Journal of Environmental Management 272 (2020) 111051

Contents lists available at ScienceDirect

Journal of Environmental Management

journal homepage: http://www.elsevier.com/locate/jenvman

Research article

Predicting flood insurance claims with hydrologic and socioeconomic demographics via machine learning: Exploring the roles of topography, minority populations, and political dissimilarity

James Knighton [a,*], Brian Buchanan [b], Christian Guzman [c], Rebecca Elliott [d], Eric White [e], Brian Rahm [f]

[a] The National Socio-Environmental Synthesis Center, Annapolis, MD, USA
[b] New York State Department of Environmental Conservation, NY, USA
[c] University of Massachusetts Amherst, MA, USA
[d] Department of Sociology, London School of Economics, UK
[e] Coastal Protection and Restoration Authority of Louisiana, LA, USA
[f] Water Resources Institute of New York, NY, USA

ARTICLE INFO

Keywords:
Flooding insurance claims
Random forest
Vulnerability
Socio-hydrology
Flooding
FEMA
LIS-FLOOD

ABSTRACT

Current research on flooding risk often focuses on understanding hazards, de-emphasizing the complex pathways of exposure and vulnerability. We investigated the use of both hydrologic and social demographic data for flood exposure mapping with Random Forest (RF) regression and classification algorithms trained to predict both parcel- and tract-level flood insurance claims within New York State, US. Topographic characteristics best described flood claim frequency, but RF prediction skill was improved at both spatial scales when socioeconomic data was incorporated. Substantial improvements occurred at the tract-level when the percentage of minority residents, housing stock value and age, and the political dissimilarity index of voting precincts were used to predict insurance claims. Census tracts with higher numbers of claims and greater densities of low-lying tax parcels tended to have low proportions of minority residents, newer houses, and less political similarity to state level government. We compared this data-driven approach and a physically-based pluvial flood routing model for prediction of the spatial extents of flooding claims in two nearby catchments of differing land use. The floodplain we defined with physically based modeling agreed well with existing federal flood insurance rate maps, but underestimated the spatial extents of historical claim generating areas. In contrast, RF classification incorporating hydrologic and socioeconomic demographic data likely overestimated the flood-exposed areas. Our research indicates that quantitative incorporation of social data can improve flooding exposure estimates.

1. Introduction

In the US, extreme rainfall and riverine flooding events are dominant environmental mechanisms of economic loss, averaging 3.3 billion USD annually (NCDC, 2019). Environmental risk is the confluence of environmental hazards (e.g. floods), exposure, and vulnerability, summed across all levels of hazard (Kron, 2005). Global climate change (Hirabayashi et al., 2013), land cover change (e.g. Wheater and Evans, 2009), human population migration (e.g. Donner and Rodriguez, 2008), and socioeconomic conditions (e.g. Dixon et al., 2017) shift in ways that modify riverine flooding hazards, exposure, and vulnerability. Established methods to accurately quantify riverine flooding risks center strongly on accurate representations of the physical mechanisms by which floods are generated (i.e. hazards), but frequently neglect or de-emphasize the role of human-flood interactions necessary for translating hazard into exposure and risk (e.g. Metin et al., 2018; Elliott, 2018, 2019; Koks et al., 2015).

Physically-based hydrologic models allow us to carry forward our prior knowledge of the physics of water movement in the landscape (e.g. mass and energy balances) to place important constraints on hazard estimates. Historically, across the Contiguous United States (CONUS) there has been disparity in the methodologies, resolution, and

* Corresponding author. The National Socio-Environmental Synthesis Center, Annapolis, MD, USA.
E-mail addresses: jknighton@sesync.org (J. Knighton), bb386@cornell.edu (B. Buchanan), cdguzman@umass.edu (C. Guzman), R.Elliott1@lse.ac.uk (R. Elliott), Eric.White@la.gov (E. White), bgr4@cornell.edu (B. Rahm).

https://doi.org/10.1016/j.jenvman.2020.111051
Received 11 February 2020; Received in revised form 26 May 2020; Accepted 3 July 2020
Available online 15 July 2020

uncertainty in established flood hazard maps, possibly a result of the resources required to develop these estimates and the potential social outcomes of redrawing hazard boundaries (Kousky, 2018; Elliot, 2018; Pralle, 2019; Blessing et al., 2017; Nance, 2015). Spatially continuous maps of flooding hazards eliminate issues of availability and heterogeneous methodology. For example, Wing et al. (2018) developed a 30 m 100-year riverine inundation map of CONUS through 2-dimensional surface flood routing, thereby establishing national coverage with a uniform methodology. Similar spatial datasets of discharge and inundation have been developed at coarser resolutions (e.g. Knighton et al., 2019a; Zheng et al., 2018; Dottori et al., 2016; Hirabayashi et al., 2013). Despite these advances in physical hazard mapping, there remain challenges to accurately identifying at-risk properties. In the US, hydrologic model-derived FEMA Flood Insurance Rate Maps (FIRMs; FEMA, 2019a) have been identified as inadequate representations of flood insurance claim-generating areas (e.g. Kousky, 2018; Highfield et al., 2013; Burby, 2001).

The unstructured nature of machine learning algorithms potentially reduces the problematic structural biases, specific data-needs, and calibration challenges of hydrologic models. A variety of machine learning techniques have been used to map between widely available hydrologic and atmospheric features and flooding hazards (e.g. Wang et al., 2019; Chen et al., 2019; Bui et al., 2019; Souissi et al., 2019; Khosravi et al., 2019, 2018; Knighton et al., 2019a; Hong et al., 2018a, b; Shafizadeh-Moghadam et al., 2018; Woznicki et al., 2019; Ngo et al., 2018; Ahmadlou et al., 2018; Giovannettone et al., 2018; Chapi et al., 2017). Together these studies demonstrate that this broad family of methodologies can facilitate rapid riverine flood hazard estimates.

Data-driven mapping approaches may bypass the traditional hydrometeorological requirements of hydrologic models (e.g. continuous data for rainfall, stream discharge, air temperatures, solar radiation, soil textures), however, there may be a stronger reliance on large datasets defining historically flooded locations. Another possible limitation of data-driven methodologies is that they may inadvertently carry forward conceptual limitations encoded in the training data. For example, Woznicki et al. (2019) and Giovannettone et al. (2018) present machine learning techniques trained to reliably estimate the Special Flood Hazard Area (SFHA), with promise for generating SFHA maps in previously unstudied regions. When established FIRMs are used as training data, algorithms risk learning many of the same biases of the hydrologic modeling methodologies employed to establish the original inundation extent. Finally, focusing exclusively on prediction of inundated area neglects that hazards alone do not describe risks, nor are hydrometeorological characteristics the only predictors of flooding loss (Di Baldassarre et al., 2018).

Political-ecology theory states that any environmental change will initiate a socioeconomic upheaval followed by an uneven redistribution of losses (Blaikie, 2008). This conceptual model has been used to describe cycles of flooding, loss, and recovery (Bolin and Kurtz, 2018). Uneven flooding losses across socioeconomic groups can reinforce existing systems of disparity, perpetuating flooding vulnerabilities. The consequences of extreme floods are often felt most by underrepresented portions of the population, which frequently align with race, class, and health in the US (Hale et al., 2018; Rufat et al., 2015). For example, lower income households in the US that cannot afford flooding insurance or mitigation measures may experience greater losses during floods. These losses can necessitate a reliance on federal assistance and charities for recovery, leading to increased economic vulnerability to future floods (FEMA, 2018; Dixon et al., 2017). Identification of knowledge gaps between physical flooding mechanisms and the socioeconomic consequences of these events has prompted recent calls to re-center studies of water resources around human-water interactions (Di Baldassarre et al., 2019; Vorogushyn et al., 2018) and national flood insurance programs (e.g. FEMA, 2018) around risk.

Advances in the study of flooding that account for social and economic dimensions (e.g. Edelenbos et al., 2017; Merz et al., 2010) could facilitate both stronger risk mitigation policies and clearer risk communication (Aerts et al., 2018). Socioeconomic demographics may provide information on which properties are flood-exposed (e.g. Burton and Cutter, 2008; Boyce et al., 2006) and which residents are likely to generate insurance claims following loss events. In residential areas, publication of new flooding insurance products can reshape local perceptions of risk (Elliott, 2018, 2019) and influence housing prices (Dixon et al., 2017; Indaco et al., 2019). Reduced property value can limit residents' ability to relocate or borrow against their home, both of which possibly lead to continued exposure to floods (Siders et al., 2019) and possibly increased claims. Conversely, Elliott (2015) and Cutter et al. (2018) found that flooding in the US induced more migration among minority residents, leaving flood-prone areas inhabited primarily by white residents with the economic means for recovery. Geographic variations in property value may provide market evidence of a history of flood exposure (e.g. Indaco et al., 2019). Case studies of flood-prone regions have found both race (Atreya et al., 2015) and income (Dixon et al., 2017) to be predictive of the willingness or ability of residents to participate in NFIP. Finally, histories of segregation along lines of race, class, or beliefs have possibly clustered individuals with demographic similarities into areas of similar environmental risk.

NFIP flood insurance claim records provide parcel- and tract-level information on historical flooding hazards and possibly hazard-risk relationships (e.g. Czajkowski et al., 2017; Kousky and Michel-Kerjan, 2017; Zhou et al., 2013). Participation in the NFIP has risen steadily over the past several decades to approximately 5 million homes, with an average of 60,000 flooding insurance claims (2008–2018) filed annually (FEMA, 2019b). Inaccurate FIRMs (Kousky, 2018), economic barriers (Dixon et al., 2017), and the risk perceptions of homeowners (Elliott, 2018) can reduce participation in NFIP, which may limit the use of insurance claims as an unbiased picture of historical hazards and risk. We introduce a novel approach to map riverine flood insurance claims through random forest regression and classification trained directly on a state-wide dataset of US National Flood Insurance Program (NFIP) claims data within New York State (NYS), US. This new approach is used to address the following research questions:

- Are hydrologic conditions and social demographics predictive of the rate at which flood insurance claims are generated from census tracts and individual parcels in NYS?
- Can a risk-based classification algorithm incorporating social demographics identify the spatial distribution of flooding claims more reliably than a classic hydrodynamic modeling-based approach which focuses exclusively on hazards?

2. Methodology

2.1. Study region

NYS, located in the northeastern US, experiences riverine and coastal flooding with average annual residential insurance claims totaling 110 million USD. Flooding loss claims are spatially distributed across the state with the highest density of claims centered on urban areas (Fig. 1d). Despite population density variations within NYS, reports of historical flash flooding events were found to be unbiased by density (Marjerison et al., 2016). Regional, extreme runoff events are initiated by several dominant mechanisms: tropical moisture export derived intense precipitation in the late summer and fall seasons (Frei et al., 2015; Huang et al., 2018), localized convective rainfall in the summer, and regional extratropical winter and spring precipitation and snowmelt often on saturated soils, leading to a seasonally bimodal flooding regime (Knighton et al., 2017; Villarini, 2016). The dominant mode of runoff generation across NYS is saturation-excess (Buchanan et al., 2018).

We selected two nearby catchments within the Hudson River Watershed (NYS, US), the Moodna and Hackensack catchments, for comparison of flood risk estimation methodologies. Moodna and

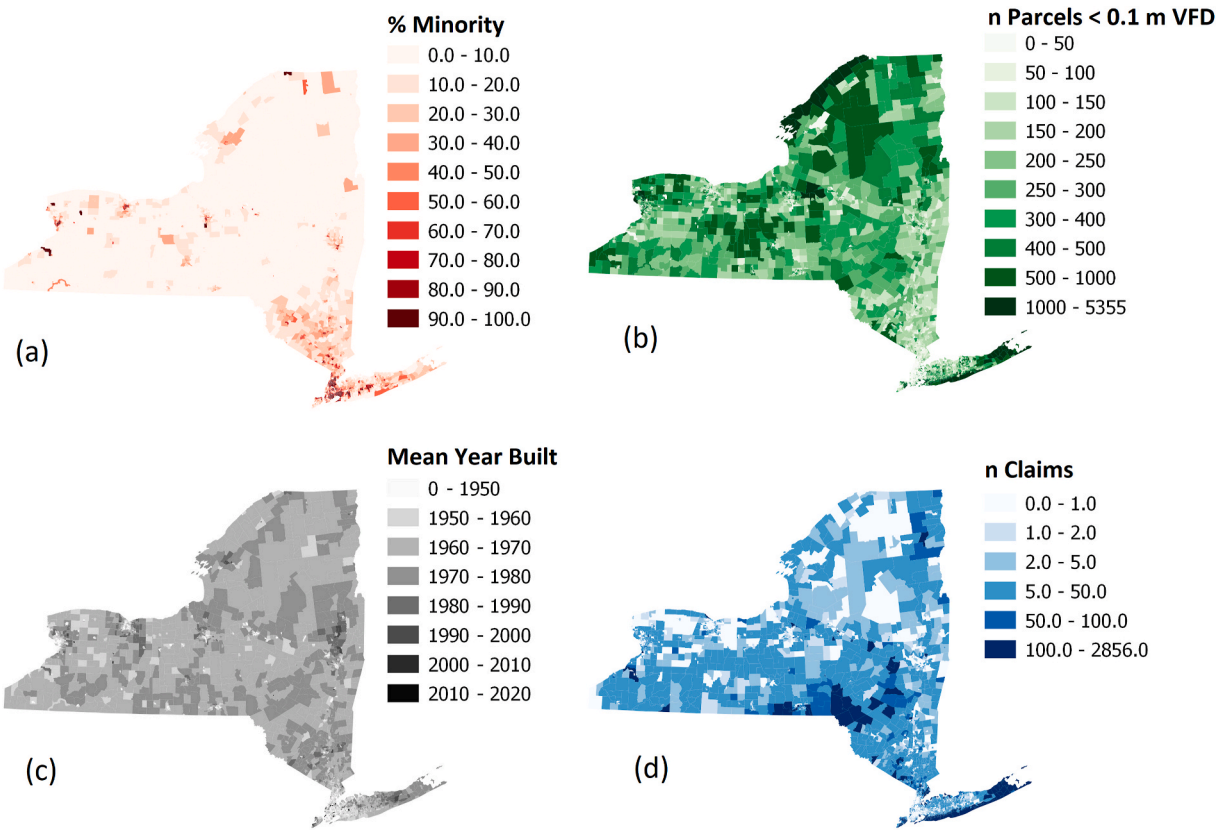

Fig. 1. New York State (NYS) census tracts showing a) percentage of minority residents, b) number of parcels less than 0.1 m elevation above nearest stream, c) mean year built of houses, and d) number of flooding insurance claims (1978–2018).

Hackensack are 484 km^2 and 80 km^2, respectively. Both catchments exist in a temperate climate region and receive about 130 cm of precipitation annually (Xie et al., 2010). The predominant soil class is Wethersfeld gravelly silt loam (saturated hydraulic conductivity [K_{SAT}] = 1.5–5 mm^1hr^{-1}, available water capacity [AWC] = 86 mm) (USDA, 2019). The catchments are composed of contrasting land uses, where the Moodna is 17% developed land, 75% forest and wetland, 8% pasture, and the Hackensack is 62% developed land, 33% forest, and 5% pasture (Fry et al., 2011). Land use change was relatively static from 2001 to 2016 within each catchment (USGS, 2016). Both catchments are dense with parcels designated as repetitive flood loss locations (Moodna = 74, Hackensack = 93). The CDC Social Vulnerability Index is a composite metric for environmental vulnerability. The SVI represents vulnerability across four subcategories (socioeconomic conditions, household composition, minority resident composition, and housing availability) on a scale of 0–1 (where 1 indicates the highest vulnerability) which are then aggregated to the composite SVI (Flanagan et al., 2011). The average SVI scores for Moodna and Hackensack were 0.21 (low) and 0.38 (moderate) respectively. Moodna and Hackensack residents had similar socioeconomic stability (both ~ 0.07), but Hackensack had a greater proportion of aging (0.74 vs 0.50) and minority residents (0.52 vs. 0.31) with less stable housing options (0.57 vs. 0.38) (Flanagan et al., 2011).

2.2. Parcel- and census tract-level flood insurance claim records

We collected available parcel-level NYS NFIP flood insurance claim records (3947 parcels) covering January 1, 1975 through December 31, 2018 (FEMA, 2019c). The location of parcel-level claims was determined by matching reported property addresses to tax parcel centroids (NYS, 2019a,2019b). Flood insurance claims used for analysis included 2946 repetitive loss properties (i.e. at least two claims exceeding 2000 USD) and 1001 non-repetitive loss properties.

We computed the rate parameter for flooding insurance claim generation at the parcel level, λ, as the number of reported claims per year (Fig. 2a). The starting date for the claim duration (over which λ was computed) was defined as the maximum of January 1, 1975 and the year the structure was built as defined in the NYS tax parcel database (NYS, 2019a,2019b). The ending date was defined as December 31, 2018, unless a flooding claim property was "mitigated" where the date of the last flood was used instead. We estimated the median λ for all claim-generating properties as 0.15 $claims^1year^{-1}$ (6.7 year claim return period). We compiled a dataset of non-claim generating parcels where we assumed the rate parameter for flood claim generation at the parcel level, λ, was 0. To do so, we randomly selected 1889 tax parcels that were at least 200 m from the location of all existing flooding claim properties (Fig. 1). This approach likely underestimates λ for some properties that did not experience hydrologic extremes within the study period (1975–2018), but would submit insurance claims during less frequent extreme events. In total, the full dataset of claim-generating and non-claim generating parcels totaled 5836 records.

We collected available tract-level NYS NFIP flood insurance claim records (4906 census tracts) covering January 1, 1975 through December 31, 2018 (FEMA, 2019c). The tract-level dataset documents 166,942 flood claims.

2.3. Hydrologic and social demographic data

Variables included in the parcel-level random forest regression and classification models are presented in Table 1 (all variables considered in model development are presented in Table S1). Following the results of Woznicki et al. (2019), Khosravi et al. (2018), and Chapi et al. (2017), among others, we include several metrics describing the topographic position of each cell (horizontal flow distance [HFD], vertical flow

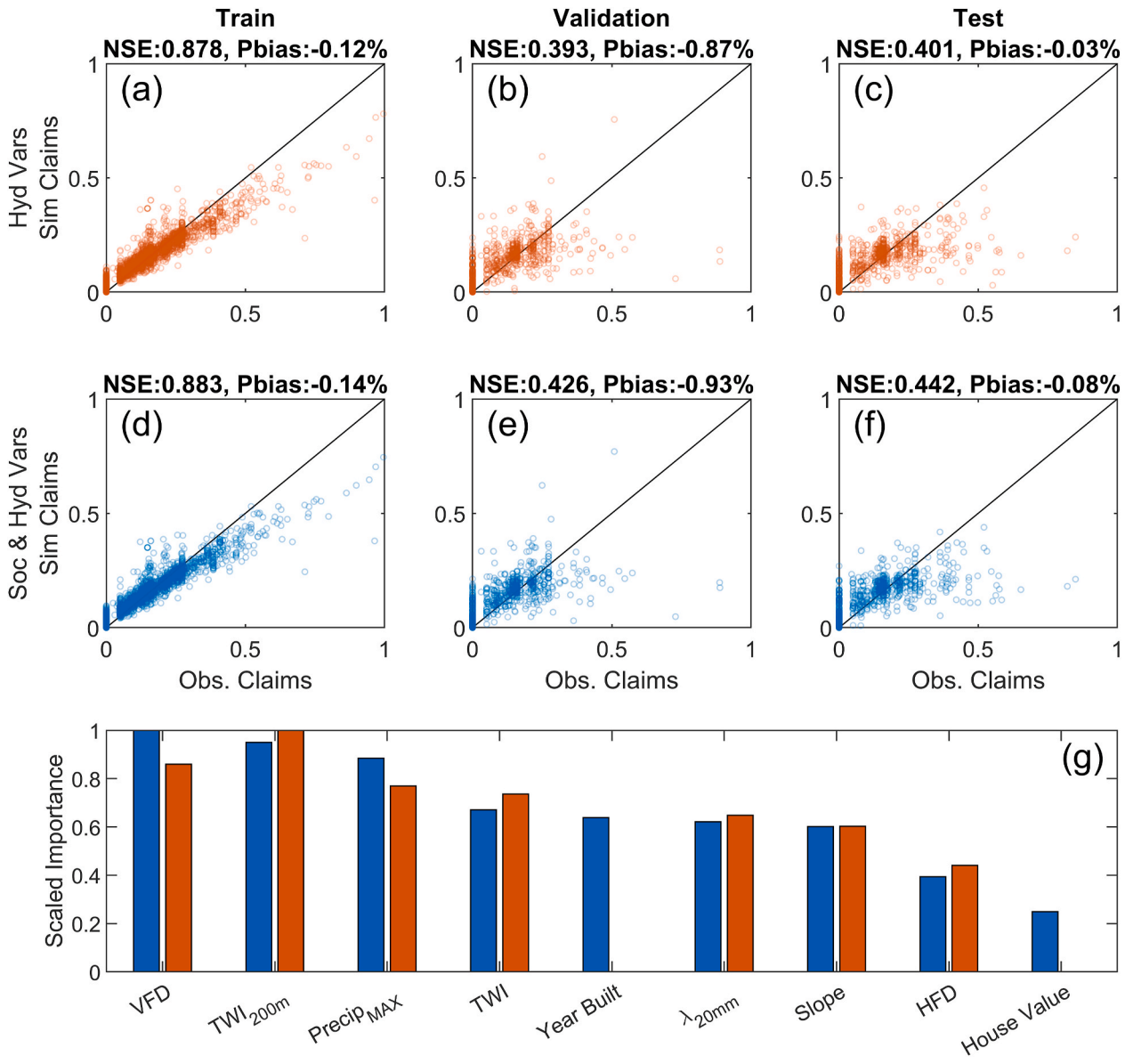

Fig. 2. Parcel-level random forest simulated versus observed flood claim frequency (λ), showing Nash Sutcliffe Efficiency (NSE) and Percent Bias (Pbias) of residuals for a-c) hydrologic variables and (d–f) hydrologic and socioeconomic variables, and g) variable importance scores (orange – hydrologic variables only; blue – hydrologic and social variables). Black lines indicate 1:1. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 1
Parcel-level hydrologic and socio-economic random forest predictor variable descriptions and references.

Layer	Type	Description	Primary Data
House Value	Social	Assessed house value	NYS (2019a, 2019b)
Year Built	Social	Year house was built	NYS (2019a, 2019b)
$Precip_{Max}$	Hydrologic	Maximum observed precipitation (1978–2018)	Xie et al. (2010)
λ_{20mm}	Hydrologic	Frequency of daily precip > 20 mm	Xie et al. (2010)
HFD	Hydrologic	Horizontal flow distance from nearest channel (m)	USGS (2019)
VFD	Hydrologic	Vertical flow distance from nearest channel (m)	USGS (2019)
TWI	Hydrologic	Topographic Wetness Index	USGS (2019)
TWI_{200m}	Hydrologic	Maximum TWI within 200 m	USGS (2019)
Slope	Hydrologic	Local land surface slope (m^1m^{-1})	USGS (2019)

distance [VFD], topographic wetness index [TWI; $TWI = ln\left(\frac{a}{tan(b)}\right)$, where a is upslope contributing area and b is the local land slope], maximum TWI within 200 m of property [TWI_{200m}], and slope; descriptions in Table 1) computed in the System for Automated Geoscientific Analyses (SAGA; Conrad et al., 2015) from a 20 m digital elevation model (USGS, 2019). Socioeconomic characteristics available at the parcel-level (assessed home value and year built) were collected from the NYS tax records.

Variables included in the tract-level random forest regression models are presented in Table 2 (all variables considered in model development are presented in Table S2). Tract-level socio-economic demographic data was collected from the USCB (2020) and CDC (2018), summarized in Table 2. Hydrologic variables were computed at a 20 m horizontal resolution and aggregated to the tract level. We considered the possibility that participation in the NFIP was related to political geography and shared political capital (Pigg et al., 2013; Emery and Flora, 2006). Within the US, adoption of revised floodplain maps was significantly

Table 2
Tract-level hydrologic and socio-economic random forest predictor variable descriptions and references.

Layer	Type	Description	Reference
%minority	Social	% minority residents within tract	CDC (2019)
Year $Built_{mean}$	Social	Mean year built of parcels within tract	USCB (2020)
House Value	Social	Mean value of parcels within tract	USCB (2020)
Similarity	Social	Shared political capital (i.e. similarity) determined from 2010 gov. election	Ansolabehere and Rodden (2011)
$Precip_{max}$	Hydrologic	Maximum observed precipitation (1978–2018)	Xie et al. (2010)
λ_{20mm}	Hydrologic	Frequency of daily precip > 20 mm	Xie et al. (2010)
TWI_{max}	Hydrologic	Maximum TWI within 200 m	USGS (2019)
VFD_{mean}	Hydrologic	Vertical flow distance from nearest channel (m)	USGS (2019)
HFD_{mean}	Hydrologic	Horizontal flow distance from nearest channel (m)	USGS (2019)
n $parcels_{<0.1m}$	Hydrologic	n parcels within tract < 0.1 m above nearest stream	USGS (2019)

correlated with county Democratic political lean (Wilson and Kousky, 2019), possibly indicating a greater likelihood of NFIP participation by groups with unified political capital or a sense of access to organizations, connection to resources, and power brokers (Emery and Flora, 2006). States affected by flooding seem to be seeking assistance from the federal government across the political spectrum (Flavelle, 2020), yet political segregation may create polarization that would result in attitudes within a state as differing from the governing dominant political party (Dottle, 2019). This then leads to either shared sense of cultural and political capital with the governance structure or a non-shared sense of cultural and political capital (Emery and Flora, 2006). While political capital could have the connotation of negative effects (Kostovetsky, 2015), here it is used to indicate the sense with which a populace seeks to engage with its state governance for positive effects (Emery and Flora, 2006). Based on this, we included a metric estimating a property's similarity/dissimilarity to its local state government (based on cultural and political capital) to capture differences in the rate of flood insurance claim submittals stemming from shared community beliefs (Emery and Flora, 2006). Similarity was estimated from the 2010 NYS gubernatorial voting results aggregated to voting precincts (Ansolabehere and Rodden, 2011). The similarity metric was computed as the sum of all Democratic votes divided by the sum of all Democratic and Republican votes within each precinct (independent candidate votes were removed from consideration) and then aggregated to tracts by area.

2.3.1. Random forest regression and classification analysis

We developed several random forest algorithms for prediction of λ and total claims, at the parcel- and tract-levels respectively, given a collection of common hydrologic indicators of flooding potential and social vulnerability. The training, validation, and testing splits were approximately 70%, 15%, 15%. All random forest computations were performed with the h2o package (Candel et al., 2016) in R version 3.6.1. RF training was performed to optimize Mean Square Error. For all regression models we present Nash-Sutcliffe Efficiency (NSE) and percent bias (Pbias), scale independent objective functions that allow for comparison across training, validation and testing datasets. Cross-fold validation was used with 10 folds to reduce the effects of over-fitting.

Inclusion of non-informative predictor variables can reduce RF performance. An iterative process was used to screen out predictor variables (Tables 1 and 2) that did not improve RF predictive skill: 1) RF model training was performed, 2) test data objective functions and variable importance scores were computed, 3) the lowest importance variable and those variables with ranked correlations ($\rho > 0.8$) to more predictive variables were progressively removed from the analysis until test objective functions stabilized. All variables evaluated for inclusion in random forest models are presented in Tables S1 and S2.

Following selection of the optimal set of variables, RF hyperparameters (number of trees [2–1000], maximum depth of individual trees [5–60]) were determined through a Monte Carlo sampling procedure where the RF model was fit 1000 times with randomly selected parameter values. Monte Carlo simulations suggested that objective function values modestly improved up to approximately 100 trees and a max depth of 40 splits in all models. We therefore used these values for all models.

Socioeconomic demographic data is more widely available at the tract-level than for individual parcels, though all datasets are aggregated to a coarser scale. Thus, four different RF regression models for predicting claims were developed:

- Regression prediction of parcel-level claims (λ) using only hydrologic predictors
- Regression prediction of parcel-level λ using hydrologic and socio-economic predictors
- Regression prediction of tract-level claims using only hydrologic predictors
- Regression prediction of tract-level claims using hydrologic and socio-economic predictors

We compared the capability of the classic methodology based on hydrologic and surface routing models to an alternative approach based on random forest classification to identify claim-generating properties within these study catchments.

- Classification of parcel-level λ using hydrologic and socio-economic predictors

For parcel-level regression analysis, the random forest was trained directly on λ defined continuously. For classification, λ was encoded as a binary response (0 – a property generated no claims within the date range of claims, 1 – a property generated at least one claim). Finally, we computed Spearman's ranked correlation (ρ) between several predictor variables and λ for all parcels and total claims at the tract level to examine correlations among predictors.

2.4. Flood hazard mapping via a physically based surface routing model

For both the Moodna and Hackensack catchments, we compared the random forest generated flood claim predictions to a methodology which closely followed that of US FEMA FIRMs (FEMA, 2016), as well as being similar to the approach used in several recent studies which established high-resolution flooding hazard products (Wing et al., 2018; Quinn et al., 2019). Though FIRMs exist within each catchment, they are defined only for a subset of streams. We computed maximum water surface elevations resulting from the 100-year return period design storm with the physically-based two-dimensional pluvial flood routing model LISFLOOD-FP (model development described in Bates et al., [2010], Neal et al., [2012], and de Almeida et al., [2013], among others).

LISFLOOD-FP (code release 5.9.6) models were developed at a 20 m horizontal grid resolution. The land surface was derived by coarsening the 10-m DEM (USGS, 2019; vertical error RMSE = 1.55 m) to a 20 m resolution. Both LISFLOOD-FP models were initialized at a 1-s computational time step, which was decreased dynamically to maintain numerical stability. A minimum depth threshold of 0.001 m was set for

hydraulic computations. Hydrologic surface losses were modeled with a constant loss rate of K_{SAT} as derived from SSURGO soils data (USDA NRCS, 2019), neglecting the influence of subsurface stormwater collection systems. Each LISFLOOD-FP model was forced with the 24-h 100-year return period precipitation depth applied to the SCS Type-2 hyetograph (NOAA, 2019).

LISFLOOD-FP simulations were validated against projected 100-year discharge rates (USGS, 2020) and through visual comparison of flooding extents defined in available FEMA FIRMs (FEMA, 2019b) for the Moodna and Hackesnsack catchments (model validation is discussed in more detail in supplemental Section S3).

3. Results

3.1. Prediction of parcel- and tract-level claims using random forest regression

The random forest regression analysis shows some ability to estimate λ when trained against all claims data across NYS using only hydrologic predictors (Fig. 2a,b,c; test NSE = 0.401). When NYS tax parcel derived socioeconomic predictors were included, the testing data set calibration improved slightly (Fig. 3f; test NSE = 0.442). Predictive variables include hydrologic (VFD, TWI_{200m}, $Precip_{MAX}$, λ_{20mm}, and HFD) and socioeconomic predictors (Year Built and House Value) (Fig. 2g). All other variables (Table 1) were less predictive and were removed from analysis.

At the tract level, incorporation of socio-economic predictors produced an NSE score (Fig. 3f; NSE = 0.625) substantially higher than that obtained using only hydrologic variables (Fig. 3c; NSE = 0.536). The

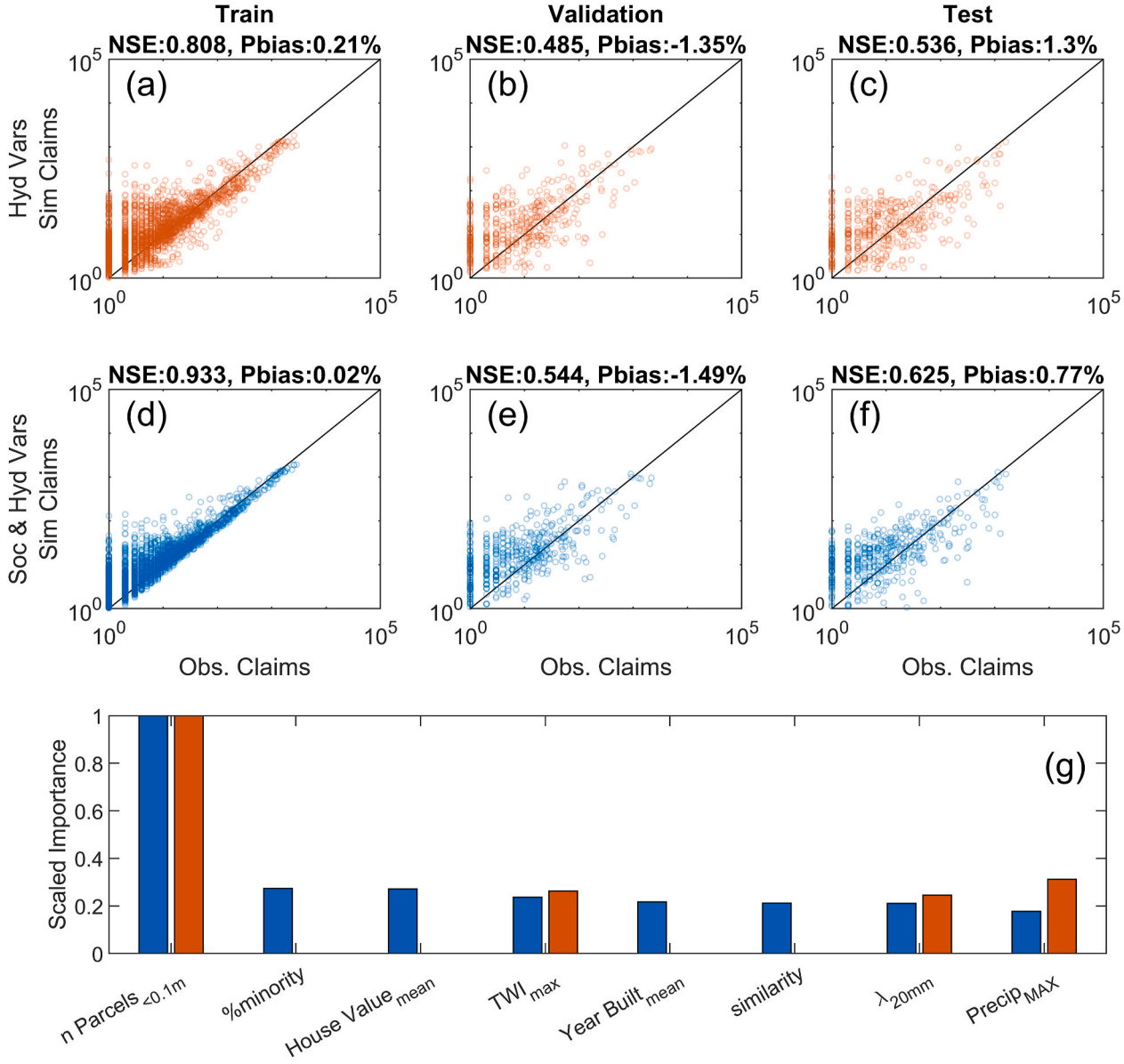

Fig. 3. Census tract-level random forest simulated versus observed flood claim frequency (λ), showing Nash Sutcliffe Efficiency (NSE) and Percent Bias (Pbias) of residuals for a-c) hydrologic variables and (d–f) hydrologic and socioeconomic variables, and g) variable importance scores (orange – hydrologic variables only; blue – hydrologic and social variables). Black lines indicate 1:1. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

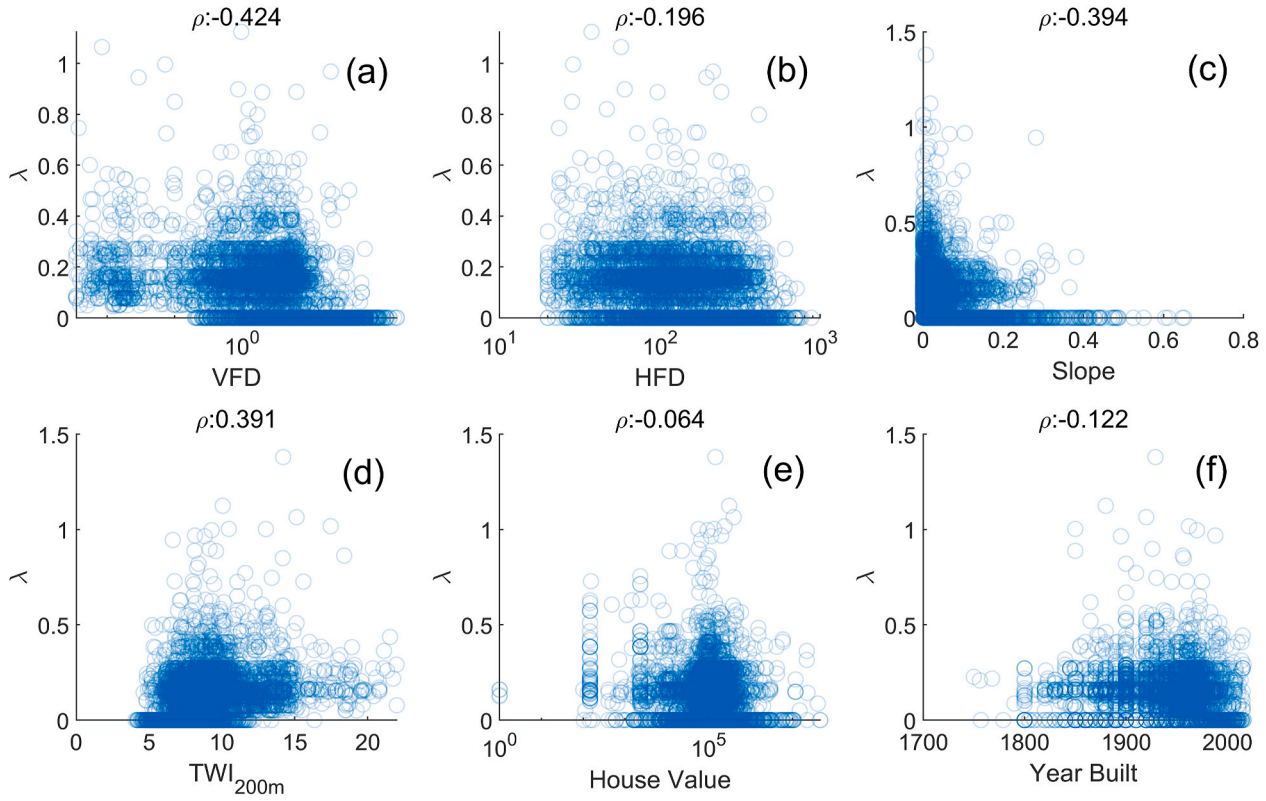

Fig. 4. Correlations between Parcel-level variables and λ. ρ indicates Spearman's ranked correlation values.

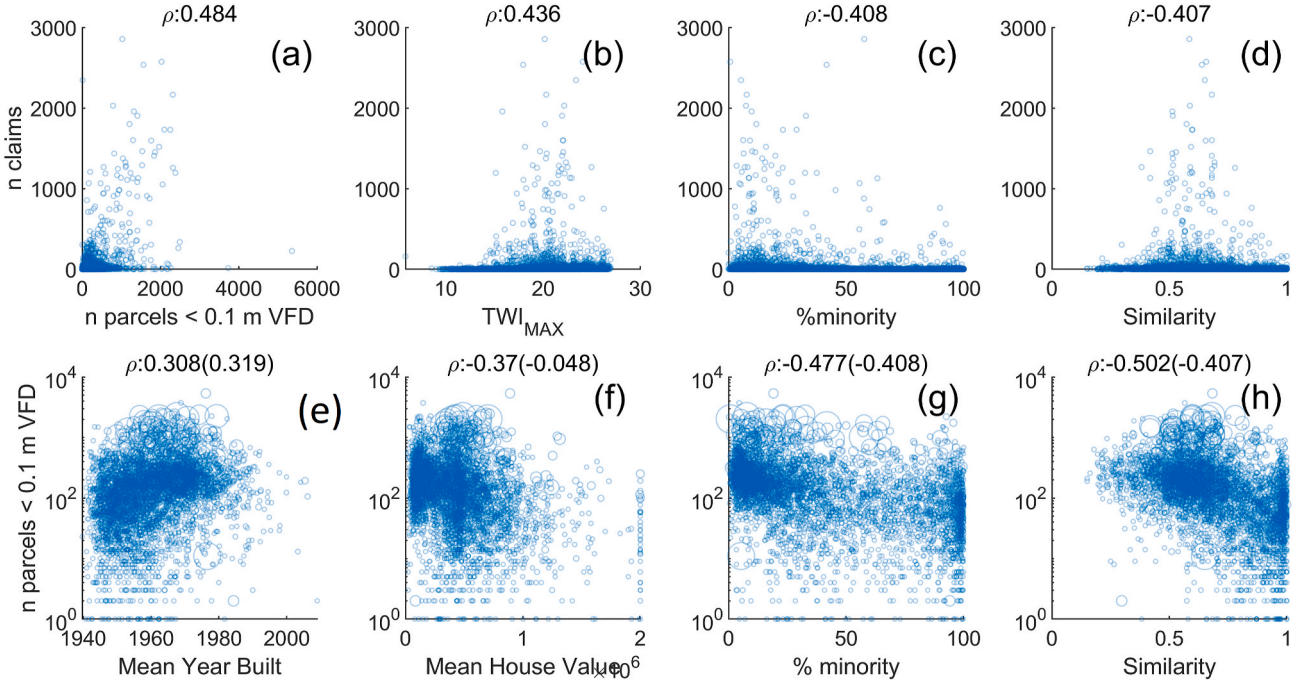

Fig. 5. Correlations between Parcel-level variables and claims (a–d) and the number of parcels within a tract within 0.1 m elevation of the nearest stream (e–f) (size of circles is proportional to n claims). ρ indicates Spearman's ranked correlation values between x and y variables. Values in parenthesis are correlation values between x variable and n claims.

number of low-lying parcels ($n\ parcels_{<0.1m}$) was the most predictive variable, followed by several hydrologic (TWI_{MAX}, $Precip_{MAX}$, λ_{20mm}) and socioeconomic variables (*%minority*, *house value*, *year built*, *similarity*) (Fig. 3g). As with the parcel-level analysis, all other variables (Table 2) were found to be less predictive and therefore removed from the analysis.

3.2. Prediction of parcel -level claims using random forest classification

The random forest model was trained to classify locations using all claims data and all hydrologic and social vulnerability predictors (Table 3). Performance of the random forest classification is presented in Table 2. The overall testing error rates (accuracy = 0.962; sensitivity = 0.966; specificity = 0.954; F1 = 0.972; AUC = 0.989; AUCPR = 0.995) indicated the model was slightly more likely to generate a false positive than false negative (i.e. over-estimation of claims).

3.3. Hydrologic and socioeconomic predictors of claims

We computed Spearman's ranked correlation (ρ) between several parcel-level predictor variables and λ (Fig. 6). Hydrologic variables VFD (ρ = −0.424), HFD (ρ = −0.196), and Slope (ρ = −0.394) all have negative rank correlation with λ, indicating the frequency of claims is higher in flat, low lying areas, adjacent to waterbodies, highlighting the importance of local hydrologic conditions. Similarly, TWI_{200m} (ρ = 0.391) is positively rank correlated with λ, indicating that parcels adjacent to areas of high flow accumulation tend to generate more claims. The socio-economic variables of house value (ρ = −0.064) and year built (−0.122) were weakly rank correlated with λ, possibly suggesting that older less expensive properties are more likely to generate flood claims.

Analysis at the tract-level indicates similarly important hydrologic variables to those of parcel-level analysis: $n\ parcels_{<0.1m}$ (ρ = 0.484), TWI_{MAX} (ρ = 0.409) (not shown are $Precip_{MAX}$ [ρ = 0.391], λ_{20mm} [ρ = 0.391]) (Fig. 5). The socio-economic variables %minority (ρ = −0.408) and similarity (ρ = −0.407) were both negatively correlated with the number of claims (Fig. 5c and d), indicating that flooding claims are generated from tracts with predominantly white populations and political views dissimilar from state level government. Social demographic conditions could align with the availability of housing options more or less exposed to hazards, therefore we present rank correlations between predictive socio-economic predictors and $n\ parcels_{<0.1m}$ (Fig. 5e–h). Correlations indicate that tracts with higher numbers of low-lying tax parcels tended to have low proportions of minority residents (ρ = −0.477), newer houses (ρ = 0.308), lower home values (ρ = −0.370), and higher political dissimilarity (ρ = −0.502).

3.4. Comparison of hydrologic modeling hazard and random forest classification risk predictions

The physically-based LISFLOOD-FP hazard approach to estimating the FHA (i.e. 1% annual exceedance inundated area) overlaps with 57% and 45% of historical insurance claims within Hackensack and Moodna, respectively (Fig. 6a, c). The random forest classification of flood generating claims captures 80.6% and 93.2% of claims within Hackensack and Moodna, respectively (Fig. 6b, d). Within the existing SFHA, both LISFLOOD-FP and the random forest classification agree well. Outside of the SFHA, the random forest algorithm possibly over-estimated flood claim-generating locations.

Table 3
Random forest parcel-level classification training, validation, and prediction error rates.

		Predicted Claim	Predicted No Claim	Error Rate
Training n = 4052	Observed Claim	2795	0	0.00%
	Observed No Claim	0	1257	0.00%
Validation n = 846	Observed Claim	554	23	3.99%
	Observed No Claim	26	243	9.67%
Testing n = 865	Observed Claim	565	20	3.42%
	Observed No Claim	13	267	4.64%

4. Discussion

4.1. Data-driven flood claim model development

Flood insurance claim records can serve as a useful, but a possibly imperfect, proxy for flooding hazards and exposure. Similar to the conclusions of previous studies focusing on prediction of the inundated extent (e.g. Khosravi et al., 2019; Woznicki et al., 2019), topography alone was a strong predictor of claim frequency at the parcel- (Fig. 2g) and tract-levels (Fig. 3g). All regression models exhibited a bias towards over-estimating the frequency of claims from properties where there were no observed claims (Fig. 2a–f and 3a-f), possibly due to household-level heterogeneity in NFIP participation or differences in the selected level of coverage (Royal and Walls, 2019; Kousky and Michel-Kerjan, 2017). Uninsured homes outside of the SFHA may still submit flood claims to receive assistance in the form of the Individual and Households Program (IHP), though this support is more limited than economic relief provided to insured properties. Underreporting of flooding losses outside of the SFHA could occur because of a lack of insurance and federal support provided to these areas. Further, individual perceptions of risk or shared political capital (Pigg et al., 2013) may lead to decisions to decline optional flooding insurance (Royal and Walls, 2019), which may result in a weaker understanding of NFIP in communities located beyond the SFHA. It is possible that RF residuals reflect decisions by some homeowners to purchase tail-loss coverage, which only covers infrequent extreme loss events (Kousky and Michel-Kerjan, 2017) rather than coverage for smaller more frequent events. Application of data-driven techniques well suited to handling zero-inflated datasets (e.g. Savage et al., 2015) could help to remedy this issue.

The exposure-focused model presented here highlights the importance of dimensions beyond hydrometeorological land surface responses that influence the frequency of insurance claim generation. In a study of urban regions in Iran, Darabi et al. (2019) demonstrate that accounting for population density and building quality improved flooding risk estimates. Similarly, Li et al. (2019) show the importance of incorporating land use information (i.e. cultivated lands), economic development areas, and population for translating hazards into risks. Metin et al. (2018) proposed that changes to flooding vulnerabilities (changes in land use, asset values, and the role of "precaution") can potentially outweigh external changes such as climate change causing thermodynamic and dynamic shifts in extreme rainfall delivery mechanisms. Our results demonstrated that local topography was generally more predictive of exposure than demographic information in NYS (Figs. 2g & 3g), though incorporation of socio-economic data at the tract level substantially improved RF claim prediction skill (Fig. 2f & f).

4.2. Population demographics of flooding insurance claims in NYS

Parcel-level analysis demonstrated that inclusion of socio-economic data can improve the prediction of the flood claim frequency of individual parcels (Fig. 2c, f), though available demographic data was limited at this spatial scale. Analysis with data aggregated to the tract-level that included a broader suite of social demographic predictors (housing stock age and value, the proportion of minority residents, and community capital similarity) resulted in a larger improvement in prediction skill (Fig. 3c, f). Census tracts with higher numbers of claims tended to have low proportions of minority residents, newer houses, and less political similarity to state level government.

Mean house value was negatively correlated with the number of flood-exposed parcels within a tract (Fig. 5f). Depressed home values in low-lying areas could be the result changing market perceptions of risk

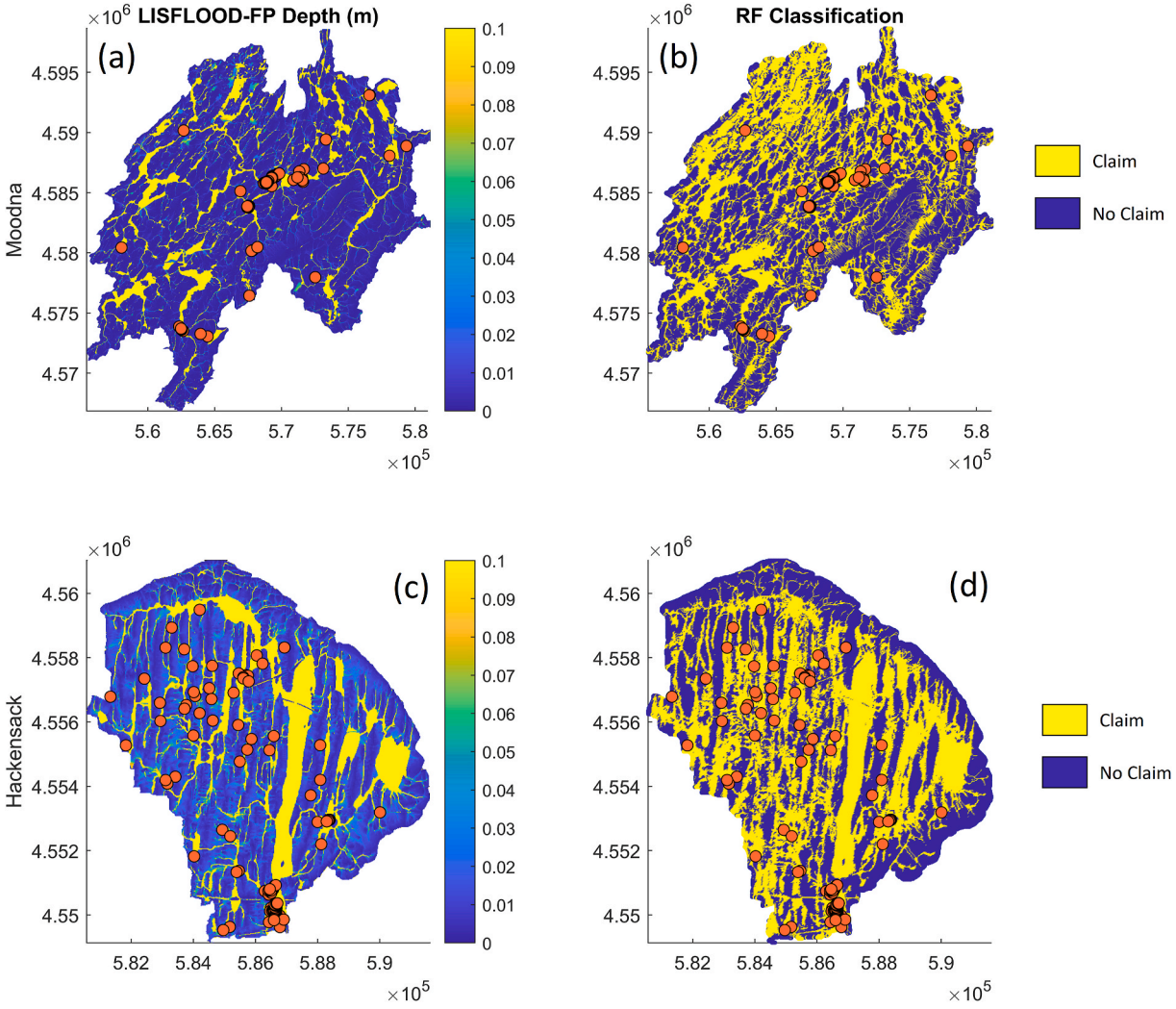

Fig. 6. Inundation maps derived from LISFLOOD-FP and random forest classification of claim generating areas for the Moodna and Hackensack catchments. Historical claim locations are shown as orange circles. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

and subsequent reductions of at-risk house values (Siders et al., 2019). Decreases in home value could also be related to NFIP reforms (Biggert-Waters Act of 2012 and Grimm-Waters Act of 2013), which placed a greater proportion of the economic burden of insurance on at-risk properties, inducing a gradual loss of floodplain property market value (Indaco et al., 2019). The positive correlation between mean home age, exposure, and number of claims (Fig. 5e) may indicate survivor bias of buyout programs. Conversely, this could also represent new development within floodplains. Despite the goal of NFIP to encourage depopulation of floodplains, federal subsidies for insurance may be reducing the economic incentive to migrate from at-risk areas (King, 2009; Michel-Kerjan, 2010). Parcel-level analysis also identified house age and value as predictors of flood claim frequency, but indicated opposite correlations where individual properties with older houses generated more frequent claims (Fig. 4f). These relationships, visible in the parcel-scale data, possibly indicate that the demographics of repetitive-loss properties differ from those aggregated to the tract.

Census tracts with higher proportions of minority residents generated insurance claims at lower rates within NYS (Fig. 5c). We also observed a clear negative correlation between the proportion of minority residents within a tract and the number of flood-exposed parcels [$n\ parcels_{<0.1m}$] (Fig. 5g), possibly indicating that race identifies differences in exposure rather than differences in the willingness or ability to participate in NFIP across NYS. Our results are similar to those of Elliott (2015) and Cutter et al. (2018) who found that flood-exposed properties were predominantly white, as minorities with limited economic means for recovery were more likely to relocate following a disaster. Hale et al. (2018) surveyed residents living within floodplains in the Wasatch Front, Utah US, and found that the population was predominantly white, and that racial minority residents were disproportionately economically impacted by floods. Our observations might also reflect a history of racial segregation across NYS where minority populations become clustered (Besbris and Faber, 2017) within locations that also happen to be at low risk of flooding.

Previous studies have found that flooding hazards disproportionately affect lower income communities in the US, who may also struggle to participate in the NFIP (FEMA, 2018; Dixon et al., 2017), likely imparting some demographic bias on flooding insurance claims records. Census tract per capita income was not predictive of the rate of claim generation, nor was income alone predictive of topographic exposure to hazards, despite positive correlation between the proportion of minority residents and social vulnerability.

Shared political capital may be another driving motivation for flood claim submissions. Similar to race, we found similarity to be negatively correlated with both claims and the number of flood-exposed parcels (Fig. 5h). Strother (2018) argues that NFIP, a previously apolitical

program, became somewhat politicized after the passage of the Biggert-Waters Act of 2012 which proposed economic reform of NFIP through rate increases. Biggest-Waters, as well as several largely publicized flooding events impacting densely populated regions (e.g. Hurricanes Sandy [2012] and Harvey [2017]), likely increased the public awareness of the NFIP program and made US politicians sensitive to the potential electoral implications of NFIP reform (Strother, 2019). The Grimm-Waters Act of 2013, which largely preserved NFIP subsidies, was strongly supported by Democratic Senators (97% in favor) but only by some Republicans (53% in favor). In contrast, we found that claims were generated with greater frequency by less Democratic leaning census tracts across NYS (Fig. 5h), possibly suggesting limited influence of shared political capital. Rather, the slightly lower predictive skill of similarity (Fig. 3g) may point to an issue of multi-collinearity, reflecting correlation with the proportion of minority residents ($\rho = 0.72$).

4.3. Re-conceptualizing FEMA FIRMs and the SFHA with social demographic data

In the US, FEMA Flood Insurance Rate Maps (FIRMs) delineate floodplains at specific frequencies of inundation to define insurance risk zones (FEMA, 2019c). The 1% annual exceedance probability inundation area (i.e. 100-year floodplain) is designated as the SFHA which demarcates the boundary within which: 1) the purchase of flooding insurance is mandatory for properties with mortgages from federally backed or insured lenders, and 2) flooding insurance premium rates are increased. Identification of high-risk properties is critical for sustainable implementation of insurance programs to hedge against flooding losses, plan future development, and to limit public exposure to flooding hazards (FEMA, 2019b). FIRMs are developed through the NFIP, a collaboration between local and federal agencies. Modern FIRMs are derived from simulations of surface runoff from a calibrated hydrologic model which is routed over the land surface and stream channel via a one- or two-dimensional hydraulic model for the prediction of maximum riverine water surface elevations and the inundated extent. Despite very specific FIRMs mapping guidance (FEMA, 2016), much of the CONUS remains unmapped, or is mapped with outdated hydrologic techniques (Kousky, 2018).

The FIRMs methodology distills riverine hazard down to one loss mechanism (i.e. inundation below the static water surface elevation defining the floodplain under a specific discharge frequency). This approach neglects that riverine flood losses may be related to other hazards such as: high overland flow velocities on steep slopes inducing erosion, deposition of suspended material on low slopes, or localized intense precipitation overwhelming natural and built water conveyance systems (e.g. roadside ditches, gutters) far from receiving waters (e.g. Knighton et al., 2018; Merz et al., 2010). In addition, hydrologic models may be developed around improper assumptions that limit their utility as unbiased predictors of flooding hazards including the misrepresentation of the dominant modes of surface runoff (Buchanan et al., 2018), overly simplified representations of vegetation (Knighton et al., 2019b; Hwang et al., 2018), or inadequate parameterizations related to model calibration challenges (e.g. Schoups and Vrugt, 2010).

Hydrologic modeling leverages our prior knowledge of physical processes that generate peak discharge and land inundation (e.g. runoff-infiltration partitioning, land surface gradient-based flood routing, surface depression storage, backwater effects from infrastructure, interactions with stormwater collection systems). Hydrologic models can provide a well-supported lower bound on the area within which properties are likely to generate flood claims through accurate delineation of flood plains (Fig. 6a, c). Data driven machine learning approaches, employed here as a random forest classification, can possibly uncover more complex relationships between hazards and risks that are not explicitly simulated in hydrologic models. Further, we have demonstrated that machine-learning techniques can readily incorporate non-traditional hydrologic datasets, such as social demographics, that could provide a more nuanced view of flooding and loss.

4.4. Methodology limitations and opportunities

There are several aspects of our methodology and the underlying datasets which potentially limited our results and subsequent discussion. Here we present a review of these aspects of our methodology and discuss possible impacts on our analysis and opportunities for future research.

First, our estimate of the durations over which λ was computed introduced some uncertainty that is difficult to quantify. We assumed that the date of building construction from the tax parcel database was reliable. Within NYS, clerical errors in the tax parcel data are acknowledged and handled through a formal process (NYS, 2019a, 2019b). Many tax records contained null values, necessitating the assumption that the year these properties were built preceded that starting date of claim collection. For mitigated properties, it was assumed that the date of the final reported flood was a reasonable ending date for computation of λ. Mitigation should indicate a decrease in either the flood hazards and/or risks posed. As demonstrated by Kousky and Michel-Kerjan (2017), mitigated properties show a reduced frequency of claim generation; however, they do occasionally generate claims, which would cause us to underestimate the duration of exposure (and overestimate λ) for mitigated properties (4% of all claim generating properties). More refined estimates of the exposure duration could refine estimates of λ.

Second, differences in local collection and storage of NFIP claims data can lead to errors and underreporting of economic losses (Gall, 2017). Structural issues, such as spatial discrepancies in the prices of insurance (Dixon et al., 2017; Royal and Walls, 2019) and misinterpretations of risk information (Bell and Tobin, 2007) likely lead to underreporting of flooding losses through NFIP. Estimates of near-stream flood hazard areas are often readily available within CONUS, though there is no such centralized database of low flooding risk locations. Our approach to providing information on non-claim generating locations assumed that properties randomly selected within NYS experienced no flooding if no insurance claim had been submitted within the study period. Development of a spatial database of known low-risk locations should be a priority to relax our reliance on the assumptions underpinning existing FIRMs and to prevent future methodologies from becoming over-conditioned by the data available on flooded properties.

Third, we assumed the random forest inputs (Tables S1 and S2) were reliable predictors of hazards and risks. The RF model adequately captures flooding claims as a binary classification, but possibly additional predictors, such as spatially distributed riverine discharge return periods, could improve estimates of λ and tract-level claims. As our hydrologic predictors were chosen following other studies which successfully reproduced riverine flooding hazard frequency (e.g Woznicki et al., 2019), we assume this was not a major limitation. Analysis incorporating several demographic predictors improved the model performance, suggesting future studies aiming to improve flood loss estimates should consider more refined social vulnerability data. A large proportion of flooding claims are generated by pre-FIRM properties (Kousky and Michel-Kerjan, 2017). Prediction of λ could be improved by examining the sequence of existing FIRM map development and building construction. Prediction might also be improved by including information of active NFIP policies, though this data is presently available only for the past decade (FEMA, 2019).

Finally, we assumed that the duration over which claims were generated (43 years) was adequate to properly capture the stochastic nature of flood hazard mechanisms and insurance claim generation. The frequency of flooding from intense landfalling tropical storms for the NYS region decreases from approximately 0.5 $year^{-1}$ along the Atlantic coast (Czajkowski et al., 2017) to 0.05 $year^{-1}$ in Central New York (Knighton et al., 2017). The low frequency of the dominant extreme

rainfall delivery mechanism possibility introduced some uncertainty into the estimates of λ and claim totals for census tracts related to the period of record.

5. Conclusions

Existing methodologies for identifying properties at risk of experiencing flooding losses frequently center on well-defined hydrologic hazards with less emphasis on defining exposure and vulnerability to these hazards. We developed random forest regression models to predict the historical rates of parcel- and census tract-level flooding insurance claim submittals across New York State (NYS) US with both hydrologic and socioeconomic predictors. The frequency of flooding claims was best predicted by a combination of hydrologic (vertical distance to the nearest stream, topographic wetness index) and social demographic (percentage of minority residents, housing stock age and value, capital dissimilarity) predictors.

Census tracts with higher numbers of claims and greater densities of low-lying tax parcels tended to have low proportions of minority residents, newer houses, and less political similarity to state level government. Socioeconomic demographic variables correlated with the number of low-lying parcels in census tracts across NYS, suggesting demographic data may be predictive of exposure rather than a willingness or ability to participate in NFIP.

Our research broadly supports the concept that quantitative incorporation of socio-economic data can produce refined estimates of flooding risks. Historical flooding insurance claim records in NYS appear to be reliable datasets that should be further analyzed to understand hydrologic and social variations in flooding claim submittals. Future research should investigate higher resolution demographic information to refine flooding risk estimates and to better understand the pathways by which communities are vulnerable to flooding.

Declaration of competing interest

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

CRediT authorship contribution statement

James Knighton: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing. **Brian Buchanan:** Conceptualization, Methodology, Software, Investigation, Writing - review & editing, Supervision. **Christian Guzman:** Writing - review & editing. **Rebecca Elliott:** Methodology, Writing - review & editing. **Eric White:** Methodology, Writing - review & editing. **Brian Rahm:** Conceptualization, Supervision, Project administration, Funding acquisition.

Acknowledgements

This research was prepared for the New York State Water Resources Institute (WRI) and the Hudson River Estuary Program of the New York State Department of Environmental Conservation, with support from the New York State Environmental Protection Fund. This work was supported [in part] by the National Socio-Environmental Synthesis Center (SESYNC) under funding received from the National Science Foundation DBI-1639145.

References

Aerts, J.C., Botzen, W.J., Clarke, K.C., Cutter, S.L., Hall, J.W., Merz, B., et al., 2018. Integrating human behaviour dynamics into flood disaster risk assessment. Nat. Clim. Change 8 (3), 193.

Ahmadlou, M., Karimi, M., Alizadeh, S., Shirzadi, A., Parvinnejhad, D., Shahabi, H., Panahi, M., 2018. Flood susceptibility assessment using integration of adaptive network-based fuzzy inference system (ANFIS) and biogeography-based optimization (BBO) and BAT algorithms (BA). Geocarto Int. 1–21.

Ansolabehere, S., Rodden, J., 2011. New York data files. Harvard database. https://doi.org/10.7910/DVN/AWE39N, V1.

Atreya, A., Ferreira, S., Michel-Kerjan, E., 2015. What drives households to buy flood insurance? New evidence from Georgia. Ecol. Econ. 117, 153–161.

Bates, P.D., Horritt, M.S., Fewtrell, T.J., 2010. A simple inertial formulation of the shallow water equations for efficient two-dimensional flood inundation modelling. J. Hydrol. 387 (1–2), 33–45.

Bell, H.M., Tobin, G.A., 2007. Efficient and effective? The 100-year flood in the communication and perception of flood risk. Environ. Hazards 7 (4), 302–311.

Besbris, M., Faber, J.W., 2017, December. Investigating the relationship between real estate agents, segregation, and house prices: Steering and upselling in New York State. Sociological Forum 32 (4), 850–873.

Blaikie, P., 2008. Epilogue: towards a future for political ecology that works. Geoforum 39 (2), 765–772.

Blessing, R., Sebastian, A., Brody, S.D., 2017. Flood risk delineation in the United States: how much loss are we capturing?. Natural Hazards Review, 18(3), 04017002.

Bolin, B., Kurtz, L.C., 2018. Race, class, ethnicity, and disaster vulnerability. In: Handbook of Disaster Research. Springer, Cham, pp. 181–203.

Boyce, J., Wright, B., Bullard, R., Pastor, M., Fothergill, A., Morello-Frosch, R., 2006. In the Wake of the Storm: Environment, Disaster and Race after Katrina. Russell Sage Found/Soc. Sci. Res. Counc, New York.

Buchanan, B., Auerbach, D.A., Knighton, J., Evensen, D., Fuka, D.R., Easton, Z., et al., 2018. Estimating dominant runoff modes across the conterminous United States. Hydrol. Process. 32 (26), 3881–3890.

Bui, D.T., Hoang, N.D., Pham, T.D., Ngo, P.T.T., Hoa, P.V., Minh, N.Q., et al., 2019. A new intelligence approach based on GIS-based Multivariate Adaptive Regression Splines and metaheuristic optimization for predicting flash flood susceptible areas at high-frequency tropical typhoon area. J. Hydrol. 575, 314–326.

Burby, R.J., 2001. Flood insurance and floodplain management: the US experience. Global Environ. Change B Environ. Hazards 3 (3), 111–122.

Burton, C., Cutter, S.L., 2008. Levee failures and social vulnerability in the Sacramento-San Joaquin Delta area, California. Nat. Hazards Rev. 9 (3), 136–149.

Candel, A., Parmar, V., LeDell, E., Arora, A., 2016. Deep learning with H2O. H2O. ai Inc.

Centers for Disease Control and Prevention/Agency for Toxic Substances and Disease Registry/Geospatial Research, Analysis, and Services Program (CDC), 2016. Social vulnerability index 2016 database New York. data-and-tools-download.html. (Accessed 20 August 2019).

Chapi, K., Singh, V.P., Shirzadi, A., Shahabi, H., Bui, D.T., Pham, B.T., Khosravi, K., 2017. A novel hybrid artificial intelligence approach for flood susceptibility assessment. Environ. Model. Software 95, 229–245.

Chen, W., Hong, H., Li, S., Shahabi, H., Wang, Y., Wang, X., Ahmad, B.B., 2019. Flood susceptibility modelling using novel hybrid approach of reduced-error pruning trees with bagging and random subspace ensembles. Journal of Hydrology.

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., Böhner, J., 2015. System for automated geoscientific analyses (SAGA) v. 2.1.4. Geosci. Model Dev. (GMD) 8, 1991–2007. https://doi.org/10.5194/gmd-8-1991-2015. Download.

Cutter, et al., 2018. Flash flood risk and the paradox of urban development. Nat. Hazards Rev. 19 (1) https://doi.org/10.1061/(ASCE)NH.1527-6996.0000268.

Czajkowski, J., Villarini, G., Montgomery, M., Michel-Kerjan, E., Goska, R., 2017. Assessing current and future freshwater flood risk from North Atlantic tropical cyclones via insurance claims. Sci. Rep. 7, 41609.

Darabi, H., Choubin, B., Rahmati, O., Haghighi, A.T., Pradhan, B., Kløve, B., 2019. Urban flood risk mapping using the GARP and QUEST models: a comparative study of machine learning techniques. J. Hydrol. 569, 142–154.

de Almeida, G.A., Bates, P., 2013. Applicability of the local inertial approximation of the shallow water equations to flood modeling. Water Resour. Res. 49 (8), 4833–4844.

Di Baldassarre, G., Nohrstedt, D., Mård, J., Burchardt, S., Albin, C., Bondesson, S., et al., 2018. An integrative research framework to unravel the interplay of natural hazards and vulnerabilities. Earth's Future 6 (3), 305–310.

Di Baldassarre, G., Sivapalan, M., Rusca, M., Cudennec, C., Garcia, M., Kreibich, H., et al., 2019. Socio-hydrology: scientific challenges in addressing a societal grand challenge. Water Resour. Res.

Dixon, L., Clancy, N., Miller, B., Hoegberg, S., Lewis, M.M., Bender, B., et al., 2017. The Cost and Affordability of Flood Insurance in New York City. RAND Corporation, Santa Monica, CA.

Donner, W., Rodríguez, H., 2008. Population composition, migration and inequality: the influence of demographic changes on disaster risk and vulnerability. Soc. Forces 87 (2), 1089–1114.

Dottle, R., 2019. Where Democrats And Republicans Live In Your City. FiveThirtyEight. https://projects.fivethirtyeight.com/republicans-democrats-cities/.

Dottori, F., Salamon, P., Bianchi, A., Alfieri, L., Hirpa, F.A., Feyen, L., 2016. Development and evaluation of a framework for global flood hazard mapping. Adv. Water Resour. 94, 87–102.

Edelenbos, J., Van Buuren, A., Roth, D., Winnubst, M., 2017. Stakeholder initiatives in flood risk management: exploring the role and impact of bottom-up initiatives in three 'Room for the River'projects in The Netherlands. J. Environ. Plann. Manag. 60 (1), 47–66.

Elliott, J.R., 2015. Natural hazards and residential mobility: General patterns and racially unequal outcomes in the United States. Soc. Forces 93 (4), 1723–1747.

Elliott, R., 2018. The sociology of climate change as a sociology of loss. Eur. J. Sociol./Archives Européennes de Sociologie 59 (3), 301–337.

Elliott, R., 2019. Scarier than another storm': values at risk in the mapping and insuring of US floodplains. Br. J. Sociol. 70 (3), 1067–1090.

Emery, M., Flora, C., 2006. Spiraling-up: mapping community transformation with community capitals framework. Community Dev. 37 (1), 19–35. https://doi.org/10.1080/15575330609490152.
Federal Emergency Management Agency (FEMA), 2019a. FEMA flood map service center. Available Online. https://msc.fema.gov/portal/home. (Accessed 8 July 2019).
FEMA, 2016. Guidance for flood risk analysis and mapping: general hydraulic considerations. Guidance document, 52 Available Online. https://www.fema.gov/media-library-data/1559218975056-a8d9146102a59f74c1442af50644e30a/General_Hydraulics_Guidance_Nov_2016.pdf. (Accessed 8 July 2019).
FEMA, 2018. An affordability framework for the national flood insurance program. Available Online. https://www.fema.gov/media-library/assets/documents/163171. (Accessed 8 July 2019).
FEMA, 2019b. The national flood insurance program (NFIP). Available Online. https://www.fema.gov/national-flood-insurance-program. (Accessed 8 July 2019).
FEMA, 2019c. Policy & claims statistics for flood insurance: claim information by state (1978 – current month). Available Online. https://www.fema.gov/policy-claim-statistics-flood-insurance. (Accessed 8 July 2019).
Flanagan, B.E., Gregory, E.W., Hallisey, E.J., Heitgerd, J.L., Lewis, B., 2011. A social vulnerability index for disaster management. J. Homel. Secur. Emerg. Manag. 8 (1).
Flavelle, C., 2020. Conservative States Seek Billions to Brace for Disaster. (Just Don't Call it Climate Change.). New York Times. https://www.nytimes.com/2020/01/20/climate/climate-change-funding-states.html.
Frei, A., Kunkel, K.E., Matonse, A., 2015. The seasonal nature of extreme hydrological events in the northeastern United States. J. Hydrometeorol. 16 (5), 2065–2085.
Fry, J., Xian, G., Jin, S., Dewitz, J., Homer, C., Yang, L., Barnes, C., Herold, N., Wickham, J., 2011. Completion of the 2006 national land cover database for the conterminous United States. Photogramm. Eng. Rem. Sens. 77 (9), 858–864.
Gall, M., 2017. Direct and insured flood damage in the United States. Flood Damage Survey and Assessment: New Insights from Research and Practice 228, 53.
Giovannettone, J., Copenhaver, T., Burns, M., Choquette, S., 2018. A statistical approach to mapping flood susceptibility in the Lower Connecticut River Valley Region. Water Resour. Res. 54 (10), 7603–7618.
Hale, R.L., Flint, C.G., Jackson-Smith, D., Endter-Wada, J., 2018. Social dimensions of urban flood experience, exposure, and concern. JAWRA J. Am. Water Resour. Assoc. 54 (5), 1137–1150.
Highfield, W.E., Norman, S.A., Brody, S.D., 2013. Examining the 100-year floodplain as a metric of risk, loss, and household adjustment. Risk Anal.: Int. J. 33 (2), 186–191.
Hirabayashi, Y., Mahendran, R., Koirala, S., Konoshima, L., Yamazaki, D., Watanabe, S., et al., 2013. Global flood risk under climate change. Nat. Clim. Change 3 (9), 816.
Hong, H., Tsangaratos, P., Ilia, I., Liu, J., Zhu, A.X., Chen, W., 2018a. Application of fuzzy weight of evidence and data mining techniques in construction of flood susceptibility map of Poyang County, China. Sci. Total Environ. 625, 575–588.
Hong, H., Panahi, M., Shirzadi, A., Ma, T., Liu, J., Zhu, A.X., et al., 2018b. Flood susceptibility assessment in Hengfeng area coupling adaptive neuro-fuzzy inference system with genetic algorithm and differential evolution. Sci. Total Environ. 621, 1124–1141.
Huang, H., Winter, J.M., Osterberg, E.C., 2018. Mechanisms of abrupt extreme precipitation change over the Northeastern United States. J. Geophys. Res.: Atmosphere 123 (14), 7179–7192.
Hwang, T., Martin, K.L., Vose, J.M., Wear, D., Miles, B., Kim, Y., Band, L.E., 2018. Nonstationary hydrologic behavior in forested watersheds is mediated by climate-induced changes in growing season length and subsequent vegetation growth. Water Resour. Res. 54 (8), 5359–5375.
Indaco, A., Ortega, F., Taşpınar, S., 2019. The effects of flood insurance on housing markets. Cityscape 21 (2), 129–156.
Khosravi, K., Pham, B.T., Chapi, K., Shirzadi, A., Shahabi, H., Revhaug, I., et al., 2018. A comparative assessment of decision trees algorithms for flash flood susceptibility modeling at Haraz watershed, northern Iran. Sci. Total Environ. 627, 744–755.
Khosravi, K., Shahabi, H., Pham, B.T., Adamowski, J., Shirzadi, A., Pradhan, B., et al., 2019. A comparative assessment of flood susceptibility modeling using Multi-Criteria Decision-Making Analysis and Machine Learning Methods. J. Hydrol. 573, 311–323.
Knighton, J., Steinschneider, S., Walter, M.T., 2017. A vulnerability-based, bottom-up assessment of future riverine flood risk using a modified peaks-over-threshold approach and a physically based hydrologic model. Water Resour. Res. 53 (12), 10043–1.
Knighton, J.O., Tsuda, O., Elliott, R., Walter, M.T., 2018. Challenges to implementing bottom-up flood risk decision analysis frameworks: how strong are social networks of flooding professionals? Hydrol. Earth Syst. Sci. 22 (11), 5657–5673.
Knighton, J., Pleiss, G., Carter, E., Lyon, S., Walter, M.T., Steinschneider, S., 2019a. Potential predictability of regional precipitation and discharge extremes using synoptic-scale climate information via machine learning: an evaluation for the eastern continental United States. J. Hydrometeorol. 20 (5), 883–900.
King, R.O., 2009. National Flood Insurance Program: Background, Challenges, and Financial Status. Congressional Research Service. Library of Congress.
Knighton, J., Conneely, J., Walter, M.T., 2019b. Possible increases in flood frequency due to the loss of Eastern Hemlock in the northeastern US: observational insights and predicted impacts. Water Resour. Res.
Koks, E.E., Jongman, B., Husby, T.G., Botzen, W.J., 2015. Combining hazard, exposure and social vulnerability to provide lessons for flood risk management. Environ. Sci. Pol. 47, 42–52.
Kostovetsky, L., 2015. Political capital and moral hazard. J. Financ. Econ. 116 (1), 144–159.
Kousky, C., 2018. Financing flood losses: a discussion of the national flood insurance program. Risk Manag. Insur. Rev. 21 (1), 11–32.
Kousky, C., Michel-Kerjan, E., 2017. Examining flood insurance claims in the United States: six key findings. J. Risk Insur. 84 (3), 819–850.
Kron, W., 2005. Flood risk= hazard• values• vulnerability. Water Int. 30 (1), 58–68.
Li, W., Lin, K., Zhao, T., Lan, T., Chen, X., Du, H., Chen, H., 2019. Risk assessment and sensitivity analysis of flash floods in ungauged basins using coupled hydrologic and hydrodynamic models. J. Hydrol. 572, 108–120.
Marjerison, R.D., Walter, M.T., Sullivan, P.J., Colucci, S.J., 2016. Does population affect the location of flash flood reports? J. Appl. Meteor. Climat. 55 (9), 1953–1963.
Merz, B., Kreibich, H., Schwarze, R., Thieken, A., 2010. Review article" Assessment of economic flood damage. Nat. Hazards Earth Syst. Sci. 10 (8), 1697–1724.
Metin, A.D., Dung, N.V., Schröter, K., Guse, B., Apel, H., Kreibich, H., et al., 2018. How do changes along the risk chain affect flood risk? Nat. Hazards Earth Syst. Sci. 18 (11).
Michel-Kerjan, E.O., 2010. Catastrophe economics: the national flood insurance program. J. Econ. Perspect. 24 (4), 86–165.
Nance, E., 2015. Exploring the impacts of flood insurance reform on vulnerable communities. Int. J. Disaster Risk Reduct. 13, 20–36.
National Oceanic and Atmospheric Administration (NOAA), 2019. Precipitation frequency data server (PFDS). Available Online. http://hdsc.nws.noaa.gov/hdsc/pfds/.
NCDC, 2019. Billion dollar weather and climate disasters. NOAA/NCDC. https://www.ncdc.noaa.gov/billions/.
Neal, J., Schumann, G., Bates, P., 2012. A subgrid channel model for simulating river hydraulics and floodplain inundation over large and data sparse areas. Water Resour. Res. 48 (11).
New York State (NYS), 2019a. Administrative correction of records. Real property tax law, article 5, title 3. Available online. https://www.tax.ny.gov/pdf/publications/orpts/coe_presentation.pdf.
New York State (NYS), 2019b. New York state tax parcel: NYS tax parcel centroids. Available Online. http://gis.ny.gov/parcels/. (Accessed 23 July 2019).
Ngo, P.T., Hoang, N.D., Pradhan, B., Nguyen, Q., Tran, X., Nguyen, V., et al., 2018. A novel hybrid swarm optimized multilayer neural network for spatial prediction of flash floods in tropical areas using Sentinel-1 SAR imagery and geospatial data. Sensors 18 (11), 3704.
Pigg, K., Gasteyer, S.P., Martin, K.E., Keating, K., Apaliyah, G.P., 2013. The community capitals framework: an empirical examination of internal relationships. Community Dev. 44 (4), 492–502.
Pralle, S., 2019. Drawing lines: FEMA and the politics of mapping flood zones. Climatic Change 152 (2), 227–237.
Quinn, N., Bates, P.D., Neal, J., Smith, A., Wing, O., Sampson, C., et al., 2019. The spatial dependence of flood hazard and risk in the United States. Water Resour. Res. 55 (3), 1890–1911.
Royal, A., Walls, M., 2019. Flood risk perceptions and insurance choice: do decisions in the floodplain reflect overoptimism? Risk Anal. 39 (5), 1088–1104.
Rufat, S., Tate, E., Burton, C.G., Maroof, A.S., 2015. Social vulnerability to floods: review of case studies and implications for measurement. Int. J. Disaster Risk Reduct. 14, 470–486.
Savage, S.L., Lawrence, R.L., Squires, J.R., 2015. Predicting relative species composition within mixed conifer forest pixels using zero-inflated models and Landsat imagery. Rem. Sens. Environ. 171, 326–336.
Schoups, G., Vrugt, J.A., 2010. A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. Water Resour. Res. 46 (10).
Shafizadeh-Moghadam, H., Valavi, R., Shahabi, H., Chapi, K., Shirzadi, A., 2018. Novel forecasting approaches using combination of machine learning and statistical models for flood susceptibility mapping. J. Environ. Manag. 217, 1–11.
Siders, A.R., Hino, M., Mach, K.J., 2019. The case for strategic and managed climate retreat. Science 365 (6455), 761–763.
Souissi, D., Zouhri, L., Hammami, S., Msaddek, M.H., Zghibi, A., Dlala, M., 2019. GIS-based MCDM-AHP modeling for flood susceptibility mapping of arid areas, southeastern Tunisia. Geocarto International 1–25 just-accepted.
Strother, L., 2018. The National Flood Insurance Program: a case study in policy failure, reform, and retrenchment. Pol. Stud. J. 46 (2), 452–480.
United States Department of Agriculture (USDA) National Resource Conservation Service (NRCS), 2019. Web soil survey. Available online. http://websoilsurvey.sc.egov.usda.gov/App/HomePage.htm.
United States Geological Survey (USGS), 2016. National land cover change index (CONUS). Available Online. https://www.mrlc.gov/data/nlcd-land-cover-change-index-conus.
United States Geological Survey (USGS), 2019. National elevation dataset. Available Online. http://nationalmap.gov/elevation.html.
USGS. 2020. StreamStats. Available online: https://streamstats.usgs.gov/ss/. Accessed on: 7/8/2020.
Villarini, G., 2016. On the seasonality of flooding across the continental United States. Adv. Water Resour. 87, 80–91.
Vorogushyn, S., Bates, P.D., de Bruijn, K., Castellarin, A., Kreibich, H., Priest, S., et al., 2018. Evolutionary leap in large-scale flood risk assessment needed. Wiley Interdisciplinary Reviews: Water 5 (2), e1266.
Wang, Y., Hong, H., Chen, W., Li, S., Panahi, M., Khosravi, K., et al., 2019. Flood susceptibility mapping in Dingnan County (China) using adaptive neuro-fuzzy inference system with biogeography based optimization and imperialistic competitive algorithm. J. Environ. Manag. 247, 712–729.
Wheater, H., Evans, E., 2009. Land use, water management and future flood risk. Land Use Pol. 26, S251–S264.
Wilson, M.T., Kousky, C., 2019. The long road to adoption: how long does it take to adopt updated county-level flood insurance rate maps? Risk Hazards Crisis Publ. Pol.

Wing, O.E., Bates, P.D., Smith, A.M., Sampson, C.C., Johnson, K.A., Fargione, J., Morefield, P., 2018. Estimates of present and future flood risk in the conterminous United States. Environ. Res. Lett. 13 (3), 034023.
Woznicki, S.A., Baynes, J., Panlasigui, S., Mehaffey, M., Neale, A., 2019. Development of a spatially complete floodplain map of the conterminous United States using random forest. Sci. Total Environ. 647, 942–953.
Xie, et al., 2010. CPC unified gauge-based analysis of global daily precipitation. 24th Conf. on Hydrology, Atlanta, GA, Amer. Meteor. Soc 2. Preprints.
Zheng, X., Maidment, D.R., Tarboton, D.G., Liu, Y.Y., Passalacqua, P., 2018. GeoFlood: large-scale flood inundation mapping based on high-resolution terrain analysis. Water Resour. Res. 54 (12), 10–13.
Zhou, Q., Panduro, T.E., Thorsen, B.J., Arnbjerg-Nielsen, K., 2013. Verification of flood damage modelling using insurance data. Water Sci. Technol. 68 (2), 425–432.