Understanding and Predicting Internal Displacement from Global Floods

Robert Meyer, Leanna Chraghchian, Felicia Liu, Brendan Foo

Abstract

Our research aims to address critical gaps in current predictive modeling methods for flood-induced displacements amid escalating global climate change. Using a unified dataset from various sources, including the UN Global Internal Displacement Database, our study employs different robust machine learning models such as random forests for classification and support vector machine for regression. Challenges of scattered and unstandardized data are noted, limiting our analysis to ~60% of the countries. Our findings emphasize the need for improved data quality and standardization for more reliable predictive modeling of flood-induced displacement.

Introduction

Climate change is one of the biggest threats facing humanity today. From constantly changing weather patterns that will eventually threaten our food production, to rising sea levels that will increase the risk of catastrophic flooding, climate change will affect every individual in the future if action is not taken sooner rather than later. According to the National Oceanic and Atmospheric Administration, the 10 warmest years in the historical record have all occurred since 2010.[1] This suggests that climate change, and in turn, people displaced from their homes due to climate change, is one of the most pressing issues of the century that we must address.

These displaced people are usually referred to as internally displaced persons or IDPs. These are "persons or groups of persons who have been forced or obliged to flee or to leave their homes or places of habitual residence, in particular as a result of or to avoid the effects of armed conflict, situations of generalized violence, violations of human rights or natural or human-made disasters, and who have not crossed an internationally recognized border."[2]

In 2022, an Intergovernmental Panel on Climate Change (IPCC) report found that approximately 3.5 billion people were living in countries with a high vulnerability to climate change, and 31.8 million people were internally displaced within the borders of their country due to weather-related hazards. Of the 31.8 million displacements, 60% were due to floods.[3] Warmer global temperatures as a result of climate change cause more water to evaporate, which subsequently increases the size and frequency of precipitation. When the size and frequency of precipitation increase, the size and frequency of floods increase as well. In addition, other climate change-related factors

also contribute to increased flooding and higher water levels globally. These include but are not limited to, timing in snowmelt, changes in streamflow, and amount of snowpack accumulation in winter months.

Our project focuses on internally displaced people within-country as a result of floods. We learned early on through our research and interviews with domain experts, however, that current data and research in the area of flood-induced internally displaced people is incredibly patchwork and borderline nonexistent. Even more surprisingly, no predictive technologies currently exist to identify or protect those impacted by climate change. Motivated by the lack of research in this field, our project aims to predict displacement in a specific region of the world if a flood were to occur. We believe that such a model would be highly beneficial in terms of serving as an early warning sign for at-risk populations and also as a guide for resource allocation for humanitarian aid groups.

Current State of Data on IDPs and Floods

Currently, data on IDPs and floods are gathered independently by different organizations and groups without a centralized or uniform system. During our search for appropriate datasets, we came across a wide variety of data, from Facebook's Data For Good Movement Range Maps to climate data produced by the Intergovernmental Panel on Climate Change. While these datasets contained useful information, they all utilized differing definitions, standards, and metrics to describe floods and movement. We will discuss the manifestations of these limitations in a later section. With such disparate data sources available for flood-induced displacement, we carefully selected and merged a few sources that seemed the most fitting for our project's end goal.

Our dataset combines four datasets (GIDD, ND-GAIN, Flood Archive, U.S. Census) into one that encodes, for each flood event, the time and place, magnitude, causes, and features of the country such as population measures and various factors related to preparedness for natural disasters. GIDD refers to the UN Global Internal Displacement Database, which is the base set of flood events. ND-GAIN refers to the Notre Dame Global Adaptation Initiative, and it provides descriptions of how well each country is prepared for flood events based on factors including infrastructure, economy, and more. Flood Archive is from the University of Colorado and supplements our data with information describing the floods, such as a measure of magnitude. This data was found manually by professors at the university through news reports and articles. Lastly, U.S. Census Bureau data provides population counts and population density for each country.

This combination of the four data sources only provided complete data, with no missing values from any of the sources, for approximately 60% of the countries in the world. Even then, the number of flood events covered was extremely limited. Our final combined dataset contained 2,746 rows and 23 columns, and these columns included:

'ISO3'
'IDPs from Event'
'Economics'
'Governance'
'Social'
'Capacity'
'Ecosystem'
'Exposure'
'Food'
'Habitat'
'Health'
'Infrastructure'
'Sensitivity'
'Area'
'Began'
'Ended'
'MainCause'
'Severity'
'Duration'
'Magnitude'
'Population'
'Population Density (People per Sq. Km.)'
'Net international migrants, both sexes'

Modeling

After compiling our dataset from various sources, we performed feature engineering to prepare it for modeling purposes. One of the biggest challenges in this modeling task is the incredibly intense skew of the target variable. As you can see in Figure 1, our target variable has a long right tail and massive outliers. To help account for this, one key feature engineering step we took was to logarithmically transform the target variable. We also did this for the product of the duration of the flood event, the severity of the flood event, and the area of the flood event) to give us a 'magnitude' score, and other heavily skewed features. For some models, such as linear regressions, we normalized all the input features to make sure they were all on the same scale. We also explored both regression and classification. Below are more details on the model selection process.

Figure 1: Distribution of Target Variable

	Minimum	**Median**	**Mean**	**Standard Deviation**	**Maximum**
IDP count	1	500	48,076	223,352	4,095,280

Linear Regression

This was our initial prototype model where we ran some simple multivariate linear regressions between the number of internal displacements and key input variables such as magnitude score and ND-GAIN features. However, this model did not perform well as it does not capture any nonlinear relationships between variables.

Neural Network

Neural networks, on the other hand, can capture these nonlinear complex relationships. This would be appropriate for our purpose due to the number of input variables from our dataset. However, this does typically require larger amounts of labeled data which is a challenge in this field and will be discussed later. Neural networks are also considered black boxes as the underlying decision-making process is not interpretable.

Support Vector Machine

Support vector machines are versatile and can be used for both classification and regression tasks. Their effectiveness extends to high-dimensional spaces, where they seek to identify a hyperplane for class separation within the feature space. This hyperplane, constituting the decision boundary, results from a linear combination of the input features. This would make support vector machines generally more interpretable than other complex neural network architectures. Nonetheless, they can still be considered black boxes when nonlinear kernel functions are used and the inner workings behind finding the optimal hyperplane may not easily be interpretable.

Random Forest Classification

Random forests also exhibit versatility which is useful in both classification and regression tasks. Since it is an ensemble learning method that builds multiple decision trees and combines their predictions, it is more robust and less sensitive to noise in the data. Similarly, it performs well in high-dimensional spaces and can capture non-linearity from our dataset. Random forests also provide a measure of feature importance which can be valuable in understanding the contribution of different input features to making a prediction. However, the combined decision of multiple trees may make it difficult to interpret and is more dependent on hyperparameter tuning.

With these model considerations in mind, we employed them for both classification and regression tasks with different types of input specifications. Figures 2 and 3 show our results summary where Figure 2 shows our findings for multi-classification and Figure 3 shows our findings for regression.

Figure 2: Multi-Classification Model Results Summary

Model Type	# of Classes	Recall	Precision	Test Accuracy

NN	3	**H** 0% | **M** 0% | **L** 100%	**H** 0% | **M** 0% | **L** 32%	33.90%
NN [Log + Norm]	3	**H** 60% | **M** 8% | **L** 76%	**H** 53% | **M** 38% | **L** 43%	47.20%
NN [Log + Norm]	2	**H** 56% | **L** 67%	**H** 67% | **L** 68%	66.30%
RF [Log]	3	**H** 50% | **M** 43% | **L** 62%	**H** 61% | **M** 42% | **L** 54%	52.00%
SVM [Norm]	3	**H** 38% | **M** 30% | **L** 53%	**H** 48% | **M** 34% | **L** 40%	40.59%

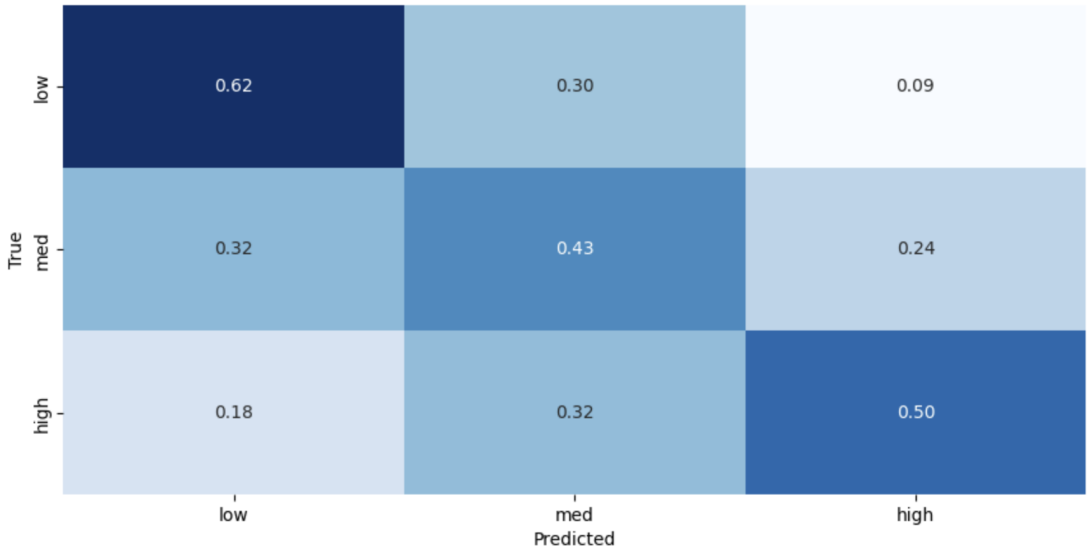
Figure 3: Regression Model Results Summary

Model Type	RMSE
NN [Log + Norm]	167,168
RF [Log]	258,013
SVM [Log + Norm]	225,414

The two models that we focused on were random forest classification and support vector machine regression.

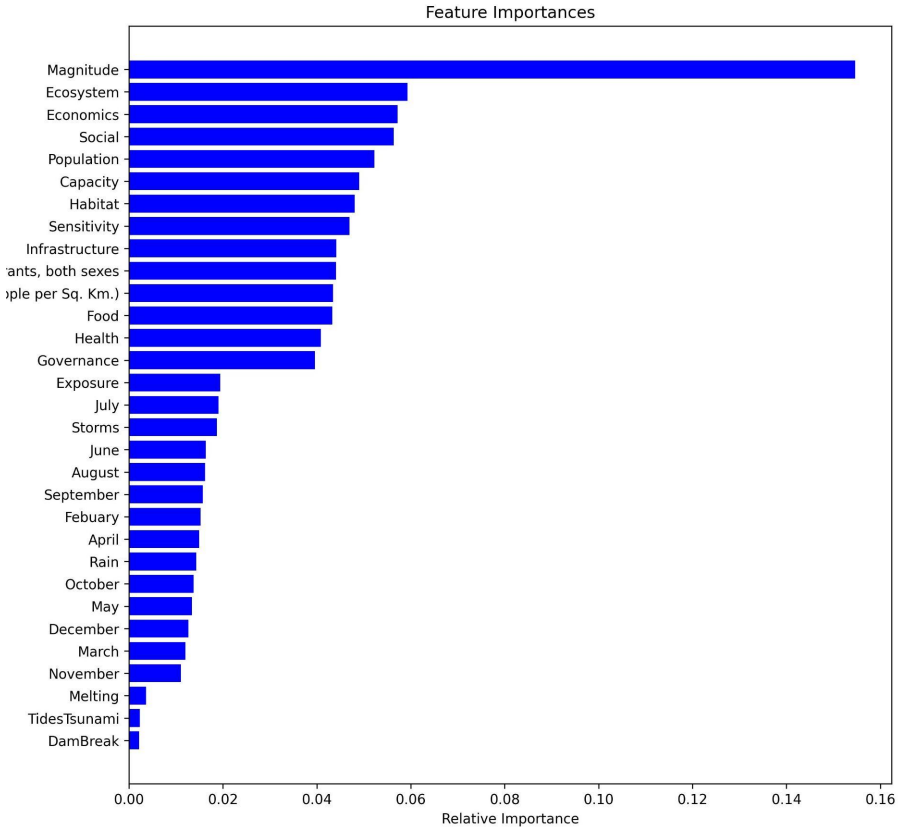
For the random forest classification model, we binned our dataset based on the logarithmic transformation of the IDP count column into three quantiles: low, medium, and high. Other data pre-processing steps included dropping null values, one-hot-encoding the month, dropping specific features (country, area, severity, and duration) as it was simplified into the flood magnitude, and applying a logarithmic transformation to certain skewed features. We also simplified the column denoting the “Main Cause” of the flood. This column was a hand-entered description. We simplified it to five columns, one for each of the most common causes such as “Storms” or “Snowmelt”. Each column had a 1 if that cause was relevant to that flood, and 0 otherwise. As part of the hyperparameter tuning for the model, we performed a random search with 100 estimators and 3-fold cross-validation. The suggested hyperparameters created very large trees to capture complexity. Ultimately, our random forest classifier achieved an overall test accuracy of ~52% and its top features included the magnitude of the flood and other ND-GAIN indicators (i.e. infrastructure, social, economic, etc.). A confusion matrix, shown in Figure 4, showed that the classifier performed well for the ‘low’ class. Class recall scores showed that the model did best among all models for the ‘medium’ class. The ‘high’ class is generally the most challenging to predict, being the large right tail in the distribution of the data with extreme outliers.

Figure 4: Random Forest Classification Confusion Matrix



As we can see in Figure 4, this model does seem to be detecting a signal in this relatively rough data. While classification (compared to regression) generalizes results, we believe that our random forest classification model is able to differentiate events with drastically different displacement effects. Furthermore, using a random forest for classification allows us to see the feature importances, providing more insight into the root causes of displacement crises.

Figure 5: Relative Importance on Features



From these feature importances in Figure 5, we see that the Magnitude score and other ND-GAIN features such as economics, social, habitat, and infrastructure are relatively important in making classification predictions. The magnitude feature is a combination of flood duration, area, and severity, all of which were calculated by the Flood Archive. These features were estimated manually for each event, using descriptions in news sources as references. The fact that this feature was so predictively powerful despite rough estimation and unstandardized data collection is extremely promising, as a more robust technique would be likely to increase the power of our models greatly. The importance of the ND-GAIN features also gives us a way to prioritize efforts to prepare countries for floods. For example, ecosystem is the most important ND-GAIN feature. This means that the state of a country's ecosystem largely determines the level of displacement, indicating humanitarian efforts should focus on countries with high ND-GAIN ecosystem vulnerability scores, and adjust for things included in these scores such as reducing ecological footprint or reliance on natural capital.

For the support vector machine regression model, we used normalized input features and logarithmically transformed IDP counts. We also dropped any missing values. These steps were taken because support vector machines, as stated earlier, are

sensitive to noisy data and outliers. As part of the hyperparameter tuning process, we performed a grid search with three-fold cross-validation and used the suggested hyperparameters. While this type of modeling performs well in high dimensional spaces, the true underlying patterns might not have been fully captured as we had a large number of input variables. To accommodate this, we used Principal Component Analysis for dimensionality reduction to simplify the complexity while retaining patterns within our data. Looking at Figure 3, our RMSE for this support vector machine with PCA was overall fairly high and greater than the RMSE for the neural network. However, the model's forecasts fell within a 3000-unit range of the true number of internally displaced persons ~72% of the time, which is higher than the other regression models. This behavior leads us to believe that this model is more practical in its outputs. The high RMSE values of all our models are reflective of a set of heavily skewed data, as shown in Figure 1. While our classification models help us to understand the data and root causes of displacement, humanitarian aid agencies will likely find the raw estimate provided by regression strategies like this more useful than a classification at a general level. The section that follows will go into more detail about the challenges in the data.

Data Challenges

The setbacks we encountered during our modeling are indicative of this larger problem of unstandardized definitions and collection techniques. For example, GIDD records the IDP count but overgeneralizes many related floods into one event. GIDD also does not record any features of the events besides the time and place. To make any sort of model, we need features about the flood, such as a severity score. To include this, we need another dataset like the Flood Archive. However, the Flood Archive has a different definition of what counts as a flood, so we needed to make some generalizations (such as simplifying the time to a month-year) to merge these two datasets. Even after these generalizations, there is still a small overlap in events, only about ⅔ of the flood events included in GIDD overlapped, meaning we sacrificed many examples to create a useful model. Once this was done, we had several flood events from the Flood Archive with the same large IDP estimate from GIDD, which had grouped those events into one. We therefore scaled down these IDP estimates in proportion to the area covered. This should result in more accurate guesses for the IDP counts, but also is a generalization that makes these further rough estimates. Considering all this, we remember that the Flood Archive's flood details are found manually from the news, and are therefore not particularly robust. We investigated other datasets with flood details, however, these had even less if any overlap with the

base GIDD. These issues highlight the need for a standardized data-gathering strategy, reporting strategy, and flood definition.

Our predictions are also limited to the country level by data availability. While some data sources are recorded at a sub-country level, no single source has sufficient information for a model like this, and we were unable to create a suitable merge of individual sources. For example, different data sets record similar and complementary information using different mapping systems that do not overlap or merge (GADM and BING tile systems for mapping). Data gathered on this topic at a more local level would enable these same predictions to be made at that level which would be more useful for humanitarian agencies allocating resources. Finally, similar data collected in a more timely fashion, such as recordings taken at multiple points during a flood, would be incredibly useful to see movement trends over time.

Strategies for Improving Data Quality

Improving data quality in predicting internal displacement involves several key strategies: standardizing data, collecting granular data, collaborating with various partnerships, using more advanced technology, and leveraging crowd-sourcing. First, standardizing data collection across various sources is crucial for consistency and accuracy. This includes creating uniform metrics for defining and evaluating flood impacts, displacement data, and demographic data. Additionally, collecting data at more localized levels, such as regional or community levels, can provide more detailed and actionable insights for more accurate and targeted relief efforts.

Next, the collaboration between researchers and various public entities – including government agencies and international organizations – is essential for broadening data accessibility. This is where policy regulations regarding mandated data collection and reporting would make a huge impact as many countries are keeping data away from public access. This secrecy will harm everyone by withholding valuable information models can use to predict a refugee crisis. While partnerships that have policy-influencing power are a priority, private partnerships with companies specializing in data analytics and GIS technologies can also introduce advanced tools for displacement data collection and analysis.

Moreover, integrating advanced technologies like remote sensing, GIS, and satellite imagery can enhance data gathering and analysis when assessing displacement and flood impacts over large areas. Giving more agency to third parties could help avoid long bureaucratic decision-making and dishonest government reporting. Finally, crowd-sourcing data through community engagement via mobile apps and online platforms offers real-time, ground-level insights during flood events. These strategies collectively

contribute to improving the quality of data and can complement traditional data collection methods.

These strategies can significantly enhance the quality of data, making predictive models more accurate and reliable. By addressing the challenges of scattered, unstandardized, and withheld data, these measures can contribute to better preparedness and response to flood-induced displacements.

Ethical Considerations

Fellow MIDS student Jared Feldman conducted a third-party privacy audit. His recommendations have been accounted for in the following summary of our ethical considerations regarding this project.

Transparency is an important part of fairness in the field of data science. To this end, we have included our data sources and any of their relevant privacy policy documentation in the Appendix. All of the data used for this project is publicly available and aggregated far above the individual level, containing no personally identifiable information. Information on how these sources are used and their limitations is detailed above.

We recognize that some groups may respond to indications of imminent displacement by closing borders or restricting movement. To this point, we would like to state that our research is intended to foster greater awareness and understanding of the effects of floods on displacement, as well as strengthen global resilience to flood events through this understanding. We do *not* intend for this information to prompt any actors to restrict the movement of displaced people due to fear of financial or political burden.

Furthermore, we recognize the immense challenge of gathering consistent data on events that are chaotic by nature on a global scale. Our use of specific datasets and recommendations for improvements on data gathering are not intended to tarnish the reputation of the data sources used here in any way. We recognize that these sources gather data to answer specific questions different from but related to the ones asked in this paper. Our work is intended to aid organizations like these in improving data collection in targeted ways to benefit their cause and show what can be done with current resources.

Our output includes a mock-up of a front-end website to display our work and increase accessibility. At the moment, this is not a functioning website. However, we intend that a published version would require no login or user information whatsoever, mitigating privacy risks. Considering that this data is highly aggregated and reveals no

individual information that we know of, we plan on continuing to use all historical data available from our current sources moving forward and see no reason for a deletion plan.

Conclusion

The overall mission of our project is to shed light on the lack of complete and quality data in this field and show the potentially life-changing impact of improving current collection methods. Studies predict that by 2050, 15% of the world's total population will be displaced due to climate change and natural disasters. Current models in the industry are reactive and only make short-term predictions after an event has occurred to allocate resources. While the models highlighted above have limited accuracy due to inconsistent data, they provide rough displacement estimates and relative risk and serve to inform decision-making regarding policy and data gathering.

Additionally, our contribution to this space is unique in that it investigates the root causes of movement to help prepare for future crises, and is also based on population density and other specific environmental factors. By taking steps to improve the current data collection processes and policies, we hope that we will be able to use data science to aid those who will most likely be hit the hardest by climate change-related displacement. We have the technology to create a successful model that can accurately predict displacement from a crisis, and we just need a uniform, updated, and centralized system for data collection. We hope that this project is the very start of these efforts, as our work shows the life-saving potential of technology within this field if the right steps are taken to create a comprehensive and high-quality data environment.

References

1. NCEI.Monitoring.Info@noaa.gov. (n.d.). Annual 2022 global climate report. Annual 2022 Global Climate Report | National Centers for Environmental Information (NCEI). https://www.ncei.noaa.gov/access/monitoring/monthly-report/global/202213
2. International standards | OHCHR - UN human rights office. (n.d.). https://www.ohchr.org/en/special-procedures/sr-internally-displaced-persons/international-standards
3. Environmental migration. Migration data portal. (n.d.). https://www.migrationdataportal.org/themes/environmental_migration_and_statistics

Appendix

1. Global Internal Displacement Database (GIDD), Internal Displacement Monitoring Centre
    a. GIDD Privacy Policy
    b. GIDD Terms of Service
2. Notre Dame Global Adaptation Initiative (ND-GAIN)
    a. Data sources and collection methods are detailed in section V of Technical Documentation.
3. International Database, United States Census Bureau
    a. Privacy Policy
4. Global Active Archive of Large Flood Events, 1985-Present (Flood Archive), Dartmouth Flood Observatory, University of Colorado.