Predicting Flood Damages and Property Risks in the US

Team 152 – Allen Yen Chen, Yi-Chun Chen, Kevin Hsu, Parth Patel, David Harold Thompson

Introduction

In recent years, the effects of climate change have become more disastrous, as the frequency and severity of flooding events have increased. These events cause significant financial losses for property owners, as damage to their homes is often unavoidable. Accurate predictions of such losses are crucial to homeowners and potential property buyers, so adequate prevention and protection measures can be set up to minimize these costs.

Problem Definition

Property owners have been lacking a reliable and personalized flood risk assessment. Studies are often performed to only provide general information about the levels of flood risk in a given area, but personalized information about each individual's property damage is missing. The objective of this project is to analyze the relationship between floods and economic damage to real property in the United States. Specifically, the analysis will allow users to utilize the characteristics of their property to predict the damage their property may incur in the event of a flood.

Literature Survey

Climate change has led to a significant increase in sea level, with the rate doubling from 1993 to 2009 (Mimura, 2013). In addition, climate change has increased the intensity and frequency of extreme weather events (Ebi et al., 2021). These papers' findings suggest that the climate is constantly changing, but their research is limited to analyzing economic damage at the regional level. Both fail to provide personalized information about financial damage experienced by individual property owners.

It is estimated that the population in the US exposed to a "100-year flood", a flood that has a 1% chance of being exceeded in any given year, would increase by 20 million by 2049 (Swain et al., 2020). Stults (2017) states that by 2050, over 40% of Federal Emergency Management Agency (FEMA) flood maps could be outdated. Not only could the unknown magnitude of future floods be a concern, but there is a lack of research on the stability of flood control infrastructure as it ages, leaving even controlled areas vulnerable to unknown risks (Salman et al., 2018). These papers suggest there could be higher flood risks than anticipated, but neither one recommends a solution for property owners to protect their properties.

Historical trends show property exposure to floods has a tremendous impact on flood insurance payouts. A $1 billion increase in property exposure raises the annual National Flood Insurance Program (NFIP) payout by $4.8 million (Bhattacharyya et al., 2024). In fact, areas with the highest flood hazards typically have higher property values (Yiannakoulias et al., 2018). This is an extreme burden on insurance companies, who may choose to transfer this burden to insurers by increasing insurance premiums. Year-to-year variations in flood peaks cause fluctuation in household insurance premiums (Nguyen et al., 2024). The papers only conclude that higher flood risk can impact property owners financially, therefore our analysis serves as a valuable reference for them to plan for future spikes in insurance costs.

Although federal insurance programs may cover a portion of the damage caused by floods, there is ongoing discussion about the shift in liabilities from the government to individual property owners. Hurricane Katrina raised the question about whether the government or property owner is responsible for the losses (Bergsma, 2019). Governments in Canada have already begun to transfer the liabilities away from publicly funded disaster assistance programs to homeowners due to increasing amounts of unavoidable floods (Thistlethwaite et al., 2018). These papers show that beyond just paying higher premiums, homeowners may have to bear the cost of flood damage completely in the future, causing the cost of property ownership to increase as governments change their policies. To bridge the gap between the findings from these papers and the potential flooding each homeowner may face, our analysis helps them quantify the damage to be expected given the characteristics of their properties. In the event that damages are no longer covered by insurance, our analysis provides value to homeowners in determining whether additional protection is worth investing in to keep their property safe.

However, it is worth noting the difficulty in making such estimates, as Quinn et al. (2019) show that recent floods have shown increasing variability in footprint size, even in areas with high-standard flood defenses. Wei et al. (2004) attempt to estimate the level of vulnerability a region is to natural disasters but using only the region's population and economic activity as input, which they acknowledge is insufficient. Current research has not found an optimal model in projecting the economic losses resulting from floods (Kaito et al., 2021). Although our analysis considers unique property characteristics, it should be approached with caution as there is no proven way of predicting the outcome of such events.

Past research has been reviewed to assess if similar flood damage analysis has been conducted. Yang et al. (2021) predict the number of flood property insurance claims in the US with strong results for county-level predictions. Armal et al. (2020) estimate the average NFIP payout in New Jersey using state data over a 40-year period, which resulted in extremely high accuracy. Multiple papers focus on this topic with similar findings but only provide insight into losses at the regional level. None predicts damage like this analysis does, which provides estimates tailored to each individual's property.

Proposed Method

For the purposes of helping homeowners make personalized damage estimates, the original dataset selected for this project includes over 2 million NFIP claim transactions since 1979. Since only residential property claims are of interest, approximately 1.2 million records and the 73 features that describe each residential property's characteristics are analyzed. Statistical models to be considered for predicting property damage include random forest, regression, gradient boosting, and k-nearest neighbors, all to be built using Python. Linear regression is used as a baseline to assess whether a linear relationship exists and is expected to perform significantly worse than the other three due to the complex relationship among the variables. The random forest algorithm utilizes predictions of numerous decision trees to improve the accuracy of its predictions. k-NN calculates the average of the *k* closest training data points to form a prediction. Lastly, the gradient boosting model is capable of learning and updating its predictions. These model characteristics suggest that more accurate predictions will be made than the linear regression model. The use of a large number of individual claim transactions will make the estimates more personalized than research conducted in the past that only provided general information.

As part of the innovations of this project, users of this analysis will be able to:

1. input the characteristics of their properties into the optimal model, which will estimate the potential building damage during a flood, and
2. visualize the level of predicted property damage at the state level based via an interactive Tableau map.

The interactive Tableau map is especially useful for home buyers interested in analyzing numerous properties. It uses the predictions generated from the optimal model and aggregates them at the state level using the average. Sliders can be adjusted to further analyze the predicted damage at a more granular level based on the feature values, allowing potential homebuyers to evaluate predicted damage of their potential homes in various dimensions.

Due to the large number of variables, many of which were suspected to be closely correlated with each other, only a subset of the 73 variables were kept. Some variables were excluded, such as "ID", because they were deemed unnecessary and provided no value to the analysis. Other variables such as "Net Building Payment Amount" were removed as this information is only available after flooding and insurance payments occur, which would not be available yet when predicting future damage amounts. The "Reported Zip Code" variable was removed as it provided similar geographic information of the property as the "Longitude" and "Latitude" variables.

Next, exploratory data analysis was performed on the dataset and a correlation plot was created. For easier visualization, the table below includes only variables with strong correlation.

Feature 1	Feature 2	Correlation Value
buildingReplacementCost	buildingPropertyValue	0.951
lowestFloorElevation	baseFloodElevation	0.813
constructionYear	postFIRMConstructionIndicator	0.633
obstructionType	elevatedBuildingIndicator	0.525
year	floodZoneCurrent	-0.495

Figure #1: Correlation Table

To confirm if any of the remaining variables needed to be removed, a forward selection stepwise regression was performed. The results showed that all the 36 variables remaining had statistical significance. Therefore no additional variables were removed.

Date-related variables were separated into Month and Year columns to provide better model interpretability. Most missing values for variables were filled using state-level medians to preserve regional trends and minimize bias introduced by null values. However, data points missing critical values that were unsuitable to be filled, such as "Latitude" and "Original Construction Date", were completely removed.

For model building, the "Building Damage Amount" was selected as the target variable. The data was split into training and testing sets by the date the flood damage occurred, as indicated by the "Date of Loss" variable. The dataset was split into a training set that consisted of records with a "Date of Loss" prior to September 2019 and a testing set with records after September 2019, resulting in a 90-10 ratio. Since the project's objective is to estimate near-future damage amounts, this split ensured that the performance of the models, although trained on old data, would have their performance be evaluated

using more recent data. It is reasonable to assume that models with better prediction performance on recent data are more likely to make better predictions on near-future data, which is the goal of this project.

The performance of each model was measured using metrics such as the Coefficient of Determination (R^2), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) on both training and test sets, in order to assess model accuracy, error rates, generalization capabilities, and provide insights into any overfitting or underfitting issues.

Experiment/Evaluation

Questions this experiment answers include:

1. Can any of the four models accurately predict potential flood damage?
2. What are the most important variables in making predictions?

The random forest and boosting model results answered the first question, as they produced relatively high R^2s and the lowest RMSEs and MAEs. Due to the complexity of the random forest algorithm and size of the dataset, the model had to use a graphics processing unit (GPU) to train the model on a local machine. Tuning of the parameters consisted of using a grid search, which included the maximum number of leaf nodes, maximum depth, etc. The model with the optimal parameters converged to have a 0.30 R^2. the lowest RMSE, and second lowest MAE for the test data.

The gradient boosting model achieved similar results, with the highest R^2, second lowest RMSE and the lowest MAE on the testing set. The optimal parameters were selected to produce the lowest RMSE. Although its results were similar to the random forest model's, the computational resources that the boosting model demanded were significantly less than what the random forest model required. With such a large number of features, it was expected that random forest and gradient boosting would outperform the other models, as they are generally capable of handling large, high-dimensional datasets due to their bagging and learning capabilities, respectively.

The linear regression model produced a rare negative R^2 value for both training and testing sets, which indicates that it was capturing almost no variance in the target variable and suggests that linear relationships may not be sufficient to model the complexity of this data. Additionally, the high RMSE and MAE values imply a relatively large error margin in predicting the building damages amount.

The k-Nearest Neighbors (k-NN) model produced extremely poor results. Despite experimenting with different values of k, the model consistently failed to achieve a positive R^2. As such, this model was deemed unsuitable for predicting property damage given a property's characteristics. The poor performance of the k-Nearest Neighbors model can be attributed to how the data was split into training and testing sets. The training set includes records from 1979 to 2019, while the testing set spans 2019 to 2024. Consequently, the nearest neighbors used for the testing set predictions may have relied on outdated data that do not account for inflation and evolving factors influencing property damage.

The table below shows the performance of each model by the metrics identified earlier.

	R^2 (Train)	R^2 (Test)	RMSE (Train)	RMSE (Test)	MAE (Train)	MAE (Test)
Linear Regression	0.01	-67.01	597,209	414,503	23,631	50,810
Random Forest	0.60	0.30	123,142	39,293	14,224	15,826
Boosting	0.37	0.32	80,240	42,571	13,179	13,316
KNN (k=5)	0.26	-278.25	515,902	839,939	15,983	36,485
KNN (k=10)	0.15	-189.94	555,232	694,555	17,510	37,637
KNN (k=15)	0.09	-236.39	573,517	774,427	18,250	44,515

Figure #2: Model results

An observation made when analyzing the results of the gradient boosting and random forest model is that many of the same variables were ranked highly important for each respective model. These include the “Building Property Value”, “Longitude”, “Latitude”, and “Construction Year” variables. Although both models took different approaches when making estimates, the overlap of the most important variables shows that predicting property damage amounts depends highly on certain characteristics of the house. The information from such variables is valuable such that using either the random forest model or boosting can most likely generate similar results.

The boosting and random forest model had the best performance out of all the models evaluated in this project. Although their performance is comparable, the boosting model was selected as the optimal model for making personalized damage estimates due to its fast run time. An additional user interface was created as part of this project to allow users to input information of the 36 features of their property into this gradient boosting model, which will then calculate the predicted damage amount.

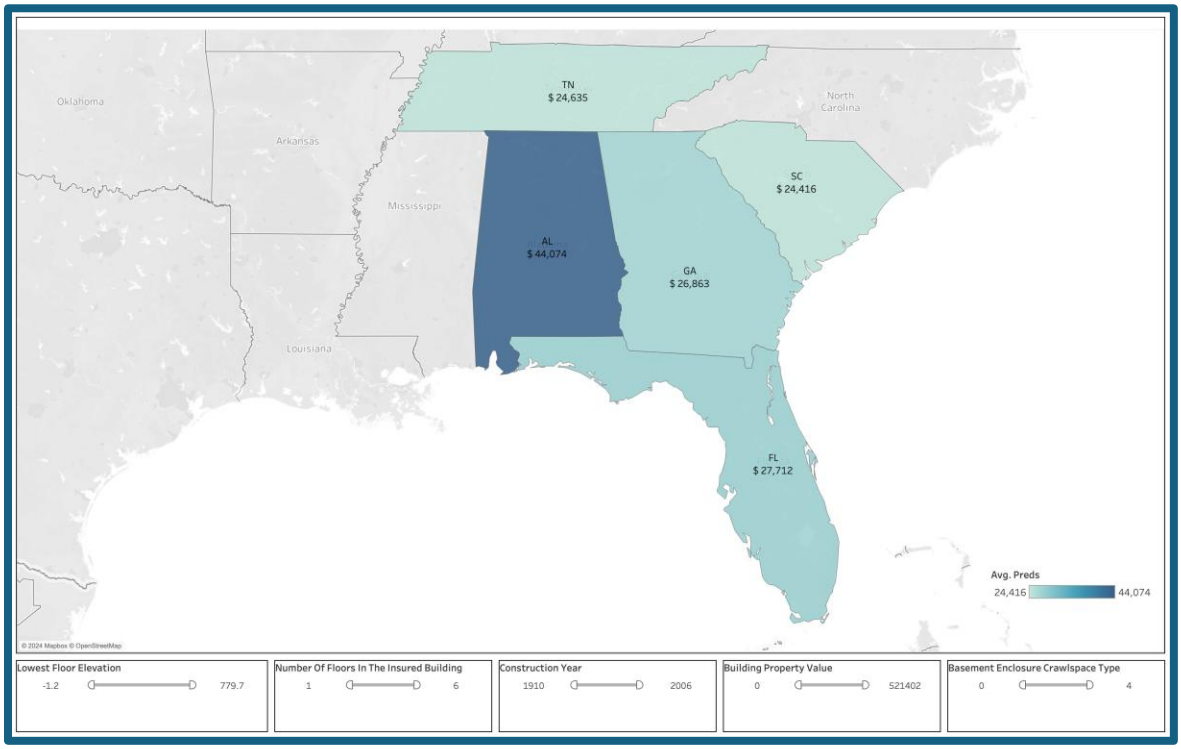

Figure #3: Average predicted damage for selected properties in TN, AL, GA, SC, FL

To demonstrate how the interactive map complements the predictions made by the boosting model, a few properties from the testing set were chosen for the map below to mimic how a hypothetical homebuyer would use this analysis. Typically, homebuyers compare multiple properties first before purchasing one. An initial analysis can incorporate the boosting model to estimate the potential flood damage to these homes. The map then presents the average of such predictions at the state level. Adjusting the sliders (only five features shown in Figure #3 as an example) lets the user filter certain values for each selected feature, which would simultaneously change the average predicted damage on the corresponding states. Potential homebuyers can assess how flood risk changes in the state when comparing certain characteristics of the houses of interest. The Tableau map is a powerful tool for potential homebuyers to understand the flood risk damage they may face when selecting the appropriate home for them.

Conclusions and Discussions

The gradient boosting and random forest models performed relatively well, with the highest R^2 and lowest RMSE and MAE, but the boosting model's efficiency outperformed. K-Nearest Neighbors was also not an appropriate model to use for this dataset, which can be attributed to how the data was split between the training and testing set. If another method was adopted for splitting the data, where both the training and testing set contained recent data, the performance of k-NN is expected to be much better.

The next step of this project is to explore the performance of these models with just a subset of variables. Such detailed information about each residential property may not always be available, so it is important to identify which variables are sufficient to provide robust predictions without sacrificing too much on accuracy. Another area to investigate is the effect of inflation on the estimates. Levels of inflation varied from 1979 to 2024, which could have caused the predictions to be inaccurate. Adjusting monetary values in the dataset to real values prior to splitting the dataset may lead to more accurate predictions.

All team members contributed a similar amount of effort.

Bibliography

Armal, S., Porter, J., Lingle, B., Chu, Z., Marston, M., & Wing, O. (2020). Assessing Property Level Economic Impacts of Climate in the US, New Insights and Evidence from a Comprehensive Flood Risk Assessment Tool. *Climate*, *8*(10), 116. https://doi.org/10.3390/cli8100116

Bergsma, E. (2019). The development of flood risk management in the United States. *Environmental Science & Policy*, *101*, 32–37. https://doi.org/10.1016/j.envsci.2019.07.013

Bhattacharyya, A., & Hastak, M. (2024). Empirical causal analysis of flood risk factors on U.S. flood insurance payouts: Implications for solvency and risk reduction. *Journal of Environmental Management*, *352*, 120075. https://doi.org/10.1016/j.jenvman.2024.120075

Ebi, K. L., Vanos, J., Baldwin, J. W., Bell, J. E., Hondula, D. M., Errett, N. A., Hayes, K., Reid, C. E., Saha, S., Spector, J., & Berry, P. (2021). Extreme weather and climate change: Population health and health system implications. *Annual Review of Public Health*, *42*(1), 293–315. https://doi.org/10.1146/annurev-publhealth-012420-105026

Kaito Kotone, Taniguchi, K., Nakamura, K., & Takayama, Y. (2021). Estimation of Potential Economic Losses Due to Flooding Considering Variations of Spatial Distribution of Houses and Firms in a City. *Journal of Disaster Research*, *16*(3), 329–342. https://doi.org/10.20965/jdr.2021.p0329

Mimura, N. (2013). Sea-level rise caused by climate change and its implications for society. *Proceedings of the Japan Academy, Series B*, *89*(7), 281–301. https://doi.org/10.2183/pjab.89.281

Nguyen, V. D., Aerts, J., Tesselaar, M., Botzen, W., Kreibich, H., Alfieri, L., & Merz, B. (2024). Exploring the use of seasonal forecasts to adapt flood insurance premiums. *Natural Hazards and Earth System Sciences*, *24*(8), 2923–2937. https://doi.org/10.5194/nhess-24-2923-2024

Quinn, N., Bates, P. D., Neal, J., Smith, A., Wing, O., Sampson, C., Smith, J., & Heffernan, J. (2019). The Spatial Dependence of Flood Hazard and Risk in the United States. *Water Resources Research*, *55*(3), 1890–1911. https://doi.org/10.1029/2018wr024205

Salman, A. M., & Li, Y. (2018). Flood Risk Assessment, Future Trend Modeling, and Risk Communication: A Review of Ongoing Research. *Natural Hazards Review*, *19*(3), 04018011. https://doi.org/10.1061/(asce)nh.1527-6996.0000294

Stults, M. (2017). Integrating climate change into hazard mitigation planning: Opportunities and examples in practice. *Climate Risk Management*, *17*, 21–34. https://doi.org/10.1016/j.crm.2017.06.004

Swain, D. L., Wing, O. E. J., Bates, P. D., Done, J. M., Johnson, K. A., & Cameron, D. R. (2020). Increased Flood Exposure Due to Climate Change and Population Growth in the United States. *Earth's Future*, *8*(11). https://doi.org/10.1029/2020ef001778

Thistlethwaite, J., Minano, A., Blake, J. A., Henstra, D., & Scott, D. (2018). Application of re/insurance models to estimate increases in flood risk due to climate change. *Geoenvironmental Disasters*, *5*(1). https://doi.org/10.1186/s40677-018-0101-9

Wei, Y.-M., Fan, Y., Lu, C., & Tsai, H.-T. (2004). The assessment of vulnerability to natural disasters in China by using the DEA method. *Environmental Impact Assessment Review*, *24*(4), 427–439. https://doi.org/10.1016/j.eiar.2003.12.003

Yang, Q., Shen, X., & Yang, F. (2021). Predicting Flood Property Insurance Claims over CONUS, Fusing Big Earth Observation Data. *Bulletin of the American Meteorological Society*, *103*(3). https://doi.org/10.1175/bams-d-21-0082.1

Yiannakoulias, N., Darlington, J. C., Elshorbagy, A., & Raja, B. (2018). Meta-analysis based predictions of flood insurance and flood vulnerability patterns in Calgary, Alberta. *Applied Geography*, *96*, 41–50. https://doi.org/10.1016/j.apgeog.2018.05.007