Global Flood Prediction: A Multimodal Machine Learning Approach

Cynthia Zeng [1] Dimitris Bertsimas[1,2]

[1]Operations Research Center, MIT [2]Sloan School of Management, MIT

Abstract

Flooding is one of the most destructive and costly natural disasters, and climate changes would further increase risks globally. This work presents a novel multimodal machine learning approach for multi-year global flood risk prediction, combining geographical information and historical natural disaster dataset. Our multimodal framework employs state-of-the-art processing techniques to extract embeddings from each data modality, including text-based geographical data and tabular-based time-series data. Experiments demonstrate that a multimodal approach, that is combining text and statistical data, outperforms a single-modality approach. Our most advanced architecture, employing embeddings extracted using transfer learning upon DistilBert model, achieves 75%-77% ROCAUC score in predicting the next 1-5 year flooding event in historically flooded locations. This work demonstrates the potentials of using machine learning for long-term planning in natural disaster management.

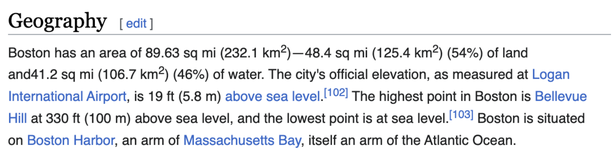
Contributions

This work presents a multimodal machine learning approach combining for global multi-year flood prediction. To the best of our knowledge, this is the first machine learning flood prediction model at the global scale and on a multi-year horizon. In addition, it is the first time text-based data has been applied to flood prediction. Our main contributions are three-fold:

1. A novel multimodal framework to incorporate text-based geographical information to complement time-series statistical features for global flood prediction. We employ state-of-the art large natural language processing techniques, including fine-tuning and transfer learning on pre-trained BERT models.
2. Our experiments show strong results for multi-year flood risk forecasting, with the strongest model achieving 75%-77% ROCAUC score in the next 1-5 year flooding prediction. In addition, we show that multimodal models, combining text with statistical data, outperform single-modal models using only statistical data.
3. Our framework can be generalised to other natural disaster forecasting tasks such as the wildfires, earthquakes, droughts, and extreme weather events. Thus, this works suggests a promising direction in long-term preparation for natural disaster management.

Data

Historical Flood Data. We use the Geocoded Disasters (GDIS) dataset, which includes geocoded information on 9,924 unique natural disasters occurred globally between 1960 and 2018 [4]. In addition, we linked this dataset with the EM-DAT dataset to add additional economic information such as damage estimation [2]. In this project, we restrict forecasting locations to those with historical flooding event. We use the date, latitude, longitude, location (given as the name of the location), and if available, damage cost from this dataset. We divide the earth into 1° by 1° grid, corresponding to about 100km by 100km squares. Using the latitude and longitude information, we compute a 'grid id' for each natural disaster from the GDIS dataset. Overall, there are 2852 unique grid locations in the dataset with a recorded historical natural disaster.



Geography [edit]

Boston has an area of 89.63 sq mi (232.1 km^2)—48.4 sq mi (125.4 km^2) (54%) of land and41.2 sq mi (106.7 km^2) (46%) of water. The city's official elevation, as measured at Logan International Airport, is 19 ft (5.8 m) above sea level.[102] The highest point in Boston is Bellevue Hill at 330 ft (100 m) above sea level, and the lowest point is at sea level.[103] Boston is situated on Boston Harbor, an arm of Massachusetts Bay, itself an arm of the Atlantic Ocean.

Figure 1. Example 'Geography' section of the Boston Wikipedia page.

Geographical Data. To incorporate the geographical information of each location, we use open-source Wikipedia website's Geographical section, which contain text-based geographical description of certain areas, as shown in Figure 1 as an example for the 'Boston' Wikipedia page. To obtain the geographical information, we use the 'location' data from the GDIS dataset for each grid id, then use the Wikipedia-API to obtain the text from the Geographical section for each location [1]. To deal with the noise in the data, since some locations have different names on Wikipedia, we search over synonyms for each location. For those location Wikipedia pages without Geography section, we use the Summary section. Among 2852 unique grid ids, we collected text-based information for 2775 grid ids, and fill the remainder grid ids as 'missing'.

Methodology

The overall goal is to predict next 1 to 5 years of flood risk using a multimodal approach. The framework adopts a three-step approach to combine distinct data formats and sources. Figure 2 illustrates the overall three-step framework. More details of the training and testing protocol can be found in the Appendix.

1. We gather different sources and modalities of data, which are a) tabular-based historical natural disaster data and b) text-based geographical data from Wikipedia pages.
2. We perform feature processing individually for each data modality, and obtain a one-dimensional feature representation (embeddings) respectively.
3. We concatenate feature embeddings from different modalities and perform feature sections, before making next-N-year flood event predictions using gradient boosted tree (XGBoost) models for binary classification task. Prediction target 1 indicates a flood in the next N years, 0 otherwise.

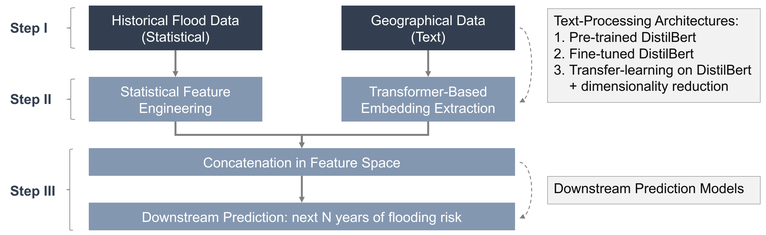


Figure 2. Three-step framework to combine statistical data with text-based data. The transformer-based text data embedding extraction contains three types of architectures.

Feature Processing

Statistical Feature Processing

We use the GDIS dataset to process historical statistics of natural disasters. In particular, for each grid id, we aggregate statistical features into yearly basis uisng only the current year's natural disaster statistics. In particular, we summarize the 'count' 'binary' and 'damage cost' feature during the year for each natural disaster: 'flood', 'storm', 'earthquake', 'extreme temperature', 'landslide', 'volcanic activity', 'drought', 'mass movement (dry)'. The 'damage cost' feature corresponds to the insurance amount claimed by the natural disaster, which is intended as a proxy to reflect the severity of the natural disaster. In total, the statistical features contain 24 features. Additionally, we record the 'year' feature as numerical feature.

Text Feature Processing

For each location, we use the Geography section from the Wikipedia page using the location name. This information is given as text, and each location is associated with a paragraph of geographical information description. Under the scope of this work, we experiment with pre-trained large language model DistilBert, a distilled version of the BERT model, which offers good performance whilst faster to train and fine-tune [5]. The two main challenges are: a) DistilBert model is trained on a large set of generic texts, whilst we would like to adapt it to encode geographical information specifically; b) feature extraction is performed on a token-by-token basis, whilst we require embeddings corresponding to a paragraph of sentences. In summary, we experiment with three distinct architectures.

Results

Table 1 contains out-of-sample binary classification performance from various models for the next 1,2,5 year flood prediction horizon on the selected 818 grid locations. In summary, a multi-modal approach demonstrates the strongest performance, achieving 70% - 75% ROCAUC score. Training and testing sets are randomly selected at 70% and 30%, and more details on the training protocols can be found in the Appendix.

We construct a deterministic baseline model which predicts the next N years of flood outcome as the same current year flood outcome. This approach aims to mark previously flooded region as high risk, which is similar to the flood risk mapping procedure employed by agencies such as FEMA.

Metric	Baseline	Single-Modal	Multimodal		
		Statsistical	DistillBert	Finetune (N=795)	Transfer (N= 61)
1-year horizon					
Class imbalance: 0.063					
rocauc	0.544	0.742	0.734	0.758	**0.772**
f1	0.545	0.519	0.527	0.554	**0.558**
acc	**0.895**	0.707	0.747	0.783	0.783
acc balanced	0.544	**0.681**	0.640	0.664	0.675
2-year horizon					
Class imbalance: 0.064					
rocauc	0.534	0.726	0.724	0.756	**0.764**
f1	0.536	0.502	0.525	0.559	**0.560**
acc	**0.889**	0.664	0.742	0.782	0.781
acc balanced	0.534	**0.676**	0.627	0.664	0.668
5-year horizon					
Class imbalance: 0.067					
rocauc	0.539	0.715	0.726	0.749	**0.767**
f1	0.541	0.501	0.522	0.545	**0.557**
acc	**0.892**	0.668	0.724	0.758	0.764
acc balanced	0.539	0.664	0.641	0.658	**0.682**

Table 1. Out-of-sample performance for the next 1,2,5 years of flood risk prediction task. Baseline model predicts the same outcome as current year outcome. Multimodal models employs statistical features and text embeddings extracted using various architectures. We record the number of total features employed in each approach given in brackets. We report ROCAUC score, accuracy, F1 score, and balanced accuracy.

Conclusion

This work presents a multimodal machine learning framework for global flood risk forecasting combining statistical natural disaster dataset with text-based geographical information. This work demonstrates strong results for multi-year flood risk forecasting globally, enabling potentials for long-term planning in natural disaster management.

References

[1] Wikipedia-API.

[2] EM-DAT The International Disaster Dataset, 2021.

[3] Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. On the sentence embeddings from pre-trained language models. *arXiv preprint arXiv:2011.05864*, 2020.

[4] E. Rosvold and H. Buhaug. Geocoded Disasters (GDIS) Dataset, 2021.

[5] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.