
Flood Prediction in Kenya - Leveraging Pre-Trained Models to Generate More Validation Data in Sparse Observation Settings

Alim Karimi
Purdue University
West Lafayette, IN, USA
karimia@purdue.edu

Hammed A Akande
Concordia University
Montreal, QC, Canada
akandehammedadedamola@gmail.com

David Quispe
University of Toronto
Toronto, ON, Canada
david.quispe@mail.utoronto.ca

Valerie Brosnan
Barcelona, Spain
valerie.brosnan@mitigasolutions.com

Nicole Mong'are
Bricev Limited
Nairobi, Kenya
nicolenm800@gmail.com

Asbina Baral
Ministry of Education, Science and Technology
Kathmandu, Nepal
asbina.baral@gmail.com

Abstract

Kenya has lacked a national flood risk management framework and also has sparse flood observation data, which makes developing deep learning flood prediction models on a national scale challenging. Flood prediction models are critical to operationalize Early Warning Systems (EWS). We propose two different models to feed into an EWS. The first model will leverage statistical machine learning approaches to predict flood or no flood events on a 0.25×0.25 degree scale (approximately 30 km x 30 km in Kenya) using ERA5 features as well as land cover and Digital Terrain Model (DTM) data. This first model will also be used to create a baseline prediction benchmark across the entire country of Kenya. The second model will leverage pre-trained remote sensing based models to generate segmented flood or no flood data on a fine spatial scale. This will increase the number of validation points by a factor of 1000, which opens the door to deep learning approaches to predict flood or no-flood events on a 30 meter x 30 meter spatial scale. We hope that this approach of leveraging pre-trained models to generate fine scale validation data can eventually be used widely in other extreme climate event forecasting scenarios given the scarcity of historical extreme climate events compared to normal weather events.

1 Introduction

Kenya's vulnerability to natural disasters, particularly flooding, has always been a significant challenge. In the first half of 2020 alone, floods claimed over 200 lives, displaced over 100,000 individuals, caused billions of Kenyan shillings in damage, and affected 36 out of 47 counties [1]. Similarly, in 2024, severe flooding again devastated the country, displacing thousands and causing widespread economic disruption in 42 counties. Despite these events, Kenya still lacks a comprehensive national

flood risk management framework [2], making this national scale modeling even more impactful compared to focusing modeling on countries with robust disaster management systems.

The demand for more accurate and timely flood forecasts has prompted the development of advanced machine learning approaches. These leverage a wide variety of predictors such as historical ERA5 data, river flow measurements, and geological conditions to enhance flood prediction capabilities.

However, the use of these technologies in Kenya's specific environmental and climate setting is relatively unexplored. This paper intends to fill this gap by proposing approaches customized for Kenya's sparse and coarse data. By focusing on local data and conditions, this project seeks to improve the ability of municipalities and authorities to predict and respond to floods more effectively. Through improved forecasting, this initiative supports broader climate adaptation efforts and contributes to the resilience of the most vulnerable Kenyans.

2 Related Approaches

Flood prediction using machine learning has been a well-explored topic. Some promising models for predicting the occurrence of floods based on predictors like precipitation and temperature in South East Asia include Support Vector Machines, Naive Bayes, K-Nearest Neighbor approaches, and Fully Connected Neural Networks [3]. A review paper from 2018 established a trend of deep learning based approaches and also noted that leveraging physically based models required significant expertise yet failed at short term predictions [4]. Statistical approaches have also shown promising results in flood prediction along the Ganga river in India. [5].

Significant work has also been done on segmenting flood, permanent water, or no-water pixels on flooded areas using earth observation satellites. Synthetic Aperture Radar (SAR) data from Sentinel-1 and optical data from Sentinel-2 were used to create the Sen1Floods11 dataset, enabling deep learning-based flood modeling [6]. IBM's Prithvi-100M, a geospatial foundation model, was fine-tuned on Sentinel-2 data for flood segmentation. [7].

These approaches show potential, but challenges remain, particularly in data scarcity to accurately predict at a national scale in Kenya.

3 Proposed Approach

The main contributions of this proposal paper are:

- To develop a performance benchmark at the spatial scale of ERA5 on one time step ahead flood prediction in Kenya.
- Demonstrate that existing geospatial foundation models can be used to generate extensive fine-scale validation data compared to current flood databases so that the number of validation points approaches the number of model parameters in deep learning prediction methods.

The Dartmouth Flood Observatory (DFO) from the University of Colorado has over 3,000 flood events recorded worldwide between 1996 - 2018. Of those, 50 flood events are in Kenya. Observations of flood events in Kenya specifically are therefore spatially and temporally sparse. They are denoted in the DFO as a single latitude and longitude point. At a high level, we first plan to extract relevant predictors and use that to predict flood or no flood events on a coarse 0.25 x 0.25 degree grid one time step ahead. The results from this first approach will be used as a performance baseline in Kenya and also to understand performance gains brought from adding additional complexity into the modeling pipeline. A 0.25 x 0.25 degree grid translates to roughly 900 square kilometers in Kenya, which is a large area to operationalize mitigation and adaptation measures from flood or no-flood predictions. Given the coarse grid, step two will consist of leveraging a pre-trained model to segment the spatial extent of flood events on a much finer grid using remote sensing data. That segmentation output will contain many more validation data points, which can then be leveraged in deep learning prediction approaches.

3.1 Phase One - Building Baseline Performance Metrics on a Coarse Spatial Grid

Phase one will focus on building a model that predicts flood events on a coarse spatial scale with minimal re-sampling for predictors. From a technical standpoint, important predictors from ERA5 will be precipitation, temperature, soil type, land surface temperature, and leaf area index. These are already on a coarse 0.25×0.25 degree scale. Other predictors like land use, elevation, and slope are already on a finer spatial resolution than 0.25×0.25 degree. These will be down sampled in the spatial domain to 0.25×0.25 degrees. Since the flood observations from DFO are recorded in a single latitude and longitude coordinate, we will construct a 0.25×0.25 degree cell grid system which contains either flood or no-flood events based on the latitude and longitude. Given the relatively small number of true positive flood observations, statistical models like a Random Forest (RF) will be used to classify cells as a flood or no flood on a testing split of the data. See details in Figure 1 for a detailed description of the validation data in this phase.

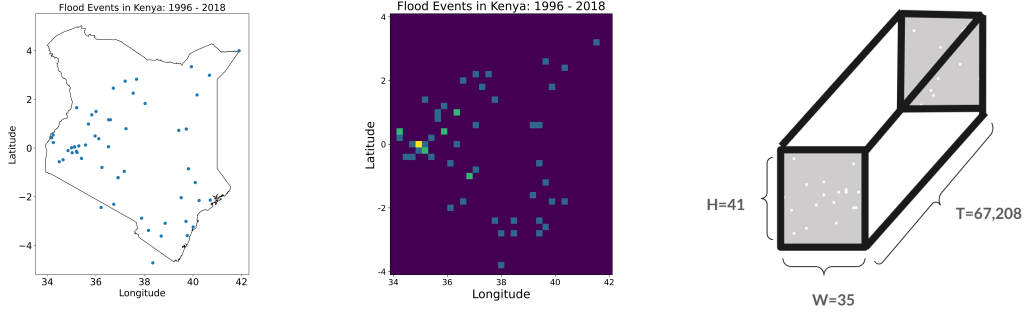


Figure 1: Ground truth validation events visualization

Left: Flood events from the DFO database in Kenya plotted over map of Kenya. **Middle:** Flood observations aggregated from 1996 - 2018 on a 0.25×0.25 degree spatial grid. Note how there is a flooding hot spot on the western side of Kenya around the 0 degree latitude line. **Right:** The data cube for ground truth events is a tensor in $\mathbb{R}^{T \times H \times W}$, where T is the number of observations in time. H and W are discretized Height (Latitude) and Width (Longitude). The cube's real world coordinates ranges from -5 to 5 latitude and 34 to 42.5 longitude. The cube's temporal frequency is every 3 hours, as this is the frequency of ERA5 observations that will be used. This means that from 1996 - 2018 there are $67,208 \times 41 \times 35$ ground truth points in the tensor and of those 6,216 are true positives (i.e., flood events that overlapped during a time step). It is possible for a flood event to span multiple days).

3.2 Phase Two - Generating Ground Truth Observations on a Fine Spatial Grid

The second phase will create a finer ground truth dataset to increase the number of validation points. Pre-trained models capable of segmenting Sentinel-1 or Sentinel-2 satellite images into categories such as flood, no flood, or permanent water will be used to generate a ground truth grid at a 30-meter spatial scale. This increases the spatial resolution from approximately 30 km to 30 meters - a factor of 1,000. While there will still be a large class imbalance in the ratio of pixels containing floods versus those without, there will be significantly more than 6,216 observations available to train a deep learning model to predict a flood event.

In order to find relevant images to segment, we plan to use an Application Programming Interface (API) to search for relevant Sentinel-1 or Sentinel-2 data overlapping with the latitude, longitude, and the nearest Sentinel-1 or Sentinel-2 observation after the flood event. We will establish an upper limit on the number of days an image can be used after a flood event. With this approach, we aim to create a validation data cube on the same temporal scale ($T = 67,208$), but with a significantly increase in spatial resolution - from 41×35 to approximately $41,000 \times 35,000$.

Note that the predictors from ERA5, such as precipitation and temperature, will need to be re-sampled to this finer grid. Other predictors, including land use, elevation, and slope, will also require re-sampling but because they natively exist at a spatial resolution much closer to 30 meters than 30 kilometers, their predictive potential can be utilized. Downscaling ERA5 predictors is another

very promising approach [8, 9] that can be used in tandem with the increased number of validation points but is currently beyond the scope of this proposal.

3.3 Modeling Approaches and Challenges: Sparsity, Models, and Metrics

As noted above, flood observations are sparse. In phase one, with only 6,216 true positive flood events available, we plan to use statistical machine learning classification models such as XGBoost and Random Forests. These models will take a vector of predictors from a cell on the spatial grid at a discrete time step to predict either a flood or no-flood event. We believe that deep learning models will have too many parameters compared to the number of validation points in this first phase to be effective.

Another approach to mitigate the sparse true positive observations is to train on the hotspot shown in Figure 1 and evaluate its generalization to the remaining sparser areas in Kenya. We also plan on using K-fold validation approaches and use that to understand the variance of accuracy metrics. Low variance in accuracy metrics across folds would increase our confidence in the results while having fewer true positives ground truth observations.

As outlined in the previous section, phase two generates additional validation points, enabling the application of deep learning approaches. Promising deep learning models for time series prediction include Long Short Term Memory (LSTM) or Gated Recurrent Unit (GRU) models, which can process a time series of observations over a spatial location to predict flood or no-flood events at each time step. To achieve this, the LSTM or GRU model requires a learned linear projection from the hidden state to a prediction layer at each time step. [10]. Transformer-based models are also suitable for this use case but the context window size will be an important design decision [11].

F1, Precision, and Recall will be metrics that we aim to measure. Additionally, it is crucial to clearly understand how many false negatives are produced because false negatives in the case of flood prediction can lead to inadequately mobilized resources in an EWS.

4 Pathway to Impact

In order to deploy this model, we plan to work with the Kenyan Water Resources Authority, Kenyan Meteorological Department, National Disaster Operations Center. These agencies are already responding to flooding disasters and we would like to keep them in the loop with iterative updates and modeling breakthroughs. We also do not plan to deploy this model directly to Kenyan residents. Rather, we plan to give the model outputs to the agencies above and let them decide on the appropriate distribution and communication channels as they already have expertise in this area.

Ethical considerations will also be taken into account. Specifically, locations in the training data that contain flood observations may already have mitigation measures in place. For instance, it is entirely possible that flooding events from 1998 may have prompted local authorities to build levees, reducing the likelihood of a future flood in that same spatial location. In this case, if we treat the training data as current, it is quite likely that similar predictor patterns might incorrectly predict a flood in that area. However, with the mitigation in place, a flood is unlikely to occur. Failing to account for this could potentially cause inefficient resource allocation by the authorities and organizations using the model outputs. Furthermore, if this data is used for insurance pricing, it is possible that an area with existing mitigation measures in place might be overpriced, potentially causing economic harm to local residents.

References

- [1] A. Kiptum, G. Mwangi, E. Otieno, A. Njogu, M. Kilavi, Z. Mwai, D. MacLeod, J. Neal, L. Hawker, T. O'Shea, H. Saado, E. Visman, B. Majani, and M. C. Todd. Advancing operational flood forecasting, early warning and risk management with new emerging science: Gaps, opportunities and barriers in kenya. *Journal of Flood Risk Management*, 2022.
- [2] A. Kiptum, Mwangi, E. Otieno, A. G. Njogu, M. Kilavi, Z. Mwai, D. MacLeod, J. Neal, L. Hawker, T. O'Shea, Visman Saado, H., E. Majani B., and M. C. Todd. Advancing operational flood forecasting, early warning and risk management with new emerging science: Gaps, opportunities and barriers in kenya. *Journal of Flood Risk Management*, 2023.

- [3] Suresh Sankaranarayanan, Malavika Prabhakar, Sreesta Satish, Prerna Jain, Anjali Ramprasad, and Aiswarya Krishnan. Flood prediction based on weather parameters using deep learning. *Journal of Water and Climate Change*, December 2020.
- [4] A. Mosavi, P. Ozturk, and K. Chau. Flood prediction using machine learning models: Literature review. *Water*, 2018.
- [5] Zaher Mundher Yaseen. Flood hazards and susceptibility detection for ganga river, bihar state, india: Employment of remote sensing and statistical approaches. *Results in Engineering*, 2024.
- [6] D. Bonafilia, B. Tellman, T. Anderson, and E. Issenberg. Sen1floods11: a georeferenced dataset to train and test deep learning flood algorithms for sentinel-1. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020.
- [7] Johannes Jakubik, Sujit Roy, C. E. Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniela Szwarcman, Carlos Gomes, Gabby Nyirjesy, Blair Edwards, Daiki Kimura, Naomi Simumba, Linsong Chu, S. Karthik Mukkavilli, Devyani Lambhate, Kamal Das, Ranjini Bangalore, Dario Oliveira, Michal Muszynski, Kumar Ankur, Muthukumaran Ramasubramanian, Iksha Gurung, Sam Khallaghi, Hanxi, Li, Michael Cecil, Maryam Ahmadi, Fatemeh Kordi, Hamed Alemohammad, Manil Maskey, Raghu Ganti, Kommy Weldemariam, and Rahul Ramachandran. Foundation models for generalist geospatial artificial intelligence, 2023.
- [8] Matthias Bittner, Sanaa Hobeichi, Muhammad Zawish, Samo Diatta, Remigious Ozioko, Sharon Xu, and Axel Jantsch. An lstm-based downscaling framework for australian precipitation projections. In *NeurIPS 2023 Workshop on Tackling Climate Change with Machine Learning*, 2023.
- [9] Giorgos E. Ntagkounakis, Panagiotis T. Nastos, and Yiannis Kapsomenakis. Statistical downscaling of era5 reanalysis precipitation over the complex terrain of greece. *Environmental Sciences Proceedings*, 26(1), 2023.
- [10] C. A. Toledo, M. Crawford, and T. Vyn. Maize yield prediction based on multi-modality remote sensing and lstm models in nitrogen management practice trials. *2022 12th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, 2022.
- [11] Claudia Toledo, Melba Crawford, and Mitchell Tuinstra. Integrating multi-modal remote sensing, deep learning, and attention mechanisms for yield prediction in plant breeding experiments. *Frontiers in Plant Science*, 2024.

A Appendix: Preliminary Results on Coarse Spatial Grid Prediction

In phase one, the objective was to establish a baseline understanding of flood prediction performance using a coarse grid of 0.25 x 0.25 degree input features and validation data.

Models were trained and tested for performance using only ERA5 features as well as ERA5 features augmented with elevation data. The DTM data was downsampled from 3 arcseconds to match the ERA5 grid. Specifically, a 300 x 300 pixel subsection was averaged at each quarter-degree to generate the downsampled DTM, as shown in Figure 2. Although the proposal mentioned incorporating land use and slope data, this remains an interesting avenue for future work but is not included in the results at the time of this paper’s publication.

For model training, one million true negative examples (no flood) features were randomly selected from the entire spatial-temporal cube. All 6,216 true positives were also selected, resulting in a total of 1,006,216 feature vectors. These training points were split into a 70 percent training and 30 percent test dataset. The results for the approximately 300,000 testing data points are shown in Table 1.

It is noteworthy that appending DTM data to the feature vectors improved the recall by more than 20 percent.

Table 1: Phase one results - Flood or no flood prediction on a coarse spatial scale

Model	F1	Precision	Recall
XGBoost - no DTM	0.63	0.66	0.54
Random Forest - no DTM	0.67	0.74	0.61
XGBoost - DTM	0.70	0.70	0.69
Random Forest - DTM	0.74	0.75	0.74

Models with no DTM are trained only on the ERA5 features as described in section 3.1. Models with DTM incorporated add one dimension to the features used for predicting a flood. Appending DTM as another dimension to the feature vector provides a significant boost the recall which implies a decrease of classifying true flood events as non-flood events. Code to reproduce results available at https://github.com/alimkarimi/flood_prediction

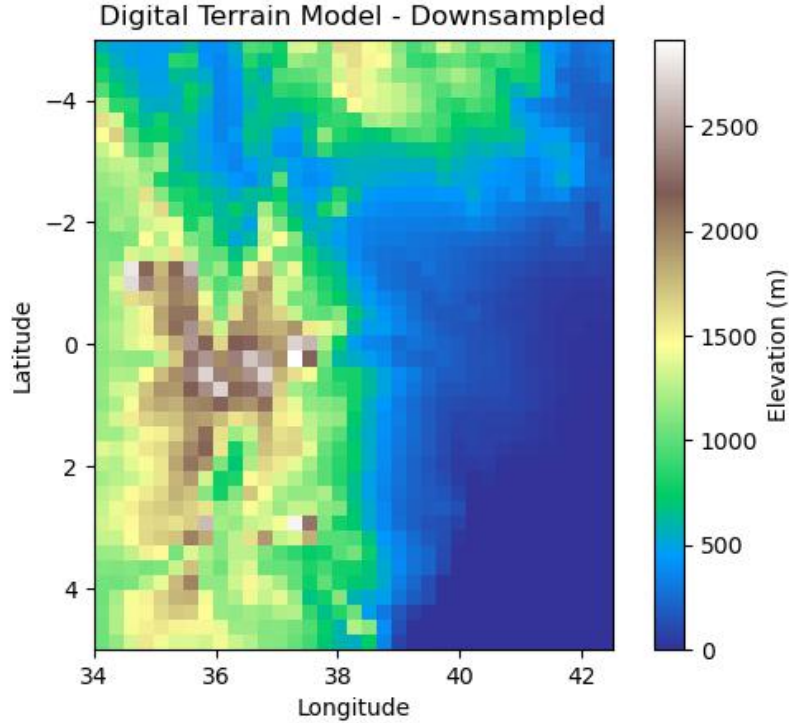


Figure 2: DTM Downsampling results. Note that the maximum elevations in the region is decreased because of downsampling. No filters were applied to the original DTM before the averages were computed on each cell.

B Appendix : Further details on foundation model for generating more validation data

To generate additional validation data, we propose leveraging a foundation model, Prithvi-100M which segments Sentinel-2 images into water or no water regions on a 30 meter pixel grid. From these segmentation results, floods can be inferred by techniques such as background subtraction. For instance, non-flood event water segmentation results can be used to compute a binary mask of regions where water is typically expected. By leveraging the temporal bounds of a flood event from the DFO

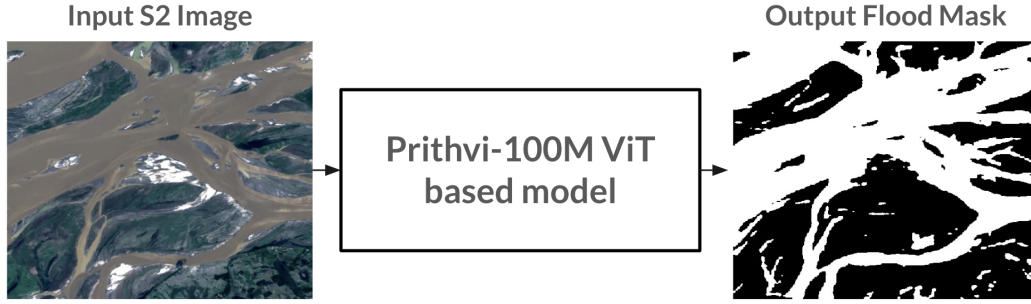


Figure 3: Example water and no water segmentation result

database, regions with water remaining after subtracting the expected water regions can be identified as flooded areas.

This approach dramatically increases the amount of data that can be generated for validation. A single 0.25 by 0.25 degree grid cell, approximately 30 kilometers by 30 kilometers, can be represented by 1,000 by 1,000, or 1 million flood or no-flood pixels. This increase in resolution is illustrated in Figure 4.

With this increase in the amount of validation data, other adjacent techniques, such as downscaling climate features, can be used to develop flood prediction models at a much finer spatial scale. This finer spatial scale is important in operationalizing early warning systems, as floods may not affect an entire 0.25 by 0.25-degree area.

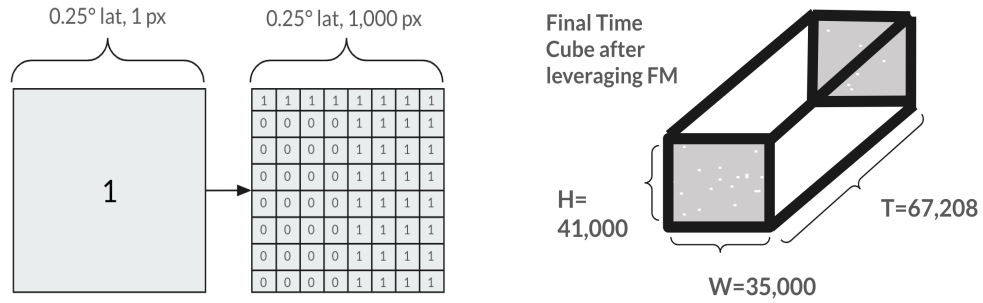


Figure 4: Visual of validation data increase and approximate resulting time cube. FM stands for Foundation Model. **Left:** Increased data visualization. One pixel, which represented an area of 0.25 by 0.25 degrees in phase one now contains 1 million pixels when the flood segmentation model is leveraged. **Right:** The spatial-temporal cube now has many more data points in the spatial dimension.