.1186/1472-6785-9-8, 2009.

Woznicki, S. A., Baynes, J., Panlasigui, S., Mehaffey, M., and Neale, A.: Development of a spatially complete floodplain map of the conterminous United States using random forest, Sci. Total Environ., 647, 942–953, 2019.

Nat. Hazards Earth Syst. Sci., 21, 807–822, 2021

https://doi.org/10.5194/nhess-21-807-2021