Maya Barron

Min Chen

Forest & Wildlife Ecology 458

8 May 2025

Investigating the Prevalence of and Fatalities Caused by Tsunamis

GitHub

https://github.com/mrbarron3/fwe458/tree/main/final_project

Introduction

Tsunamis are large natural weather events that often hold great potential to cause tragedy due to their unpredictable nature and mere hours of warning time between detecting the event and the tsunami making landfall. Somewhat recent advancements in technology around the globe have allowed for mid-ocean earthquakes, and the resulting tsunamis, to be detected only five to seven minutes after they form. The focus of more recent warning and mitigation efforts has centered around the Indian Ocean as a result of the tragedy that took place there in 2004. These efforts have included the implementation of 1,400 sea level monitoring stations and 75 DART buoys tracking pressure changes on the seafloor in this ocean alone (Frost).

At this point, tsunamis are just about as predictable before they form as earthquakes are before they strike, likely because most tsunamis are initiated by earthquakes. Scientists know which earthquakes and other events are likely to cause tsunamis, but they are not able to predict an event before it occurs. Tsunami Warning Centers currently use data on underwater earthquake occurrences to determine if a tsunami warning message should be issued to the general public. The key pieces of data they look at immediately are earthquake location, depth, and magnitude, which are tracked by the sea level monitoring stations and the DART buoys. Warning messages are updated using real-time data collected from those sensors and by using simulated scenarios mapping out possible ocean movements and estimated coastal impacts. Tsunamis initiated by other events (landslides, volcanoes, etc.) are much harder to detect early on and can thus make landfall with little to no warning (National Tsunami Warning Center).

The motivation behind this project is to get a better understanding of this devastating natural disaster that impacts so many individuals across a wide range of countries and continents. This project will aim to predict the deadliness of any given tsunami, which could be useful for evacuation efforts. Additionally, this project will investigate how and why the overall number of tsunamis that form is so different from year to year. The goal of this project is to help those with the potential to be hit by a tsunami stay up to date so they can make informed decisions about mitigation efforts and evacuation plans.

Methods

This project makes use of three different datasets to visualize and model the previous points. The global historical tsunami database details tsunamis occurrences from 2000 B.C. to 2020 A.D.. This dataset comes from an offshoot of NOAA’s dataset that can be found on Kaggle. The ocean heat content dataset has information on the yearly change from 1957 in the ocean heat content for the entire ocean, as well as just the northern hemisphere and southern hemisphere. This dataset comes straight from NOAA. Finally, the world population dataset has information about the top 1000 worldwide cities in terms of population. This data comes from Peter Hull on Infocartography. I combined all three of these datasets to use in my machine learning algorithms. Each tsunami in the historical tsunami dataset had a latitude and longitude listed indicating where the natural disaster struck, which I used to find the closest city out of the top 1000 cities by population to help better predict the casualties one of these events could cause. I also added in the change in ocean heat content compared to 1957 for all tsunamis that occurred from 1958 onwards to better understand the role that climate change is playing in the prevalence of tsunamis. I used the latitude column again here to determine whether the tsunami struck in the northern or southern hemisphere so I could add in the most applicable change in ocean heat content data for the given year. The final resulting dataset was sorted by year ascending for ease of readability, namely in determining where missing values were coming from. The resulting dataset had a significant amount of missing values, which can be seen in Figure 1 by examining the yellow segments which each represent one piece of missing data. Opposingly, each purple bar represents a non-missing piece of data. All missing values were not immediately dropped from the dataset; instead, I removed rows containing missing values after selecting the most important features for my machine learning models. After looking through the data, I decided to only keep tsunamis that were listed as "probable" or "definite" events to avoid as many inaccuracies in the data as I could while also keeping enough data to model and predict tsunamis on.

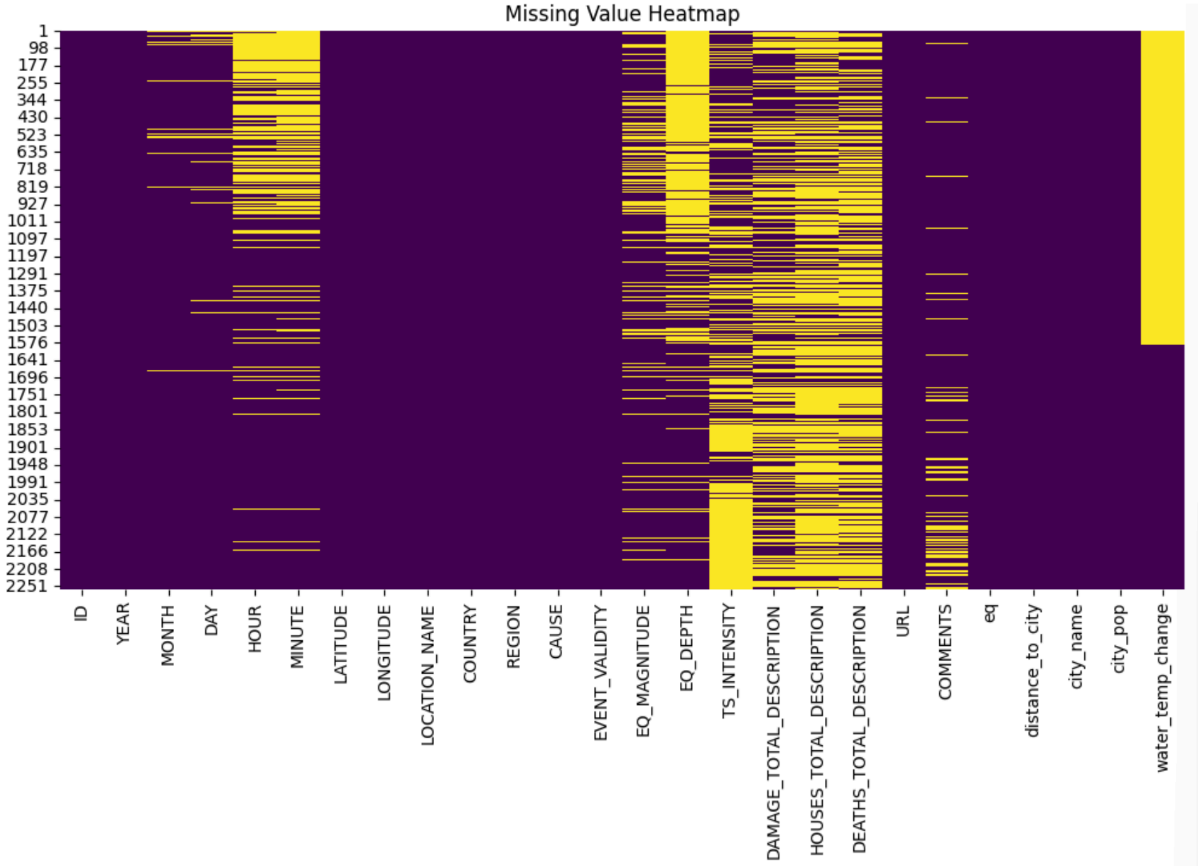

Figure 1. Heatmap highlighting missing values in the dataset.

Model set 1:

I began my process of investigating the number of fatalities caused by tsunamis by mapping out the location of each tsunami and coloring each strike by the category of fatalities the event caused (Fig. 2) to give me a better understanding of the data behind the question I was looking to answer. For my first set of models, I wanted to predict the fatality category a given tsunami would fall into based on some features of the tsunami, surrounding population, and ocean heat content in the hemisphere the tsunami struck in. I removed the non-descriptive features in the dataset including URL, COMMENTS, and ID, and scaled and encoded the remaining features. I used a k-nearest neighbors model and permutation feature importance and found that the two most important predictive features were the distance to the nearest big city and the population of that nearest city. I added in a few more features that I personally felt might aid in predicting the number of casualties a tsunami causes, including the month, year, latitude, and longitude of the event, the magnitude of the earthquake if the tsunami was caused by an earthquake, and the change in ocean heat content from 1957 in the hemisphere the tsunami struck in (either northern or southern). I kept all of my selected features the same within this first set of models, as well as between the second set. I split my dataset into random training and testing sets with 20% of the data in the testing set. I used k-nearest neighbors, decision tree, and random forest classification models to predict the fatalities a tsunami would cause.

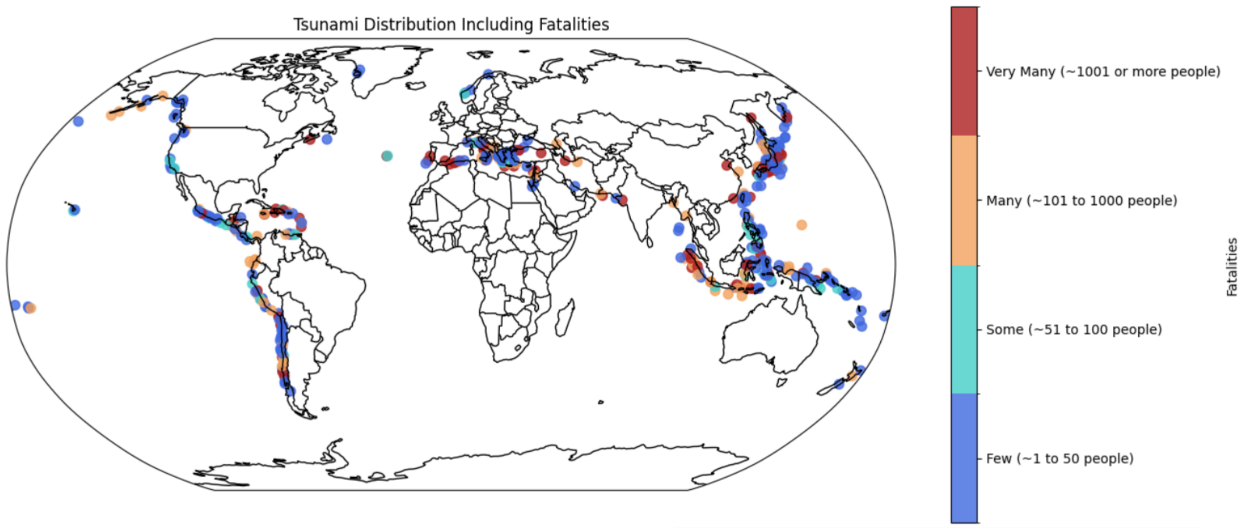
Tsunami Distribution Including Fatalities
Very Many (~1001 or more people)
Many (~101 to 1000 people)
Some (~51 to 100 people)
Few (~1 to 50 people)
Fatalities

Figure 2. Robinson world projection detailing the location of tsunami strikes and the category of casualties inflicted by each event.

Model set 2:

In an effort to produce a more accurate model, I changed my problem into a Boolean classification problem by creating a new column in the DataFrame detailing the death severity of the event where “many” and “very many” fatalities were assigned a 1 (101+ fatalities, “severe”) and “few” and “some” fatalities were assigned a 0 (1-100 fatalities, “less severe”). All of the same features from the previous set of models were used to model here as well, and the dataset was split so a random 20% of the data was in the testing dataset. I resampled the training data so there was an even split of “severe” (1) and “less severe” (0) events so the models were less likely to just predict “less severe” (0) for all tsunamis. I used k-nearest neighbors, decision tree, random forest, and logistic classification models to predict whether a tsunami would cause a “less severe” number of fatalities (0 marker) or a “severe” number of fatalities (1 marker).

Model set 3:

I then moved away from this question and examined how the prevalence of tsunamis has changed over time. I made a new DataFrame with the yearly count of tsunamis, and I joined this DataFrame with the DataFrame detailing the yearly average changes in ocean heat content relative to 1957 for the entire ocean, just the northern hemisphere, and just the southern hemisphere. Because the ocean heat content change dataset only had data from 1958 onwards, I omitted all older tsunami counts from my model. I also eliminated tsunamis that were not marked as “probable” or “definite” events. I split the DataFrame into training and testing sets by holding out the last 10 years in the data (2011-2020) to be used as testing, with everything else being used as training. I used the northern hemisphere and southern hemisphere ocean heat content features mentioned above, as well as year, in all of my regression models to predict how many tsunamis would be formed in a given year. The overall world ocean heat content change feature I mentioned above is simply the addition of the northern and southern hemisphere ocean heat content changes and is thus not providing any new information to the regression (Fig. 3). I used linear regression, extreme gradient boosting, decision tree, and random forest models to predict how many tsunamis would form in a given year. I also extended my predictions to predict how many tsunamis were formed in 2021 and 2022 since my ocean heat content dataset provided information for those years as well.

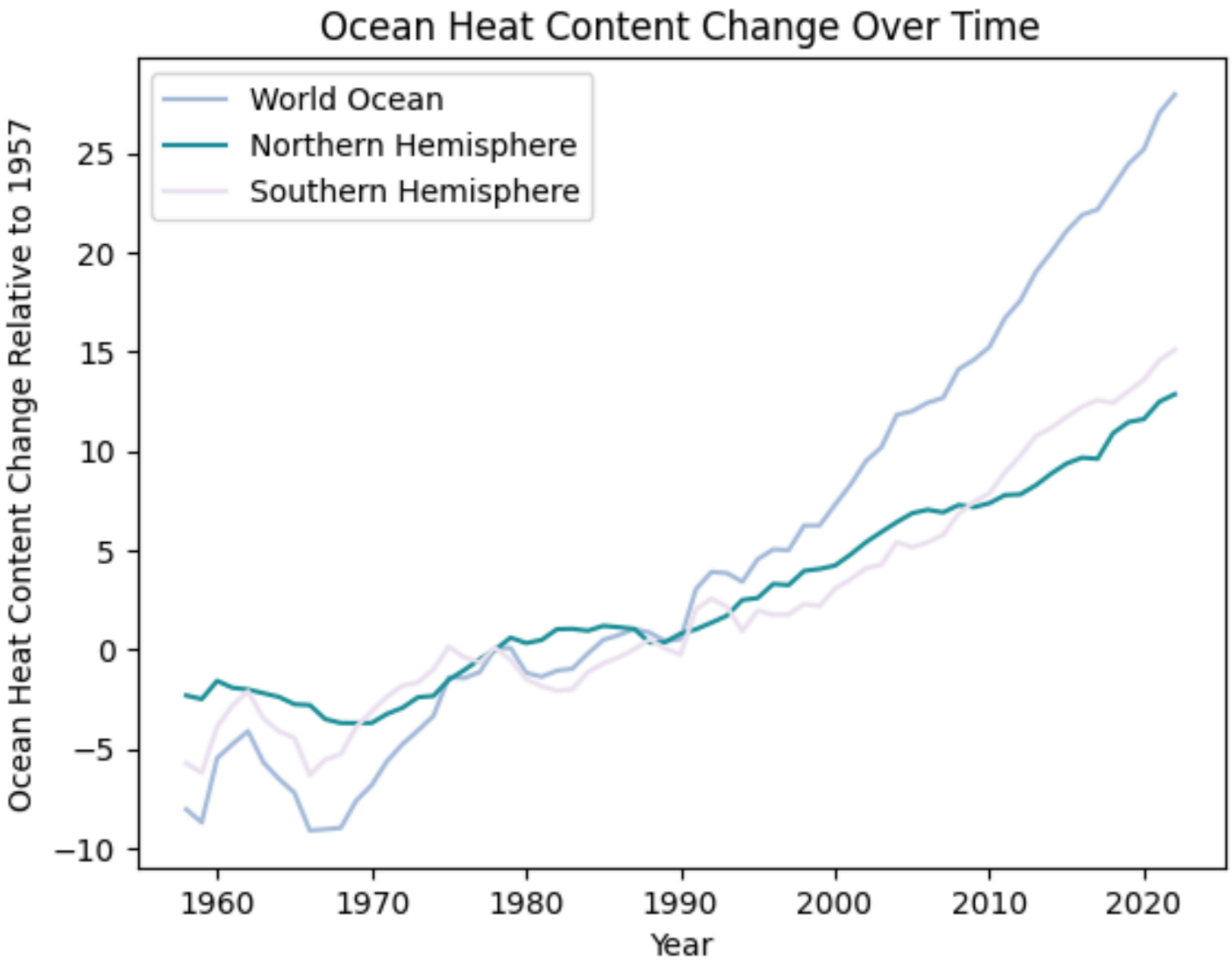

Figure 3. Line plot illustrating the change in ocean heat content relative to 1957 for the entire ocean, just the northern hemisphere, and just the southern hemisphere.

Results

Model set 1:

The features that had the most influence on these models were the distance to the nearest large city and the population of that city (Fig 4), but neither of these features were solely responsible for the decision making of these models, and thus the accuracy scores. These models performed worse than I wanted, which led me to the decision to turn this 4-class classification problem into the 2-class classification problem seen under model set 2. Overall, the random forest model performed the best on average across the board, with the k-nearest neighbors model edging it out in both the accuracy and recall metrics (Table 1).

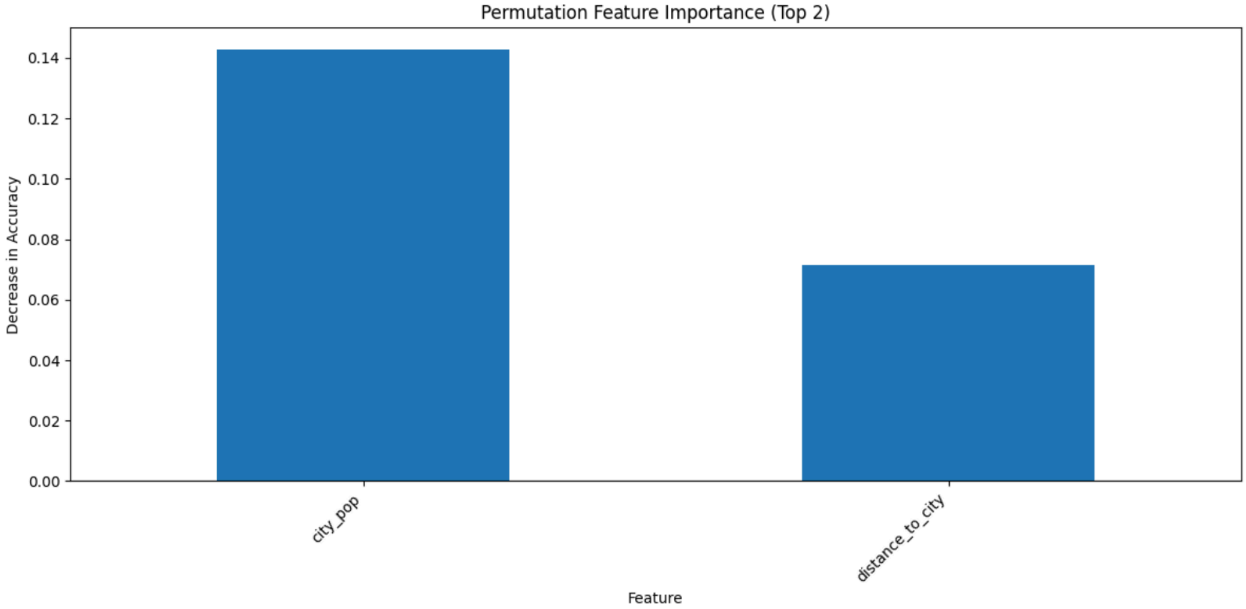

Figure 4. Bar plot showing the two most important features by decrease in accuracy for predicting the fatalities caused by a tsunami.

Model	Accuracy	F1-score	Precision	Recall
K-nearest neighbors	0.58	0.42	0.33	0.58
Random forest	0.55	0.48	0.48	0.55
Decision tree	0.45	0.44	0.43	0.45

Table 1. Performance of models predicting the classification of tsunami fatalities into 4 classes: “Very Many (~1001 or more people”, “Many (~101 to 1000 people)”, “Some (~51 to 100 people)”, and “Few (~1 to 50 people)”.

Model set 2:

Even with oversampling to make the training predicted death severity variable an even split between “severe” (1) and “less severe” (0) events, the logistic classification model always predicted that a given tsunami would fall under the “less severe” classification. I predicted that these models would all outperform the previous models since the classification problem was simplified, but I was mistaken in my judgement (Table 2). Overall, the decision tree model performed the best out of the 4 models, and it was the only one that outperformed all of the model set 1 machine learning algorithms in a majority of the performance metrics listed.

Model	**Accuracy**	**F1-score**	**Precision**	**Recall**
K-nearest neighbors	0.55	0.4	0.31	0.56
Logistic classification	0.73	0.00	0.00	0.00
Random forest	0.73	0.18	0.50	0.11
Decision tree	0.79	0.59	0.63	0.56

Table 2. Performance of models predicting the classification of tsunami fatalities as “severe” (101+ fatalities), indicated by a 1 in the dataset, or “less severe” (1-100 fatalities), indicated by a 0 in the dataset.

Model set 3:

With the change in ocean heat content since 1957 rising yearly since 1990 with no indication of stopping (Fig. 3), the number of tsunamis every year on average is predicted to increase by all of my models. None of my models performed quite as well as I was expecting or hoping for, even after altering the variables given to the models. Always predicting the average number of tsunamis would have yielded better results than any of the models I used, as evidenced by the negative R-squared scores (Table 3). Because of this, all results from these models should be backed up with further research before drawing any definitive conclusions. Overall, the linear regression model performed the best out of the four models, although it didn't perform well. My decision tree and random forest regression models performed the worst and exactly the same as each other, which was quite a surprising result.

Model	**R-squared**	**Mean squared error**
Linear regression	-0.10	16.74
Extreme gradient boosting	-0.14	17.34
Decision tree	-0.41	21.50
Random forest	-0.41	21.50

Table 3. Performance of models predicting the yearly number of tsunamis formed.

Discussion

With so many missing variables throughout the combined dataset, it was difficult to make full use of every dataset involved in this project. I stand by my decision to not immediately omit all rows with missing values from the working DataFrame, as that move would have significantly reduced the number of rows I had to work with. I believe that further data exploration could have uncovered some better correlations between the features and the variables of interest, especially the original (model set 1) fatality classification variable of interest. Working off of that point, it's possible that the best features were not selected for each individual model in this project, as I chose to keep the selected features the same within and between all of model set 1 and model set 2, as well as the same within just model set 3 by itself. I believe some of these models could have performed better if they were able to select and use the features that were more important to them. I believe this would have mainly applied to the logistic classification model from model set 2, which ended up always predicting that a tsunami would inflict “less severe” fatality numbers.

Based on the results from my three different sets of models, I believe that more research is needed before drawing any useful and definitive conclusions, especially any conclusions concerning tsunami prevalence over time. A wider range of variables would be helpful in predicting both the fatalities a tsunami may cause as well as the number of tsunamis that are likely to form in a given year. Going from here, I propose that making use of the data from the same systems used to detect tsunamis would be a good next step. With there being 75 DART buoys in each ocean as well as over 1,400 sea level monitoring stations in the Indian Ocean alone, these datasets are likely very large and would require a much greater amount of time to use to model and predict. I believe that the models in this project were a great first step in tsunami exploration and prediction, but that a wider range of variables must be brought into play to understand the full scope of both what causes tsunamis to be so deadly and how tsunami prevalence is likely to change in the future. Unfortunately, research like this is quite difficult because the data we have available is quite sparse, as tsunamis don’t occur that frequently, as compared to some other natural disasters such as tornadoes, earthquakes, and wildfires. Nonetheless, additional resources/variables could definitely improve the predictions the models in this project made.

References

Frost, Rosie. “How the 2004 Indian Ocean Tsunami Became a “Wake up Call” for Early Warning Systems.” *Euronews*, Euronews, 23 Dec. 2024, www.euronews.com/green/2024/12/23/how-the-2004-indian-ocean-tsunami-became-a-wake-up-call-for-early-warning-systems. Accessed 8 May 2025.

Hull, Peter. “Top 1,000 Cities by Population.” *Infocartography*, 2022, infocartography.com/world-top1000-pop. Accessed 7 May 2025.

NASA. “Ocean Heat Content | NASA Global Climate Change.” *NASA*, NASA, Dec. 2023, climate.nasa.gov/vital-signs/ocean-warming/?intent=121.

National Tsunami Warning Center. “U.S. Tsunami Warning Centers.” *Usa.gov*, 2015, www.tsunami.gov/?page=tsunamiFAQ.

“Tsunami Dataset.” *Kaggle*, 2024, www.kaggle.com/datasets/andrewmvd/tsunami-dataset. Accessed 7 May 2025.