Science of the Total Environment 647 (2019) 942–953

Contents lists available at ScienceDirect

Science of the Total Environment

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

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

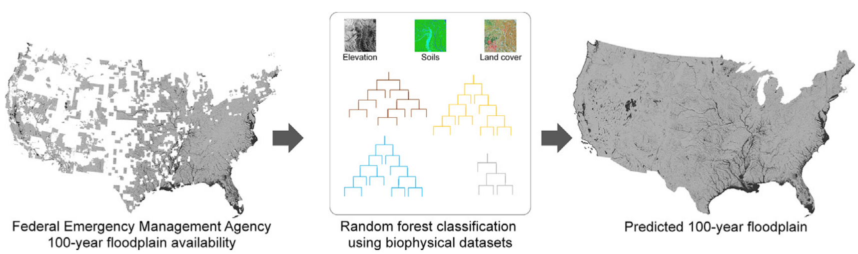
Sean A. Woznicki [a,*], Jeremy Baynes [a], Stephanie Panlasigui [b,1], Megan Mehaffey [a], Anne Neale [a]

[a] National Exposure Research Laboratory, US Environmental Protection Agency, Research Triangle Park, NC 27711, USA
[b] Oak Ridge Institute for Science and Education Research Participant Program, US Environmental Protection Agency, Research Triangle Park, NC 27711, USA

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.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 18 May 2018
Received in revised form 19 July 2018
Accepted 24 July 2018
Available online 25 July 2018

Editor: Ouyang Wei

Keywords:
CONUS
Ecosystem services
EnviroAtlas
Flood
Geographic information systems
Machine learning

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.

* Corresponding author at: US Environmental Protection Agency, 109 T.W. Alexander Drive, Research Triangle Park, NC 27711, USA.
E-mail address: woznicki.sean@epa.gov (S.A. Woznicki).
[1] Present address: Eno River Association, 4404 Guess Rd. Durham, NC 27712.

https://doi.org/10.1016/j.scitotenv.2018.07.353

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 mid-west 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 (<1% 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