Development of a spatially complete floodplain map of the conterminous United States using random forest

Sean A. Woznicki\textsuperscript{a,\,*}, Jeremy Baynes\textsuperscript{a}, Stephanie Panlasigui\textsuperscript{b,1}, Megan Mehaffey\textsuperscript{a}, Anne Neale\textsuperscript{a}

\textsuperscript{a}National Exposure Research Laboratory, US Environmental Protection Agency, Research Triangle Park, NC 27711, USA

\textsuperscript{b}Oak Ridge Institute for Science and Education Research Participant Program, US Environmental Protection Agency, Research Triangle Park, NC 27711, USA

Abstract

Floodplains perform several important ecosystem services, including storing water during precipitation events and reducing peak flows, thus reducing flooding of downstream communities. Understanding the relationship between flood inundation and floodplains is critical for ecosystem and community health and well-being, as well as targeting floodplain and riparian restoration. Many communities in the United States, particularly those in rural areas, lack inundation maps due to the high cost of flood modeling. Only 60\% of the conterminous United States has Flood Insurance Rate Maps (FIRMs) through the U.S. Federal Emergency Management Agency (FEMA). We developed a 30-meter resolution flood inundation map of the conterminous United States (CONUS) using random forest classification to fill the gaps in the FIRM. Input datasets included digital elevation model (DEM)-derived variables, flood-related soil characteristics, and land cover. The existing FIRM 100-year floodplains, called the Special Flood Hazard Area (SHFA), were used to train and test the random forests for fluvial and coastal flooding. Models were developed for each hydrologic unit code level four (HUC-4) watershed and each 30-meter pixel in the CONUS was classified as floodplain or non-floodplain. The most important variables were DEM-derivatives and flood-based soil characteristics. Models captured 79\% of the SFHA in the CONUS. The overall F1 score, which balances precision and recall, was 0.78. Performance varied geographically, exceeding the CONUS scores in temperate and coastal watersheds but were less robust in the arid southwest. The models also consistently identified headwater floodplains not present in the SFHA, lowering performance measures but providing critical information missing in many low-order stream systems. The performance of the random forest models demonstrates the method’s ability to successfully fill in the remaining unmapped floodplains in the CONUS, while using only publicly available data and open source software.

Graphical Abstract

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

\textsuperscript{*}Corresponding author at: US Environmental Protection Agency, 109 T.W. Alexander Drive, Research Triangle Park, NC 27711, USA. woznicki.sean@epa.gov (S.A. Woznicki).

\textsuperscript{1}Present address: Eno River Association, 4404 Guess Rd. Durham, NC 27712.
1. Introduction

Floods are the leading cause of natural disaster losses in the United States, with annual average flood damages in the 1990s costing $5.6 billion, jumping to almost $10 billion in the 2000s (ASFPM, 2013). The six costliest natural disasters in United States history were Hurricanes Katrina (2012), Harvey (2017), Maria (2017), Sandy (2012), Irma (2017), and Andrew (1992) which all caused extensive flooding. Hurricanes Harvey, Irma, and Maria caused widespread damage in less than one month with an estimated cost totaling about $265 billion (NCEI, 2018). Non-tropical, inland flooding can also be catastrophic, as the “Flood of 1993” damaged crops, infrastructure, homes and businesses over large parts of the Midwestern US along the Missouri River, causing over $36 billion (2017 dollars) in damages and resulting in 48 deaths (NCEI, 2018).

Although flooding is extremely costly, only 61% of the conterminous United States (CONUS) is mapped under the Federal Emergency Management Agency (FEMA) National Flood Insurance Program (NFIP). This equates to only a third of the total stream miles in the United States, leaving 2.3 million miles of streams without flood data based on the National Hydrography Dataset (ASFPM, 2013). FEMA creates Flood Insurance Rate Maps (FIRMs) that delineate the Special Flood Hazard Area (SFHA), which identifies areas that have a 1% annual chance of flooding, i.e. the 100-year floodplain, and thus require purchase of flood insurance through the NFIP. The gaps that exist in FIRM mapping are due to the costly nature of detailed modeling and mapping, at $5000–$10,000 per river mile (FEMA, 2005). The high cost usually excludes areas with limited development and low populations from being mapped. Unmapped areas are prevalent west of the Mississippi River, in the midwest and western United States, including large areas devoted to agricultural production. Although they are unmapped, these places still have considerable flood risk with respect to their local economies, and may be developing without flood maps as a guide (ASFPM, 2013). For example, agricultural losses from the “Flood of 1993” along the Missouri River totaled $6–8 billion (Rosenzweig et al., 2002), while much of the Missouri River watershed remains unmapped.

Given the investment of time and money required to develop flood maps and the large spatial gaps that exist, new methods are needed to rapidly map areas of potential flooding.
Machine learning methods are one potential approach to develop flood inundation maps over large areas. These methods have often been used in spatial hazard studies, requiring only publicly available geographic information systems (GIS) datasets. Landslide susceptibility mapping is the most common, using methods including random forest (RF) (Hong et al., 2016; Youssef et al., 2015), support vector machines (SVMs) (Pradhan, 2013), and boosted regression trees (Youssef et al., 2015). Other applications and methods include forest fire susceptibility mapping using kernel logistic regression (Tien Bui et al., 2016), wetland identification using RF (Berhane et al., 2017; Berhane et al., 2018; Maxwell et al., 2016), and mineral prospectivity modeling using artificial neural networks (ANNs), regression trees, RF, and SVMs (Rodriguez-Galiano et al., 2015).

Local and regional flood susceptibility and inundation has also been mapped with machine learning methods. Decision trees, frequency ratio, logistic regression, weight-of-evidence, and SVMs have been used in the past (Tehrany et al., 2013; Tehrany et al., 2014; Tehrany et al., 2015). SVMs have also been used to extract flooded area from Landsat satellite imagery (Ireland et al., 2015). Fernández and Lutz (2010) mapped urban flood hazard using the analytic hierarchy process, a multi-criteria decision analysis method. RF has been used to map flood susceptibility in China’s mountainous regions at a relatively coarse 0.1 decimal degree resolution (Zhao et al., 2018), while Lee et al. (2017) used similar methods to map flood susceptibility for Seoul, South Korea at 30-meter resolution based on observed flood inundation. Kourgialas and Karatzas (2017) used multi-criteria analysis and ANNs to produce a national flood hazard map for Greece. These studies typically use publicly available spatial datasets (e.g. land cover, soil characteristics, geology, topography, and stream networks) at varying resolutions for small watersheds to regional river basins. On a larger scale, Sangwan and Merwade (2015) used GIS and soil attributes (flood frequency, soil taxonomy, water bodies, and geomorphic description) from the United States Department of Agriculture (USDA) Soil Survey Geographic Database (SSURGO) to map floodplain extents for the CONUS. Most recently, Wing et al. (2017) developed a physically based, spatially complete flood hazard model for fluvial and pluvial floods in the CONUS, validated against SFHA maps, while Jafarzadegan and Merwade (2017) used a DEM based thresholding approach to classify floodplains in North Carolina.

We sought to build on these efforts by using RF to map the 100-year coastal and fluvial floodplains for the CONUS at 30-meter resolution using open-source tools and publicly available data. The method presented here builds upon previous soil-based (Sangwan and Merwade, 2015) and DEM-based (Jafarzadegan and Merwade, 2017) studies. By using a rapid computational method such as RF, we hypothesized that we can simultaneously map both coastal and fluvial floods, while developing models that are uniquely tailored to varying physiographic regions across the CONUS. Our objectives were to (1) test the applicability of RFs for large scale flood inundation mapping using only publicly available national-coverage spatial datasets and open-source tools and (2) develop a spatially complete and publicly available 100-year flood inundation map for the CONUS. The methods used in this study are designed to be easily adapted to include updated datasets, improvements in spatial resolution, and potential extension to data-scarce regions beyond the CONUS.
2. Materials and methods

2.1. Data

2.1.1. Study area—Models were developed at the level-4 hydrologic unit code (HUC-4) watershed, of which there are 202 in the CONUS (Fig. 1). Modeling the floodplain at this scale was a compromise between a large area model that captured more available FEMA data necessary for model training and testing versus smaller area models allowing for more specialization based on local physiography.

The FIRMs consist of both areas within the SFHA and areas with minimal flood hazard (b1% annual chance of flooding). Fig. 1 demonstrates the large areas of the CONUS that do not have mapped 100-year floodplains. Unmapped areas are particularly extensive in the west, e.g. HUCs 10, 11, 14, and 17. The percentage of mapped area varies widely among the HUC-4s. The mean percentage of area mapped across the HUC-4s is 57% (standard deviation: 31%). Eight HUC-4s are completely mapped, and two are unmapped. Fifteen HUC-4s have <10% FIRM coverage, while 43 HUC-4s have >90% coverage. The percentage of mapped area that is classified as the SFHA by FEMA is most important, because these data form the basis for training and testing the RF models of floodplain extent. The mean percentage of mapped area in the SFHA is 12% (standard deviation: 10%) for the HUC-4s. There are 105 watersheds in which the SFHA makes up <10% of the total mapped area.

2.1.2 Response variable—The response variable was the FEMA National Flood Hazard Layer (NFHL) (FEMA, 2017) 100-year floodplain (one-percent chance of occurring in any given year). There are several zone designations that make up the 100-year floodplain; these zones were reclassified into a binary classification where all pixels in the 100-year floodplain were given a value of one, and all other pixels were given a value of zero (Table 1). The FEMA NFHL is comprised of several smaller scale studies of individual reaches and watersheds that are merged to form a coverage of the CONUS. Detailed flood studies in the NFHL are performed differently depending on the type of flooding: riverine flooding, lacustrine flooding, coastal flooding due to hurricanes or storms, and shallow flooding, ponding, and sheet flow (FEMA, 2005). Detailed studies are typically available for developed communities and areas experiencing rapid growth, while approximate studies that use aerial imagery, soils mapping, high water profiles, and topographic maps are prevalent in areas with little development (FEMA, 2005). We included both riverine and coastal areas as training data in the modeling process, and data from both detailed and approximate studies.

2.1.3. Predictor variables—Several landscape datasets were used as predictor variables (Table 2). The variables were publicly available land cover, soils, and topography datasets, or derivatives of those datasets. Complete CONUS coverage and 30-meter resolution was required for each variable.

Land cover data from the 2011 National Land Cover Database (NLCD) (Homer et al., 2015) were used to incorporate water and wetland classes into the modeling process. NLCD classes were reclassified to six classes: water, developed, forest, planted/cultivated, wetlands, and an additional class that included barren, shrub/scrub, grassland/herbaceous, and pasture/
hay. All subsequent predictor variables and the response variable were converted to 30-meter raster to match the spatial extent and resolution of the NLCD.

Soil data were obtained from SSURGO and the USDA Digital General Soil Map of the United States (STATSGO2). The higher resolution SSURGO dataset was used where available (see https://websoilsurvey.sc.egov.usda.gov/DataAvailability/SoilDataAvailabilityMap.pdf), and missing information was filled using the coarser resolution STATSGO2 dataset. Spatial soil information related to flood frequency and fluvial soil taxonomy classes were obtained from these datasets. These attributes were identified according to Sangwan and Merwade (2015).

Flood frequency information in the SSURGO/STATSGO variable floodfreqdcd (flood frequency, dominant condition) is based on the dominant flood frequency class for a map unit. The flood frequency classes are none (<0.2% annual chance of flooding), very rare (0.2–1% annual chance), rare (1–5% annual chance), occasional (5–50% annual chance), frequent (>50% annual chance of flooding but <50% chance of flooding in all months of any year), and very frequent (>50% chance of flooding in all month of any year). These were classified into a categorical raster dataset with six classes corresponding to the flood frequencies above. This variable was named fldfreq.

The presence of fluvial soils is an indicator of past flooding (Sangwan and Merwade, 2015). Therefore, the taxonomic subgroup (taxsubgrp) of each map unit was used to identify fluvial soils, or those containing “fluv” in their taxonomy. The fluvial soils dataset was reclassified into a three-class raster with the following categories: non-fluvial soils, soils of fluvial origin, and water. This variable was named fluvclass.

Elevation data from the National Elevation Dataset (NED) digital elevation model (DEM) at 30-meter resolution were used to derive topographic datasets describing slope, topographic variability, topographic wetness, and horizontal and vertical distances to water bodies. The vertical accuracy of the DEM is 2.44 m (root mean square error), but this varies geographically (Gesch et al., 2014).

The standard deviation of a five-by-five pixel moving window was calculated on the DEM. This variable (DEM5x5) was used to evaluate the local variability of the elevation at each pixel as opposed to the absolute elevation, as the absolute elevation is not as meaningful across large regions such as HUC-4s.

Topographic wetness was included in the form of the compound topographic index (CTI) because it “parameterizes first-order controls on water movement from topographic information” (Hjerdt et al., 2004). In addition, a modified CTI was used in a previous study as the sole identifier of flood-prone areas (Manfreda et al., 2011), indicating its potential utility in identifying floodplains. CTI is calculated ln(α / tanβ), where α is flow accumulation multiplied by the contributing area and β is the slope in radians. A slope raster and flow accumulation raster derived from the DEM were used to calculate CTI.

Using System for Automated Geoscientific Analyses (SAGA GIS) (Conrad et al., 2015), we derived gridded datasets for four variables that measure each pixel’s distance to the nearest
channel. These variables were vertical overland flow distance (VOFD; vertical distance traveled along the flow path), horizontal overland flow distance (HOFD; horizontal distance traveled along the flow path), overland flow distance (OFD; distance to the nearest channel considering horizontal and vertical travel along the flow path), and vertical distance to channel (VDC; difference in elevation between pixel and nearest channel, not considering flow path).

To develop the flow distance variables, the DEM and a channel network raster were required. Sinks in the DEM were filled using the Planchnon and Darboux (2002) approach in SAGA GIS. A 30-meter flow accumulation raster was derived from the DEM using GIS to fill sinks, calculate flow direction, then determine flow accumulation. From the flow accumulation raster, we created a channel network raster by setting a threshold of 5000 pixels for a pixel to be considered a channel, per the catchment definition used by the USGS Elevation Derivatives for National Applications (EDNA; https://lta.cr.usgs.gov/edna). A universal threshold was used for alignment with other DEM variables and consistency across the CONUS, although the channel threshold varies with differences in climate and geomorphology. The modules used a D8 flow algorithm (eight discrete flow angles with each pixel having a single flow direction) during processing. Coastlines were considered channels for the purposes of calculating the distance variables. The DEM boundary (i.e., the boundary of all non-null DEM values) along coastlines was extended by one pixel and this one pixel extension was identified as the coastline channel.

2.2. Modeling process

2.2.1. Random forest—Random forest (Breiman, 2001) was used as a binary classifier to classify each 30-meter pixel’s membership in the 100-year floodplain using the SFHA as training data. RF is an ensemble learning method that uses multiple independently constructed decision trees each using a unique bootstrap sample of the data set (Liaw and Wiener, 2002).

The RF algorithm has two levels of randomization in each tree. The first is bootstrap aggregation (“bagging”), where a random subset consisting of two-thirds of the data are sampled with replacement for training and the remaining third of the data (“out-of-bag” observations) are excluded for validation. The second level of randomization is at each node of the individual decision tree. Here, a random subset of variables is selected and the variable with the best split is selected to grow the tree at that node. The best binary split at each node is identified using Gini impurity (Breiman et al., 1984), with the goal of minimizing impurity to achieve homogeneous subgroups of the data. This process is done for each individual tree in the RF. Each tree growth can be stopped by achieving a final depth or by defining a stopping criteria based on training performance. The final prediction class for each pixel is based on which class received the most votes, with each decision tree contributing one vote.

RF processing was done on a 12 core Intel Xeon E5–2630 at 2.30 GHz with 64 GB RAM running Red Hat Enterprise Linux 7.4. Data were processed using Python 2.7.12 with H2O 3.10.4.8 (H2O. ai, 2017), GDAL 2.1.0, and NumPy 1.13.0 libraries. H2O was selected for
building the RF models for its ability to inherently process continuous and categorical input data. All modeling tools and datasets are open-source.

2.2.2. Model setup—A unique model was developed for each HUC-4 with sufficient SFHA coverage (200 total) using the H2O Python module distributed RF command. At the HUC-4 level, only two watersheds (HUC 0903 – Rainy River and 0904 – Saskatchewan River) were without mapped FIRMs, meaning models were not developed based on data in those watersheds. The average model predictions from their neighboring HUC-4s were used.

If the entire FIRM dataset was divided into training and testing datasets for one CONUS-scale model it would contain over 4.6 billion pixels. The average number of pixels in each HUC-4 is about 23 million. Due to computational limitations, a random sample of 500,000 pixels classified as either [0, 1] from the processed FIRM (Table 1) along with the ten corresponding predictor variables (Table 2) was used to train a RF model for each HUC-4. An additional 500,000 completely random pixels were used for independent validation for each HUC-4.

About 90% of mapped pixels in the CONUS are not located in a flood zone. The proportion of floodplain pixels varies by HUC-4, with the maximum proportion of flood pixels at 66% and the minimum at 0.5%. This class imbalance in the response variable required special consideration. Four sampling schemes were tested: completely random sampling (using the floodplain/nonfloodplain ratio present in the HUC), and stratified random sampling to achieve 10%, 20%, or 50% floodplain pixels in the training data. The latter three sampling schemes artificially altered the distribution of each class to adjust the representation of the classes within each tree (Chen et al., 2004). However, the target sample ratios were only used in the HUC if the pixel ratio did not meet that target (i.e. if the HUC had 15% of FEMA classified pixels in a floodplain, only the completely random, 20%, and 50% schemes would be used). Results from the methods were compared using multiple performance measures (see section 3.2.3 below) to select the final sampling approach for the RF models.

The final component of model setup was defining and tuning three RF hyperparameters: number of trees (ntrees), number of variables randomly sampled as candidates at each split (mtries), and the maximum depth of each tree (max_depth). RF models were built on combinations of select hyperparameter values and the iteration with the lowest logarithmic loss (log loss) was selected for final model development. Log loss penalizes false classifications of the flood zone, where a log loss of zero is best. Generally, log loss was static after a minimum of 51 ntrees; 51 ntrees were used to ensure adequate ntrees for all models and to eliminate the possibility of a tie vote. To define mtries and max_depth, a hyperparameter grid search was performed with values of [3, 4, 5] and [8, 10, 12, 20] respectively. Mtries = 4, and max_depth = 12 were selected based on lowest log loss. Following the model building stage, the model was implemented for its HUC-4 by classifying every pixel in its domain.

2.2.3. Model performance measures—Following model training, an additional 500,000 random validation pixels were selected to assess the accuracy of our models (i.e.,
we used the FIRM as both a training and a validation dataset. We assumed that the accuracy of our RF model outside of FIRM coverage areas is comparable to the accuracy we calculated from data within the coverage area.

The primary objective was to correctly classify pixels within the floodplain. Therefore, we examined the classification accuracy using precision (P), recall (R), and F1 score of the floodplain. Precision, also known as positive predicted value, is how often the classified product is correct when compared to the source dataset (Eq. (1)). The complement of precision is false discovery rate (1 – P), which is the rate at which the model overpredicts the floodplain. Recall, also known as hit rate, is how often the source dataset is correctly classified by the model (Eq. (2)). F1 measures classification accuracy by combining precision and recall using their harmonic mean (Eq. (3)).

\[
P = \frac{TP}{TP + FP} \quad (1)
\]

\[
R = \frac{TP}{TP + FN} \quad (2)
\]

\[
F1 = \frac{2 \times P \times R}{P + R} \quad (3)
\]

where \( TP \) is the number of true positives (number of pixels correctly classified as located in the floodplain), \( FP \) is the number of false positives (number of pixels incorrectly classified as located in the floodplain), and \( FN \) is the number of false negatives (number of pixels incorrectly classified as not located in the floodplain). The \( F1 \) score was used as the primary measure for model evaluation because it balances precision and recall.

Finally, we calculated the error bias (E, in Eq. (4)) following Wing et al. (2017). Error bias uses the ratio of false positives to false negatives to determine whether the model underpredicts (\( E < 1 \)) or overpredicts (\( E > 1 \)) the floodplain.

\[
E = \frac{FP}{FN} \quad (4)
\]

3. Results

3.1. Sampling strategy

The precision, recall, and F1 scores of the four model sampling strategies were compared to select the final sampling scheme for model application across each HUC-4. The three performance measures were compared across the four sampling schemes using boxplots summarizing all HUC-4 watersheds (Fig. 2a). As recall increases, precision is sacrificed (increasing false discovery rate), and vice versa. The 50% floodplain sampling scheme has the greatest recall, but consequently overpredicts the floodplain extent and has poor
precision. Conversely, the random sampling rate has the lowest recall but greater precision and often underestimates the floodplain extent with respect to the SFHA. Our goal was to balance the precision and recall, as represented by the best F1 score. The 10% and 20% sampling had comparable F1 scores (10% mean F1: 0.74, 20% mean F1: 0.73). The 20% sampling strategy was used for classification of the entire CONUS as it had a better hit rate over the 10% strategy with minimal impact on F1 score.

As the proportion of SFHA samples used in model training increases from random, about 8% floodplain sampling rate (Fig. 2b) to 10% (Fig. 2c), 20% (Fig. 2d), and 50% (Fig. 2e), the modeled floodplain extent increases. Most of the increases in extent are false positives, which decreases precision, while more of the main channel SFHA floodplain is correctly identified by the RF model, increasing recall (Fig. 2a). Fig. 2d and e show several floodplains that are not included in the SHFA, but identified by the RF. This trend is present throughout the HUC-4s.

3.2. Variable importance

RF produces a measure of variable importance based on each variable’s relative influence. In H2O, variable importance is calculated as the improvement in squared error over all trees given a node split based on a specific variable (Breiman et al., 1984; Friedman, 2001). The distribution of the importance magnitudes is presented in Fig. 3. A consistent set of variables were most important across many of the models. The DEM-derived vertical flow distance measures, VOFD and VDC, were usually ranked as first or second in importance for most HUCs (mean ranks across HUC-4s of 1.9 and 2.5, respectively). VOFD and VDC were ranked most important in 52% and 23% of the models, translating to an importance generally between 0.2 and 0.5 across the HUC4s (Fig. 3). The moderately important variables were the categorical variables related to soil and land cover; flood frequency (fldfreq) and fluvial class (fluvclass) along with the NLCD were consistently ranked in the top half of variable importance, with mean ranks of 4.1, 5.0, and 4.5, respectively. These variables were occasionally ranked as most important for a small number of HUC-4s, where their importance exceeded 0.3 (Fig. 3). Slope, DEM5×5, OFD, and HOFD were unimportant, with average ranks > 7 across all HUC-4s and consistently low values for relative importance.

3.3. Model performance

Applying the RF models to each HUC-4 in the CONUS resulted in a total 100-year floodplain area of about 980,450 km², which is 12.1% of the total land area. In contrast, FEMA has identified a total of about 572,209 km² as floodplain, although only 60% of the CONUS is included in the FIRMs. The additional 408,241 km² of floodplain identified by the RFs includes areas unmapped by FEMA and the false positives along small streams that are not included in the SFHA (see Fig. 2d for an example). Similar false positive trends were found by Wing et al. (2017). To get a more accurate measure of performance with respect to the SFHA, they used a 1-km buffer around the SFHA to focus solely on those areas that were modeled by FEMA (Wing et al., 2017).
Performance metrics for the CONUS includes both the overall calculation (comparison with every pixel in the FIRM) and the calculation using the 1-km buffer around the SFHA while excluding open water (Table 3). The latter calculation omits streams that were not modeled by FEMA but identified by the RF models, providing an improved comparison with the SHFA.

Precision was greater than recall and error bias was less than one, indicating more false negatives (incorrectly classifying a pixel as non-floodplain) than false positives (incorrectly classifying a pixel as floodplain). Therefore, the models generally underpredicted the floodplain extent. This was the case for both methods of calculating the performance metrics, although, as expected, precision increased using the 1-km buffer due to elimination of stream floodplains not identified in the SHFA. While recall decreased slightly due to omission of open water pixels, the F1 score and precision are greater because unidentified streams are not counted as false positives.

Model performance varied geographically, as demonstrated by the map of HUC-4 performance using the F1 score (Fig. 4). The greatest F1 scores occurred in coastal regions and parts of the temperate midwest, while the lowest F1 scores were primarily located in the arid southwest and western United States.

Geographic trends in model performance were parsed using the NLCD 2011 (Homer et al., 2015) (Fig. 5a), level I ecoregions of North America (CEC, 1997) that are consistent with Olson et al. (2001) Terrestrial Ecoregions of the World (Fig. 5b), and Koppen-Geiger climate type (Peel et al., 2007) (Fig. 5c). Recall is also presented for FIRM zones in Fig. 5d, including the distinction between fluvial flooding (A classes) and coastal wave action flooding (V classes). Precision and F1 cannot be calculated because there are no associated false positives when comparing FIRM zones (e.g. no coastal non-flooding class to classify).

The models perform best for water and wetland classes in the NLCD, and the worst in developed areas (Fig. 5a). This is unsurprising given that NLCD was included as a variable within the models to improve wetland and water floodplain classification. Performance in developed land only reached an F1 score of 0.56, while F1 scores for agriculture and forest were about 0.7.

Level I ecoregions show how the model varied across ecosystems of the CONUS, highlighted by strong performance in tropical and temperate areas and relatively poor performance in arid ecoregions (Fig. 5b). Tropical southern Florida (15.0 – Tropical Wet Forests), which is dominated by wetlands and much of the area is in a floodplain, had the highest precision, recall and F1. The worst ecoregions for model performance was in arid (10.0 – North American Deserts) and semi-arid (12.0 – Southern Semi-arid Highlands) areas, which was corroborated using the Koppen-Geiger climate types.

When the HUC-4s are characterized by Koppen-Geiger broad climate type (Fig. 5c), temperate and tropical zones, typically in the southeast US, match the SFHA well, with F1 scores > 0.8. Arid (western and southwestern CONUS) is the worst with an F1 score of 0.62, along with a recall of 0.7, like those trends observed using the level I ecoregions. Cold (northeast, Great Lakes, Great Plains) falls in between with a 0.74 F1 score. Polar
performance is poor, but the area of the CONUS included in this zone is negligible compared to the other classes.

We used recall in the FIRM zones (Fig. 5d) to compare coastal (V zones) versus fluvial (A zones) performance. Recall was highest for the coastal areas (0.98 for V and VE), and lowest for AO and AH (the flooding areas related to shallow flooding, sheet flow, and ponding). Recall for fluvial FIRM zones was highest for AE (0.84), those areas where base flood elevations are provided (indicative of a detailed floodplain study).

The varying model performance and mapping characteristics across diverse landscapes and climates is presented in Fig. 6. There are four images of the modeled versus FEMA floodplain extent that highlight true positives, false positives, false negatives, and model performance beyond FEMA’s mapped extent. These locations are: a coastal/urban landscape in North Carolina (Fig. 6a), a cultivated landscape in Missouri (Fig. 6b), an arid hot desert in Arizona (Fig. 6c), and a forested/cultivated landscape in Montana (Fig. 6d). Their precision, recall, and F1 scores are listed in Table 4, along with their Koppen-Geiger broad climate types, i.e. cold, temperate, and arid (Peel et al., 2007).

The coastal area in Fig. 6a shows consistency between the SFHA and the RF model. There are a small number of false positives and false negatives, resulting in reasonably good precision and recall (Table 4). These performance measures are representative of much of the HUC-4s located in coastal and temperate areas of the country. Fig. 6b shows the modeled extent in a watershed that is partially unmapped by FEMA. The model has high precision and recall (Table 4), capturing both the mainstem and tributaries of the Marais de Cygnes River, although it over-predicts in some of the headwaters. The model also models the unmapped Osage River well with respect to the Marais des Cygnes River. Precision is greater than the recall for this watershed, indicating more false positives than false negatives. Fig. 6c visualizes the poor recall throughout the arid southwestern United States. Although the model matches well with the main stem of the stream, it misses much of the intermittent headwaters that rarely flow. The lower than average model performance in arid regions when compared to the rest of the CONUS also occurred in Wing et al. (2017). Finally, Fig. 6d demonstrates a location where the mainstem of the river has been modeled by FEMA but its tributaries were likely omitted. The performance measures indicate a greater recall than precision, although the lower precision is due in part to the headwater streams captured by the RF model but missed by the SFHA.

4. Discussion

4.1. Sampling strategy

Model performance was sensitive to sampling scheme choice, particularly in the relationship between precision and recall. We selected the 20% floodplain pixels as a minimum sampling rate because of its higher recall and ability to capture tributaries unmapped by FEMA. However, it could be optimized for each HUC (Fig. 2). Specifically, the floodplain pixel sampling rate would be a calibration parameter, with the precision, recall, and F1 scores an objective function to be maximized. Depending on the application, the desired floodplain extent may call for a more conservative (higher precision) or aggressive (higher recall)
estimate. For example, the floodplain sampling rate could be calibrated to produce more false positives. This would increase the floodplain’s areal extent and identify headwater streams at a greater rate, but would result in a decline in the performance metrics used here. Local knowledge of the watershed would be paramount in this approach.

4.2. Variable selection

Given the consistency of variable importance across the RF models, some variables could be excluded for a more efficient modeling process with fewer data collection and processing requirements. The consistent importance of DEM-derived variables VDC and VOFD along with the fluvial and flood condition soil variables are critical (Fig. 4), as is the ancillary information provided by the NLCD wetland and water classes. Variables with consistently low importance, particularly OFD and HOFD, are highly correlated with VDC and VOFD. The lack of unique information added to the RF models beyond what VDC and VOFD offered likely resulted in their limited importance.

Some relevant variables were omitted from the modeling process because they do not lend themselves to the classification method used here. This includes dams, bridges, urban stormwater infrastructure, and flood defenses. FEMA’s detailed studies do account for these structures. These variables do not lend themselves to a classification task using RF because they typically do not have a spatial extent greater than the 30-meter resolution used here. However, we can consider them to be implicitly included in the RF models because they are trained using FEMA’s SFHAs that do include those structures. Their information is more suited to use in hydrodynamic flood models than in pixel-based classification. In the case of levees, the NLD contains a “Leveed Area” layer that identifies areas protected by a levee, but it often intersects the SHFA (even for those levees designed to protect the 1% annual chance of flooding), indicating it would not be a good predictor of the floodplain using RF.

4.3. Model performance

The aggregated models performed well at the CONUS scale, with an F1 score of 0.78 and a recall (hit rate) of 0.79. This indicates that we correctly classified 79% of the floodplain pixels identified in the SFHA. These measures demonstrate performance consistent with other large-scale floodplain mapping applications such as Sangwan and Merwade (2015) and Wing et al. (2017).

The model performance was geographically variable, with the best models being in temperate ecoregions and Koppen-Geiger zones and the worst in the arid southwest (Figs. 4, 5, and 6). A decline in ability to capture the SHFAs in arid areas has been documented (Wing et al., 2017). Complex topography, variable hydrology, and low rates of mapped floodplains result in poor performance for some of these HUC-4s. This complexity is evident in Fig. 6c, where the RF model captures the main river well but misses many intermittent headwater streams that rarely experience flowing water. Low soil permeability in these areas relates to the threshold of stream generation, and while these areas flood quickly during times of intense rainfall, only terminal catchments are typically identified as floodplain soils. Because rainfall intensity and permeability were not included in the RF, some arid streams remain unidentified by the model. The potential addition of a Normalized
Difference Vegetation Index (NDVI) in arid areas may help identify floodplains. NDVI has been found to be positively correlated with river flows in arid watersheds (Nguyen et al., 2015), and Wang et al. (2015) found NDVI to be moderately influential in predicting flood risk.

Modeling urban flooding poses some difficulties for the RF models, with recall of 0.52 and F1 of 0.56 (the lowest values of any land cover type) (Fig. 5a). This is due to the complexity of urban hydrology, which is not captured in a 30-meter DEM or the flow path derivatives we use here. Highly populated, urban areas are most likely to have detailed FEMA studies, which account for hydraulic structures like bridges and culverts. This level of detail is not included in the RF models or underlying flow direction and accumulation datasets used to build the distance to channel variables. In their modeling effort, Wing et al. (2017) also observed lower recall scores when compared to forested or other undeveloped land cover types.

Recall for coastal flooding zones was superior to fluvial areas (Fig. 5d). Coastal HUCs typically have greater floodplain map coverage and a greater rate of areas classified as floodplain, giving the RF more training data to use in model building. In addition, wetlands were prevalent in these low-lying areas, demonstrating the utility of both the wetland NLCD class and the vertical distance to channel/coast variables.

The most consistent errors were the model tendency to overpredict the floodplain extent, typically due to identification of tributaries not present in some FIRM studies (Fig. 6d). This is not necessarily an indicator of poor performance, but adds value by identifying floodplains of streams excluded from the SHFA. This is the most common reason for some watersheds having low precision coupled with high recall, and is a function of the flow accumulation and stream channel network delineation process in Section 2.1.3. Higher DEM resolution and/or a lower flow accumulation threshold for channel delineation would likely result in an even greater rate of capturing headwater floodplains. When accounting for this by calculating model performance around a 1-km buffered SFHA, precision did decrease through elimination of misleading false positives (Table 3).

A few HUC-4s stand out for their poor performance in otherwise well performing regions. For example, HUC 0805 near the outlet of the Mississippi River has one of the lowest F1 (0.46) and recall (0.33) scores in an otherwise well performing HUC-2 watershed. The poor performance is likely because the entirety of the watershed area is leveed according to the United States Army Corps of Engineers (USACE) National Levee Database (NLD; nld.usace.army.mil). In this case, the flood defenses throughout the watershed complicate the use of distance to channel variables for building the RF models (NLCD was identified as the most important variable). HUC 0902, containing the Red River of the North has a poor recall (0.50) and F1 (0.55) scores. The issues here lie in the complexity of the watershed, in the combination of spring snowmelt, flat topography, and the potential for ice jams. Because the river flows north, the headwaters experience snowmelt before the downstream areas, which causes ice jams and backwater flooding. These factors cannot be captured by the RF.
There are a few sources of error and uncertainty within the modeling process, mainly with respect to variable selection and the way they variables were created. The most prominent variables deemed important in the RF (i.e. VDC and VOFD) are derived from a DEM, which carries its own accuracy and uncertainty issues. Derivation of the distance to channel measures are a function of flow accumulation and the threshold at which the stream network is defined. Choice of flow direction algorithm and threshold to stream network generation influences the distance variables. For example, the D-Infinity flow direction algorithm (Tarboton, 1997) improves upon the D8 algorithm used here (Section 2.1.2) because it uses continuous flow angles and flow partitioning between neighbor pixels, allowing for better representation of flow on divergent hillslopes. Calibrating these approaches at the HUC-4 level could lead to improved floodplain predictions, particularly in arid locations where headwaters were routinely missed by the RF model.

The gaps in county-level soil information from SSURGO were filled with the coarser resolution STATSGO soils datasets, resulting in lower quality input data in some locations, mostly in the western CONUS. A more complete high-resolution soil dataset would improve floodplain predictions in areas that currently lack it, particularly poor-performing arid HUC-4s. SSURGO and STATSGO have their own errors and uncertainties (Sangwan and Merwade, 2015), and the delineation of soil map units affects the RF modeling process. Emerging soil mapping methods for the CONUS, such as the 30-meter resolution probabilistic soil data (POLARIS) by Chaney et al. (2016), have the potential to strengthen the model in areas unmapped via SSURGO, while accounting for uncertainty in delineated soil maps.

Finally, there is uncertainty in the SFHA itself. The SFHA is based on numerous individual site studies, most involving hydrodynamic modeling, but many of them are approximate. These approximate studies derive floodplains through soils maps, high water profiles, aerial photographs, and topographic maps (FEMA, 2005). The HUC-4 RF models were often trained on data that encompassed both detailed and approximate studies, of which there could be several in a watershed. There are also other flood studies that have been conducted that do not meet FEMA standards, but may be included in FEMA’s approximate floodplain areas (FEMA, 2005). The differences in RF performance across different SFHA classifications are evident in Fig. 5d, as recall was greater for those areas with detailed studies (Zone AE) versus the approximate floodplain areas (Zone A).

### 4.4. Applications

This process is scalable to higher resolution floodplain mapping. For example, the USGS National Elevation Dataset now includes a seamless 10-meter resolution DEM with full coverage in the CONUS, along with partial coverage of the CONUS at 3 m and 1 m resolution. These higher resolution datasets could be used in future mapping efforts, now that we have successfully verified the capability of RF to identify floodplains. Community-level efforts that have created DEMs with high resolution LiDAR are also available. Meanwhile, gridded SSURGO (gSSURGO) data exists at 10 m resolution. The methodology of this study is flexible for use in higher resolution applications as datasets at these scales become more prevalent. As more data are produced in places beyond the CONUS, this
method could be applied to rapidly produce moderate-resolution, screening-level floodplain information in unmapped areas for site-specific applications.

There are multiple efforts to map past flooded extents both in the United States (U.S. Flood Inundation Map Repository; USFIMR; https://sdml.ua.edu/usfimr/) and globally (Global Flood Inundation Map Repository; GloFIMR; https://sdml.ua.edu/glofimr/) using satellite information from Landsat, MODIS, Sentinel-1 synthetic aperture radar (SAR), etc., although spatial resolution and temporal challenges remain (Revilla-Romero et al., 2015) in building comprehensive flood catalogs. The Dartmouth Flood Observatory (DFO; http://floodobservatory.colorado.edu/) has also mapped many observed flooded extents. The RF models and data in this study could be used with these previously mapped flooded extents to further identify floodplain return periods other than the confines of the 100-year SFHA. As the CONUS and global catalogs of floods captured by satellites, the algorithms for identifying floodplains using physiographic characteristics will improve, especially in data-sparse regions.

This dataset provides spatial estimates of flood inundation for areas without mapped floodplain extents, whether it is used for land use planning or application in studies that require floodplain information for ecological restoration and/or ecosystems services assessments. Floodplains provide several valuable ecosystem services, including downstream flood mitigation, reducing stress on dams and reservoirs, and supporting floodplain fisheries (Opperman et al., 2009; Remo et al., 2017). This map provides a complete picture of the extent of floodplains in the CONUS, allowing for an assessment of their quality, condition, and their ability to provide ecosystem goods and services. However, it is not intended to replace FEMA’s FIRMs, nor is it applicable for scenario/change analysis. Potential applications include identifying areas for floodplain restoration and channel reconnection (Watson et al., 2016), identifying hydrologic connections between floodplains and wetland systems (Lane et al., 2017), and assessing the interplay between agricultural lands, floodplains, and water quality (Schilling et al., 2015). To that end, the data are publicly available on the EnviroAtlas (Pickard et al., 2015), a suite of interactive tools that facilitate ecosystem services decision-making at the community- and national-scales.

5. Conclusions

Almost 40% of the CONUS does not have mapped floodplains through FEMA. This research presents a method to rapidly classify landscapes as lying in the 100-year floodplain, both coastal and fluvial, to fill in these gaps. We use only nationally available public data and existing machine learning algorithms. Individual RF models were developed at 30-meter resolution for each HUC-4 watershed in the CONUS using derivatives of topography, soil, and land cover data as predictor variables, and FEMA’s SFHA as training data. The HUC-4 models were merged to create a seamless floodplain map for the CONUS. The rapid process developed here contrasts with the time-intensive and costly modeling process used by FEMA to develop FIRMs. As FIRMs are added and updated, this data can be incorporated into the RF models to create updated and improved predictions of locations lacking FIRMs.
Overall performance of the RF models was good, with an F1 score of 0.79 for the CONUS. The models were tuned to balance the rate of false positives and false negatives, resulting in some under-predictions and over-predictions of floodplain extent. Much of the false positives captured headwater streams and low-order systems typically not included in FEMA’s SFHA. These systems are critical components of the watershed ecosystem and were previously unmapped in large portions of the CONUS.

Performance at the HUC-4 level varied geographically across the CONUS, with best performance in temperate climates and coastal areas. Floodplains in the arid southwest were the most difficult to capture, likely due to their complex topography and prevalence of intermittent streams. While the models reproduced main channel floodplains, they often missed arid headwaters that are often dry. Other HUC-4s that performed poorly were likely due to information not included in the RFs, such as levees, although the National Levee Database remains incomplete.

The RF floodplain maps cannot replace FEMA’s detailed flood inundation studies in delineation of SFHAs, but provide estimates of floodplain extents in areas that have not yet been mapped. The pixel-based classification process here does not include hydrodynamic modeling, nor does it represent the physical processes that control flood flow. Therefore, it cannot be used in scenario testing or modeling of return periods other than the 1% annual chance of flooding. However, it is a valuable resource for exploratory floodplain analyses and provides a technique to rapidly map data-scarce regions. As a national map of flood extent, this product and the methods used to create it have applications with respect to floodplain restoration and ecosystems services, particularly in headwater streams.

Acknowledgements

This paper has been reviewed in accordance with the U.S. Environmental Protection Agency’s peer-review policies and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use. Statements in this publication reflect the authors’ personal views and opinions and should not be construed to represent any determination of policy of the U.S. Environmental Protection Agency. This project was supported in part by an appointment to the Internship/Research Participation Program at the Office of Research and Development, U.S. Environmental Protection Agency, administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and EPA. The final data are available for use or download on EPA’s EnviroAtlas (https://www.epa.gov/enviroatlas/enviroatlas-interactive-map), and input datasets are available upon request.

References

ASFPM, 2013. Flood Mapping for the Nation: A Cost Analysis for the Nation’s Flood Map Inventory. Association of State Floodplain Managers.

Berhane T, Lane C, Wu Q, Anenkhanov O, Chepinoga V, Autrey B, et al., 2017. Comparing pixel- and object-based approaches in effectively classifying wetland-dominated landscapes. Remote Sens. 10, 46.

Berhane TM, Lane CR, Wu Q, Autrey BC, Anenkhanov OA, Chepinoga VV, et al., 2018. Decision-tree, rule-based, and random forest classification of high-resolution multispectral imagery for wetland mapping and inventory. Remote Sens. 10, 580.

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

Breiman L, Friedman J, Stone CJ, Olshen RA, 1984. Classification and Regression Trees. CRC press.

CEC, 1997. Ecological Regions of North America: Toward a Common Perspective. Commission for Environmental Cooperation, Montreal, Quebec, Canada.
Chaney NW, Wood EF, McBratney AB, Hempel JW, Nauman TW, Brungard CW, et al., 2016. POLARIS: a 30-meter probabilistic soil series map of the contiguous United States. Geoderma274, 54–67.

Chen C, Liaw A, Breiman L, 2004. Using Random Forest to Learn Imbalanced Data. Vol. 110. University of California, Berkeley, pp. 1–12.

Conrad O, Bechtle B, Bock M, Dietrich H, Fischer E, Gerlitz L, et al., 2015. System for automated geoscientific analyses (SAGA) v. 2.1. 4. Geosci. Model Dev. 8, 1991.

FEMA, 2005. National Flood Insurance Program (NFIP) floodplain management requirements: A study guide and desk reference for local officials. FEMA 480. Federal Emergency Management Agency.

FEMA, 2017. National Flood Hazard Layer (NFHL) 20170313. Federal Emergency Management Agency.

Fernández DS, Lutz MA, 2010. Urban flood hazard zoning in Tucumán Province, Argentina, using GIS and multicriteria decision analysis. Eng. Geol. 111, 90–98.

Friedman JH, 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 1189–1232.

Gesch DB, Oimoen MJ, Evans GA, 2014. Accuracy Assessment of the US Geological Survey National Elevation Dataset, and Comparison With Other Large-area Elevation Datasets: SRTM and ASTER. US Geological Survey, p. 10.

H2O.ai, 2017. Python Interface for H2O, Python Module Version 3.10.4.8.

Hjerdt KN, McDonnell JJ, Seibert J, Rodhe A, 2004. A new topographic index to quantify downslope controls on local drainage. Water Resour. Res. 40.

Hong H, Pourghasemi HR, Pourtaghi ZS, 2016. Landslide susceptibility assessment in Lianhua County (China): a comparison between a random forest data mining technique and bivariate and multivariate statistical models. Geomorphology259, 105–118.

Ireland G, Volpi M, Petropoulos G, 2015. Examining the capability of supervised machine learning classifiers in extracting flooded areas from Landsat TM imagery: a case study from a Mediterranean flood. Remote Sens. 7, 3372–3399.

Jafarzadeegan K, Merwade V, 2017. A DEM-based approach for large-scale floodplain mapping in ungauged watersheds. J. Hydrol. 550, 650–662.

Kourgialas NN, Karatzas GP, 2017. A national scale flood hazard mapping methodology: the case of Greece–protection and adaptation policy approaches. Sci. Total Environ. 601, 441–452. [PubMed: 28575822]

Lane CR, Hall A, D’Amico E, Sangwan N, Merwade V, 2017. Characterizing the extent of spatially integrated floodplain and wetland systems in the White River, Indiana, USA. J. Am. Water Resour. Assoc. 53, 774–790. [PubMed: 33408455]

Lee S, Kim J-C, Jung H-S, Lee MJ, Lee S, 2017. Spatial prediction of flood susceptibility using random-forest and boosted-tree models in Seoul metropolitan city, Korea. Geom. Nat. Haz. Risk8, 1185–1203.

Liauw A, Wiener M, 2002. Classification and Regression by randomForest. 2. R news, pp. 18–22.

Manfreda S, Di Leo M, Sole A, 2011. Detection of flood-prone areas using digital elevation models. J. Hydrol. Eng. 16, 781–790.

Maxwell AE, Warner TA, Strager MP, 2016. Predicting palustrine wetland probability using random forest machine learning and digital elevation data-derived terrain variables. Photogramm. Eng. Remote. Sens. 82, 437–447.

NCEI, 2018. In: Information NNCfE (Ed.), U.S. Billion-dollar Weather and Climate Disasters.

Nguyen U, Glenn EP, Nagler PL, Scott RL, 2015. Long-term decrease in satellite vegetation indices in response to environmental variables in an iconic desert riparian ecosystem: the upper San Pedro, Arizona, United States. Ecohydrology8, 610–625.
Olson DM, Dinerstein E, Wikramanayake ED, Burgess ND, Powell GV, Underwood EC, et al., 2001. Terrestrial ecoregions of the world: a new map of life on earth. A new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. Bioscience 51, 933–938.

Opperman JJ, Galloway GE, Fargione J, Mount JF, Richter BD, Secchi S, 2009. Sustainable floodplains through large-scale reconnection to rivers. Science 326, 1487–1488. [PubMed: 20007887]

Peel M, Finlayson B, McMahon T, 2007. Updated world map of the Koppen-Geiger climate classification. Hydrol. Earth Syst. Sci. 11, 1633–1644.

Pickard BR, Daniel J, Mehaffey M, Jackson LE, Neale A, 2015. EnviroAtlas: a new geospatial tool to foster ecosystem services science and resource management. Ecosyst. Serv. 14, 45–55.

Planckon O, Darboux F, 2002. A fast, simple and versatile algorithm to fill the depressions of digital elevation models. Catena 46, 159–176.

Pradhan B, 2013. A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS. Comput. Geosci. 51, 350–365.

Remo J, Guida R, Secchi S, 2017. Screening the suitability of levee protected areas for strategic floodplain reconnection along the LaGrange segment of the Illinois River, USA. River Res. Appl. 33, 863–878.

Revilla-Romero B, Hirpa FA, Pozo JT-D, Salamon P, Brakenridge R, Pappenberger F, et al., 2015. On the use of global flood forecasts and satellite-derived inundation maps for flood monitoring in data-sparse regions. Remote Sens. 7, 15702–15728.

Rodriguez-Galiano V, Sanchez-Castillo M, Chica-Olmo M, Chica-Rivas M, 2015. Machine learning predictive models for mineral prospectivity: an evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geol. Rev. 71, 804–818.

Rosenzweig C, Tubiello FN, Goldberg R, Mills E, Bloomfield J, 2002. Increased crop damage in the US from excess precipitation under climate change. Glob. Environ. Chang. 12, 197–202.

Sangwan N, Merwade V, 2015. A faster and economical approach to floodplain mapping using soil information. J. Am. Water Resour. Assoc. 51, 1286–1304.

Schilling KE, Jacobson PJ, Vogelgesang JA, 2015. Agricultural conversion of floodplain ecosystems: implications for groundwater quality. J. Environ. Manag. 153, 74–83.

Tarboton DG, 1997. A new method for the determination of flow directions and upslope areas in grid digital elevation models. Water Resour. Res. 33, 309–319.

Tehrany MS, Pradhan B, Jebur MN, 2013. Spatial prediction of flood susceptible areas using rule based decision tree (DT) and a novel ensemble bivariate and multivariate statistical models in GIS. J. Hydrol. 504, 69–79.

Tehrany MS, Pradhan B, Jebur MN, 2014. Flood susceptibility mapping using a novel ensemble weights-of-evidence and support vector machine models in GIS. J. Hydrol. 512, 332–343.

Tehrany MS, Pradhan B, Mansor S, Ahmad N, 2015. Flood susceptibility assessment using GIS-based support vector machine model with different kernel types. Catena 125, 91–101.

Tien Bui D, Le K-TT, Nguyen VC, Le HD, Revhaug I, 2016. Tropical forest fire susceptibility mapping at the Cat Ba National Park area, Hai Phong City, Vietnam, using GIS-based kernel logistic regression. Remote Sens. 8, 347.

Wang Z, Lai C, Chen X, Yang B, Zhao S, Bai X, 2015. Flood hazard risk assessment model based on random forest. J. Hydrol. 527, 1130–1141.

Watson KB, Ricketts T, Galford G, Polasky S, O’Niel-Dunne J, 2016. Quantifying flood mitigation services: the economic value of Otter Creek wetlands and floodplains to Middlebury, VT. Ecol. Econ. 130, 16–24.

Wing OEJ, Bates PD, Sampson CC, Smith AM, Johnson KA, Erickson TA, 2017. Validation of a 30 m resolution flood hazard model of the conterminous United States. Water Resour. Res. 53, 7968–7986.

Youssef AM, Pourghasemi HR, Pourtaghi ZS, Al-Katheeri MM, 2015. Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir region, Saudi Arabia. Landslides 13, 839–856.
Zhao G, Pang B, Xu Z, Yue J, Tu T. 2018. Mapping flood susceptibility in mountainous areas on a national scale in China. Sci. Total Environ. 615, 1133–1142. [PubMed: 29751419]
| **HIGHLIGHTS** |
|----------------|
| - Floodplains provide many ecosystem services, but are unmapped for 40% of the US. |
| - Random forest and publicly available geospatial data used to classify floodplains. |
| - Models captured 79% of the floodplain as identified using FEMA 100-year floodplain. |
| - Floodplains in previously unmapped areas were successfully identified. |
| - The methods used can be adapted to regions lacking floodplain maps |
Fig. 1. 
HUC-4 watersheds and FEMA FIRM coverage (see Table 1 for description of FIRM zones).
Fig. 2.
Comparison of model accuracy and inundation area across sampling strategies. (a) Precision, recall, and F1 scores for each sampling scheme across all HUC-4 watersheds. Modeled floodplain extents for each sampling strategy are presented for (b) random (8% for this HUC), (c) 10%, (d) 20%, and (e) 50% sampling ratios of the floodplain class versus the SFHA. The location depicted is HUC 1025 at the confluence of the Republican River and Arikaree River near Haigler, NE.
Fig. 3.
Variable importance violins and boxplots showing distribution of the magnitude of scaled variable importance across all HUC-4 models. Dots indicate medians, lower and upper hinges correspond to first and third quartiles, and the whiskers extend no further than 1.5 times the interquartile range.
Fig. 4. F1 score by HUC-4. Darker colors indicate better F1 score.
Fig. 5.
Model performance across multiple landscape classifications: (a) NLCD 2011 Level I land use/land cover, (b) level I ecoregion, (c) Koppen-Geiger Climate, and (d) FEMA FIRM zone. Note that (d) only includes recall (R) because false positives cannot be classified for the FIRM zones. See Table 1 for FIRM zone descriptions. Level I Ecoregion classifications are as follows: 5.0 (Northern Forests), 6.0 (Northeastern Forested Mountains), 7.0 (Marine West Coast Forest), 8.0 (Eastern Temperate Forests), 9.0 (Great Plains), 10.0 (North
American Deserts), 11.0 (Mediterranean California), 12.0 (Southern Semi-arid Highlands),
13.0 (Temperate Sierras), 15.0 (Tropical Wet Forests).
Fig. 6.
Modeled versus SFHA for (a) HUC 0303, Cape Fear River near Wilmington, NC, (b) HUC 1029, the confluence of the Marais des Cygnes River and the Osage River near Rich Hill, MO, (c) HUC 1507, unnamed arroyos southwest of Phoenix, AZ, and (d) HUC 1007, Yellowstone River near Waco, MT.
### Table 1
FEMA FIRM zones and their training classification, adapted from FEMA (2005).

| FIRM zone | Description                                                                                                                                                                                                 | Training classification |
|-----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------|
| A         | The 100-year floodplain mapped by approximate methods, i.e., Base Flood Elevations (BFEs) are not determined. This is often called an unnumbered A Zone or an approximate A Zone.                                      | 1                       |
| A99       | Area to be protected from 100-year flood by levees or Federal Flood Protection Systems under construction. BFEs are not determined.                                                                            | 1                       |
| AE        | The 100-year floodplain where base flood elevations are provided.                                                                                                                                             | 1                       |
| AH        | Shallow flooding 100-year floodplain.                                                                                                                                                                          | 1                       |
| AO        | The 100-year floodplain with sheet flow, ponding, or shallow flooding.                                                                                                                                       | 1                       |
| Area not included | Area of undetermined but possible flood hazards.                                                                                                                  | No Data                |
| D         | Area of undetermined but possible flood hazards.                                                                                                                                                         | No Data                |
| Open water |                                                                                                                                                                                                             | 1                       |
| V         | The coastal area subject to a velocity hazard (wave action) where BFEs are not determined on the FIRM.                                                                                                       | 1                       |
| VE        | The coastal area subject to a velocity hazard (wave action) where BFEs are provided on the FIRM.                                                                                                               | 1                       |
| X         | Area of moderate or minimal flood hazard, usually the area above the 100-year flood level.                                                                                                                  | 0                       |
Table 2

Predictor variables used in the modeling process.

| Variable name                  | Property   | Type     | Derivative source                       |
|--------------------------------|------------|----------|-----------------------------------------|
| NLCD                           | Land cover | Categorical | National Land Cover Database 2011     |
| fldfreq (flood frequency)      | Soil       | Categorical | SSURGO/STATSGO                          |
| fluvclass (fluvial class)      | Soil       | Categorical | SSURGO/STATSGO                          |
| Slope                          | Topography | Continuous | National Elevation Dataset              |
| DEM5x5                         | Topography | Continuous | National Elevation Dataset              |
| CTI (compound topographic index) | Topography | Continuous | National Elevation Dataset              |
| OFD (overland flow distance)   | Topography | Continuous | National Elevation Dataset              |
| HOFD (horizontal OFD)          | Topography | Continuous | National Elevation Dataset              |
| VOFD (vertical OFD)            | Topography | Continuous | National Elevation Dataset              |
| VDC (vertical distance to channel) | Topography | Continuous | National Elevation Dataset              |
Table 3
Model performance metrics for the CONUS, with and without a 1-km buffer around the SHFA.

| Calculation method | Precision | False discovery rate | Recall | F1    | Error bias |
|-------------------|-----------|----------------------|--------|-------|------------|
| Overall           | 0.78      | 0.22                 | 0.79   | 0.78  | 1.11       |
| SFHA 1-km buffer  | 0.83      | 0.17                 | 0.78   | 0.81  | 0.74       |
### Table 4

Performance measures for HUC-4s depicted in Fig. 6.

| HUC (matched fig) | Koppen-Geiger class | FEMA mapped rate | Precision | Recall | F1   |
|-------------------|----------------------|------------------|-----------|--------|------|
| 0303 (Fig. 6a)    | Temperate            | 99%              | 0.77      | 0.78   | 0.78 |
| 1029 (Fig. 6b)    | Cold                 | 76%              | 0.83      | 0.85   | 0.84 |
| 1507 (Fig. 6c)    | Arid                 | 51%              | 0.73      | 0.58   | 0.65 |
| 1007 (Fig. 6d)    | Cold                 | 53%              | 0.47      | 0.90   | 0.62 |