Coastal Flood Risk Prediction Using GIS and Machine Learning in Balayan, Batangas Province

Liezheel Mynha Y. Alejandro [1], Pocholo Miguel A. De Lara [1], Joe Marie M. Buela [1], Alexis Richard C. Claridades [1]

[1] Department of Geodetic Engineering, University of the Philippines Diliman, Quezon City, Philippines - (lyalejandro, padelara, jmbuela, acclaridades)@up.edu.ph

Keywords: Climate Resilience, Coastal Flood Risk, Flood Risk Map, IPCC Risk Framework, Random Forest Model.

Abstract

Recent coastal flooding in Balayan, Batangas, driven by Typhoon Neneng (NESAT, 2022), Severe Tropical Storm Kristine (Trami, 2024), and Typhoon Carina alongside the enhanced southwest monsoon (2024) which highlights the increasing vulnerability of coastal communities to extreme weather events. With climate change accelerating, there is an urgent need for data-driven flood risk assessment methods that integrate both environmental and socio-economic factors. This study employs a Geospatial Artificial Intelligence (GeoAI) workflow leveraging the Random Forest algorithm to estimate weights for key flood risk indicators within the IPCC risk framework, which conceptualizes risk as a function of hazard, exposure, and vulnerability. The model incorporates diverse hydroclimatic and socio-environmental variables, such as elevation, slope, projected population density, proximity to mangroves and seagrasses, distance to the coastline, and sea level rise projections under RCP 8.5 that represents a high-emissions climate scenario. To ensure consistency and transferability, all input variables were normalized to a 0–1 scale via min-max scaling, and standardized risk thresholds were applied. The GeoAI workflow was trained on historical flood data spanning 2022–2024 and validated under future climate conditions to minimize subjectivity and data imbalance while enhancing robustness. The Random Forest (RF) model achieved high predictive performance, with a mean cross-validation accuracy of 98.62% and test accuracy of 98.73%, which effectively discriminates flooded from non-flooded zones. Furthermore, this workflow supports near-term planning by providing spatially explicit risk maps and quantifying potential impacts such as projected agricultural losses, thereby informing climate-resilient adaptation strategies and policymaking.

1. Introduction

Coastal zones are among the most economically important regions worldwide and support a significant portion of the global population, with approximately 40% living within 100 kilometers of the coast (Agardy, 2005). These regions serve as vital centers for fishing, agriculture, tourism, and urban development. Moreover, coastal ecosystems provide essential services, including storm buffering, carbon storage, and the preservation of biodiversity.

Coastal municipalities such as Balayan in Batangas Province have faced increasingly severe flooding in recent years. Notable events include Typhoon Neneng (NESAT) in 2022, Severe Tropical Storm Kristine (Trami) in 2024, and floods caused by Typhoon Carina and the enhanced southwest monsoon in 2024, which led to widespread displacement and economic losses. Balayan's susceptibility to flooding is heightened by its coastal location and relatively low elevation, as identified in the Batangas Provincial Development and Physical Framework Plan 2022–2030.

Despite rising coastal flood risks, existing prediction methods in the Philippines are limited. Most models emphasize physical factors such as rainfall, elevation, and proximity to coastline but often overlook socio-environmental factors like population exposure and protective ecosystems. Furthermore, many models do not incorporate future climate scenarios which are essential to assess long-term risks. Additionally, the application of GeoAI in flood risk assessment remains limited in the Philippine context, despite its proven potential to enhance model accuracy and reproducibility.

Studies worldwide have incorporated socio-environmental factors in coastal flood risk assessments. For instance, a study (Asiri et al. 2024) applied GeoAI techniques by integrating the Analytical Hierarchy Process with Support Vector Machine (AHP-SVM) and Decision Trees (AHP-DT) to assess flood susceptibility in Iran. Their AHP-DT model achieved high accuracy (AUC = 0.95), which demonstrates the effectiveness of combining expert-based weighting with machine learning. The study included variables such as population density and elevation that emphasize the role of socio-environmental factors.

Similarly, a study (Wu, 2021) highlighted the importance of socio-economic variables, including low income and education levels, in increasing flood vulnerability. This underscores the need to integrate these human dimensions for comprehensive coastal flood risk assessments.

Another study (Atmaja et al., 2024) found that the Random Forest model outperformed KNN and ANN in predicting coastal flood risk, achieving a precision of 0.86 and recall of 0.67. By including socio-economic and environmental factors, their study highlighted the importance of integrating human and natural elements in flood risk assessment.

The objective of this study is to develop a geospatial framework for coastal flood risk prediction in Balayan, Batangas, using the Random Forest (RF) model. The framework incorporates elevation, slope, population projections, proximity to mangroves and seagrasses, and future climate scenarios to improve accuracy and relevance. Flood risk maps were developed using the Random Forest method, specifically Random Forest, and compared to the Intergovernmental Panel on Climate Change (IPCC) risk framework, which defines risk as a function of hazard, exposure, and vulnerability. Two modeling approaches were applied using the IPCC risk framework: one using equal weighting of variables and another utilizing feature importance derived from RF analysis. The study also introduces a scalable

https://doi.org/10.5194/isprs-annals-X-5-W4-2025-19-2026 |

method for quantifying flood impacts, demonstrated through projected agricultural damage and losses. While agriculture is the initial focus, the methodology is designed to be adaptable for other socio-economic sectors in future research.

2. Materials and Methodology

The coastal flood risk assessment integrates geospatial indicators of vulnerability, exposure, and hazard, which follows a structured framework based on both IPCC guidelines and data-driven modeling. Figure 1 shows the workflow comprises two main approaches: (1) an IPCC-based approach utilizing weighted overlay analysis with both equal weights and Random Forest Feature Importance (RFFI) weights and (2) a machine learning-based approach using Random Forest (RF) classification to derive risk levels based on historical flood extents.

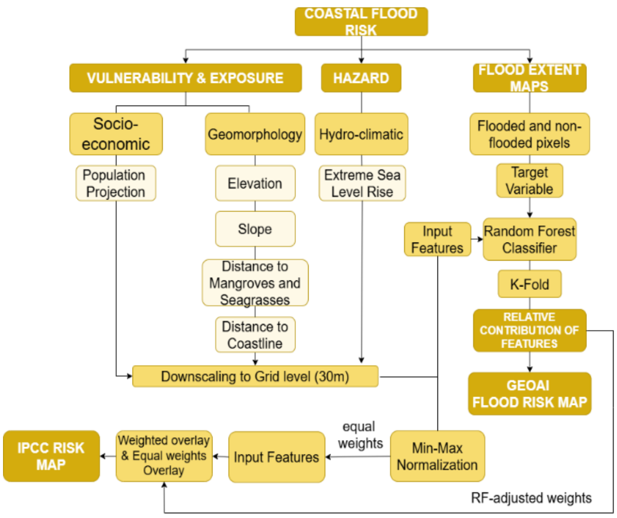

Figure 1. Workflow for the generation of Flood Risk Projections and Flood Risk using the IPCC Risk Approach

2.1 Study Area

Balayan, a coastal municipality in Batangas, lies approximately 106 kilometers from Manila, the national capital, and 48 kilometers from Batangas City, the provincial capital. Figure 2 shows the geographical boundaries of the municipality.

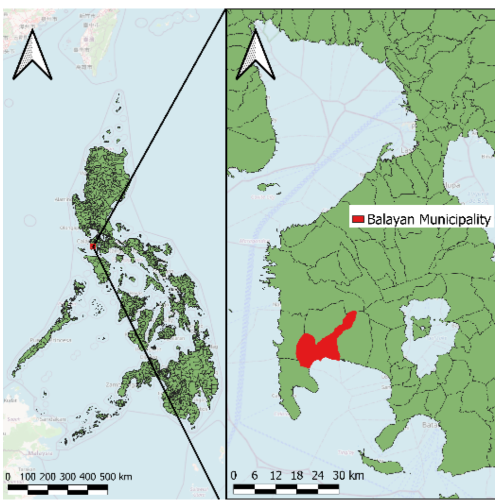

Figure 2. The geographical bounds of the Municipality of Balayan

Balayan is known for its coastal setting, but it also experiences rapid population growth. Its key agricultural products include sugarcane, rice, and corn. In recent years, the people living in Balayan, Batangas have experienced flood events caused by typhoons and tropical storms. Due to these flood events, yearly crops are destroyed, and livelihoods are disrupted. Thus, an accurate flood risk assessment is necessary to prevent and reduce the disruption of economic activity in the municipality of Balayan.

2.2 Data Sources

According to the available data for Balayan, Batangas Province, and a literature review (Atmaja et al., 2024), six key factors have been identified that potentially indicate an increased risk affecting an area's risk of coastal flooding. These factors encompass a wide range of variables, including physical, ecological, and demographic factors: Extreme Sea Level (ESL), proximity to the coastline, elevation, slope, distances from mangroves and seagrasses, and population density, which are essential for modeling coastal flood indicators and identifying high-risk zones. By integrating data from global and national sources, the model captures spatial heterogeneity and the dynamic interactions between natural and human systems that enhance the predictive capacity of the RF model. To build a reliable coastal flood risk prediction framework, this study utilized multi-source geospatial and socio-environmental datasets derived from both literature and open-access repositories.

The factors selected for this study were based on their relevance to coastal flood risk as supported by previous literature:

- Extreme Sea Level (ESL) is a key driver of coastal inundation, accounting for storm surges and sea level rise (Kirezci et al., 2020).
- Proximity to the coastline indicates exposure, with nearer areas more likely to be affected by flooding (Neumann et al., 2015).
- Elevation is critical, as low-lying zones are more susceptible to water accumulation (Park and Lee, 2020).
- Slope influences flood extent by affecting how quickly water spreads over land (Atmaja et al., 2024)
- Distances from mangroves and seagrasses were included due to their role as natural buffers that reduce wave energy and limit flood impacts (Atmaja et al., 2024)
- Population density represents the level of human exposure (Hauer et al., 2021)

Table 1 illustrates the key variables that collectively enhance the Random Forest model's ability to simulate flood risk with spatial and contextual accuracy.

Key Factor	Raw Data Format	Temporal Range	Source	Feature Encoding
Coastal Flood Occurrence	Polygon	2022 - 2024	PhilSA - HDX	Flooded and Non-flooded areas
ESL Inundation	Point (CSV format)	1980 - 2100	European Commission	Meters (vertical)
Coastline Proximity	Polygon	NA	NAMRIA	Distance in meters
Elevation	GeoTIFF	1970 - present	NASA EarthData	Meters above sea level (masl)
Slope	GeoTIFF	1970 - present	NASA EarthData	Percent slope
Mangrove and Seagrass Distance	Polygon	2020	NAMRIA	Distance (meters)
Population Density	GeoTIFF	2020	WorldPop	People per sq. km

Table 1. Key Variables and Data Sources

2.3 GeoAI Risk Approach Modelling

The GeoAI-based flood risk modeling utilizes a supervised machine learning classifier to derive spatial risk predictions from historical flood data and associated biophysical and socio-economic predictors. In this study, the Random Forest (RF) algorithm was utilized since it uses tree-like structures to select the most appropriate class for given inputs (Breiman, 2001). All the pre-processed raster data, including the labeled classes, in our case, the binary raster coastal flood occurrence, are split into training and test data in a ratio of 75-25. The scikit-learn toolkit is used to implement the Random Forest model and its Feature Importance to output the Gini Importance, which is used as the relative importance or weights of each key factor.

The RF model's performance was assessed using 10-fold cross-validation, where the dataset is split into ten parts, each serving as a test set once while the remainder forms the training set. This yields ten accuracy scores, which allows for a robust evaluation of model generalization. A classification report summarizing metrics such as precision, recall, and F1-score was also used for a detailed performance assessment.

All the data's maximum values are bounded up to Low Elevation Coastal Zones (LECZ). LECZ is defined as areas along the coast where their elevations are less than 10 meters (McGranahan and Anderson, 2007). Populations living in these areas are most prone to coastal flooding.

Key factors formatted as Vector Data (Point or Polygon) are first converted to raster for analysis. The spatial resolution used for all raster data for this analysis is 30 x 30 meters. Thus, all gathered raster data must be converted to 30 x 30 meters resolution for proper alignment and consistency of data for analysis.

The coastline was first extracted, and then a proximity raster was generated from the extracted coastline of Balayan. Mangrove and Seagrass extents were directly used to generate proximity raster.

The ESL data from the European Commission is in point vector form, representing sea level conditions at coastal locations around the world. To generate a spatial distribution of ESL across the coastline, the data are first interpolated for the baseline period, RCP8.5 2050 and RCP8.5 2100. These datasets are used to train a random forest model and generate future coastal flood risk maps in the study area. Each dataset includes geographic coordinates along with statistical values such as the median, 5th percentile, and 95th percentile of projected sea levels. For this study, the median value of each point was used. The point data are spatially interpolated to produce gridded layers at a coarse resolution of 0.54 degrees, consistent with the original data spacing. The grids are resampled to 30 x 30 meters resolution to support detailed flood risk analysis under future climate scenarios.

Elevation data derived from NASA are clipped to the study area. The spatial resolution of the elevation data is already at 30 x 30 meters. Since the Slope is derived from elevation data, the resolution and extent of the slope were inherited from the pre-processed elevation data. On the other hand, population Density gathered from the WorldPop of the School of Geography and Environmental Science of the University of Southampton was also resampled into 30 x 30 meters since the data represented is in population count per square kilometer.

Table 2 shows the coastal flooding occurrence data gathered from the Philippine Space Agency's (PhilSA) Flood Extent maps, which are derived from SAR (Synthetic Aperture Radar) images. These maps provide vector data corresponding to areas suspected to be flooded. The coastal flood occurrence vector data are then converted to raster format where flooded areas correspond to value one (1) and non-flooded areas to value zero (0). Union operation was used to combine multiple flood extent datasets, which ensured comprehensive coverage of all flooding events across the study period.

Flood Event	Date
Flooding	August 27, 2022
Typhoon Neneng (Nesat)	October 14, 2022
Flooding	July 17, 2023
Tropical Cyclone Egay (Doksuri)	July 29, 2023
Typhoon Hanna (Haikui)	September 03, 2023
Flooding	November 26, 2023
Southwest Monsoon (Habagat) and Typhoon Carina (Gaeemi)	July 23, 2024
Severe Tropical Storm Kristine (Trami)	October 24. 2024
Super Typhoon Pepito (Man-yi)	November 20, 2024

Table 2. Flood Events gathered from PhilSA and their corresponding dates

2.4 IPCC Risk Approach Modelling

The IPCC-based risk approach adopts a deductive multi-criteria evaluation (MCE) framework, in which flood risk is conceptualized as a function of hazard, exposure, and vulnerability, consistent with the principles established by the Intergovernmental Panel on Climate Change. To generate the IPCC-based flood risk maps, all key factors were first normalized to a common, unitless scale ranging from 0 to 1. This standardization was accomplished using min-max normalization, expressed mathematically as:

$$X_{i,norm} = \frac{X_i - X_{min}}{X_{max} - X_{min}}, \quad (1)$$

$$X_{i,norm} = \frac{X_{max} - X_i}{X_{max} - X_{min}}, \quad (2)$$

where X_i = individual data point to be normalized
X_{min}= lowest value for a factor
X_{max} = highest value for a factor
$X_{i,\,norm}$ = normalized data point

Equation (1) is applied to factors that are directly proportional to flood risk, where higher values indicate greater risk, while Equation (2) is used for factors with an inverse relationship, where higher values indicate reduced susceptibility to flooding.

Using the normalized factors, two flood risk maps were created through overlay analysis: one with equal weights, assuming all factors contribute uniformly, and another weighted by Random Forest-derived feature importance scores. For slopes and population density, classifications from the Bureau of Soils and Water Management (BSMW) and Degree of Urbanisation (DegUrba) from the United Nations were used for classifying the said factors according to vulnerability to coastal flooding. Elevation, and proximity key factors (coastline and mangrove and seagrass) are assumed only to be effective for areas under LECZ. Thus, their proximities were only bounded by the furthest area under 10 meters of elevation which was measured to be 3821 meters from the coast. Table 3 summarizes the key factors, their thresholds, and their assumed relationships to coastal flood risk.

Key Factor	Thresholds		Relation to Coastal Flood	Source
	Min	Max		
Elevation	<= 0 m	> 10 m	Indirect	LECZ
Slope	<= 3%	> 50%	Indirect	Classification from BSWM
Coastline Proximity	0	> 3821 m	Indirect	Max distance of LECZ from Coast
Mangroves and Seagrass proximity	0	> 3821 m	Direct	Max distance of LECZ from Coast
Population Density	0 person per sqm	1500 person per sqm	Direct	DegUrba Classification
ESL Inundation		9.22 m	Direct	Maximum ESL from RCP8.5 2100

Table 3. Summary of thresholds used for data normalization of factors for IPCC Risk approach flood risk map

2.5 Flood Risk Projection

The model generated from training the Random Forest model was used to predict Flood risk for 2050 and 2100 based on the IPCC's Representative Concentration Pathway (RCP) 8.5. RCPs are defined as scenarios of time series or projections of greenhouse gas (GHG) concentrations (Moss et al., 2008). It explores different scenarios for GHG and its effect on climate depending on our response as a society. For the Flood risk projection generation, the ESL RCP8.5 is used to generate the flood risk for 2050 and 2100.

For the projection of population density, the Compound Annual Growth Rate (CAGR) method was used and described as follows:

$$CAGR = (\frac{V_f}{V_0})^{1/t} - 1, \quad (3)$$

where $CAGR_i$ = Compound Annual Growth Rate
V_f= final population count
$V_{0;}$ = initial population count
t = years between the initial and final population

2.5.1 Projected Crop Damage from IPCC Risk Approach

Coastal Flood Risk Maps: To provide a comparison of projected damage to crops for areas under High and Very High risk in both Feature Importance and Equal Weights flood risk maps, an estimate of projected damage to crops is done by intersecting the high and very high-risk areas with Annual and Perennial Crop areas from land cover map from NAMRIA. The area of intersection is then computed to estimate the value of damage to crops.

The Department of Agriculture – Disaster Risk Reduction Management Operation Center's (DA-DRRM-OC) Facebook page posts bulletins for information dissemination of damages to crops and areas affected by disasters, particularly flooding. For this study, DA-DRRM-OC's Bulletin No. 13, dated November 4, 2024, was used to estimate the value of crop damage. To estimate the value of crop damage per area, it is computed using:

$$Value\ Loss\ per\ Area = \frac{Value\ Loss\ (Rice, HVC, Corn)}{Area\ (ha)}, \quad (4)$$

The data from the said bulletin is collated and summarized in Table 4. Using equation 3, the estimated Value Loss per Area is 46,054.35 PHP/ha.

Name	Area Affected (ha)	Value Loss in PHP
Corn	2,123	7,405,000
Rice	107,820	4,460,000,000
High-Value Crops	7,290	865,090,000
TOTAL	**117,233.00**	**5,399,090,000.00**

Table 4. Summary of Value Loss and Area of Crops Affected by Flooding as reported in the DA-DRRM-OC's Bulletin No.13 for Severe Tropical Storm Kristine (Trami)

3. Results and Discussions

Coastal flood risk is influenced by physical, ecological, and demographic factors such as ESL rise, proximity to the coast, elevation, slope, natural buffers like mangroves & seagrasses, and population density. These factors influence not only the probability of flooding but also the scale of potential damage to both people and infrastructure.

In this study, a combination of Geographic Information Systems (GIS) and Python-based data processing techniques was employed to quantify and spatially analyse these variables. Flood extent maps from the PhilSA which captures flood occurrences from 2022 to 2024 were used as the primary reference for identifying and classifying flood-prone versus non-flooded areas across the municipality, as shown in Figure 3.

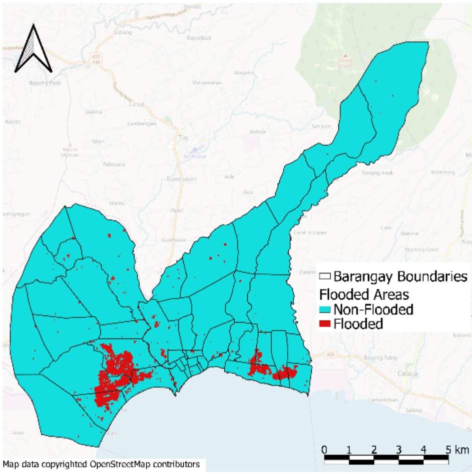

Figure 3. Spatial distribution of 3127 flooded locations between 2022 - 2024 and Balayan, Batangas barangay boundary.

3.1 Coastal Flood Risk Using the IPCC Risk Approach for Overlay Analysis

Flood risk maps for Balayan, Batangas were generated using the IPCC Risk approach with two weighting methods: equal weights and weights based on Feature Importance scores from a Random Forest model.

Figure 4 shows the Gini importance scores for the 2020 baseline, which highlights elevation as the most influential factor, followed by population density, mangrove/seagrass distance, and coastal proximity. Slope had a moderate impact, while ESL had minimal influence. These scores informed the adjusted IPCC risk weights, which will provide a data-driven alternative to equal weighting.

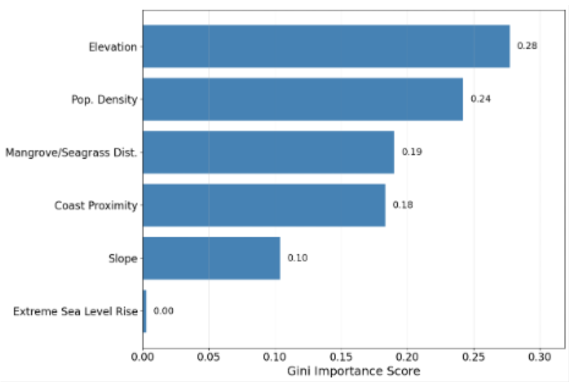

Figure 4. Gini Feature Importance

Figure 5 shows the optimal threshold for the IPCC RFFI method was 0.4300, while the threshold for the IPCC Equal Weights method was 0.4340. These thresholds indicate that areas with risk scores equal to or greater than the specified value are classified as flood risk. The small difference between the two thresholds suggests that both methods resulted in similar risk score distributions. Applying these thresholds facilitates the generation of interpretable and actionable binary flood risk maps that can support effective decision-making in flood mitigation.

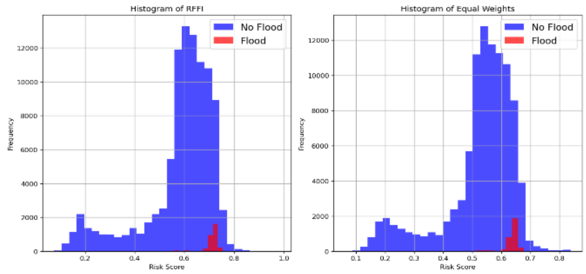

Figure 5. Coastal flood risk distribution from the IPCC flood risk maps overlaid with coastal flood occurrence data.

The two weighing methods for the IPCC flood risk maps were evaluated using the sourced coastal flood occurrence. The dataset used for evaluation exhibited a highly imbalanced distribution with 96,293 instances labeled as non-flood and only 3,127 instances labeled as flood, which results in a flood rate of just 3.1%. To determine the effectiveness of the IPCC risk maps in identifying flood-risk areas, optimal threshold values were identified for converting continuous risk scores (ranging from 0 to 1) into binary flood predictions.

3.2 Comparison of the Random Forest Model and the IPCC Flood Risk maps

A comparative evaluation was conducted between the Random Forest (RF) model and two IPCC risk-based flood risk mapping approaches using normalized performance metrics, specifically precision, recall, F1-score, accuracy, and AUC, to ensure fairness despite differing dataset sizes. The RF model was trained on 2020 data comprising 24,855 observations, while the IPCC-based models used a larger dataset of 99,420 observations. Table 5 summarizes the confusion matrix components for each model, showing true negatives (TN), false positives (FP), false negatives (FN), and true positives (TP).

Metric	IPCC RFFI Weights	IPCC Equal Weights	Random Forest (GeoAI)
TN	78,800	78,844	23,984
FP	17,493	17,449	89
FN	211	274	227
TP	2,916	2,853	555
Total Observation Data	99,420	99,420	24,855

Table 5. Confusion Matrix Summary for Flood Risk Classification Using IPCC and GeoAI Models

Table 6 presents the normalized evaluation metrics for the models. The RF model achieved a precision of 0.8618, indicating that 86% of its flood predictions accurately classify flooded areas. Its recall score of 0.7097 reflects the model's ability to correctly identify 71% of actual flood-affected areas, while an F1-score of 0.7784 shows balanced performance between precision and recall. The RF model also recorded a high overall accuracy of 98.73% and an AUC score of 0.9862, indicating excellent classification performance.

In contrast, the IPCC RFFI and Equal Weights models had much higher false positive rates (17,493 and 17,449), resulting in low precision (~0.14) despite very high recall scores (0.93 and 0.91). Their F1 scores (~0.24) were considerably lower than the RF model, indicating an imbalance in predictive power. These IPCC models effectively flagged most flood-prone areas but tended to overpredict, leading to many false alarms.

Metric	IPCC RFFI Weights	IPCC Equal Weights	Random Forest (GeoAI)
Accuracy	0.8219	0.8217	0.9873
Precision	0.1429	0.1405	0.8618
Recall	0.9325	0.9124	0.7097
F1 Score	0.2478	0.2435	0.7784
AUC Score	0.8887	0.889	0.9862

Table 6. Evaluation Metrics for Flood Risk Classification Across IPCC and GeoAI Models

Both evaluation metrics show that while the IPCC models prioritize sensitivity (recall), the RF model balances sensitivity and specificity by reducing both false positives and false negatives. This makes the RF model more suitable for practical flood risk management, where avoiding overprediction is critical to efficient resource allocation and maintaining public trust in risk alerts.

3.3 Coastal Flood Risk Projections

The Random Forest (RF) model was applied to simulate flood risk under both current and future climate and population scenarios. The baseline reflects conditions as of 2020, while projections for 2050 and 2100 integrate anticipated increases in Extreme Sea Level (ESL) under the RCP8.5 high-emissions pathway. Population estimates for these future years were derived using Equation 3.

3.3.1 Predictive Performance: Figure 6 shows that although the model maintains high overall accuracy (above 98%), its performance metrics reveal a declining ability to identify flood-prone areas over time, which indicates potential limitations for long-term risk assessment.

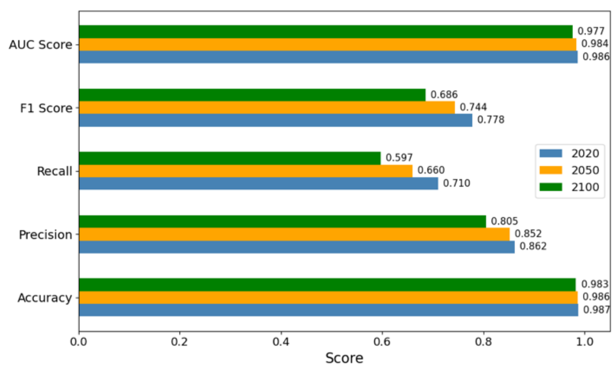

Figure 6. RF Model Evaluation Report

Figure 6 shows that although the model maintains high overall accuracy (above 98%), its performance metrics reveal a declining ability to identify flood-prone areas over time, which indicates potential limitations for long-term risk assessment.

The recall performance declined from 71% in 2020 to around 60% by 2100. This trend suggests the model becomes more conservative over time, which reduces false positives but increasingly fails to detect flood-prone areas. Such limitations may hinder long-term disaster preparedness. These findings highlight the need to enhance model sensitivity, potentially by incorporating additional environmental variables or using hybrid machine-learning approaches. While the RF model offers a solid framework for flood risk assessment, refinements are needed for future-oriented accuracy.

3.3.2 Spatial Flood Risk Trends: The 2050 and 2100 spatial projections show predicted coastal flood risks under the RCP8.5 scenario and population growth. Compared to the baseline and historical data, these two future scenarios indicate a decrease in coastal flood occurrences, contrary to typical climate change expectations (see Figures 7 and 8).

The baseline year (2020) recorded the highest projected number of coastal flood pixels at 644 out of 24,855 (2.59%). This declined to 606 pixels (2.44%) in 2050 and further to 580 pixels (2.33%) in 2100, representing a total reduction of 0.26 percentage points from the 2020 baseline, as shown in Figure 7. The consistent decrease across future scenarios, most notably in 2100, contradicts typical climate change expectations, which suggests possible model limitations or shifts in environmental factors influencing flood susceptibility.

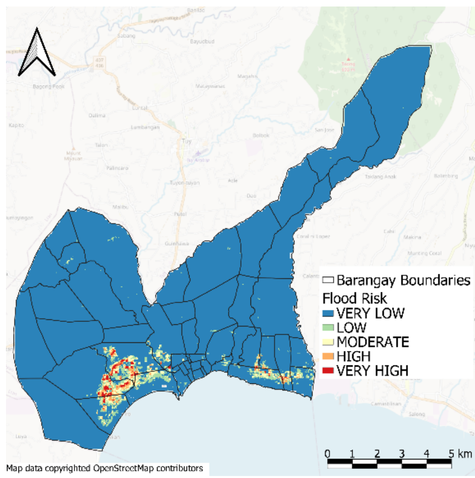

Figure 7. Flood Risk Projection in 2050

Figure 8 shows that by 2100, the projected number of coastal flood areas was the lowest between the two flood projections, totaling 580 out of 24,855 points, or 2.33%. This represents a decrease of 0.26% compared to the baseline year 2020, which had 644 flood points (2.59%). The consistent reduction observed from 2020 to 2100 suggests a downward trend in projected coastal flood occurrences. However, this trend is contrary to typical climate change expectations where increased sea level rise and extreme events would likely result in greater flood exposure. The 2100 results may reflect model limitations or shifts in environmental and demographic factors reducing future flood risk.

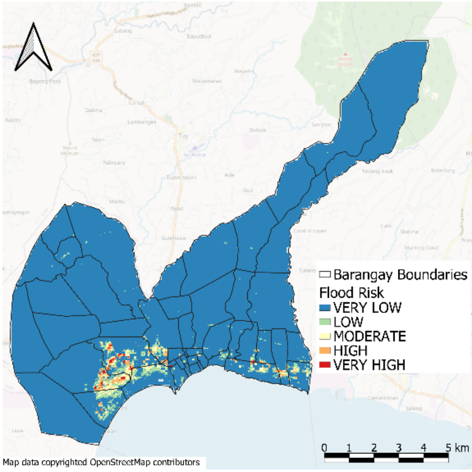

Figure 8. Flood Risk Projection in 2100

One of the possible reasons for the downward trend observed for the predicted flood risks in Balayan may be due to the flood events used to train the Random Forest model. They are derived from satellite images and still need to be validated on the ground as noted by the data source, PhilSA. Flooded areas were mapped using radar satellite images. Radar satellite images, when used for flood mapping, generally will have difficulty distinguishing between flooded areas and paddy fields. Thus, may result in specific areas to be always flooded and do not correlate with the data used in the study.

3.4 Estimate of Crop Value Loss due to Flooding from IPCC Flood Risk Maps

This study compares two IPCC flood risk maps, one with equal weights and one with Random forest-based weights, to assess flood impacts on annual and perennial crops from NAMRIA's Land Cover Map, as shown in Figure 9. It estimates vulnerable crop areas and potential economic losses, offering key insights for agricultural planning and risk management.

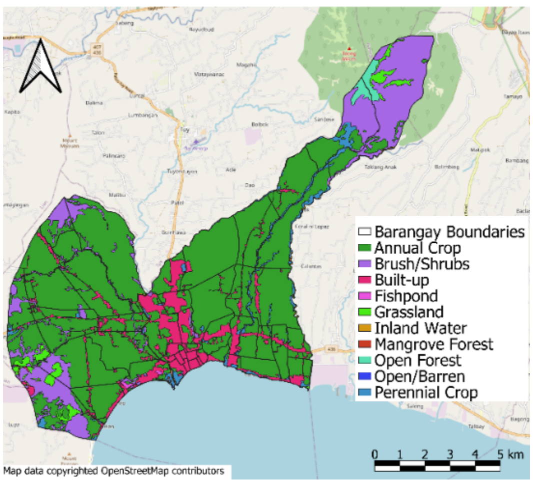

Figure 9. Land Cover Map of Balayan from NAMRIA's Geoportal

3.4.1 Comparison of Weighted Risk Maps: In the RF-weighted IPCC flood risk map, elevation data was assigned with the highest feature importance with a score of 0.28. At the same time, population density was assigned a high feature importance score of 0.24. the weights for both key factors for equal weights is 17%. This indicates that very low elevations and densely populated areas are more strongly associated with higher flood risk, than the equal weights IPCC flood risk map which contributes to the larger crop areas identified as vulnerable in this model. Figure 10 shows the large agricultural areas at high flood risk identified by the RF-weighted IPCC flood risk map.

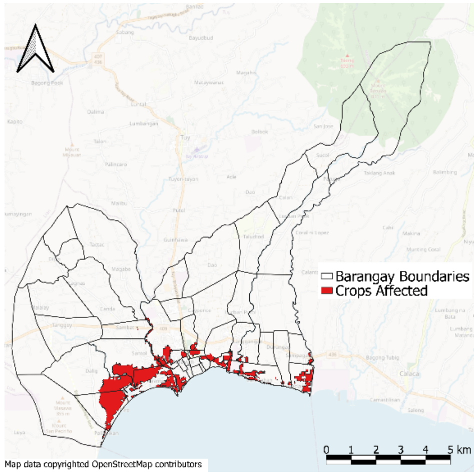

Figure 10. Flood-affected crops based on the IPCC Flood risk map using Random Forest feature importance as weights

In contrast, Figure 11 illustrates the flood-affected crop areas identified using equal weights for all key flood risk factors. The spatial extent of vulnerable crops is considerably smaller compared to the RF-weighted map, which suggests that this approach may underestimate flood impacts on agriculture.

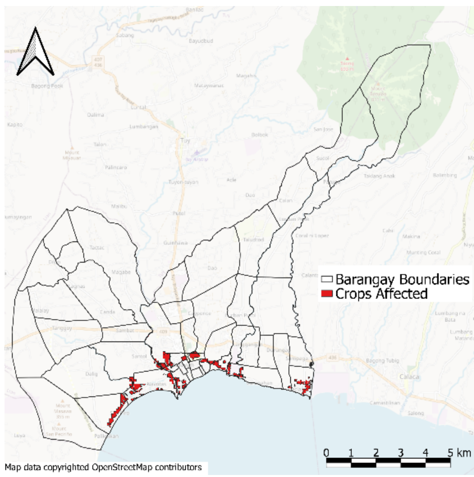

Figure 11. Flood-affected crops based on the IPCC Flood risk map using equal weights for all key factor

3.4.2 Economic Impact on Agriculture: Table 7 quantifies this difference, which shows the RF-weighted map estimates 433.8 hectares of crops at risk and PHP 19.98 million in losses, while the equal weights map estimates 126.3 hectares and PHP 5.82 million. The 343.89% difference highlights the value of data-driven weighting in flood risk assessment.

	Affected Area (ha)	Total Value Loss (PHP)
Random Forest Feature Importance	433.827	19,979,620.50
Equal Weights	126.264	5,815,006.45

Table 7. Comparison of crops affected and potential value loss for each IPCC Flood risk maps

These findings suggest that the higher importance given to population density and elevation in the RF model increases the identification of vulnerable agricultural areas, providing a more realistic and comprehensive assessment of flood risk impacts. This supports better-informed planning and resource allocation for flood risk reduction.

4. Conclusions and Recommendations

This study applied GIS and machine learning, using the Random Forest model, to assess coastal flood risk in Balayan, Batangas by incorporating physical, ecological, and demographic factors along with climate projections. This study demonstrates a location-specific application of these methods within the Philippine context, aligned with local disaster risk reduction and climate change adaptation (DRR-CCA) efforts. By combining spatial hazard modeling with machine learning at the municipal scale, the study offers a practical and adaptable framework for coastal communities facing climate variability.

The study generated flood risk predictions for the years 2050 and 2100 by integrating future climate scenarios and current population growth trends in the municipality. While the 2050 projection yielded promising results, the longer-term forecast for 2100 indicated a lower estimated flood risk. This discrepancy reflects limitations in the model and the current lack of detailed, time-specific data. Since flooding is a highly time-bound phenomenon, the absence of event-specific observations during the flood reduces prediction accuracy.

The study also has limitations that can be improved through future studies by incorporating more physical factors like rainfall data during the coastal flooding occurrence, storm surge data, and vertical land motion across the LECZ. Physical factors, particularly vertical land motion, can exaggerate the effects of extreme sea level rise as a subsiding area can be more prone to retreat of coastal areas. Using updated datasets and incorporating additional socio-economic factors (e.g., median barangay income), exposure indicators (e.g., number or density of roads), and vulnerability measures (e.g., proximity to evacuation centers) could enhance the flood risk prediction capability of the RF model.

Future studies can apply the framework used in this study for other coastal municipalities with different data. Random Forest can be used with both physical and socio-economic data that can be categorized as Hazard, Exposure, Vulnerability to check for their correlation with flooding events. The resulting correlation or model can be used to predict future flood risk given projection of data used to create the prediction model.

References

Agardy, T., Dayton, J., Curran, S., Kitchingman, A., Wilson, M., Catenazzi, A., Restrepo, J., Birkeland, C., Blaber, S., Saifullah, S., Branch, G., Boersma, D., Nixon, S., Dugan, P., Davidson, N., Vorosmarty, C., 2005: Coastal Systems. *Ecosystems and Human Well-being: Current State and Trends*. *Island Press*., 513–549.

Asiri, M.M., Aldehim, G., Alruwais, N., Allafi, R., Alzahrani, I., Nouri, A.M., Assiri, M., Ahmed, N.A., 2024: Coastal Flood risk assessment using ensemble multi-criteria decision-making with machine learning approaches. *Environmental Research*., Volume 245, 118042. doi.org/10.1016/j.envres.2023.118042.

Atmaja, T., Setiawati, M.D., Kurisu, K., Fukushi, K., 2024: Advancing Coastal Flood Risk Prediction Utilizing a GeoAI Approach by Considering Mangroves as an Eco-DRR Strategy. *Hydrology*., 11, 198. doi.org/10.3390/hydrology11120198.

Breiman, L., 2001: Random Forests. *Machine Learning*., 45, 5–32. doi.org/10.1023/A:1010933404324.

Hauer, M.E., Hardy, D., Kulp, S.A., Mueller, V., Wrathall, D.J., Clark, P.U., 2021: Assessing population exposure to coastal flooding due to sea level rise. *Nature Communications*., 12, 6900. doi.org/10.1038/s41467-021-27260-1.

Kirezci, E., Young, I.R., Ranasinghe, R., Muis, S., Nicholls, R.J., Lincke, D., Hinkel, J., 2020: Projections of global-scale extreme sea levels and resulting episodic coastal flooding over the 21st Century. *Scientific Reports*., 10, 11629. doi.org/10.1038/s41598-020-67736-6.

McGranahan, G., Balk, D., Anderson, B., 2007: The rising tide: assessing the risks of climate change and human settlements in low elevation coastal zones. *Environment & Urbanization*., 19(1), 17–37. doi.org/10.1177/0956247807076960.

Moss, R., Babiker, M., Brinkman, S., Calvo, E., Carter, T., Edmonds, J., Elgizouli, I., Emori, S., Lin, E., Hibbard, K., Jones, R., Kainuma, M., Kelleher, J., Lamarque, J.-F., Manning, M., Matthews, B., Meehl, J., Meyer, L., Mitchell, J., Zurek, M., 2008: Towards New Scenarios for Analysis of Emissions, Climate Change, Impacts, and Response Strategies. IPCC Expert Meeting Report, Intergovernmental Panel on Climate Change.

Neumann, B., Vafeidis, A.T., Zimmermann, J., Nicholls, R.J., 2015: Future Coastal Population Growth and Exposure to Sea-Level Rise and Coastal Flooding - A Global Assessment. *PLOS ONE*., 10(6). doi.org/10.1371/journal.pone. 0118571.

Park, S.-J., Lee, D.-K., 2020: Prediction of coastal flooding risk under climate change impacts in South Korea using machine learning algorithms. *Environmental Research Letters*., 15, 094052. doi.org/10.1088/1748-9326/aba5b3.

Wu, T., 2021: Quantifying coastal flood vulnerability for climate adaptation policy using principal component analysis. *Ecological Indicators*., 129, 108006. doi.org/10.1016/j.ecolind.2021.108006.