Predicting Flood Damages using Machine Learning and National Flood Insurance Program Data

Azara Boschee[1], Tom Corringham[2], Weiming Hu[3]

[1]St. Cloud State University, St. Cloud, MN, [2]Center for Western Weather and Water Extremes, Scripps Institution of Oceanography, La Jolla, CA, [3]James Madison University, Harrisonburg, VA

CW3E · UC San Diego · SCRIPPS INSTITUTION OF OCEANOGRAPHY · ST. CLOUD STATE UNIVERSITY

Introduction

- Predicting when, where, and how much flood damage may occur can allow for better damage mitigation and response efforts by governments (Ghaedi et al. 2022).
- Federal Emergency Management Agency (FEMA) flood insurance rate maps (FIRMs) can estimate where damage is probable to occur overall; however, FIRMs are outdated and are potentially under-representative of where flood damage is likely to occur (Kousky 2018).
- This project assesses the ability of random forest models to predict flood damages reported to the National Flood Insurance Program (NFIP) at a 1-degree resolution in locations across California.
- Random forest is used because it is easy to visualize and can handle datasets with complex relationships between variables while maintaining comparable accuracy and efficiency to other models (Yang et al. 2022).

Data and Methods

- Daily data from 1978 – 2011 across the contiguous United States (CONUS) was available for the predictive variables shown in **Table 1.** This allowed for 26 years of data for model training (1978-2004) and 6 years of testing (2005-2011).
- Two models were trained using the randomForest package in R: a regression and binary classification. All parameters were kept to their defaults due to simplicity and time restraints.
- For the classification, a value of '1' was assigned to cases with damage and '0' to cases without damage.
- Since flood damage is infrequent, most cases are '0', which may cause the model to rely on the percentage of occurrences rather than predictors. Random undersampling with the ROSE package in R was utilized to address this. As a result, the model was only trained on 2,560 cases out of the possible ~158,000 cases.

Variable	Units	Grid Cell Aggregation	Original Resolution	Data Source
Real Damage[1]	**US Dollars**	**Total in grid cell**	**1/10th Degree**	**FEMA**
Runoff	mm/day	Maximum in grid cell	1/16th Degree	Variable Infiltration Capacity (VIC) hydrological model data from Livneh et al. (2013)
Baseflow	mm			
Snow Water Equivalent (SWE)	mm			
Soil Moisture	mm			
Precipitation	mm			
Wind (gust)	m/s			
Urban Land Cover	%	Percent of area in grid cell	30 meters	CEC NALCMS (2023)
Insurance Policies		Total in grid cell	1/10th Degree	NFIP FOIA Request

Table 1. Information about variable to be modelled (Real Damage in bold) and predictive variables.
[1]Cost adjusted for 2022 Quarter 4 Inflation using data from US BEA

Results: Regression Model

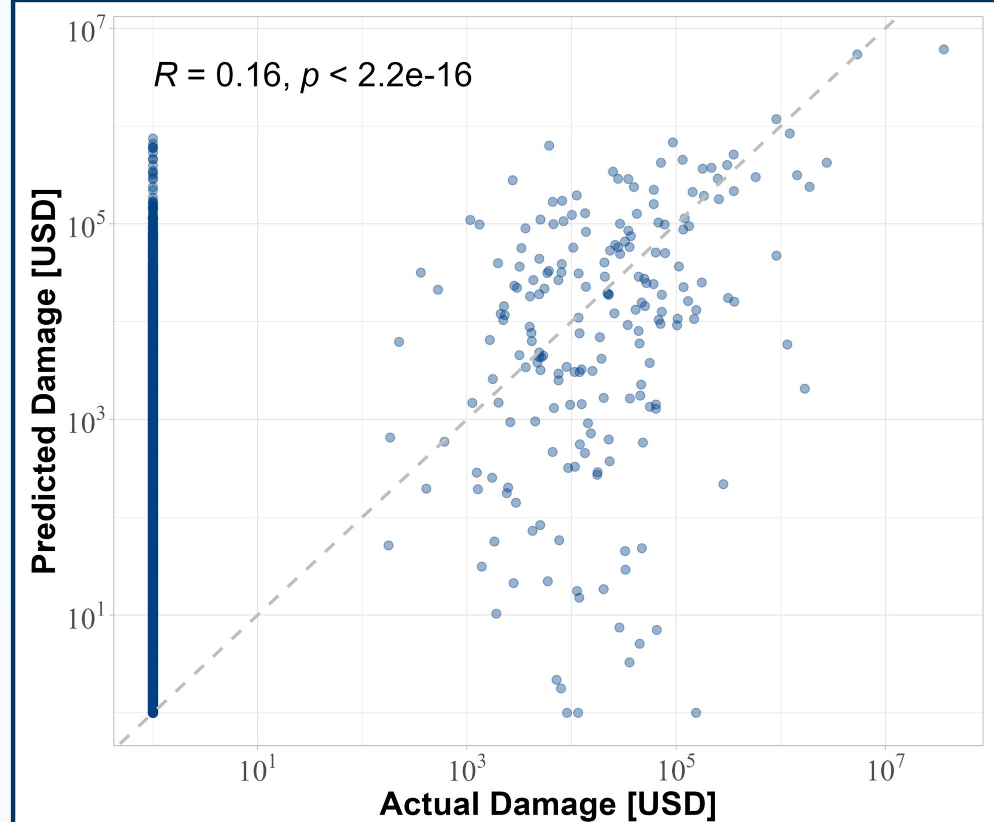

Figure 1. *(above) Scatter plot compares predicted & actual damage January 2005 - December 2011 in CA. Damage costs on Log10 scale. Grey dashed line shows where they are equal.*

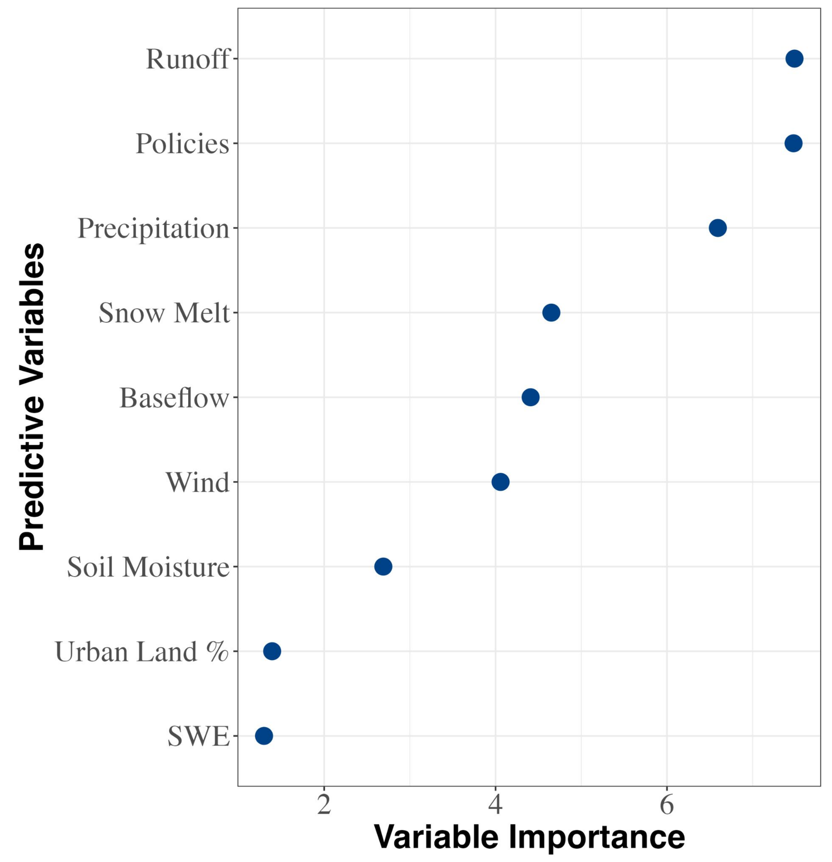

Figure 2. *(above) RF regression model variable importance plot, measured as % Increase in Mean Square Error.*

Root Mean Squared Error	150,800 USD
Mean Absolute Error	2,100 USD

Table 2. (above) RF regression model statistics

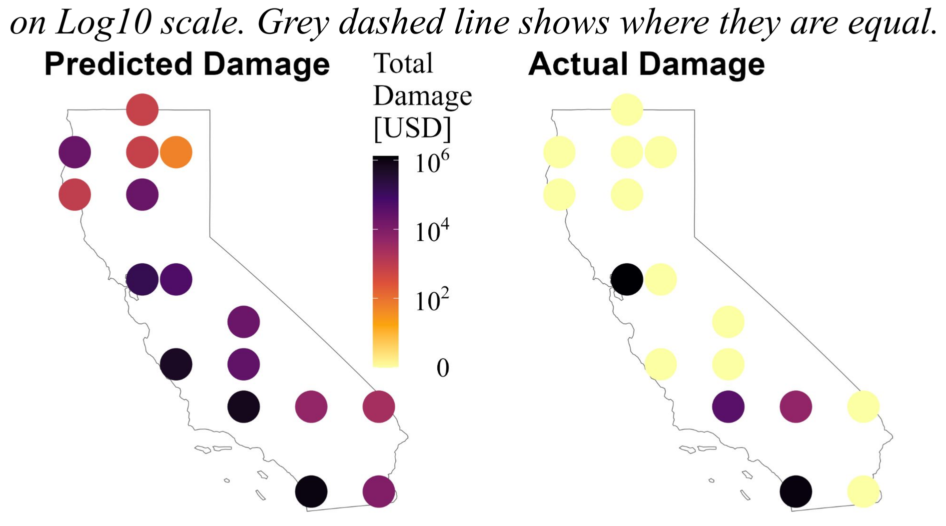

Figure 3. *(left) Maps of California comparing total predicted & actual damage during the January 1st - 15th, 2005 Los Angeles County Flood. Damage on Log10 scale.*

Results: Classification Model

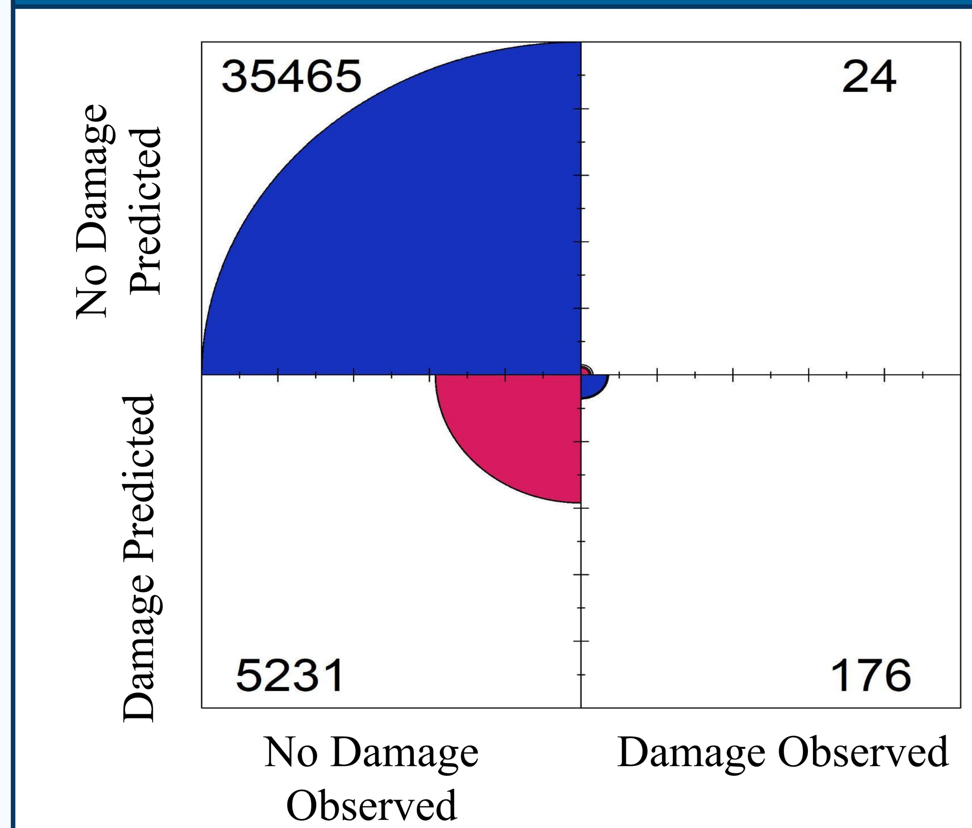

Figure 4. *(above) Four fold plot showing the confusion matrix for classification model results of whether damage occurred or not for January 2005 - December 2011 in California.*

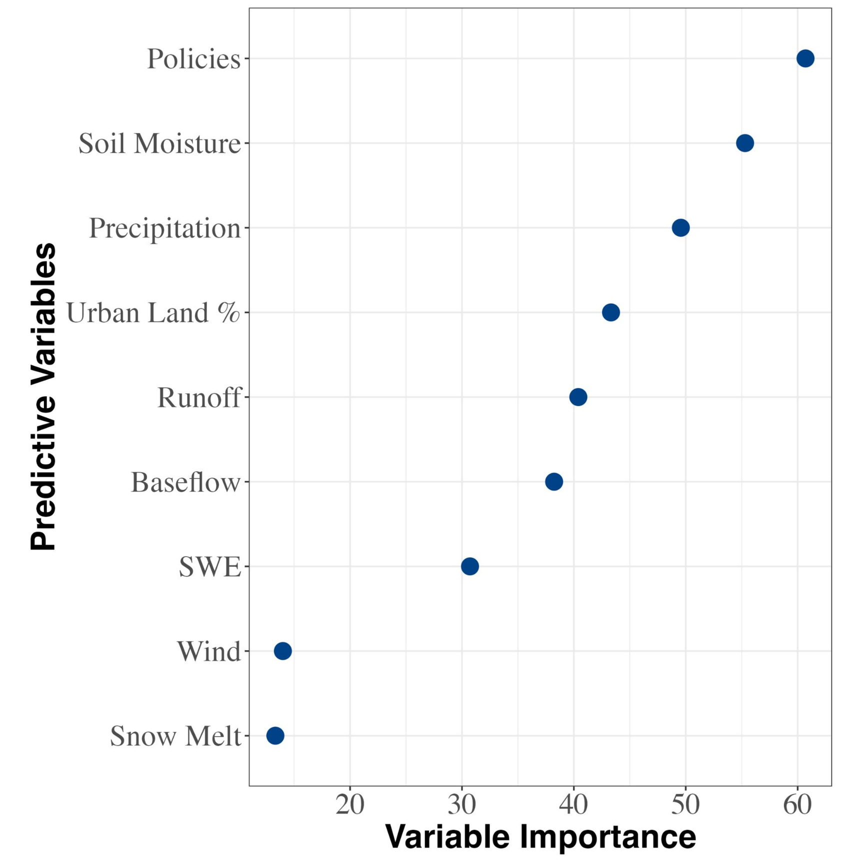

Figure 5. *(above) RF binary classification model variable importance plot, measured as Decrease in Mean Accuracy.*

Accuracy	87.15%
F1 Score	93.1 %

Table 3. (above) RF classification model statistics

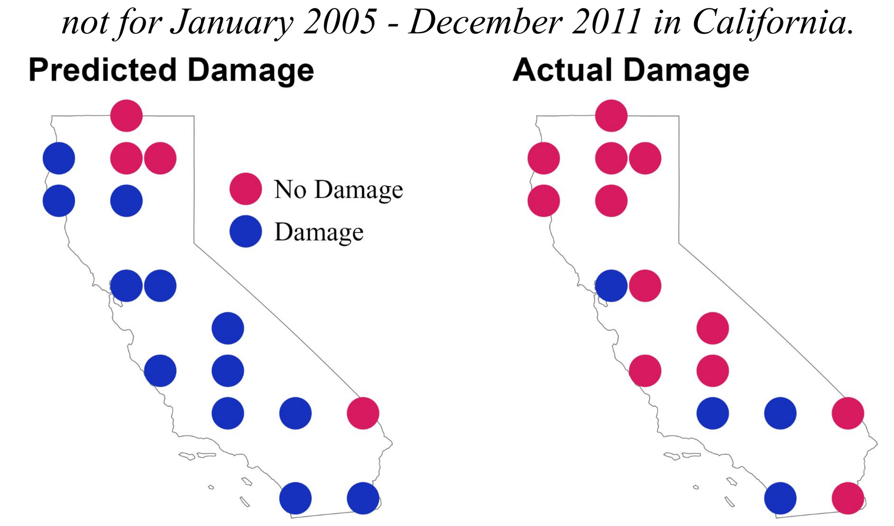

Figure 6. *(left) Maps of California comparing where damage was predicted & actually occurred during the January 1st - 15th, 2005 Los Angeles County Flood.*

Discussions and Limitations

- The regression model overpredicts ~90% of the test data where most are cases in which no damage actually occurs. It also cannot predict values that are $0 but predicts many values less than $1.
- The classification model performs better than the regression model with an accuracy of 87%.
- Variable importance differs between models due to different importance metrics. However, runoff, number of insurance policies, soil moisture, and precipitation affected model accuracy the most while snow melt and SWE were the least impactful.
- NFIP policies are distributed unevenly across the US preventing the models from predicting damage in locations without policies. Policy distribution is most unevenly distributed in the Western US, including California. (**Figure 7**)

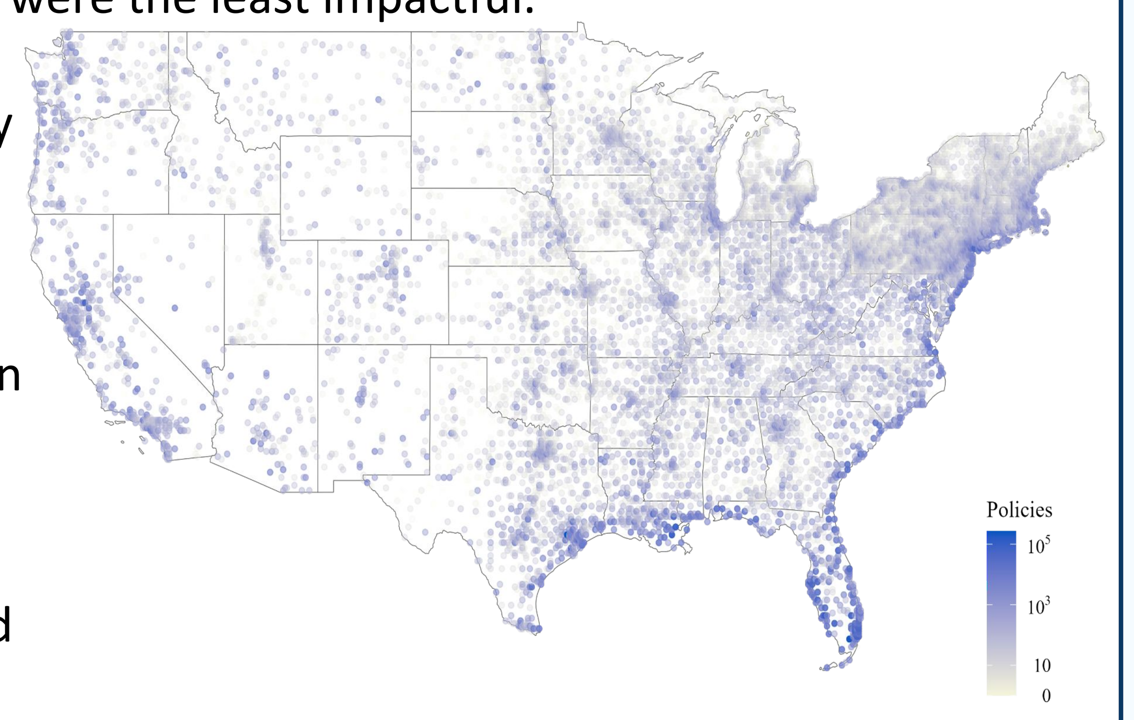

Figure 7. *Map showing the number of policies across CONUS in 2011 on a 0.1 degree resolution. The number of policies is on a log10 scale.*

Future Work

- This project is continuing by changing the domain and resolution of the machine-learning model to predict over CONUS at 0.1 degrees.
- This model will use a binary classification that predicts cases in which damages occur and then estimates the damage costs with a regression model as in Ghaedi et al. (2022).
- The current models' parameters were set to the defaults. These models could be improved with hyperparameter optimization.
- The current model uses data from VIC reanalysis data and only contains data through 2011. This will be replaced with the National Water Model retrospective dataset which allows for analysis through 2020.

References

CEC/NALCMS, 2023: North American Land Cover, 2020 (Landsat, 30m), Accessed 17 July 2023, http://www.cec.org/north-american-environmental-atlas/land-cover-30m-2020/

FEMA: Damage and Claim data from the OpenFEMA FIMA NFIP Redacted Claims - v2 Dataset. Subset used: January 1978 - December 2011, accessed 5 June 2023, https://www.fema.gov/openfema-data-page/fima-nfip-redacted-claims-v2#

This product uses the FEMA OpenFEMA API, but is not endorsed by FEMA. The Federal Government or FEMA cannot vouch for the data or analyses derived from these data after the data have been retrieved from the Agency's website(s).

Ghaedi, H., A. C. Reilly, H. Baroud, D. V. Perrucci, and C. M. Ferreira, 2022: Predicting flood damage using the flood peak ratio and Giovanni Flooded Fraction. *PLOS ONE*, **17**, e0271230, https://doi.org/10.1371/journal.pone.0271230.

Kousky, C., 2018: Financing Flood Losses: A Discussion of the National Flood Insurance Program. *Risk Manage. Insur. Rev.*, **21**, 11–32, https://doi.org/10.1111/rmir.12090.

Livneh, B., E.A. Rosenberg, C. Lin, B. Nijssen, V. Mishra, K.M. Andreadis, E.P. Maurer, and D.P. Lettenmaier, 2013: A Long-Term Hydrologically Based Dataset of Land Surface Fluxes and States for the Conterminous United States: Update and Extensions, *J. Climate*, **26,** 9384–9392.

Livneh daily CONUS near-surface gridded meteorological and derived hydrometeorological data data provided by the NOAA PSL, Boulder, Colorado, USA, from their website at https://psl.noaa.gov

US Bureau of Economic Analysis (BEA): Personal Consumption Expenditures (Implicit Price Deflator) (DPCERD3Q086SBEA): Subset used: Q1 1978 - Q4 2022, accessed 5 June 2023, https://fred.stlouisfed.org/series/DPCERD3Q086SBEA

Yang, Q., and Coauthors, 2022: Predicting Flood Property Insurance Claims over CONUS, Fusing Big Earth Observation Data. *Bull. Amer. Meteor. Soc.*, **103**, E791–E809, https://doi.org/10.1175/BAMS-D-21-0082.1.