Bowers, C., Serafin, K. A., and Baker, J. W. (2024). "Uncovering Drivers of Atmospheric River Flood Damage Using Interpretable Machine Learning." Natural Hazards Review, 25(3), 04024022. *https://doi.org/10.1061/NHREFO.NHENG-1995*

Uncovering Drivers of Atmospheric River Flood Damage using Interpretable Machine Learning

Corinne Bowers, Ph.D.[1,2], Katherine A. Serafin, Ph.D.[3], and Jack W. Baker, Ph.D.[1]

[1]Civil and Environmental Engineering Department, Stanford University, Stanford, CA.

[2] Currently employed at U.S. Geological Survey, Reston, VA. Email: cbowers@usgs.gov.

[3]Department of Geography, University of Florida, Gainesville, FL.

ABSTRACT

The intensity of an atmospheric river (AR) is only one of the factors influencing the damage it will cause. We use random forest models fit to hazard, exposure, and vulnerability data at different spatial and temporal scales in California to predict the probability that a given AR event will cause flood damage, as measured by National Flood Insurance Program (NFIP) claims. We first demonstrate the usefulness of data-driven models and interpretable machine learning to identify and describe drivers of AR flood damage. Hazard features, particularly measures of AR intensity such as total precipitation, increase the probability of damage with increasing values up to a threshold point, after which the probability of damage saturates. While hazard is generally the most important risk dimension across all models, exposure and vulnerability contribute up to a third of the explanatory power. Exposure and variability features generally increase the probability of damage with increasing values, apart from a few instances which can be explained by physical intuition, but tend to affect the probability of damage less for the largest AR events. Comparisons between random forest models at different spatial and temporal scales show general agreement. We then examine limitations inherent in publicly available exposure, vulnerability, and loss data, focusing on the difference in temporal resolution between variables from different risk dimensions and discrepancies between NFIP claims and total flood losses, and describe how those limitations may

affect the model results. Overall, the application of interpretable machine learning to understand the contributions of exposure and vulnerability to AR-driven flood risk has identified potential community risk drivers and strategies for resilience, but the results must be considered in the context of the data that produced them.

INTRODUCTION

Flooding is the most common and costly natural disaster that Americans face. Climate change has already increased the frequency and severity of floods; of the 41 billion-dollar flood disasters in the United States since 1980, 18 have occurred in the past decade (NOAA NCEI 2023). Floods become disasters based on not only the intensity of the hazard, but also interactions with the landscape, infrastructure, and communities at a particular location. In order to characterize flood risk, we rely on the well-established definition that risk is the product of three dimensions: hazard, exposure, and vulnerability. Hazard includes the intensity of the atmospheric event as well as environmental factors that could mediate or amplify flooding, such as impervious land cover or wet antecedent conditions. Exposure represents the people and buildings who experience the hazard, and vulnerability quantifies the ability of those people and buildings to withstand the hazard. All three risk dimensions must be accounted for in order to build models of flood damage that are both accurate (able to predict the magnitude of damage expected from a given storm event) and interpretable (able to determine which risk factors contributed most to damage during that event).

Flood damage can be modeled using either a process-based or a data-driven approach. Process-based models start with the driving hydroclimatic conditions and simulate the physical processes from streamflow and inundation to damage and loss (e.g., FEMA (2006)). Most process-based models in the literature stop at loss prediction, though, and do not extend their analysis to quantify the drivers of loss. Data-driven models approach the problem from the opposite direction, starting with instances when impacts were observed and working backwards to empirically estimate the factors that are most predictive of impact (Solomatine and Ostfeld 2008). A data-driven model can take many forms, ranging from ordinary least squares regression to complex machine learning (ML) and artificial intelligence (AI) models. ML models have increased in popularity for assessment

of flood hazard (Sadler et al. 2018; Mobley et al. 2021) and flood damage (Wagenaar et al. 2017; Szczyrba et al. 2021) because of their high predictive accuracy and the increased availability of data for fitting the models.

However, there are two main limitations in existing ML models for flood risk assessment. First, the increased predictive accuracy of more complex model forms comes at a price of decreased model interpretability. Interpretability is especially important in a flood risk context where knowing *why* the model produced a certain outcome matters as much as or more than *what* the outcome was. This limitation has been partially addressed through a suite of tools that fall under the umbrella of "interpretable ML" (Molnar 2023), all developed with the goal of improving the degree to which a human can understand the cause of a data-driven model decision. While some researchers have incorporated interpretable ML results in their flood risk assessments (e.g., Stein et al. (2021)), the practice is not widespread. Second, previous research has shown that data-driven models for flood risk assessment are sensitive to the spatial resolution of the data (Komolafe et al. 2018; Pollack et al. 2022) and that differences in temporal resolution across predictor variables can skew results (Mobley et al. 2021). Very few studies have examined the effect of spatial or temporal resolution on model interpretability results, and to our knowledge none have considered both factors in tandem.

In this paper, we build random forest (RF) classification models to predict the likelihood of flood damage due to atmospheric rivers (ARs) in California at different spatial and temporal scales. ARs are the primary drivers of flood risk in the western US, associated with extreme precipitation (Lamjiri et al. 2017), hydrologic floods (Konrad and Dettinger 2017), and economic impacts (Corringham et al. 2019). We create an extensive dataset with over forty predictor variables representing hazard, exposure, and vulnerability. We then use interpretable ML to explore the contributions of these variables to the prevalence of insurance claims from the National Flood Insurance Program (NFIP). Our RF models quantify the value of including information about community-level exposure and social and infrastructural vulnerability in models of AR-driven flood damage and identify nonlinear threshold points and variable interactions that can guide potential resilience strategies. This paper also makes a more general methodological contribution

to the literature on ML in flood risk by comparing predictive accuracy and model interpretability results from RF models created at multiple spatial and temporal scales. We offer a perspective on the benefits and limitations of existing publicly available exposure, vulnerability, and loss data, and conclude by proposing avenues of work to improve future data-driven models of both flood risk and flood risk drivers.

DATA

Response Variable

The response variable of interest is a binary indicator of whether or not an AR storm caused flood damage in a specific geographic unit (county or census tract). We define a damaging AR event as one that causes flood insurance claims to be submitted by a policyholder in the Federal Emergency Management Agency (FEMA) National Flood Insurance Program (NFIP) (FEMA 2023b). We include denied claims and claims below the policy deductible (zero payout) in our analysis, assuming a filed claim indicates that the policyholder experienced some negative consequence due to an AR event. We do not distinguish between pluvial, fluvial, coastal, or indirect flood effects. The number of claims needed to qualify an AR event as damaging depends on the spatial resolution. At the census tract level, the threshold is one claim. At the county level, large differences in population between counties mean that more populous counties have more policyholders and are consequently more likely to have at least one claim filed somewhere in the county during the AR event; we therefore define the threshold as $(N_{county})_i/N_{state}$, where $(N_{county})_i$ is the number of NFIP policies in county i and N_{state} is the statewide median of policies per county. This policy-based threshold corrects the population bias at the county level and more evenly distributes damaging events across the state.

NFIP claims are often used as a proxy for flood impacts because claims are available at the census tract scale and tagged to a specific date of loss, which allows for a granular examination of flood impacts. Multiple other researchers have used NFIP claims to fit data-driven models of flood hazard (Mobley et al. 2021) and loss (Czajkowski et al. 2017; Knighton et al. 2020). However, NFIP policyholders are not a representative sample of California residents. Only about 2% of eligible

homeowners are insured (FEMA 2023b), and due to a combination of both self-selection within risky areas and the mandatory purchase requirement for homes with federally backed mortgages within the 100-year floodplain, NFIP policyholders are more likely than the general population to live in high-risk areas with a history of flooding (Bradt et al. 2021). Insurance take-up rates are also influenced by education (Atreya et al. 2015), community flood protection investment (Zahran et al. 2009), and income and home value (Darlington and Yiannakoulias 2022), among other factors.

Previous works have addressed the representation bias of the NFIP in a number of ways, from applying correction factors (Smith and Katz 2013; Corringham and Cayan 2019) to modeling damage at uninsured properties (Thomson et al. 2023). We address it here by limiting our analysis to classification rather than regression. Focusing on damage versus no-damage and neglecting claim payout values avoids issues arising from differences in coverage limits between policyholders within and outside of the 100-year floodplain, coverage limit changes over time, and concerns about overrepresentation of higher-valued properties. However, it means our results will only show the underlying drivers of damage probability, which may or may not be the same as drivers of damage magnitude (Rözer et al. 2019). There also still remain demographic and socioeconomic differences at the intra-county level between who is insured and who is not.

Predictor Variables

Table 1 lists all of the predictors in the model by risk dimension and by concept, where concepts represent groups of related variables. Table 1 also includes references to the data source for each variable as well as references that support that variable's potential connection to the response. Spatial variation of the data is at the census tract scale or smaller (T), at the county scale (C), or constant (–). Temporal variation is at the event level (E), monthly (M), yearly (Y), or constant (–).

We provide additional context around the variables chosen to represent each risk dimension, starting with hazard. Each record in our dataset represents one AR event. To identify ARs, we used the Rutz et al. (2014) algorithm, which defines ARs as contiguous areas greater than 2,000 km in length and with integrated water transport (IVT) values over 250 kg/m/s. IVT was calculated

from the Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2; Gelaro et al. (2017)). MERRA-2 reports data at a resolution of $0.5° \times 0.625°$ (~50km × 50km) from 1980 to present. We recorded the maximum IVT and duration of each AR event and used the Ralph et al. (2019) scale to categorize the intensity of each AR from 1 (mostly beneficial) to 5 (mostly hazardous). Antecedent conditions are measured with three variables: total precipitation in the 3 days prior to the AR event, total precipitation in the 14 days prior to the AR event, and average soil moisture over the 3 days prior to the AR event. Large-scale climate modes such as ENSO and PDO capture time periods when flood risk increases or decreases over a broad geographic range, and land surface variables capture on-the-ground conditions that can amplify storm effects.

For exposure, we focus on variables related to population and housing. The large majority of NFIP policies cover residential buildings, so NFIP claims are more representative of housing exposure than other types of infrastructure. We do not measure other types of exposed assets such as roads, critical infrastructures, crops and livestock, and cultural heritage sites. We also include variables identifying specific geographies associated with higher NFIP insurance takeup rates.

For social vulnerability, we rely on constructed indices, particularly the Centers for Disease Control (CDC) Social Vulnerability Index (SVI) (Flanagan et al. 2011) and CalEnviroScreen 4.0, a statewide screening tool for vulnerability to environmental hazards (August et al. 2021). The CDC SVI was chosen because of its long time record, with values extending back to 2000, and its ability to explain both recorded damages and fatalities in an empirical validation exercise (Bakkensen et al. 2017). We included the four components of SVI (socioeconomic status, household characteristics, racial & ethnic minority status, and housing type & transportation) as separate predictor variables. The CalEnviroScreen metrics (population characteristics, pollution burden, and disadvantaged communities) were chosen because of their calibration to California, their relevance in statewide planning decisions, and their focus on environmental justice. One drawback of constructed indices, though, is that they are designed for comparison over space rather than over time. While the indices are updated regularly, each data generation is normalized such that the values at any given time represent only the relative ranking of one census tract or county, so the values do not necessarily

capture absolute changes in vulnerability over time (Bakkensen et al. 2017). Therefore we include median household income, percent of the population as non-Hispanic white, and percent of the population as working age (18–64) as standalone metrics of socioeconomic vulnerability common to many indices that have physical meaning. Note that while increasing index values signify increasing vulnerability, increases in these standalone metrics signify decreasing vulnerability.

Infrastructural vulnerability, similar to exposure, includes metrics relevant to housing, such as building age and construction type. Lastly, we include two metrics of flood experience, number of federally declared disasters in the past three years and county-level participation in the Community Rating System (CRS) program. The CRS program is a mechanism to incentivize insurance uptake and increase community-level flood resilience through community-wide policy discounts offered in exchange for flood risk reduction actions. 70% of NFIP policyholders nationwide live in participating communities, and 26 out of California's 58 counties have participated in the CRS at some point since its inception, with 24 of those counties still in the program today (FEMA 2023a).

Spatiotemporal Resolution

To analyze the effects of spatial and temporal resolution on predictive accuracy and model interpretability results, we fit RF models at two spatial and two temporal scales, for an overall total of four models. The two spatial scales are at the county-level across all of California and at the census tract-level across Sacramento County. Sacramento County was chosen because of its history of flood events (James and Singer 2008) and its significant investment in flood mitigation, particularly its commitment to the CRS program. The two temporal scales are 1981–2021 and 2009–2021. While most hazard variables are available starting in 1981, many exposure and vulnerability variables are not available until later. For the models starting in 2009, the variables footnoted with (b) or (c) in Table 1 no longer require extrapolation beyond the range of their record and the variables footnoted with (d) are allowed to vary annually rather than remain constant.

Nearly all of the hazard variables have higher temporal resolutions than the exposure and vulnerability variables, even in the 2009–2021 timeframe. Most hazard variables are recorded at the event scale, meaning that each record in the dataset will have different values based on the

characteristics (maximum IVT, 3-day antecedent soil moisture, etc.) of that particular AR event. The exposure and vulnerability variables, though, are recorded at the annual scale. This is logical in a physical sense; for example, the total number of housing units does not change day-to-day like the weather does. From a data perspective, though, this means that there is no intra-annual variation to exploit for the RF model, and RF models are known to preferentially split on features with higher variance (Strobl et al. 2007). The significantly higher resolution of the hazard data is analogous to the disproportionate focus on hazard in the physical model space (Merz et al. 2010). Nevertheless, comparison of the 1981–2021 and 2009–2021 models provides some insight into the usefulness of collecting additional exposure and vulnerability data for the earlier years in the historical record.

METHODOLOGY

Dataset Preparation

We generated four datasets of AR events spanning the forty-year time period from 1981–2021. Each record included the AR characteristics of the event plus the additional hazard, exposure, and vulnerability variables documented for that specific time and place. If an AR passed over multiple geographic units, it was tabulated as multiple records (one per county/tract) in order to capture the effect of differences in exposure and vulnerability between different locations. Table 2 shows the total number of records in each of the four datasets.

We sampled 80% of the data using stratified random sampling by county/tract for training and validation, reserving the remaining 20% to test the performance of the final model. Stratified random sampling ensures that there are counties (in the statewide model) or tracts (in the Sacramento model) the model has never seen before, which in turn ensures that the goodness-of-fit metrics calculated on the test set are more faithful representations of the model's true performance. We then implemented the Synthetic Minority Oversampling TEchnique (SMOTE) (Chawla et al. 2002) on the training data to fix the class imbalance in Table 2. With highly imbalanced classes, it is difficult to train an ML model that accurately captures damaging events; simply put, a naive model that predicts no damage every time would be correct approximately 95% (statewide models) or 99.5% (Sacramento models) of the time. SMOTE increases (oversamples) the number of records in the

minority class, in our case the damaging AR events, by creating synthetic records based on the distribution of historical events, and decreases (undersamples) the number of non-damaging AR events by a corresponding amount to achieve an even class balance. Combined over/undersampling on imbalanced data improves the predictive accuracy of ML models across a range of contexts (e.g., Estabrooks et al. (2004)).

After implementing SMOTE on the training set for each model, we performed feature selection to remove highly collinear variables. While feature collinearity does not affect the accuracy of RF models, it does adversely impact model interpretability, which is the main goal of this paper. We therefore first identified clusters of correlated variables using principal component analysis (PCA), then calculated the Akaike information criterion (AIC) of each variable in the cluster and kept only the ones with the highest predictive power. We repeated this two-step process of clustering and consolidating until the maximum variable inflation factor (VIF) fell below 10 and the maximum Pearson correlation coefficient fell below 0.8 (James et al. 2013). Despite the stochastic nature of the SMOTE algorithm, the clusters were very stable, and our process removed a consistent subset of the variables every time. Finally, we added one additional feature with uniform random noise, which serves as a check for our feature importance and impact analyses; if a given feature is less important than random noise, it is discarded.

Model Training

An RF model has three hyperparameters determined by the user: the number of trees in the forest, the depth of each tree, and the number of predictors selected to fit each tree. The number of trees in the forest was held constant at $n = 1,000$, consistent with other RF applications in similar contexts (e.g., Alipour et al. (2020)), and each tree was allowed to reach its maximum possible depth (1 data point at each leaf). The number of predictors selected to fit each tree was tuned between 1 and 10. We fit all models using 10-fold cross-validation, fitting the model on 90% of the training data and calculating accuracy metrics on the remaining 10%, then repeating that process across all the folds of the data. The best-fit model was chosen based on accuracy, which is the number of correct predictions divided by the total number of predictions.

Performance Evaluation

We compared each of the fitted RF models against their respective withheld test sets. The test data had the same class imbalance as the original data, so we utilized three performance evaluation metrics appropriate for imbalanced data: area under the Receiver Operating Characteristic curve (ROC-AUC), area under the precision-recall curve (PR-AUC), and balanced accuracy. All three are derived from the confusion matrix, which summarizes the four potential outcomes for each prediction: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). True positives and true negatives occur when the model correctly predicts a damaging or non-damaging AR event, respectively. False positives occur when the model incorrectly labels an AR as a damaging event, and false negatives occur when the model incorrectly labels an AR as a non-damaging event. From the confusion matrix, we derive secondary performance metrics, as shown in Equation 1. Precision measures correctly predicted positives out of all predicted positives. Recall, or sensitivity, measures correctly predicted positives out of all observed positives; the two names come from different disciplinary conventions, so we use both here in their respective contexts. Specificity measures correctly predicted negatives out of all observed negatives.

$$Precision = \frac{TP}{TP + FP}, \quad Recall\ (Sensitivity) = \frac{TP}{TP + FN}, \quad Specificity = \frac{TN}{TN + FP}. \quad (1)$$

Sensitivity and specificity together are used to plot the ROC curve (Fig. 1a), a scale-invariant model diagnostic frequently used for imbalanced data (Kotsiantis et al. 2006). Precision and recall together are used to plot the PR curve (Fig. 1b). PR curves have been put forward as an even more informative tool for imbalanced datasets (Saito and Rehmsmeier 2015). The plots in Figure 1 measure the respective values of precision, recall (sensitivity), and specificity at varying detection thresholds for each of the four RF models. Model-optimized detection thresholds are marked with a dot. Finally, the balanced accuracy is defined as the average of sensitivity and specificity and calculated based on the model-optimized detection threshold. The values of the three metrics for each model are reported in Table 3. The fitted RFs at all spatial and temporal scales clearly

outperform the null models, signalling that they are far more informative about damaging events.

MODEL INTERPRETATION

SHapley Additive exPlanations (SHAP)

We use SHapley Additive exPlanations (SHAP; Lundberg and Lee (2017)), which utilize Shapley values to assess local and global feature importance, for interpretation of our RF models. Shapley values are a game-theory approach for fairly apportioning a "prize" between multiple "players." In the context of ML, the "players" are predictor variables and the "prize" is the difference between the overall mean prediction (expected value) and the prediction for a specific observation (observed value). SHAP conceptualizes the calculation of Shapley values as an additive feature attribution model and estimates the feature's contribution to the difference between the expected and observed value for every record in the dataset. We prefer SHAP over other interpretable ML techniques for several reasons. First, it is the only method to satisfy the three statistical properties (local accuracy, missingness, and consistency) that are necessary and sufficient to ensure a fair apportionment of the overall contribution among the various predictors (Lundberg and Lee 2017). Second, it provides one coherent framework for examining both the magnitude and direction of a feature's effect and for building from local to global importance (Molnar 2023). Many other interpretable ML methods only apply to one of these use cases, leading to an analysis that relies on unrelated tools with different baseline assumptions. Third, comparisons between SHAP and other interpretable ML techniques showed agreement at every step of the analysis.

Feature Importance

We first focus on global feature importance, which is calculated as the mean of the absolute value of SHAP values across all observations. Figure 2 shows the overall importance of features in each model, grouped by risk dimension and concept and normalized to a total of 100%. Across the four models, hazard features account for approximately 70% of the model's predictive power, exposure features account for approximately 10%, and vulnerability features comprise the remaining 20%. The statewide 1981–2021 model is an outlier with higher-than-average exposure and vulnerability

contributions. Figure 2 also separates the risk dimensions by the concepts defined in Table 1. We notice some patterns; for example, social vulnerability is the most important vulnerability concept in the statewide models, and flood experience is more important in the 1981 models than the 2009 models. Population exposure is the most important exposure concept in the Sacramento models, while housing exposure is more important in the statewide models. Factors affecting insurance takeup only appear as important in the 1981 models. The relative contributions of the hazard concepts are more stable across the spatial and temporal scales, but climate modes and land surface variables are slightly more important in the 2009 models.

We move to comparing the rankings of individual features, noting common trends across all four models and exploring differences. In all cases, the top five predictors are related to the hazard dimension, mostly from either the AR characteristics concept or the antecedent conditions concept. Total precipitation and maximum IVT always occupy the top two positions, and total precipitation is the most important predictor in three out of four models. Lagged cumulative precipitation and lagged average soil moisture appear frequently, which highlights the important contribution of antecedent conditions to AR-driven flood risk in California. Two features related to exposure are noteworthy: percentage of the population living in the FEMA 100-year floodplain appears in the top ten in all four models, and total number of housing units appears in the top ten in both statewide models. For vulnerability, CRS score occupies the tenth position in the 1981 statewide model and median housing unit age occupies the tenth position in the 1981 Sacramento model. There are no features related to the vulnerability dimension in the top ten of either of the 2009 models.

There are more noticeable shifts in feature rank across space than across time. While it is important in all cases, total precipitation has a larger SHAP feature contribution in the statewide models. The contributions are more evenly spread among the top five to ten features in the Sacramento models. AR category is more important in the 1981 models and ENSO climate index is more important in the 2009 models, but otherwise there are no clear temporal patterns. While our methodology does not necessarily allow for a direct comparison between the two, this may suggest that spatial scale has a larger effect on model results than temporal scale.

Feature Impact

Feature impact plots visualize the direction and magnitude of a particular feature's influence on the model's prediction. We use accumulated local effects (ALE) curves (Apley and Zhu 2020) paired with SHAP values to analyze feature impact. The atomic unit of an ALE curve is the local effect, or the difference between the prediction at x and the prediction at some perturbed value of x within a small interval δ. ALEs are the sum of the local effects for all observations falling within $x \pm \delta$, calculated for each x in the feature domain. ALE curves show marginal contributions, not conditional contributions, so the effects of correlated features are not separated; however, they are more robust to collinearity than the more commonly used partial dependence plots (Stein et al. 2021), so they are well suited for our analysis. We pair the ALE curves, which are average metrics, with random samples of 1,000 SHAP values from individual records. The SHAP values provide an understanding of the variance and show where the ALE curve is based on more or less data.

Figures 4a–c plot the SHAP values and ALE curves for the top three hazard features in the 1981 statewide model: total precipitation (4a), AR maximum IVT (4b), and AR duration (4c). The behavior of all of the hazard features follows a similar pattern, where probability of damage increases with increasing feature values until some threshold point. At a certain point, the magnitude of the hazard becomes so large that damage becomes the probable outcome. For total precipitation (Fig. 4a), the threshold point is roughly 75mm, or approximately 15% of California's mean annual total precipitation. The threshold point for maximum IVT (Fig. 4b) is about 750mm, which would be a Category 2–3 AR event, and the threshold point for AR duration (Fig. 4c) is about 30 hours.

Figures 4d–f plot the SHAP values and ALE curves for the top three exposure features in the 1981 statewide model: total number of housing units (4d), percent population living in the FEMA 100-year floodplain (4e), and percent housing stock as single family homes (4f). The plot of total housing units (Fig. 4d) largely shows increasing probability of damage with increased housing stock. More people and more buildings at risk imply more chances for damage, so this matches intuitive reasoning. Less intuitive is the influence of the percent population living in the 100-year floodplain (Fig. 4e). We would expect more people in the floodplain to increase the

likelihood of damage; instead, it appears to have a negative influence on damage probability. The distributions of the SHAP values in these panels provide more information about the discrepancy. Statewide, the median percent population in the floodplain is 8.1%, which means about half of all counties fall on the portion of the ALE curve in Figure 4b that is increasing. There could also be a confounding relationship with county-level flood resilience; the counties with the highest percentage of population living in the floodplain, say 20% or more, may be more prepared for flooding and thus less likely to sustain damage from an AR event. Another counterintuitive relationship is the negative influence of the percent of single family homes (Fig. 4f). NFIP policyholders disproportionately live in single family homes, so more of this housing type would mean more opportunity for claims. But residents of single family homes also tend to be more likely to invest in individual-level flood mitigation efforts, so the negative relationship might indicate that this particular feature is capturing more vulnerability than exposure. Percent single family homes is also negatively correlated with total population, so counties with higher percentages likely have fewer opportunities to submit insurance claims that would be recorded as damage.

Figures 4g–i plot the SHAP values and ALE curves for the top three vulnerability features in the 1981 statewide model: CRS score (4g), median household income (4h), and CalEnviroScreen pollution burden score (4i). Lower numbers indicate more significant flood resilience investment. 1 is the best possible score score and 10 means that a county has not engaged with the CRS. The perhaps-surprising trend of higher damage probabilities at CRS scores of 7–9 compared to those at a CRS score of 10 may be because counties with a history of damaging flood events are more likely to invest time and money into joining the CRS program. Multiple studies have found that CRS-participating counties and communities see significant reductions in flood loss (Highfield and Brody 2017; Gourevitch and Pinter 2023), and the SHAP values in Figure 4g suggest that achieving a CRS score of 6 or better does pay off in terms of flood risk reduction; however, Sacramento County is the only county in California that has achieved a rating of 4 or better, so scores beyond this point are not necessarily representative of the entire state. The median household income (Fig. 4h) and the CalEnviroScreen pollution burden score (Fig. 4i), both measures of social vulnerability,

seem to indicate that the probability of damage decreases with increasing vulnerability. This may be indicative of limitations in the link between NFIP claims and damage sustained by all members of the population, suggesting that broad interpretations for flood risk should be made with caution. Further considerations for researchers using NFIP data are included in the Discussion.

Feature Interactions

Figure 5 plots interactions between hazard, as measured by the Ralph et al. (2019) intensity category, and exposure and vulnerability. In most cases, exposure and vulnerability variables become less useful predictors of damage likelihood as hazard intensity increases. Category 5 ARs have historically almost always caused some amount of damage (Corringham et al. 2019), so while exposure and vulnerability variables may still impact the severity of damage, they no longer have any influence on the probability. For example, in both the total number of housing units (Fig. 5a) and the percentage of housing stock as single-family homes (Fig. 5c), the observed trend is strongest for Category 1 events and weakest for Category 5 events, when probability of damage is elevated and individual-level characteristics are more likely to be overwhelmed by the severity of the hazard. The pattern flips, though, for CRS score (Fig. 5d) and median household income (Fig. 5e). The increase in probability of damage moving from a CRS score of 10 to 9 all but disappears for the largest ARs, and the benefit of improving a county's CRS score from a 5 to a 4 or better increases with increasing intensity category. For median household income, there is a slight reduction in damage probability for the lowest incomes that only occurs at the highest hazard intensities. These figures do not provide a comprehensive list of the ways hazard, exposure, and vulnerability interact; rather, they illustrate examples of the kind of practically relevant insights from our RF models that can be used to help communities better understand their risk under different AR scenarios.

DISCUSSION

Benefits of Data-Driven Approach

Through a combination of global feature importance, feature impact plots, and feature interactions, our RF models were able to identify new connections between the risk dimensions and

AR-driven flood damage in California. The global feature importance analysis showed that hazard features, individually and collectively, had the biggest influence on flood damage. This is not surprising; flooding and flood damage are far more likely to occur on rainy days than sunny days. An example of decision-relevant information related to hazard from our models are the nonlinear thresholds in the feature impact plots. If emergency managers are confident that flood damage is likely above a certain precipitation threshold, it reduces the information burden required to make a decision and allows for quicker mobilization of resources. Another example comes from the interaction plots: changes in exposure and vulnerability only affected damage probability during Category 5 events in one out of four features shown, so a well-rounded flood resilience strategy would include elements that decrease risk across the spectrum of potential hazards.

We also showed that exposure and vulnerability explain up to a third of the model's predictions. This finding has important implications from a management perspective because exposure and vulnerability can be altered at the community level more easily than hazard. The ALE-SHAP plots are therefore useful to understand the magnitude and direction of effects and to highlight overlooked risk factors for NFIP policyholders. For example, for the percent of the population living in the 100-year floodplain (Fig. 4e), probability of damage peaks at 5–10% before starting to decrease. Values in this range might indicate potential risk hotspots, where hazard is high enough to cause damage but not high enough that counties have significantly invested in resilience efforts, and identify a subset of counties that are worthy of more in-depth local analysis.

Geographic Representation

While random forests and interpretable ML methods are powerful tools for combining disparate data sources and extracting insights, the results of the models are only as good as the data used to train, fit, and validate them. We discuss two key limitations of this work, which stem from the assumption that NFIP claims accurately represent total losses from floods and relate to (a) geographic and (b) socioeconomic differences in insurance takeup rates.

First, there are significant inter-county differences in NFIP takeup rates, ranging from almost 0% in Mariposa County to over 18% in Sutter County. We compare observed versus expected

takeup rate (see Appendix I) to identify counties with anomalously high or low numbers of NFIP policies-in-force. Sutter, Yuba, and Sacramento Counties all have more policies than expected. All three border the Sacramento River, which has an extensive history of severe floods, and all three have invested heavily in the CRS program: Yuba County has a score of 6 (top 30% of all participating communities), Sutter County has a score of 5 (top 13%), and Sacramento County has a score of 2 (top 0.5%) (FEMA 2023a). One of the activities that garners CRS points is public advertisement of the NFIP, and better community ratings lead to higher policy discounts, so CRS participation has likely increased the number of policies in these areas. Joining the CRS also requires significant upfront investment, so it is more widely adopted in the communities and counties that can afford to participate (Sadiq et al. 2020). On the other hand, Alpine, Mariposa, Tuolumne, and Imperial Counties are rural counties with relatively low populations, and all have 60 or fewer policies-in-force in 2021. The low number of policies mean that even if a flood event does cause damage or loss, it is less likely to lead to a NFIP insurance claim and therefore less likely to be labeled as damaging event in our dataset. Future work could use the expected takeup rates calculated here and determine the appropriate county-level correction factors to account for differences in takeup rates that cannot be attributed to real differences in flood hazard.

Socioeconomic Representation

In addition to the unequal representation of different counties, the demographic and socioeconomic characteristics of NFIP policyholders contribute to representativeness issues. The most socioeconomically vulnerable populations are typically most affected by floods and other disasters (e.g., Debbage (2019)), but there are several intersectional factors affecting vulnerable populations that simultaneously reduce the likelihood of NFIP participation and exacerbate flood risk. As one example, renters are particularly vulnerable to negative consequences from flooding (Heiman 2022). People of color are more likely to be renters, and renters tend to have lower incomes than homeowners (ACS 2023b). However, 80% of NFIP policyholders are single-family homeowners. While the NFIP does offer policies for renters, renters are often unaware that standard renter's insurance does not cover flood damage, and in many places landlords are not required to disclose

an apartment's history of flooding (Heiman 2022). Therefore renters are likely underrepresented in NFIP claims data, meaning that if there are factors specific to renters that affect flood risk, they will not be identified as important by our model. Using a different dataset than the NFIP as the response variable, such as remote sensing imagery (Szczyrba et al. 2021) or post-event surveys (Merz et al. 2013), could improve the representation of renters and other vulnerable populations. The predictor variables used in this study could also be enhanced through interviews or community-specific knowledge to better capture unaccounted-for resilience characteristics (Ismail-Zadeh et al. 2017). Lastly, using ML techniques to move from county- or state-level summary statistics to maps of spatially varying hazard, exposure, and vulnerability would be of great benefit for future flood mitigation investment decisions.

CONCLUSION

In this paper, we used interpretable machine learning (ML) tools to understand how the three dimensions of risk, hazard, exposure, and vulnerability, relate to AR-induced flood damage in California. We collected a large dataset of over forty predictor variables to quantify the contributions of each the three dimensions to the probability of flood damage, as measured using flood insurance claims from the National Flood Insurance Program (NFIP). We considered two spatial resolutions for the data: the county scale, modeled for all of California, and the census tract scale, modeled for Sacramento County. We also considered timescales of 1981–2021, using exposure and vulnerability data with limited temporal variation, and 2009–2021, using exposure and vulnerability data at an annual resolution. This produced a total of four random forest classification models, each of which detected true positives (AR events with NFIP claims) with a high level of accuracy in very imbalanced datasets.

We showed the power of interpretable ML to identify and investigate drivers of AR-driven flood risk given publicly available hazard, exposure, and vulnerability data. We gained insight into damage drivers by examining feature importance (how much does a given feature influence the model's predictions?) and feature impact (how does increasing or decreasing the value of the feature affect the response?) using SHapley Additive exPlanations (SHAP) as a unifying framework.

While hazard intensity features were the most important predictors of whether an AR would cause damage, exposure and vulnerability contributed up to a third of the model's explanatory power, and the overall relative contributions by risk dimension and risk concept broadly agreed across the spatial and temporal scales considered. Total precipitation was the most important predictor in three out of four models, and features related to the intensity of the hazard consistently represented the majority of the top ten. An analysis of feature impact for the top three hazard features in the county-level (statewide) model fit on data from 1981 onward revealed that increasing hazard severity increased the probability of flood damage, up to some threshold point. Above that threshold, probability of damage reached a saturation point where it was no longer sensitive to changes in precipitation; at a certain point, it became not a question of if damage will occur, but how much. In most cases, increasing exposure and vulnerability also increased the probability of damage, although the interpretation differed slightly depending on the specific feature under consideration. The physically plausible explanations of the data-driven outputs from SHAP and other interpretable ML tools support our confidence that the model is characterizing real drivers of flood risk.

We also illustrated the ramifications of the assumptions made to fit our RF models utilizing available data. Changes in the spatial and temporal resolution of the input data altered the ranking of which features were deemed significant in the analysis of global feature importance. The higher temporal scale of hazard data relative to the other risk dimensions and the differences in NFIP representativeness across geographic and socioeconomic boundaries were noted as limitations. We proposed avenues for future work that would mitigate these limitations and potentially uncover new pathways to increased resilience. Overall, our work highlights both the possibilities and pitfalls of using interpretable ML for flood risk assessment. It enhances our understanding of the relationship between individual AR events and their negative effects and broadens the discussion around AR-driven flood damage to include more explicit characterization of exposure and vulnerability. Understanding the drivers of damage improves our ability to predict and prepare for the impacts of ARs, today and in the future.

Data Availability Statement

All data used in this study is publicly available, and all code created to generate results is available in a Github repository (Bowers 2023). In particular, the datasets described in Table 1 are available as downloadable CSV files, the *reproduce_figures.html* markdown file recreates all figures and numerical results from this paper, and the *figure4.html* markdown file recreates Figure 4 for models at all spatial and temporal resolutions.

Acknowledgments

This material is based upon work supported by both the Stanford Graduate Fellowship and the National Science Foundation (NSF) Graduate Research Fellowship under Grant No. 1000265549. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the NSF. CB contributed conceptualization, data curation, methodology, formal analysis, validation, visualization, writing - original draft, and writing - review and editing. KAS and JWB contributed supervision and writing - review and editing, and JWB provided resources. We additionally thank Jenny Suckale and two anonymous reviewers for their helpful feedback that improved the quality of this work.

APPENDIX I. CALCULATION OF NFIP TAKEUP RATE BY COUNTY

We calculated the observed and expected number of NFIP policies and insurance takeup rates for each county in California to determine which counties had more or fewer policies than predicted. Observed numbers of policies were calculated based on 2021 policies-in-force (FEMA 2023b), and observed takeup rates were calculated as the number of policies divided by the number of 2021 housing units in each county.

The expected numbers of policies and takeup rates were calculated as follows. We categorized both NFIP policies and housing units as either within-floodplain or out-of-floodplain based on the FEMA National Flood Hazard Layer (NFHL). NFIP policies were determined to be in or out of the floodplain by the NFHL flood zone code included in the policy information. Housing units were determined to be in or out of the floodplain by finding the percentage of each census block group that overlapped with a NFHL spatial polygon, then dividing the housing units in that block group assuming an even distribution in space. For example, if 40% of a block group was covered by the NFHL, then 40% of the housing units were labeled as in-floodplain (HU_{in}) and 60% were labeled as out-of-floodplain (HU_{out}). We summed all policies and housing units to estimate statewide within-floodplain and out-of-floodplain insurance takeup rates. The 2021 statewide within-floodplain takeup rate was found to be 12.6% and the 2021 statewide out-of-floodplain takeup rate was found to be 0.69%. The expected number of policies by county was then calculated by aggregating over all block groups in that county, as illustrated in Equation 2. Lastly, county-level expected takeup rates were found by dividing the expected number of policies by the total number of housing units in each county.

$$\text{Expected Policies} = \sum_{bg \in \left\{\substack{\text{all block} \\ \text{groups}}\right\}} 0.126 * (HU_{in})_{bg} + 0.0069 * (HU_{out})_{bg} \tag{2}$$

REFERENCES

ACS (2023a). "DP04: Selected Housing Characteristics, 2009-2021, <https://data.census.gov/>.

ACS (2023b). "DP05: ACS Demographic and Housing Estimates, 2009-2021, <https://data.census.gov/>.

Alipour, A., Ahmadalipour, A., Abbaszadeh, P., and Moradkhani, H. (2020). "Leveraging machine learning for predicting flash flood damage in the Southeast US." *Environmental Research Letters*, 15(2), 024011.

Apley, D. W. and Zhu, J. (2020). "Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(4), 1059–1086.

Atreya, A., Ferreira, S., and Michel-Kerjan, E. (2015). "What drives households to buy flood insurance? New evidence from Georgia." *Ecological Economics*, 117, 153–161.

August, L., Bangia, K., Plummer, L., Prasad, S., Ranjbar, K., Slocombe, A., and Wieland, W. (2021). "CalEnviroScreen 4.0." *Report no.*, California Office of Environmental Health Hazard Assessment, Sacramento, CA, <https://oehha.ca.gov/media/downloads/calenviroscreen/report/calenviroscreen40reportf2021.pdf>.

Bakkensen, L. A., Fox-Lent, C., Read, L. K., and Linkov, I. (2017). "Validating Resilience and Vulnerability Indices in the Context of Natural Disasters." *Risk Analysis*, 37(5), 982–1004.

Bergstrand, K., Mayer, B., Brumback, B., and Zhang, Y. (2015). "Assessing the Relationship Between Social Vulnerability and Community Resilience to Hazards." *Social Indicators Research*, 122(2), 391–409.

Blum, A. G., Ferraro, P. J., Archfield, S. A., and Ryberg, K. R. (2020). "Causal Effect of Impervious Cover on Annual Flood Magnitude for the United States." *Geophysical Research Letters*, 47(5).

Bowers, C. (2023). "Supplemental Code Release: Uncovering Effects of Exposure and Vulnerability on Atmospheric River Flood Damage using Interpretable Machine Learning, <https://github.com/corinnebowers/damagedrivers>.

Bradt, J. T., Kousky, C., and Wing, O. E. (2021). "Voluntary purchases and adverse selection in

the market for flood insurance." *Journal of Environmental Economics and Management*, 110, 102515.

Brody, S. D., Kim, H., and Gunn, J. (2013). "Examining the Impacts of Development Patterns on Flooding on the Gulf of Mexico Coast." *Urban Studies*, 50(4), 789–806.

Brunner, M. I., Sikorska, A. E., and Seibert, J. (2018). "Bivariate analysis of floods in climate impact assessments." *Science of The Total Environment*, 616-617, 1392–1403.

Cao, Q., Gershunov, A., Shulgina, T., Ralph, F. M., Sun, N., and Lettenmaier, D. P. (2020). "Floods due to Atmospheric Rivers along the U.S. West Coast: The Role of Antecedent Soil Moisture in a Warming Climate." *Journal of Hydrometeorology*, 21(8), 1827–1845.

CDC (2022). "CDC/ATSDR Social Vulnerability Index, 2000-2020, <https://www.atsdr.cdc.gov/placeandhealth/svi/data

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). "SMOTE: Synthetic Minority Over-sampling Technique." *Journal of Artificial Intelligence Research*, 16, 321–357.

Corringham, T. W. and Cayan, D. R. (2019). "The Effect of El Niño on Flood Damages in the Western United States." *Weather, Climate, and Society*, 11(3), 489–504.

Corringham, T. W., Ralph, F. M., Gershunov, A., Cayan, D. R., and Talbot, C. A. (2019). "Atmospheric rivers drive flood damages in the western United States." *Science Advances*, 5(12).

Cutter, S. L. (2016). "The landscape of disaster resilience indicators in the USA." *Natural Hazards*, 80(2), 741–758.

Czajkowski, J., Villarini, G., Montgomery, M., Michel-Kerjan, E., and Goska, R. (2017). "Assessing Current and Future Freshwater Flood Risk from North Atlantic Tropical Cyclones via Insurance Claims." *Scientific Reports*, 7(1), 41609.

Darlington, J. C. and Yiannakoulias, N. (2022). "Experimental Evidence for Coverage Preferences in Flood Insurance." *International Journal of Disaster Risk Science*, 13(2), 178–189.

Debbage, N. (2019). "Multiscalar spatial analysis of urban flood risk and environmental justice in the Charlanta megaregion, USA." *Anthropocene*, 28, 100226.

DeFlorio, M. J., Pierce, D. W., Cayan, D. R., and Miller, A. J. (2013). "Western U.S. extreme

precipitation events and their relation to ENSO and PDO in CCSM4." *Journal of Climate*, 26(12), 4231–4243.

Dewitz, J. and USGS (2021). "National Land Cover Database (NLCD) 2019 Products (ver. 2.0, June 2021).

Erlingis, J. M., Rodell, M., Peters-Lidard, C. D., Li, B., Kumar, S. V., Famiglietti, J. S., Granger, S. L., Hurley, J. V., Liu, P., and Mocko, D. M. (2021). "A High-Resolution Land Data Assimilation System Optimized for the Western United States [Dataset]." *Journal of the American Water Resources Association*, 57(5), 692–710.

Estabrooks, A., Jo, T., and Japkowicz, N. (2004). "A multiple resampling method for learning from imbalanced data sets." *Computational Intelligence*, 20(1), 18–36.

FEMA (2006). "Hazus Flood Model Technical Manual." *Report no.*, Department of Homeland Security, Washington, DC, <https://www.fema.gov/flood-maps/tools-resources/flood-map-products/hazus/user-technical-manuals>.

FEMA (2020). "National Flood Hazard Layer (NFHL), <https://www.fema.gov/flood-maps/tools-resources/flood-map-products/national-flood-hazard-layer>.

FEMA (2023a). "CRS Participating Communities, <https://www.fema.gov/floodplain-management/community-rating-system>.

FEMA (2023b). "OpenFEMA Data Sets, <https://www.fema.gov/about/openfema/data-sets>.

Flanagan, B. E., Gregory, E. W., Hallisey, E. J., Heitgerd, J. L., and Lewis, B. (2011). "A Social Vulnerability Index for Disaster Management." *Journal of Homeland Security and Emergency Management*, 8(1).

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C. A., Darmenov, A., Bosilovich, M. G., Reichle, R., Wargan, K., Coy, L., Cullather, R., Draper, C., Akella, S., Buchard, V., Conaty, A., da Silva, A. M., Gu, W., Kim, G.-K., Koster, R., Lucchesi, R., Merkova, D., Nielsen, J. E., Partyka, G., Pawson, S., Putman, W., Rienecker, M., Schubert, S. D., Sienkiewicz, M., and Zhao, B. (2017). "The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2) [Dataset]." *Journal of Climate*, 30(14), 5419–5454.

Gourevitch, J. D. and Pinter, N. (2023). "Federal incentives for community-level climate adaptation: an evaluation of FEMA's Community Rating System." *Environmental Research Letters*, 18(3).

Heiman, E. R. (2022). "Protecting Renters from Flood Loss." *University of Pennsylvania Law Review*, 170(3), 783–809.

Highfield, W. E. and Brody, S. D. (2017). "Determining the effects of the FEMA Community Rating System program on flood losses in the United States." *International Journal of Disaster Risk Reduction*, 21(November 2016), 396–404.

Highfield, W. E., Brody, S. D., and Shepard, C. (2018). "The effects of estuarine wetlands on flood losses associated with storm surge." *Ocean & Coastal Management*, 157, 50–55.

Highfield, W. E., Peacock, W. G., and Van Zandt, S. (2014). "Mitigation Planning: Why Hazard Exposure, Structural Vulnerability, and Social Vulnerability Matter." *Journal of Planning Education and Research*, 34(3), 287–300.

Ismail-Zadeh, A. T., Cutter, S. L., Takeuchi, K., and Paton, D. (2017). "Forging a paradigm shift in disaster science." *Natural Hazards*, 86, 969–988.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer Texts in Statistics. Springer, New York, NY.

James, L. A. and Singer, M. B. (2008). "Development of the Lower Sacramento Valley Flood-Control System: Historical Perspective." *Natural Hazards Review*, 9(3), 125–135.

Jonkman, S. N. (2005). "Global perspectives on loss of human life caused by floods." *Natural Hazards*, 34(2), 151–175.

Knighton, J., Buchanan, B., Guzman, C., Elliott, R., White, E., and Rahm, B. (2020). "Predicting flood insurance claims with hydrologic and socioeconomic demographics via machine learning: Exploring the roles of topography, minority populations, and political dissimilarity." *Journal of Environmental Management*, 272, 111051.

Komolafe, A., Herath, S., and Avtar, R. (2018). "Sensitivity of flood damage estimation to spatial resolution." *Journal of Flood Risk Management*, 11, S370–S381.

Konrad, C. P. and Dettinger, M. D. (2017). "Flood Runoff in Relation to Water Vapor Transport by

Atmospheric Rivers Over the Western United States, 1949–2015." *Geophysical Research Letters*, 44(22), 456–11.

Kotsiantis, S., Kanellopoulos, D., and Pintelas, P. (2006). "Handling imbalanced datasets: A review." *GESTS International Transactions on Computer Science and Engineering*, 30.

Kousky, C. (2011). "Understanding the Demand for Flood Insurance." *Natural Hazards Review*, 12(2), 96–110.

Lamjiri, M. A., Dettinger, M. D., Ralph, F. M., and Guan, B. (2017). "Hourly storm characteristics along the U.S. West Coast: Role of atmospheric rivers in extreme precipitation." *Geophysical Research Letters*, 44(13), 7020–7028.

Lundberg, S. M. and Lee, S.-i. (2017). "A Unified Approach to Interpreting Model Predictions." *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA.

Merz, B., Kreibich, H., and Lall, U. (2013). "Multi-variate flood damage assessment: a tree-based data-mining approach." *Natural Hazards and Earth System Sciences*, 13, 53–64.

Merz, B., Kreibich, H., Schwarze, R., and Thieken, A. H. (2010). "Review article: "Assessment of economic flood damage"." *Natural Hazards and Earth System Sciences*, 10(8), 1697–1724.

Mobley, W., Sebastian, A., Blessing, R., Highfield, W. E., Stearns, L., and Brody, S. D. (2021). "Quantification of continuous flood hazard using random forest classification and flood insurance claims at large spatial scales: a pilot study in southeast Texas." *Natural Hazards and Earth System Sciences*, 21(2), 807–822.

Molnar, C. (2023). "Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, <https://christophm.github.io/interpretable-ml-book/>.

NOAA. "Multivariate ENSO Index Version 2 (MEI.v2), <https://psl.noaa.gov/enso/mei/>.

NOAA. "Pacific Decadal Oscillation (PDO), <https://www.ncei.noaa.gov/access/monitoring/pdo/>.

NOAA NCEI (2023). "U.S. Billion-Dollar Weather and Climate Disasters, <https://www.ncei.noaa.gov/access/billions/>.

PEP (2023). "Population and Housing Unit Estimates Tables, <https://www.census.gov/programs-surveys/popest/data/tables.html>.

Pollack, A. B., Sue Wing, I., and Nolte, C. (2022). "Aggregation bias and its drivers in large-scale flood loss estimation: A Massachusetts case study." *Journal of Flood Risk Management*, 15(4), 1–16.

Ralph, F. M., Rutz, J. J., Cordeira, J. M., Dettinger, M. D., Anderson, M., Reynolds, D., Schick, L. J., and Smallcomb, C. (2019). "A scale to characterize the strength and impacts of atmospheric rivers." *Bulletin of the American Meteorological Society*, 100(2), 269–289.

Rözer, V., Kreibich, H., Schröter, K., Müller, M., Sairam, N., Doss-Gollin, J., Lall, U., and Merz, B. (2019). "Probabilistic Models Significantly Reduce Uncertainty in Hurricane Harvey Pluvial Flood Loss Estimates." *Earth's Future*, 7(4), 384–394.

Rufat, S., Tate, E., Burton, C. G., and Maroof, A. S. (2015). "Social vulnerability to floods: Review of case studies and implications for measurement." *International Journal of Disaster Risk Reduction*, 14, 470–486.

Rutz, J. J., Steenburgh, W. J., and Ralph, F. M. (2014). "Climatological characteristics of atmospheric rivers and their inland penetration over the western united states." *Monthly Weather Review*, 142(2), 905–921.

Sadiq, A. A., Tyler, J., and Noonan, D. (2020). "Participation and non-participation in FEMA's Community Rating System (CRS) program: Insights from CRS coordinators and floodplain managers." *International Journal of Disaster Risk Reduction*, 48(September 2019).

Sadler, J., Goodall, J., Morsy, M., and Spencer, K. (2018). "Modeling urban coastal flood severity from crowd-sourced flood reports using Poisson regression and Random Forest." *Journal of Hydrology*, 559, 43–55.

Saito, T. and Rehmsmeier, M. (2015). "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets." *PLoS ONE*, 10(3), 1–21.

Sanders, B. F., Schubert, J. E., Kahl, D. T., Mach, K. J., Brady, D., AghaKouchak, A., Forman, F., Matthew, R. A., Ulibarri, N., and Davis, S. J. (2022). "Large and inequitable flood risks in Los Angeles, California." *Nature Sustainability*, 6(1), 47–57.

Smith, A. B. and Katz, R. W. (2013). "US billion-dollar weather and climate disasters: Data

sources, trends, accuracy and biases." *Natural Hazards*, 67(2), 387–410.

Solomatine, D. P. and Ostfeld, A. (2008). "Data-driven modelling: Some past experiences and new approaches." *Journal of Hydroinformatics*, 10(1), 3–22.

Stein, L., Clark, M. P., Knoben, W. J., Pianosi, F., and Woods, R. A. (2021). "How Do Climate and Catchment Attributes Influence Flood Generating Processes? A Large-Sample Study for 671 Catchments Across the Contiguous USA." *Water Resources Research*, 57(4), 1–21.

Strobl, C., Boulesteix, A. L., Zeileis, A., and Hothorn, T. (2007). "Bias in random forest variable importance measures: Illustrations, sources and a solution." *BMC Bioinformatics*, 8.

Szczyrba, L., Zhang, Y., Pamukcu, D., Eroglu, D. I., and Weiss, R. (2021). "Quantifying the Role of Vulnerability in Hurricane Damage via a Machine Learning Case Study." *Natural Hazards Review*, 22(3), 1–12.

Tate, E., Rahman, M. A., Emrich, C. T., and Sampson, C. C. (2021). "Flood exposure and social vulnerability in the United States." *Natural Hazards*, 106(1), 435–457.

Thomson, H., Zeff, H. B., Kleiman, R., Sebastian, A., and Characklis, G. W. (2023). "Systemic Financial Risk Arising From Residential Flood Losses." *Earth's Future*, 11(4), 86.

USGS (2023). "National Hydrography Dataset, <https://www.usgs.gov/national-hydrography/access-national-hydrography-products>.

Wagenaar, D., de Jong, J., and Bouwer, L. M. (2017). "Multi-variable flood damage modelling with limited data using supervised learning approaches." *Natural Hazards and Earth System Sciences*, 17(9), 1683–1696.

Willis, H., Narayanan, A., Fischbach, J., Molina-Perez, E., Stelzner, C., Loa, K., and Kendrick, L. (2016). "Current and Future Exposure of Infrastructure in the United States to Natural Hazards." *Report no.*, Santa Monica, CA.

Woldemeskel, F. and Sharma, A. (2016). "Should flood regimes change in a warming climate? The role of antecedent moisture conditions." *Geophysical Research Letters*, 43(14), 7556–7563.

Zahran, S., Weiler, S., Brody, S. D., Lindell, M. K., and Highfield, W. E. (2009). "Modeling national flood insurance policy holding at the county scale in Florida, 1999–2005." *Ecological*

Economics, 68(10), 2627–2636.

Fig. 1. Model performance metrics for the four RF models. The spatial scale is represented by color, where Sacramento is blue and statewide is gold; the temporal scale is represented by shading, where darker indicates 1981–2021 and lighter indicates 2009–2021. **(a)** Receiver Operating Characteristic (ROC) curves. A perfect model would reach the top-left corner of the plot. The solid lines are the results for the respective models at different detection thresholds between 0 and 1, and the points on the curves indicate the sensitivity and specificity of the model-optimized detection threshold. The dashed black line represents the ROC of random guessing. **(b)** Precision-recall (PR) curves. A perfect model would reach the top-right corner of the plot. The solid lines are the results for the respective models at different detection thresholds between 0 and 1, and the points on the curves indicate the precision and recall of the model-optimized detection threshold. The dashed lines represent the precision of the null models that predict no damage for all records. The precision is constant and equal to the class imbalance ratio of the test data.

Fig. 2. SHAP relative global feature importance by risk dimension and concept. Relative global feature importance is shown for both spatial resolutions (statewide and Sacramento County) and both temporal resolutions (1981–2021 and 2009–2021). The color represents the risk dimension, and the bolded percentages indicate the overall contribution of that risk dimension to the overall model performance. The shading, labeled on the right-hand side of the plot, represents the risk concept as defined in Table 1. Feature-level SHAP importance estimates are normalized so that the total for each model sums to 100%.

Fig. 3. SHAP feature importance rank. Featured are colored by their global SHAP feature contribution in each model, where the global SHAP feature contribution is calculated as the mean of all observation-level SHAP values. For example, a SHAP feature contribution of 5% means that the value of that feature increases or decreases the probability of damage by 5% on average. The top ten features with the largest SHAP feature contributions in each model are labeled with their rank. Features without shaded bars were either removed during feature selection or had smaller global SHAP values than the random noise variable. The H/E/V labels along the left side of the plot indicate whether each feature is related to the hazard, exposure, or vulnerability dimension.

Fig. 4. Top three hazard, exposure, and vulnerability features in the 1981 statewide model. The top three most important hazard features are **(a)** total precipitation, **(b)** AR maximum IVT, and **(c)** AR duration. The top three most important exposure features are **(d)** total housing units, **(e)** percent of the population living in the 100-year floodplain, and **(f)** percent of the housing stock as single family homes. The top three most important vulnerability features are **(g)** CRS score, **(h)** median household income, and **(i)** CalEnviroScreen pollution burden score. The left and right Y-axes represent the change in damage probability relative to the average probability of damage. Black lines represent ALE curves that show the average trend between predictor and response; ALE curves are plotted over the middle 95% of the data to reduce the visual impact of outliers. Gray points represent SHAP values for 1,000 individual observations randomly sampled from the training set.

Fig. 5. Interactions between AR category and exposure and vulnerability features in the 1981 statewide model. Exposure features are **(a)** total housing units, **(b)** percent of the population living in the 100-year floodplain, and **(c)** percent of housing stock as single-family homes. Vulnerability features are **(d)** CalEnviroScreen Pollution Burden score, **(e)** median household income, and **(f)** CRS score. ARs of different categories are colored based on the legend at the bottom. The tick marks along the bottom of each panel indicate the distribution of values for that feature. ALE curves are plotted over 95% of the data to reduce the impact of outliers, and lines are plotted with a smoothing factor to reduce visual clutter.

TABLE 1. Predictor variables. This table includes variables that were retained through feature selection in at least one of the four models. Variables are grouped by risk dimension and variable concept. Spatial resolution is at the county scale (C), census tract scale (T), or constant (–). Temporal resolution is by event (E), monthly (M), annual (A), or constant (–). The Data Source column lists a citation where data for the variable can be retrieved, the Justification column lists a citation supporting the variable's inclusion in the model.

Risk Dimension	Concept	Variable	Spatial	Temporal	Data Source	Justification
Hazard	AR characteristics	Maximum IVT (kg/m/s)	T[a]	E	Gelaro et al. (2017)	Corringham et al. (2019)
		Duration (hours)	T[a]	E	Rutz et al. (2014)	Corringham et al. (2019)
		AR category	T[a]	E	Ralph et al. (2019)	Corringham et al. (2019)
		Total precipitation (mm)	T[a]	E	Gelaro et al. (2017)	Brunner et al. (2018)
	Antecedent conditions	3- & 14-day total precipitation prior to AR event (mm)	T[a]	E	Gelaro et al. (2017)	Woldemeskel and Sharma (2016)
		3-day mean soil moisture prior to AR event (mm/m)	T	E	Erlingis et al. (2021)	Cao et al. (2020)
	Climate modes	El Niño Southern Oscillation (ENSO)	–	M	NOAA (a)	Corringham and Cayan (2019)
		Pacific Decadal Oscillation (PDO)	–	M	NOAA (b)	DeFlorio et al. (2013)
	Land surface	Impervious land cover (%)	T	Y[c]	Dewitz and USGS (2021)	Blum et al. (2020)
		Developed land cover (%)	T	Y[c]	Dewitz and USGS (2021)	Brody et al. (2013)
		Wetlands land cover (%)	T	Y[c]	Dewitz and USGS (2021)	Highfield et al. (2018)
Exposure	Population	Population density per square mile	T[b]	Y	PEP (2023)	Jonkman (2005)
		Population within FEMA floodplain (%)	T	–	FEMA (2020)	Sanders et al. (2022)
	Housing	Housing units	T[b]	Y	PEP (2023)	Willis et al. (2016)
		Single-family homes (%)	T	–[d]	ACS (2023a)	Rufat et al. (2015)

Risk Dimension	Concept	Variable	Spatial	Temporal	Data Source	Justification
	Insurance takeup	Riverside census tract indicator	T	–	USGS (2023)	Kousky (2011)
		Coastal county indicator	C	–	USGS (2023)	Kousky (2011)
Vulnerability	Socioeconomic	CDC SVI components	T	Y[c]	CDC (2022)	Bakkensen et al. (2017)
		CalEnviroScreen metrics	T	–	August et al. (2021)	Bergstrand et al. (2015)
		Median household income (2022 dollars)	T[b]	Y	ACS (2023a)	Cutter (2016)
		Non-Hispanic white population (%)	T[b]	Y	ACS (2023b)	Cutter (2016)
		Working-age population (%)	T[b]	Y	ACS (2023b)	Cutter (2016)
	Infrastructural	Housing units over 40 years old (%)	T	–[d]	ACS (2023a)	Highfield et al. (2014)
		Median housing age (years)	T	–[d]	ACS (2023a)	Knighton et al. (2020)
		Owner-occupied housing units (%)	T	–[d]	ACS (2023a)	Rufat et al. (2015)
		Mobile homes (%)	T	–[d]	ACS (2023a)	Tate et al. (2021)
	Flood experience	3-year lagged total disaster declarations	C	Y	FEMA (2023b)	Bakkensen et al. (2017)
		Community Rating System (CRS) score	C	Y	FEMA (2023a)	Highfield and Brody (2017)

[a] Values are aggregated to the tract/county level from MERRA-2 (~50×50 km grid cells) (Gelaro et al. 2017).
[b] Pre-2000 tract-level estimates are based on a weighted distribution of county/statewide values, where weights are determined as a function of the distribution of tract values in 2000.
[c] Data is only available starting from 2000 (for the CDC SVI) or 2001 (for the land surface variables), so for prior years the values are equal to those recorded in the earliest year of data.
[d] Data is only available starting from 2009, so these are considered to be temporally constant in the 1981–2021 models and annually varying in the 2009–2021 models.

TABLE 2. Dataset statistics by spatial and temporal scale.

Spatial Scale	Temporal Scale	Number of Records	Damaging Records	Class Balance
Statewide	1981–2021	17,265	863	5.0%
Statewide	2009–2021	5,659	234	4.1%
Sacramento	1981–2021	145,977	874	0.60%
Sacramento	2009–2021	47,776	114	0.24%

TABLE 3. Random forest model performance metrics. Fitted model performance (*RF*) is compared against null model performance (*Null*) for each model and each metric under consideration.

Spatial Resolution	Temporal Resolution	ROC-AUC		PR-AUC		Balanced Accuracy	
		RF	*Null*	*RF*	*Null*	*RF*	*Null*
Statewide	1981–2021	0.914	0.500	0.435	0.051	0.707	0.500
Statewide	2009–2021	0.914	0.500	0.353	0.041	0.658	0.500
Sacramento	1981–2021	0.959	0.500	0.311	0.007	0.761	0.500
Sacramento	2009–2021	0.898	0.500	0.046	0.002	0.702	0.500