Narrative Intelligence and Impact Signatures of Flash Flood Events in D.C., Maryland, and Virginia

Jessica Lin, Danielle Sitalo, Emma Zou

Project Group 4

George Washington University

June 13, 2025

Introduction

Flash flooding is becoming a growing threat to urban and suburban communities alike across the United States, especially in coastal and low-lying regions. In recent decades, the frequency as well as severity of these flood events have increased, caused by the global climate change and, thus, rising sea levels. Scientific studies have shown that even a small increase in sea level (between 10 to 20 centimeters) can more than double the frequency of extreme coastal flooding events within just a few decades [4, 5]. This significant increase can be seen across the East Coast where densely populated cities often face challenges to their infrastructure, public and private property, and public safety. As climate change progresses, understanding and adapting to flood events is crucial for cultivating resilience in Washington, D.C., and its surrounding states.

The primary goal of this project is to improve understanding of the impacts of the flash floods in Washington, D.C., and its neighboring counties in Maryland and Virginia by utilizing data science techniques. We focus on events from 1996 to the present that include narrative description, with the primary objectives to extract structured impact information from unstructured event narratives using large language models (LLMs), compute a Flood Impact Score (FIS) for each event, and analyze contextual patterns such as co-occurence of impact types, spatial clustering, and temporal trends. Ultimately, we intend to obtain findings which can be used to inform resilience strategies and emergency planning by identifying high-impact flood scenarios.

This project is motivated by the need to support local governments and emergency planners with actionable information regarding flood risks under the context of the rapidly changing climate and environment. As seen in recent research, the odds of exceeding the critical water-level thresholds in coastal regions increase exponentially with sea-level rise, meaning that once believed to be rare and extreme flooding events are becoming far more frequent and common [4]. By systematically analyzing and quantifying the impacts of floods, this project aims to help drive data-informed decision-making and effective mitigation strategies. Furthermore, as structured and detailed data is not as common as narrative data, our methods will be able to extend the application of impact-quantification to events that may only have narrative records.

Methods

Our data was procured from the National Oceanic and Atmospheric Administration's (NOAA) Storm Events Database [3], which provides historical information on U.S. weather events with enough intensity to significantly disrupt property, commerce, infrastructure, and/or human life. Each datapoint represents an event, with features such as location, time, and most notably, narrative. The narratives are split between event and episode narratives, but they essentially serve the same purpose as qualitative descriptions of each flood event. Each names specific locations, impacts, and various other data such as precipitation levels and financial loss.

We focused on flash flood data in the D.C., Maryland, and Virginia area from 1996 to February 2025, which resulted in a total of 5558 data points. Events which lacked narratives were removed, resulting in 5406 remaining rows. A total of 21 labels, consisting of 15 impact tags and 6 weather tags, were created to describe and categorize events. Impact tags describe human and societal consequences of flood events, while weather tags describe the types of weather phenomena involved in the events (Fig. 1).

Tag	Description
death	Human fatalities
injury	Physical injuries
evacuation	Any form of population displacement
rescue	Emergency or civilian rescues
car_crash	Explicit vehicle collisions or crashes
home_damage	Flooded or damaged residences
infrastructure_damage	Damage to roads (includes flooding), bridges, water plants, tunnels, public transit systems, etc.
soft_infrastructure_damage	Damage to structures hospitals, schools, community buildings (churches, libraries, etc.)
road_closure	Blocked roads or intersections
power_outage	Loss of electricity or power infrastructure
tree_damage	Damage to trees
vehicle_loss	Anything that results in the loss/damage of a vehicle
agricultural_damage	Losses in crops, farmland, barns
animal_loss	Animal deaths
campground_damage	Trail and camp erosion/damage
nor_easter	Northeastern storm
thunderstorm	Thunderstorm
hurricane	Hurricane
tornado	Tornado, twister
lightning	Lightning and/or fire
mudslide	Mudslide, landslide

Figure 1: *List of impact tags and weather tags, and their descriptions.*

We created a test dataset by hand-labeling the first 900 data points in the dataset. Some of the 22 final tags were proposed before labeling, while others were added throughout the process. Using insights gained from the labeling process, each tag was assigned a corpus of synonyms and relevant words. Labels with wider definitions tended to have longer corpora as well, while shorter ones sufficed for narrowly defined impact tags and the more self-explanatory weather tags.

Each narrative was cleaned using tokenization, stopword removal, punctuation removal, and lemmatization. While we initially used stemming rather than lemmatization, we ultimately chose the latter method because of its ability to return words to their base form yet ensure each is still a valid English word. Items in the tag corpora were cleaned using the same methods, in order to standardize the language used between narratives and corpora.

A simple classification algorithm, which assigns tags to a given row based on whether the row's associated narrative contains items in the respective corpora, was created. To test it, we ran the classification model on our labeled testing data and calculated the accuracy for each tag. The few tags whose accuracies fell below 80% were revisited, and their corpora were revised to be more descriptive and representative. After our corpora and algorithm achieved sufficient accuracy, we applied the algorithm to all remaining unlabeled data.

In order to quantify the total impact of each event in a comparable way, we created a weighted function on the tags to assign a Flood Impact Score (FIS) to each event. Each impact tag was assigned a weight between 0 and 1 (Fig. 2).

Tag	Weight
death	1
injury	0.9
evacuation	0.8
rescue	0.8
infrastructure_damage	0.75
home_damage	0.7
soft_infrastructure_damage	0.65
power_outage	0.65
car_crash	0.5
road_closure	0.4
tree_damage	0.4
agricultural_damage	0.4
campground_damage	0.3
vehicle_loss	0.2
animal_loss	0.2

Figure 2: *List of tags and weights in descending order.*

In deciding the weights, we looked to the Federal Emergency Management Agency's (FEMA) methodology of assessing damage for assistance programs [2], which prioritizes human impact (death, injury, evacuation, etc.) and personal property damage as indicators of damage level. Impacts on normal community functions—such as disruptions of education, hospitals and treatment facilities, and other public institutions—were also ranked highly.

These weights were used as coefficients in a simple linear equation. The maximum possible value without scaling was 8.65, and eventually, all FIS scores were scaled such that the values were on a scale of 1 to 10. The weighting function was applied to each datapoint, resulting in a FIS score for each flood event.

Results

Accuracy

To assess the accuracy of the labeling/tagging on the different narratives, we took the proportion of narratives correctly matched between the manually labeled data frame and the classification model data frame. From the results, the model created performs quite well, with the majority of the tags being correct 80% of the time. However, our *infrastructure_damage* tag is only accurately labeled about 35% of the time. This could be because the words in the corpus used were not entirely representative of infrastructure damage, especially with many words appearing closely with *road_closure*, causing possible misclassifications (Fig. 3).

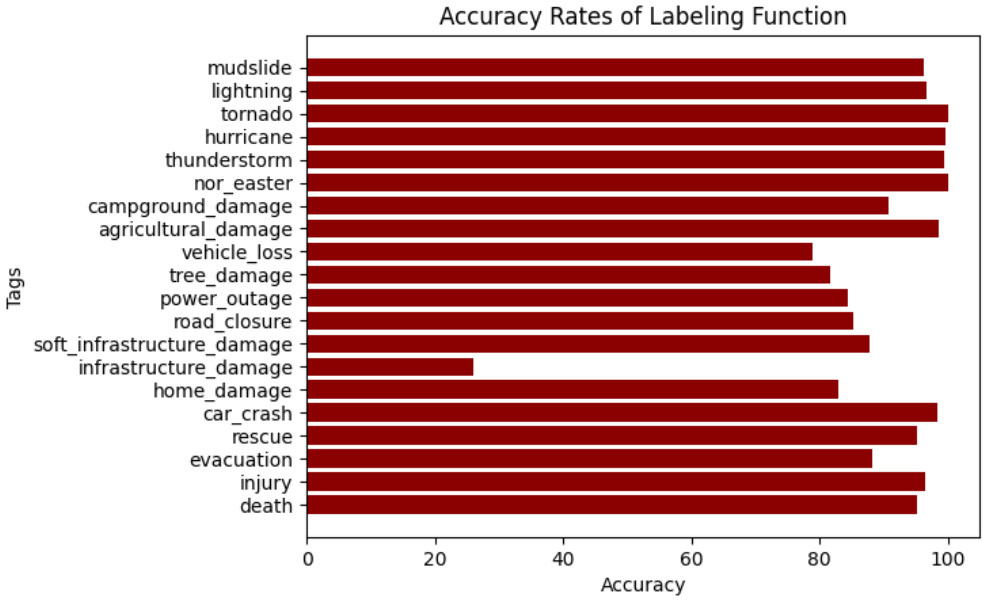

Figure 3: *Bar graph showing the accuracy (from 0 to 100 percent) of tag labels for the flash flood events.*

Frequency of Tags

Total Frequency

To determine the prioritization of different resilience solutions to flooding, we found and compared frequencies of the different impact tags using a bar chart (Fig. 4). Overall, we see that *road_closure* was the most frequently occurring impact factor, with *vehicle_loss*, *tree_damage*, and *home_damage* following it as the most commonly reported damages from a flood event.

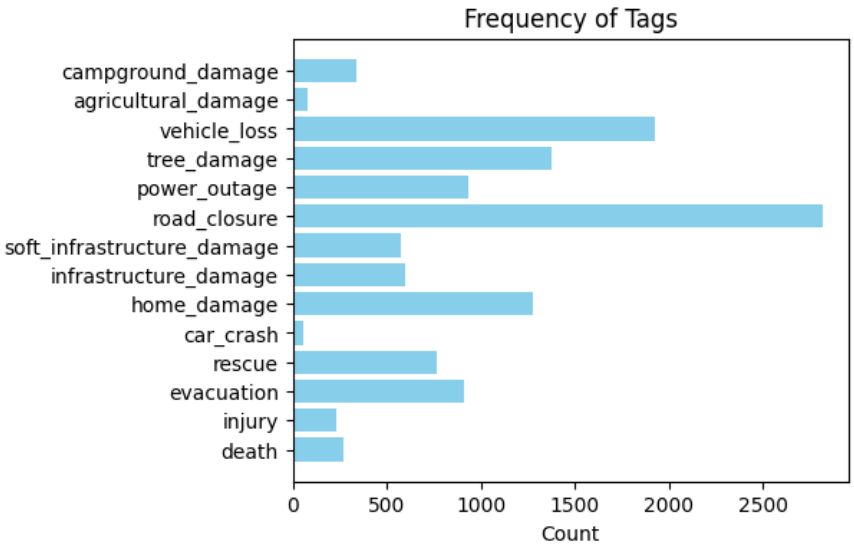

Figure 4: *Bar graph showing the frequency of impact tags across NOAA–documented flash flood events in D.C., Maryland, and Virginia.*

Similarly, we also determined which extreme weather events were most frequently reported with the flash floods (Fig. 5). From the visualization, thunderstorms appear to be the most common. An important observation is that lightning does not always occur with thunderstorms, as *lightning* was tagged less than 500 times while *thunderstorm* was tagged over 3500 times. Mudslides and hurricanes were tagged about the same amount as lightning.

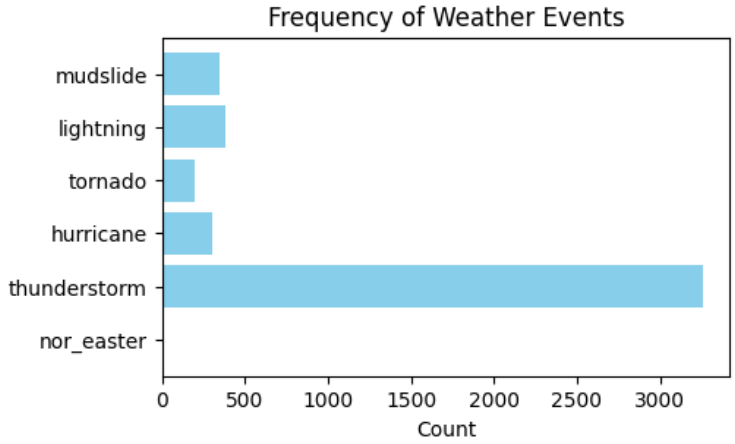

Figure 5: *Bar graph showing the frequency of impact tags across NOAA–documented flash flood events in D.C., Maryland, and Virginia. Note that the scaling of the count makes the frequency of nor_easter tags appear as 0, but there are at least two cases documented in the narrative observations.*

Region Breakdown

Following total frequency, we broke down the analysis further by region (Fig. 6).

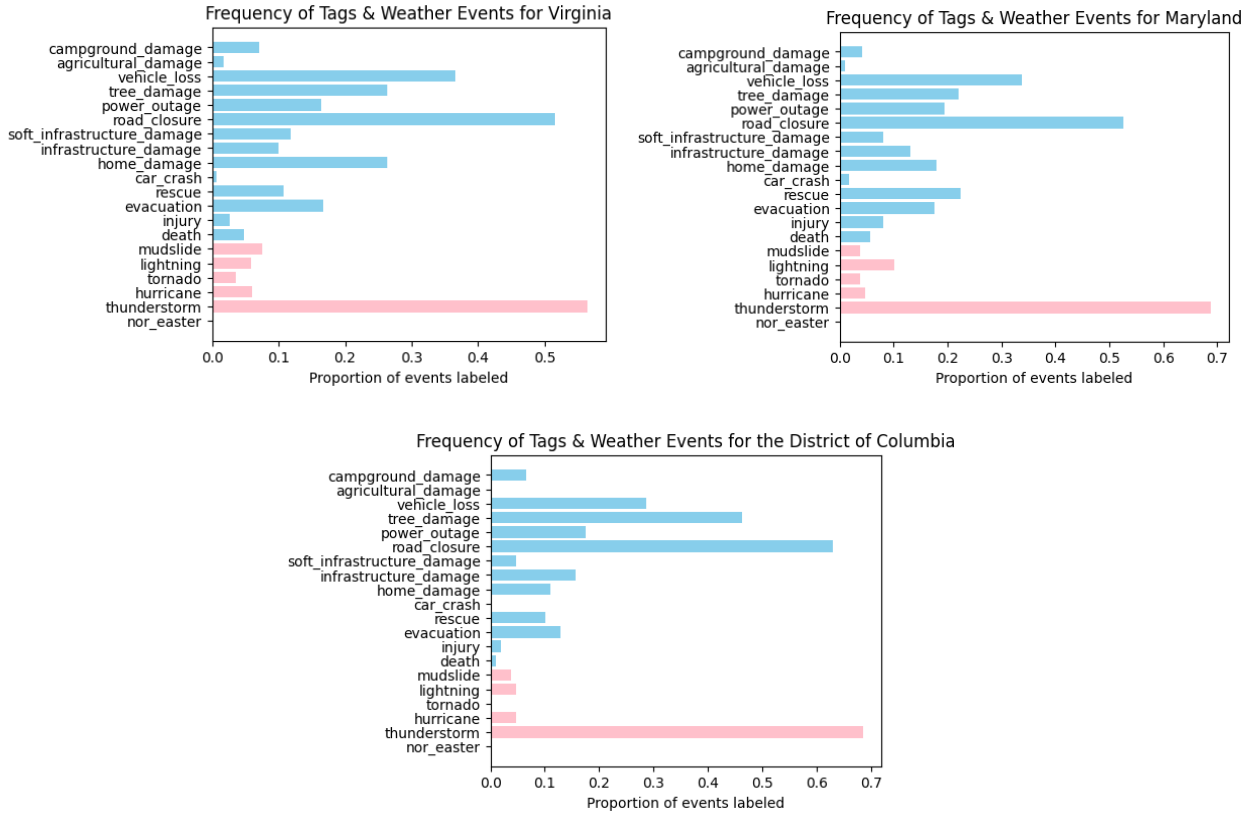

Figure 6: *Bar graphs showing the frequency of impact tags and extreme weather events across NOAA–documented flash flood events by region.*

Focusing more specifically on each region that composes the D.C., Maryland, Virginia (DMV) area, we see similar patterns of proportion to the overall damage types. *road_closure* was the most frequent, appearing at least 50% of the time for all regions. These similar patterns could reflect the similarity of weather patterns and response systems already present across the three regions. However, there seems to be a larger difference between *tree_damage* and *power_outage* in D.C. and Virginia than in Maryland. This could indicate a difference in the way that power grids are designed as to not be affected significantly by tree damage in D.C. and Virginia compared to Maryland.

In terms of weather patterns, however, Maryland experiences more lightning: 10% of Maryland events are tagged with *lightning* compared to 5% for each of D.C. and Virginia. Furthermore, D.C. reports a lack of tornadoes compared to Maryland and Virginia, most likely due

to its urban location. Virginia has a higher rate of mudslides reflecting its more wood-area based geography.

Correlation Between Tags

Another research goal was to evaluate any correlation between tags and to determine whether there were certain tags that would appear together (Fig. 7).

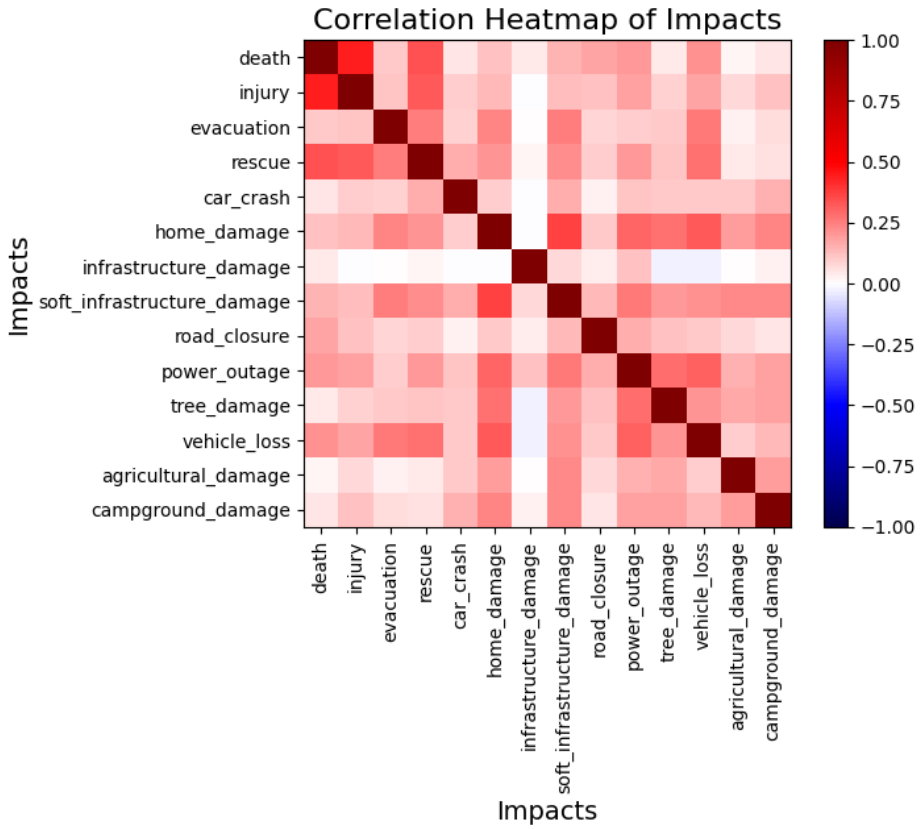

Figure 7: *Correlation heatmap of disaster impact factors. The darker the color, the stronger the correlation between the two factors. Red means positive correlation and blue means negative correlation.*

Most notably, the correlation between injury and death is the highest at over 50% co-occurrence. Meanwhile, the *home_damage* tag is highly positively correlated with *soft_infrastructure_damage*, *vehicle_loss*, and *tree_damage*. This can be explained by the fact that typical homes are surrounded by trees and/or cars and other vehicles. Furthermore, this implies that residential buildings may be generally located near government/public buildings such as schools, grocery stores, and fire departments. From these correlations, it would also make sense for these variables to correlate with power outages, as trees often topple the

power lines or other sources of power for the home. The *evacuation* tag is also strongly correlated with *vehicle_loss* as most people lose their cars in the event of a flood, whether that be through abandoning them or using them as a flotation device to get rescued.

We also created a heatmap to assess correlations between the extreme weather events and impact factors (Fig. 8).

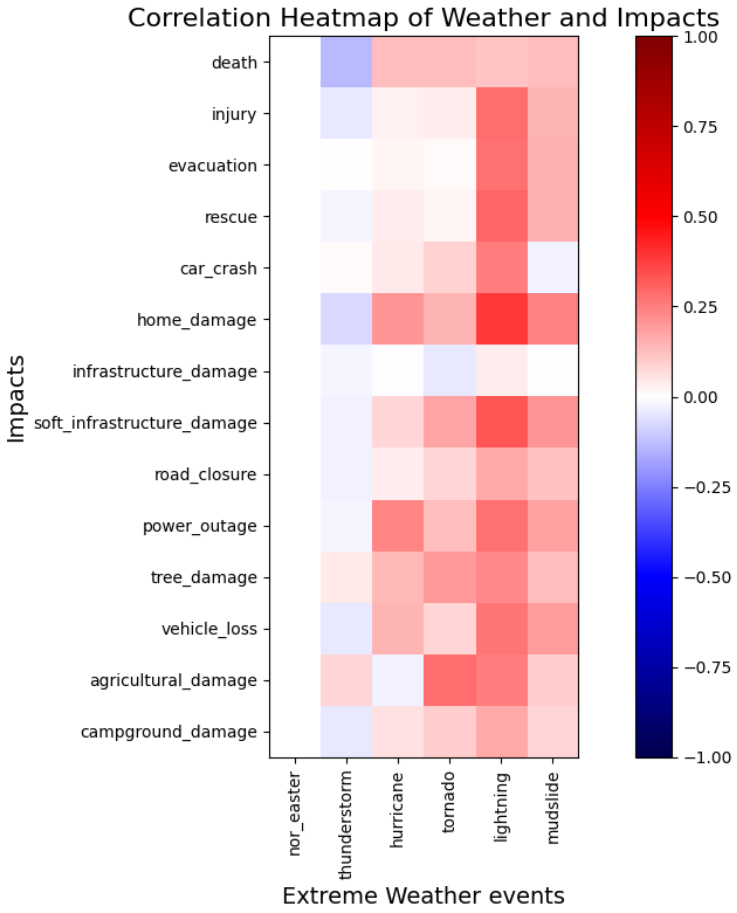

Figure 8: *Correlation heatmap of weather events and disaster impact factors. The darker the color, the stronger the correlation between the weather event and the impact factors. Red means positive correlation and blue means negative correlation.*

It can be seen that lightning, compared to the other extreme weather events, has correlation with the most impact tags. The *lightning* tag is moderately to strongly correlated with all of the impact factors except *infrastructure_damage*. This makes sense as lightning can down trees and create fires, which then affect more easily-flammable structures like buildings, homes and nature-based areas, rather than physical infrastructure with non-flammable material like roads and bridges.

Co-occurrence Network

To get a better sense of the relationships between more than two impact factors at a time, we created a network graph joining the impacts based on how frequent two impact factors simultaneously occur in one flood event (Fig. 9).

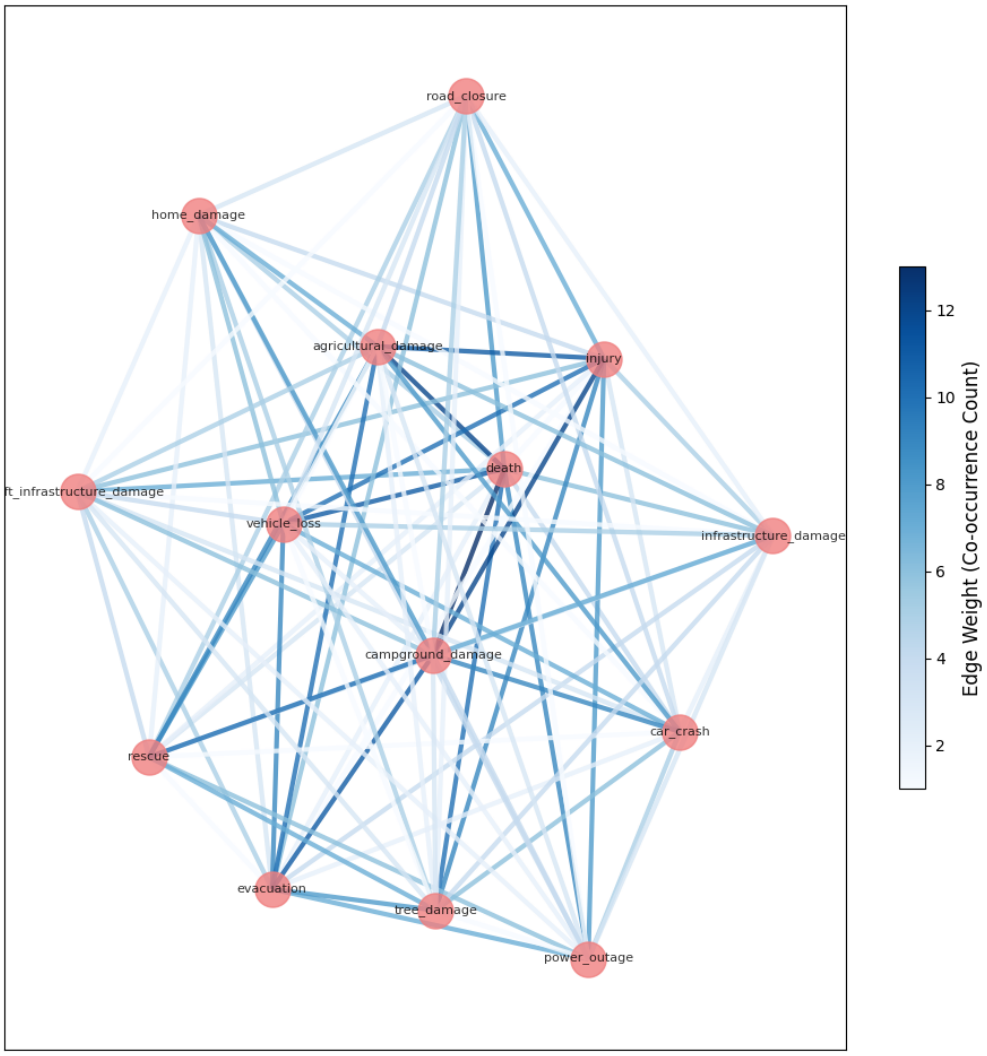

Figure 9: *Network graph of all impact factors. The darker the edge color, the stronger the frequency of the two factors co-occurring together in one event.*

Most notably, *agricultural_damage*, *injury*, and *death* all occur frequently with each other as represented by the darker and thicker edges in the graph. *campground_damage* also has strong frequency ties with most of the other impacts except for *power_outage* and *road_closures*.

Flood Impact Score (FIS) Distributions

After calculating the FIS scores for each recorded flood event, we analyzed the general, spatial, and temporal distribution of the scores (Fig. 10).

General Distribution

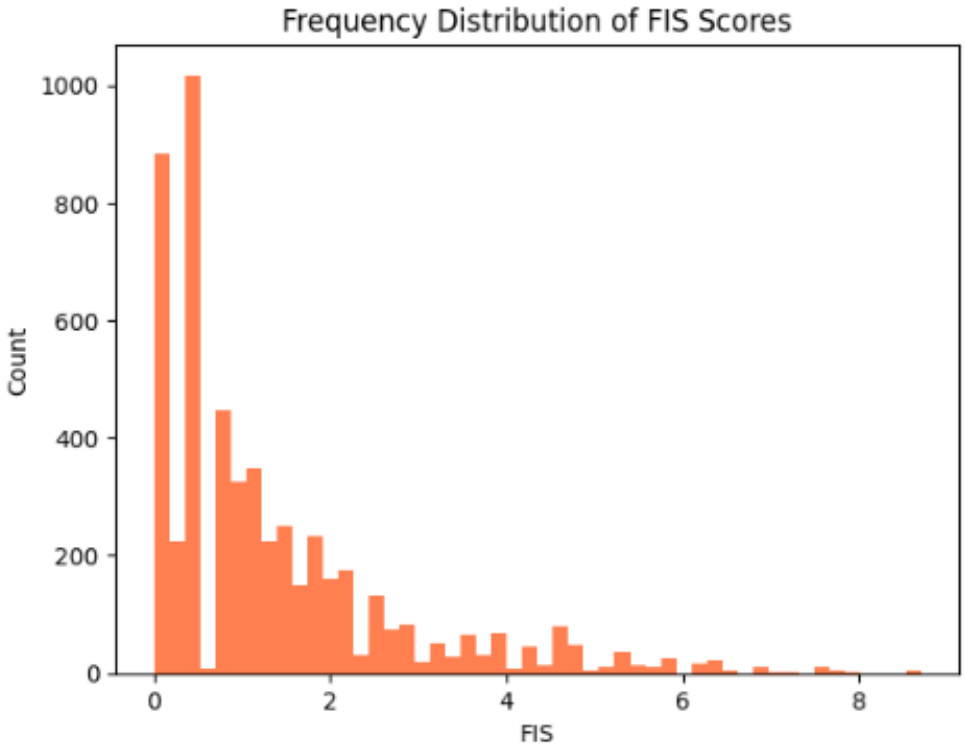

Figure 10: *Histogram of the frequency of FIS scores. Note that the maximum possible score is 10, but due to the actual distribution of scores, it is not included in the scale.*

Out of a maximum possible score of 10, the largest observed score in the dataset was 8.71. The lowest was 0, meaning there were events that received no labels. The FIS scores appear to be heavily concentrated in the lower range, with the mean score being 1.36 and the median score being 0.94. The right-skewed nature of the distribution contributes to the mean being higher than the median.

Overall, this shows that the majority of flood events have low-medium impacts, and are either a combination of fewer higher-impact tags (death, infrastructure damage, etc.) or of several lower-impact tags (animal loss, vehicle loss, etc.)

Spatial Distribution

In order to view the distribution of FIS scores across space, we calculated the average FIS score per county (where D.C. is its own county) and displayed them in a choropleth map (Fig. 11). At first glance, the county with the highest average FIS score is Fairfax County, Virginia. Other areas with high average FIS scores are those bordering the northern portion of the Chesapeake Bay, and those in northwest Virginia and southwest Maryland. However, since the size of the counties varies, and population density differs between counties as well, some of the contrast between average FIS scores may be due to those confounding variables.

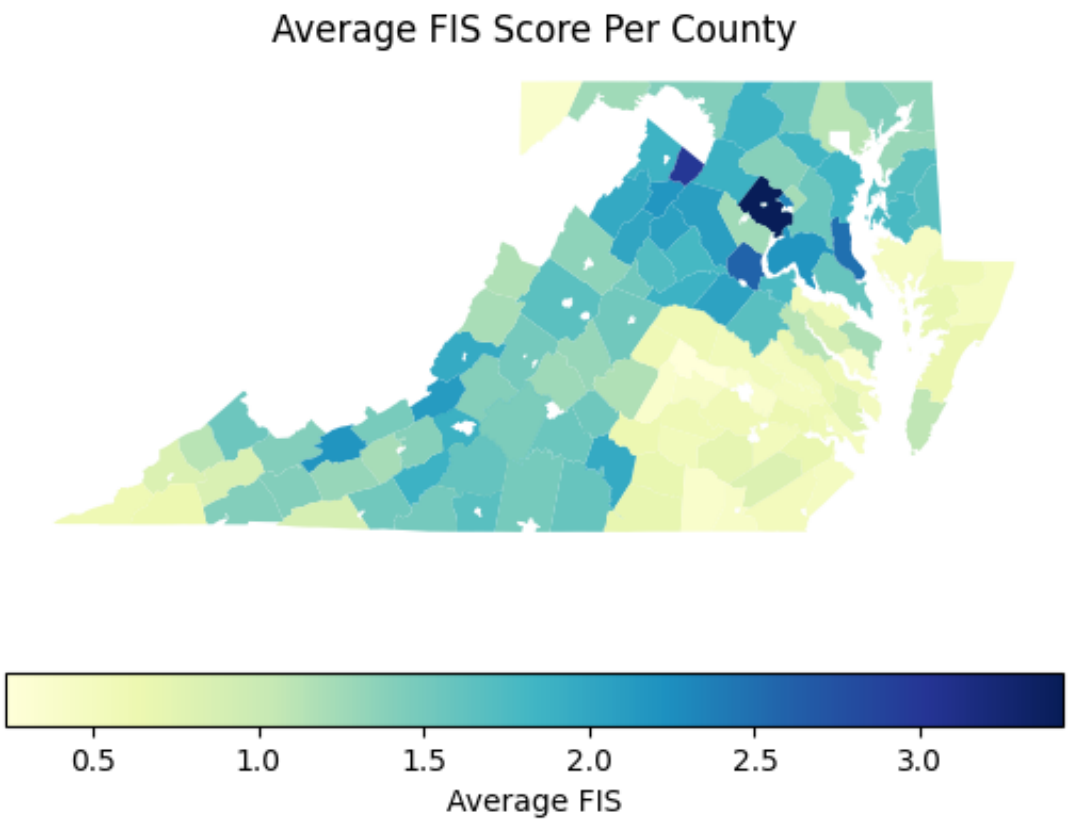

Figure 11: *Choropleth map of average FIS scores per county. Census tract divisions are also depicted, but averages were calculated by county.*

Surprisingly, counties bordering the Atlantic Ocean and the southern portion of the Chesapeake Bay appear to have lower-impact floods. Since the dataset included data on coastal flood events, this implies that either the floods in the northern Chesapeake Bay area are more intense than those in the south, or that the northern Chesapeake Bay area is more vulnerable in other ways. This could be due to differences in infrastructure, community resilience, etc.

Temporal Distribution

To track changes in FIS scores throughout the years, we calculated the average FIS score in one-year and five-year intervals for our range, 1995 to 2025 (Fig.12).

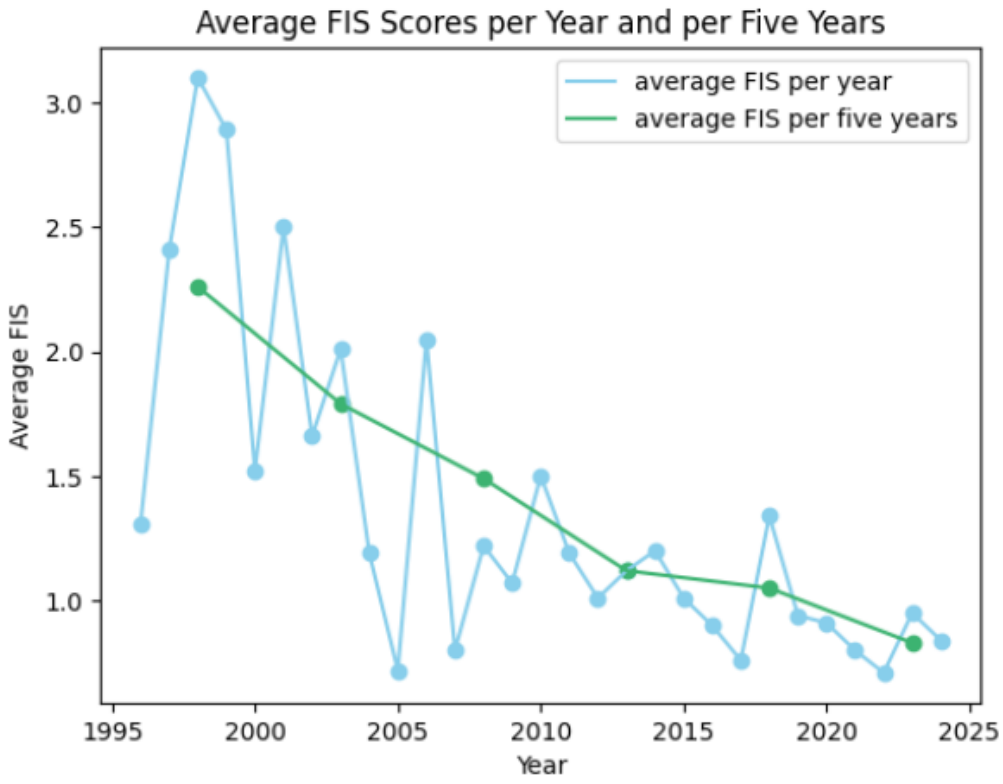

Figure 12: *Scatterplot of average FIS scores per year and per five years. Connecting lines have been drawn through the dots for the respective datasets.*

While there is high variation between the one-year FIS averages, with scores ranging from 2.5 one year to 0.8 the next, there is still a general downward trend and decrease in variation as time passes. The downward trend is more apparent in the five-year FIS averages, which show less variation and more linearity. This trend could be explained by two possible scenarios: first, it is possible that the severity and/or frequency of floods is decreasing throughout the years, resulting in lower impacts.

However, studies have shown that not only are floods increasing in frequency in recent years, but also in intensity. Alternatively, the reason could be that communities are becoming more resilient against flood events as time passes. If this is the case, communities should build on current implemented flood-relief plans and continue the upward trend in resilience.

Discussion

Key Takeaways

Firstly, our current model demonstrates a strong performance in tagging and quantifying flood-related impacts from unstructured narrative data. By utilizing a simple classification algorithm, we have developed a sufficient basic method for extracting and analyzing the various consequences of flash flood events.

Our analysis shows that certain impact factors, specifically road closures, vehicle loss, tree damage, power outages, and home damage, are most frequently observed and are highly correlated across the other aforementioned factors. These factors can often co-occur, especially alongside lightning events, indicating that storms with sufficient electrical activity can produce disruptions in infrastructure and utilities. The strong associations between lightning, downed trees, and road closures display the importance of prioritizing emergency response in areas where these conditions are more likely to occur. With targeted and informed interventions, we could mitigate some of the most disruptive effects observed.

Additionally, agricultural damage has also emerged as a growing concern as it frequently appears alongside other impact tags in our dataset, suggesting that agricultural areas are vulnerable to direct flood damage. As a result, it is important that flood resilience planning in the region include strategies for protecting agricultural crops and livestock as well as methods to support quick recovery in rural areas.

Finally, our spatial and temporal analyses indicate that there may be some meaningful difference in vulnerability and resilience across the region, specifically between the northern and southern Chesapeake Bay communities. Fortunately, we also observed a general decline in average Flood Impact Score over time, which could potentially reflect the increasing community resilience, improved infrastructure, or more effective emergency response practices across the region. Ultimately, further refinement of our impact tagging and scoring models will be essential for tracking these trends in the future and guiding flood risk mitigation while facing the evolving challenges climate change poses.

Limitations

Though our results are already quite informative, there can be more done to improve the quality of our results and future areas of study. In terms of limitations, the biggest one is our method of manual labeling. Specifically, we only manually labeled the first 900 rows

which only consisted of D.C. and Virginia-occurring events, leaving the results for Maryland based on the classification/tagging model.

Besides that, we also did not include any sort of precipitation tags, such as heavy rain or light rain because of its complexity and frequency in all of the narratives present. Additionally, because the three project members did not peer-review the labels for the narratives, some labels were tagged that may not be entirely accurate of the narrative. This is reflected in the low accuracy score of the infrastructure_damage tag, since we shifted its definition several times throughout the tagging process.

As for the FIS model, there are two limitations that come to mind. For one, we only conducted binary classification in terms of tagging – if the damage was present or it was not present. This method did not take into consideration the differences in the extent of damage that could be present between different narratives, which would have brought a more representative and possibly more accurate model for determining the Flood Impact Score of a documented event. Furthermore, the weights we used were determined based on the interpretation of the damage guide provided by FEMA which did not provide explicit rankings in the prioritization of different tags, thus limiting the actual representativeness of the score.

Future Directions

To address our study limitations, there are a couple different courses of action that can be taken. For example, for observations labeled, we have at least two reviewed and calculate their inter-rater reliability score to ensure that the tags are representative enough of people's interpretation of the tags. Additionally, we can work to improve the FIS model so that it quantifies the severity of the tag beyond its binary labeling (e.g. taking into account that 10+ deaths are more severe than just one death at an event, for example).

As for extended research, we can look at flood events that go beyond the D.C., Maryland, Virginia region to other states and perhaps other countries around the world to find continental and global patterns of the consequences of flash floods.

Acknowledgements

PIT-UN contributed funding for this work. PIT-UN is a project of the New Venture Fund (NVF), a 501(c)(3) public charity that supports innovative and effective public interest projects.

References

[1] Bureau, U. C. (n.d.). *Cartographic Boundary Files*. The United States Census Bureau. https://www.census.gov/geographies/mapping-files/time-series/geo/cartographic-boundary.html

[2] *Damage Assessment Operations Manual: A Guide to Assessing Damage and Impact*. (2016). https://www.fema.gov/sites/default/files/2020-07/Damage_Assessment_Manual_April62016.pdf

[3] *Storm Events Database — National Centers for Environmental Information*. (n.d.). Www.ncdc.noaa.gov. https://www.ncdc.noaa.gov/stormevents/ftp.jsp

[4] Taherkhani, M., Vitousek, S., Barnard, P.L. et al. (2020). Sea-level rise exponentially increases coastal flood frequency. *Scientific Reports*, 10, 6466. https://doi.org/10.1038/s41598-020-62188-4

[5] Vitousek, S., Barnard, P., Fletcher, C. et al. (2017). Doubling of coastal flooding frequency within decades due to sea-level rise. *Scientific Reports*, 7, 1399. https://doi.org/10.1038/s41598-017-01362-7

Appendix

Link to our Team's GitHub Repository

The link to our GitHub Repository can be found here. We used Python 3.10 and 3.11.9 for our code. We also drew upon specific python libraries, including Pandas, NumPy, Natural Language Tool Kit (NLTK), matplotlib, networkX, and geoPandas.