EXPLORING THE POTENTIAL OF PROBABILITY-BASED MODELING FOR URBAN FLOODS IN MANDAUE CITY, PHILIPPINES

Isabella Pauline L. Quijano*[1,2,3] and Jao Hallen L. Bañados[1,2]
[1]Smart City Solutions to Urban Flooding Program, College of Science, University of the Philippines Cebu, Gorordo Ave., Lahug, Cebu City, Philippines 6000
[2]Member, Earth and Space Sciences Division, National Research Council of the Philippines
[3]Email: ilquijano@up.edu.ph

KEY WORDS: Urban Flood Resilience, Probability Density Function, Probability-based Modeling, Random Forest, Geographic Information System

ABSTRACT: With the increasing frequency of urban floods caused by heavy precipitation in the Philippines, flood mitigation strategies to improve the city's resilience must be rigorously pursued. Recent advances in digital technology, such as probability-based modeling, have provided an effective means to identify areas in which such strategies should be implemented. This paper gives a comprehensive analysis of the potential of probability-based modeling to help anticipate urban floods in Mandaue City, Philippines using probability density function (PDF). PDF is a statistical technique used to model the probability of an event occurring based on previous data. In this study, we use GIS by combining different layers containing flooding risk factors such as elevation and land use with existing historical data on flood events in the area. This combination can then be used to create a PDF that can show the probability of an urban flood occurring at various locations, which can allow for informed decision making around flood risk management. The results of this study seeks to inform stakeholders in order to create specific strategies for urban environments that could be used to reduce flooding risks and more importantly, offers a framework for other cities to apply probability-based modeling to generate tailor-made strategies for flood resilience.

1. INTRODUCTION

It has been observed that the combined effects of swift urban expansion and the impacts of climate change have given rise to a multitude of ecological challenges and unfortunate events (Rahmati et al., 2019). Floods can result in a large number of fatalities, harm to ecosystems, and socioeconomic effects (De Silva & Kawasaki, 2020). Floods can happen in two ways: naturally, through heavy rain, snowmelt, and extended periods of rainfall; or unnaturally, through greater degradation brought on by population growth, deforestation, and urbanization (Chan et al., 2018; Şen, 2018; Wang et al., 2019; Van & Schwarz, 2020).

Numerous nations across the globe have experienced flooding, and Philippines is no different. Within the geographical context of the Philippines, an average of 20 tropical cyclones are encountered each year, with the apex of this phenomenon transpiring between July and October, constituting approximately 70% of the total typhoon occurrences (Santos, 2020).

One of the major challenges in flood mapping is providing a precise estimate of the flood extent and damage in affected areas (Esfandiari et al., 2020). To address this issue, several methods with distinct strengths and weaknesses have been developed for the purpose of creating flood prediction maps. Many studies have employed multi-criteria decision analysis and probabilistic models for flood mapping (Wang et al., 2011; Masood & Takeuchi 2012). Additionally, another study utilized random-forest and boosted-tree models to map flood susceptibility and evaluate the prediction ability of the models using metrics like the area under the curve (AUC) (Sunmin et al., 2017). Numerous research comparing several probability distributions with various parameter estimation techniques for on-site food frequency analysis have been published in the past (Hassan et al., 2019). For this analysis, we employ the Random Forest algorithm (Ho 1995; Breiman 2001). This algorithm constructs numerous decision trees and offers insights into the significance of various parameters in the decision-making process.

A significant challenge confronting researchers is the intricate nature of the multitude of conditioning factors, encompassing hydrological, topographical, and geological layers. Additionally, augmenting machine learning algorithms with additional conditioning factors holds the potential to yield improved outcomes (Esfandiari, et al., 2020). Common factors affecting flood prediction among related studies are aspect, curvature, elevation, flow accumulation, flow direction, geology, slope percentage, land use, Normalized Difference Vegetation Index (NDVI), rainfall, distance from rivers, soil group, Stream Power Index (SPI), Sediment Transport Index (STI), Terrain Roughness Index (TRI), and Topographic Wetness Index (TWI) (Aldiansyah & Wardani, 2023; Esfandiari, et al., 2020; Sunmin et al., 2017).

This research endeavour aims to investigate flood forecasting within the boundaries of Mandaue City. The approach employed for this investigation involves probability-based modeling, specifically utilizing the capabilities of the random forest algorithm.

2. METHODOLOGY

2.1 Study Area

Mandaue City, situated within the province of Cebu in the Philippines, is a bustling urban center known for its commercial and industrial activities. It is located at coordinates 10.3321° N latitude and 123.9357° E longitude. This city boasts a high-income status and is surrounded by other densely urbanized cities, as illustrated in Figure 1. In terms of climate, the region experiences a distinct wet and dry season pattern, with the rainy season typically spanning from June to November. This climate is classified as Coronas climate type 3. Mandaue City receives an average annual precipitation of approximately 1,570 millimeters, as reported by JICA in 2010.

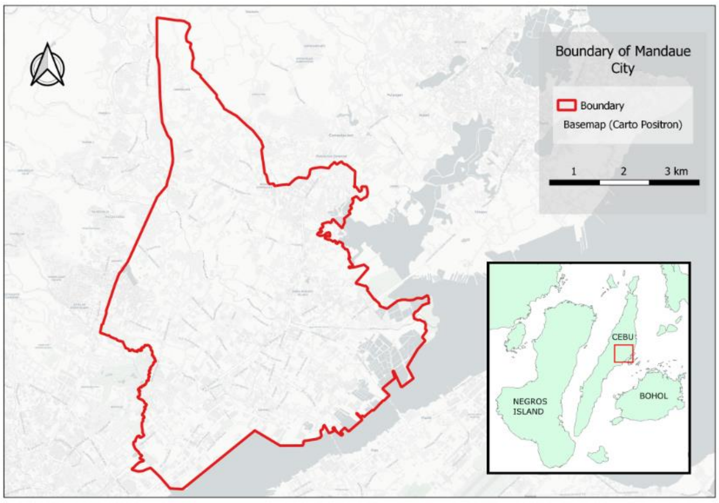


Figure 1. Mandaue City, Cebu, Philippines

2.2 Flood Occurrence Data Points

Utilizing the Flo-2D software, a generated flood hazard and depth models serve as the foundation for flood prediction, particularly focusing on the 25-year rain return model. The random flood occurrence points in the study area were generated using the random points generator tool within ArcMap 10.8. A total of 1,000 points were created to represent flood occurrences, with 444 points indicating non-flooded areas and 556 points representing flooded areas. The data is stored in a shapefile format, and the spatial distribution of these points relative to the study area is visualized in Figure 2.

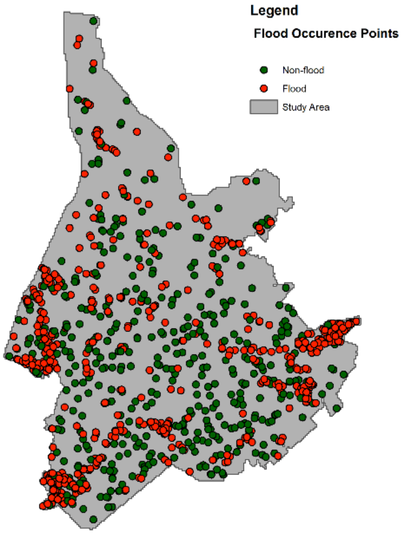


Figure 2. Flood occurrence data points within the study area

2.3 Factors Causing Flood in GIS

This study incorporates 16 variables, a common set employed in previous research (Aldiansyah & Wardani, 2023; Esfandiari et al., 2020; Sunmin et al., 2017). These variables encompass aspect, curvature, elevation, flow accumulation, flow direction, geology, slope percentage, land use, NDVI, rainfall, distance from rivers, soil group, SPI, STI, TRI, and TWI, as visually represented in Figure 3.

Table 1. Data sources

Data	Resolution	Source	Raster
Interferometric Synthetic Aperture Radar (IFSAR) DEM	30 m	National Mapping and Resource Information Authority (NAMRIA)	DEM, TWI, SPI, TRI, STI, Slope, Curvature, Aspect, Flow Direction, Floow Accumulation
Feature Dataset		Mandaue City Planning and Development Office (MCPDO), Department of Agriculture (DA)	River Network, Land Use, Soil, Geology
Landsat-8 Satellite Images (2013-2022)		Google Earth Engine	Normalized Difference Vegetations Index (NDVI)
Climate Hazards Group InfraRed Precipitation with Station Data (CHIRPS)			Rainfall

Table 1 further provides insights into the sources of each variable. Notably, the Digital Elevation Model (DEM) raster was utilized to derive aspect, curvature, elevation, slope, flow direction, flow accumulation, TWI, SPI, TRI, and STI factors. The river network, land use, soil group, and geology feature dataset were converted into rasters through the ArcMap software. The NDVI and rainfall data were sourced from the Google Earth Engine (GEE) dataset, which was derived from Landsat-8 Satellite Images spanning the years 2013 to 2022. Additionally, rainfall data was obtained from the Climate Hazards Group InfraRed Precipitation with Station Data (CHIRPS). It is worth mentioning that all rasters underwent a review process to ensure uniform resolution, coordinate system, and extent.

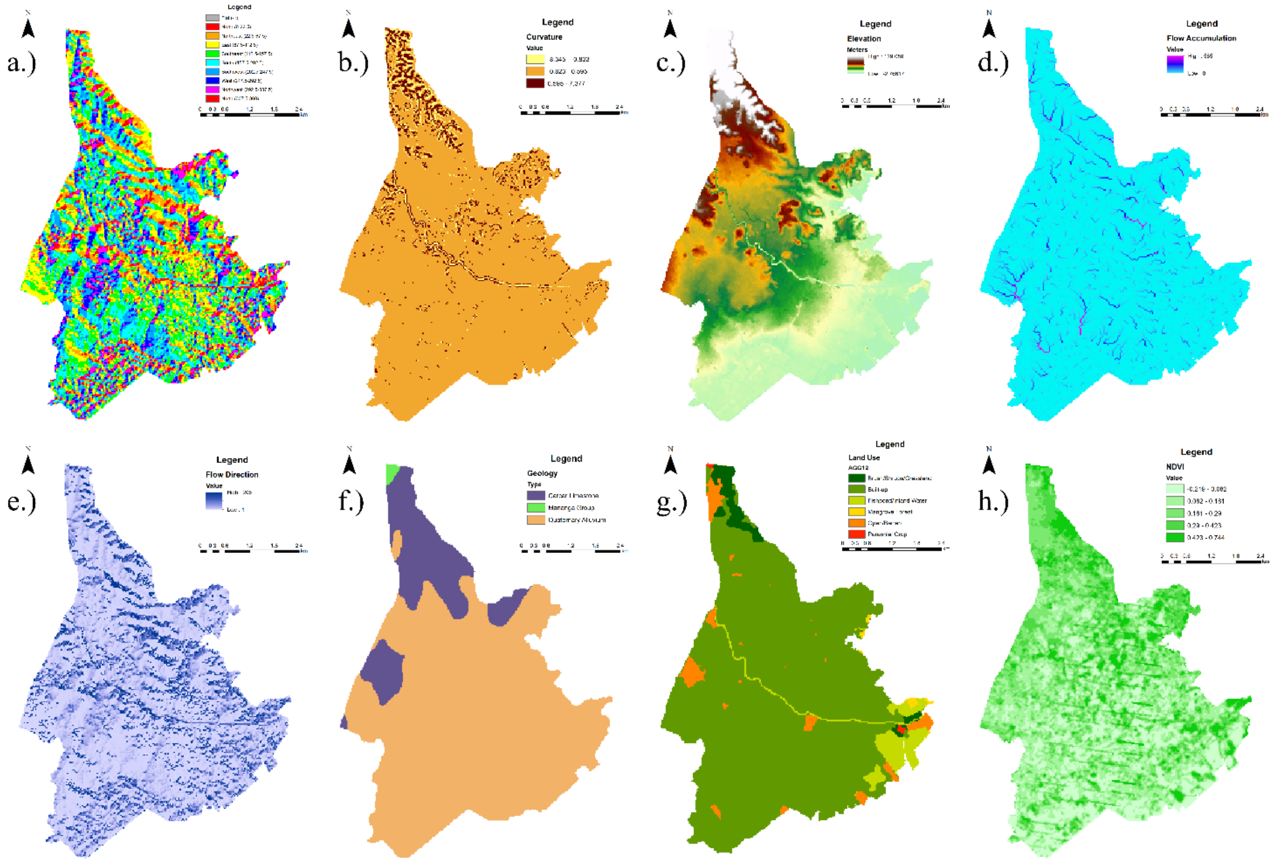


Figure 3. Factors causing flood: a.) aspect; b.) curvature; c.) elevation; d.) flow accumulation; e.) flow direction; f.) geology; g.) land use

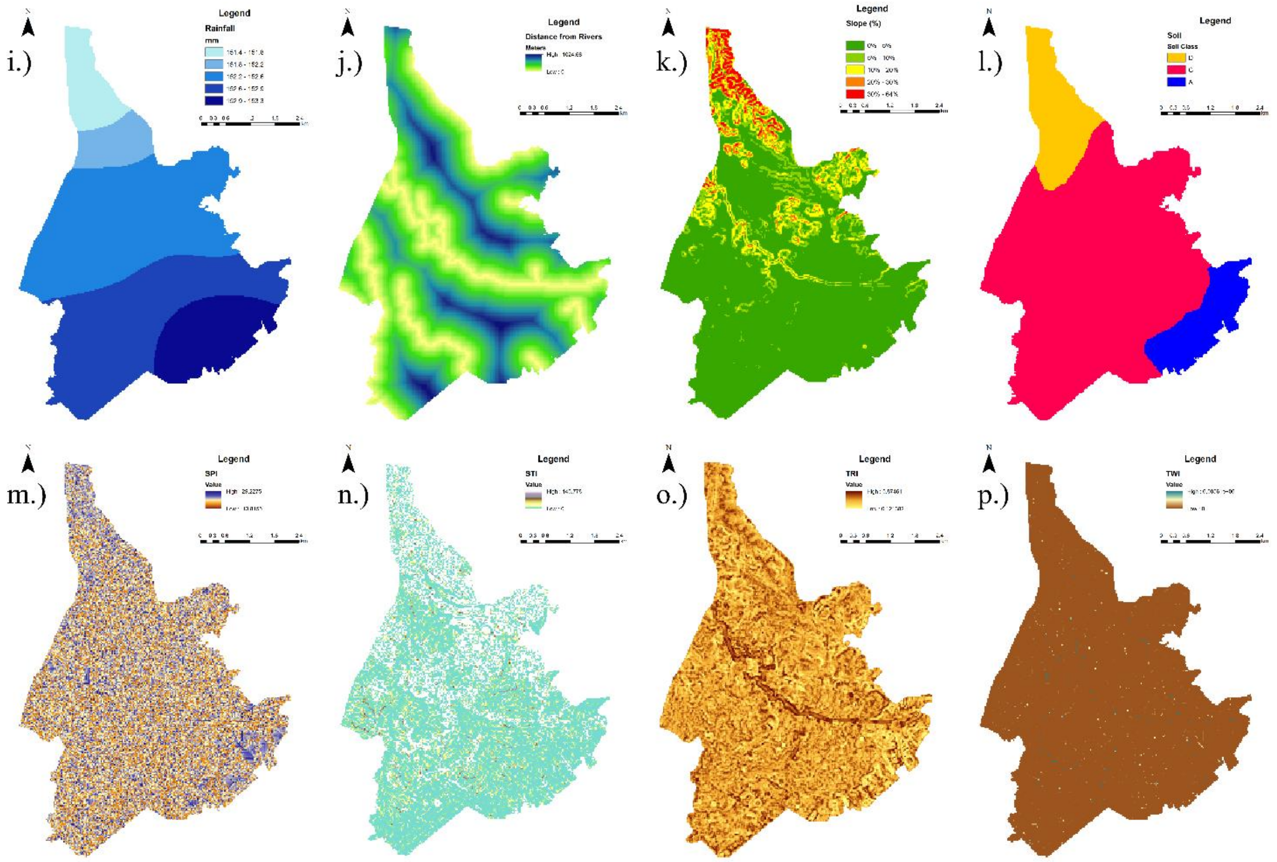


Figure 4. Factors causing flood: h.) NDVI; i.) rainfall; j.) distance from rivers; k.) slope percentage; l.) soil group; m.) SPI; n.) STI; o.) TRI; and p.) TWI

2.4 Random Forest in R

The Random Forest (RF) model employs innovative approaches to aggregate and fuse data for the creation of multiple trees used in making predictions. When selecting the predictive factor, the RF model generates a tree structure akin to a Classification and Regression Tree (CART) (Naghibi & Pourghasemi, 2015). The RF algorithm was employed using RStudio version 2023.06.2, in conjunction with the randomForest package. Raster data representing various flood condition factors were stacked, and the flood data points were divided into 70% for training and 30% for testing purposes. The primary parameters that play a pivotal role in the RF model are the quantity of trees and the number of predictive factors that determine how the decision tree is constructed to its fullest extent and subsequently left unpruned. In this study, the model was executed with 1600 different random seeds, utilizing a total of 1600 trees, and employed cross-validation as a resampling technique.

2.5 Resampling Approach

In this study, the training control method employed is cross-validation. Cross-validation is chosen to enhance the robustness of the model's performance evaluation, reducing its reliance on the random division of the dataset. This mitigates the potential for acquiring overly optimistic or pessimistic performance estimates. Additionally, cross-validation aids in maximizing the utility of the available data, which is particularly beneficial when dealing with a small dataset. In this specific study, a 10-fold cross-validation with 5 repetitions is utilized.

2.6 Accuracy Assessment

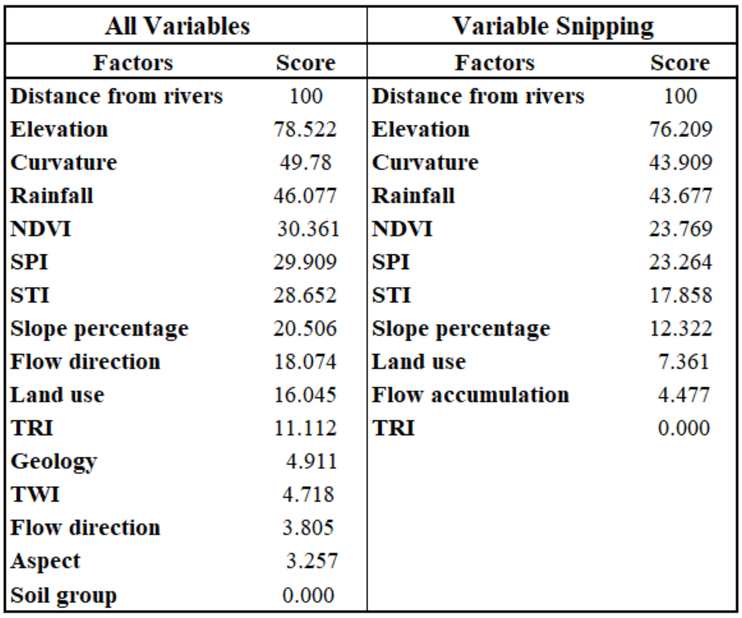
The evaluation of spatial prediction models in flood susceptibility modeling has commonly relied on the utilization of the Area Under the Receiver Operating Characteristic Curve (AUC) (Lee et al., 2017; Chen et al., 2020; Mosavi et al., 2022). AUC is employed to quantitatively appraise the performance of the comprehensive model created for this purpose. An AUC value exceeding 0.7 is considered indicative of a favorable model performance (Shabani et al., 2018). In this paper, caTools package in RStudio was used to assess the accuracy of the random forest model.

3. RESULTS

3.1 Variable Importance

Feature selection and variable trimming are essential steps in probability-based modeling prediction as they lead to more accurate, efficient, interpretable, and robust models. By focusing on the most informative variables and eliminating noise, these techniques enable better flood risk assessment and informed decision-making. Hereby, we refer to our original model containing all relevant factors as model 1 and the succeeding model with trimmed variables as model 2.

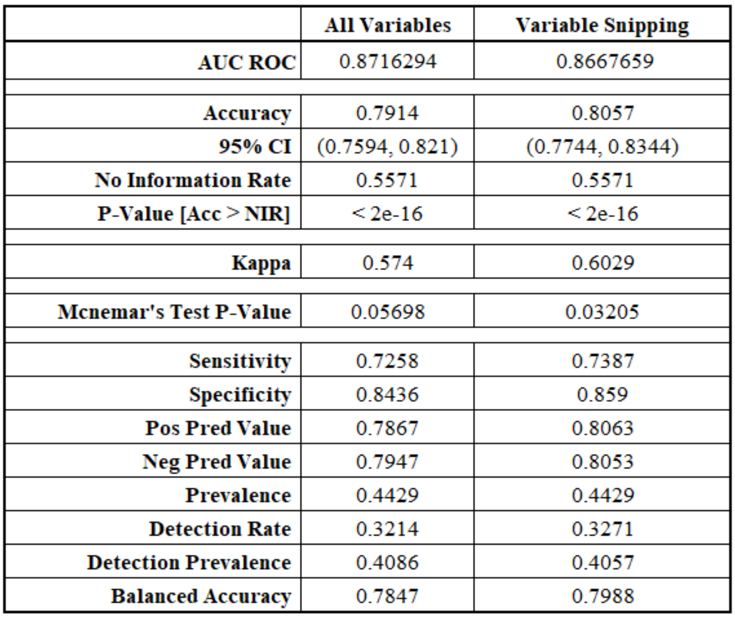
Table 2. Variable Importance of Model Runs



All Variables		Variable Snipping	
Factors	**Score**	**Factors**	**Score**
Distance from rivers	100	**Distance from rivers**	100
Elevation	78.522	**Elevation**	76.209
Curvature	49.78	**Curvature**	43.909
Rainfall	46.077	**Rainfall**	43.677
NDVI	30.361	**NDVI**	23.769
SPI	29.909	**SPI**	23.264
STI	28.652	**STI**	17.858
Slope percentage	20.506	**Slope percentage**	12.322
Flow direction	18.074	**Land use**	7.361
Land use	16.045	**Flow accumulation**	4.477
TRI	11.112	**TRI**	0.000
Geology	4.911		
TWI	4.718		
Flow direction	3.805		
Aspect	3.257		
Soil group	0.000		

Based on our initial run with 16 factors, we have variables that respond to our flood points with scores less than 10. We choose to eliminate or trim these factors and interpret these as noisy or irrelevant features which may reduce the chances of our model being misled by irrelevant patterns or outliers in these datasets. Moreover, feature selection helps in removing irrelevant or redundant features, reducing overfitting, and allowing the model to generalize better to unseen data. These are further proven after comparing both models' accuracy assessments (Table 3).

Table 3. Model Evaluations



	All Variables	**Variable Snipping**
AUC ROC	0.8716294	0.8667659
Accuracy	0.7914	0.8057
95% CI	(0.7594, 0.821)	(0.7744, 0.8344)
No Information Rate	0.5571	0.5571
P-Value [Acc > NIR]	< 2e-16	< 2e-16
Kappa	0.574	0.6029
Mcnemar's Test P-Value	0.05698	0.03205
Sensitivity	0.7258	0.7387
Specificity	0.8436	0.859
Pos Pred Value	0.7867	0.8063
Neg Pred Value	0.7947	0.8053
Prevalence	0.4429	0.4429
Detection Rate	0.3214	0.3271
Detection Prevalence	0.4086	0.4057
Balanced Accuracy	0.7847	0.7988

When comparing the performance of our model runs for flood prediction, we can see that Model 1 (All Variables) demonstrates a marginally better AUC ROC, but Model 2 (Variable Snipping) excels in terms of accuracy, Kappa, sensitivity, specificity, and balanced accuracy, making it a robust choice for flood prediction tasks. Model 1 exhibits a slightly higher AUC ROC of 0.8716294 compared to Model 2's AUC ROC of 0.8667659, indicating slightly superior discrimination between positive and negative cases based on the ROC curve. However, this difference is minimal and may not significantly impact performance. Model 2, on the other hand, excels in terms of accuracy, boasting an accuracy of 0.8057, surpassing Model 1's accuracy of 0.7914. This suggests that Model 2 makes more correct predictions overall, enhancing its reliability. Moreover, Model 2 demonstrates a higher Kappa value (0.6029) compared to Model 1 (0.574), indicating a higher level of agreement beyond chance in its predictions, further emphasizing its overall strength.

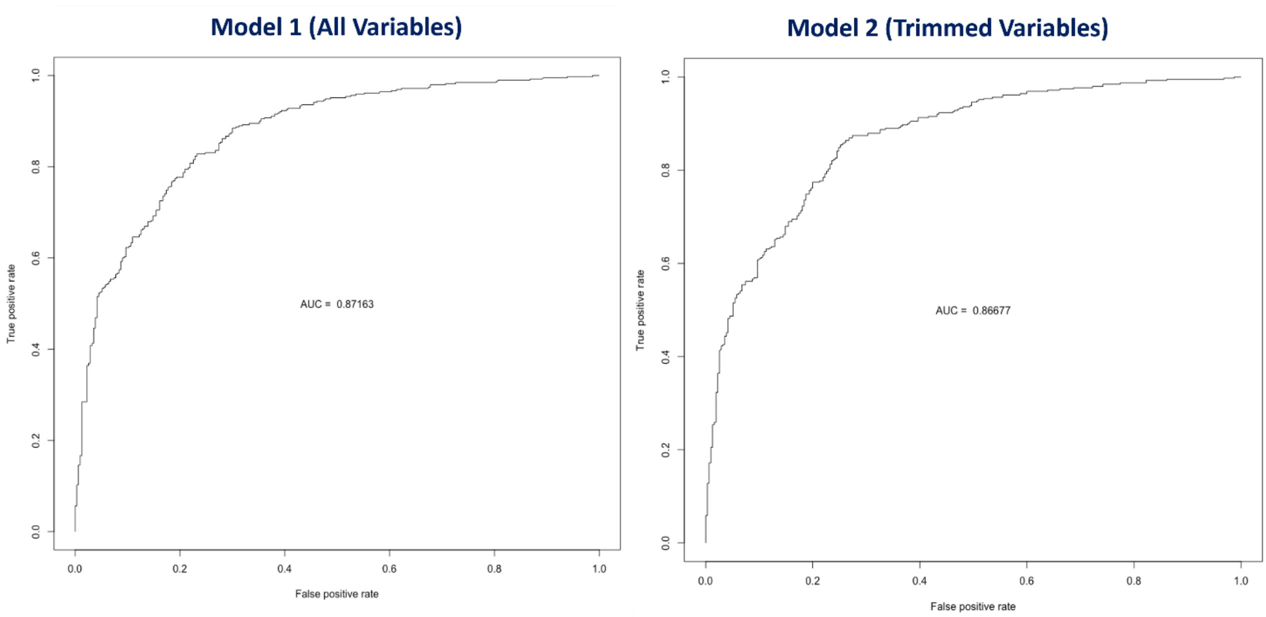


Figure 5. Model AUC ROCs

Mcnemar's Test P-Value also favors Model 2 (0.03205) over Model 1 (0.05698), implying a significant difference in performance between the two models. Sensitivity and specificity are notable strengths of Model 2, with sensitivity at 0.7387 and specificity at 0.859, surpassing Model 1's values (Sensitivity: 0.7258, Specificity: 0.8436), indicating Model 2's superior ability to correctly identify both positive and negative cases. Additionally, Model 2 exhibits a higher Positive Predictive Value (0.8063) while maintaining a similar Negative Predictive Value (0.8053) compared to Model 1 (Pos Pred Value: 0.7867, Neg Pred Value: 0.7947), signifying its slightly better performance in predicting both positive and negative instances. The balanced accuracy metric also leans in favor of Model 2, with a balanced accuracy of 0.7988 compared to Model 1's 0.7847, showcasing Model 2's superior overall performance, considering both sensitivity and specificity.

3.2 Probable Flood Prone Areas

The results of our models are shown in Figure 6 below. Because of the minute variability in the model's accuracy model 1 and 2's flood maps seem identical. Flooding in the final model is clearly seen in the vicinity of the city's major rivers, in line with the variable importance of the 'distance to rivers' factor as 100%.

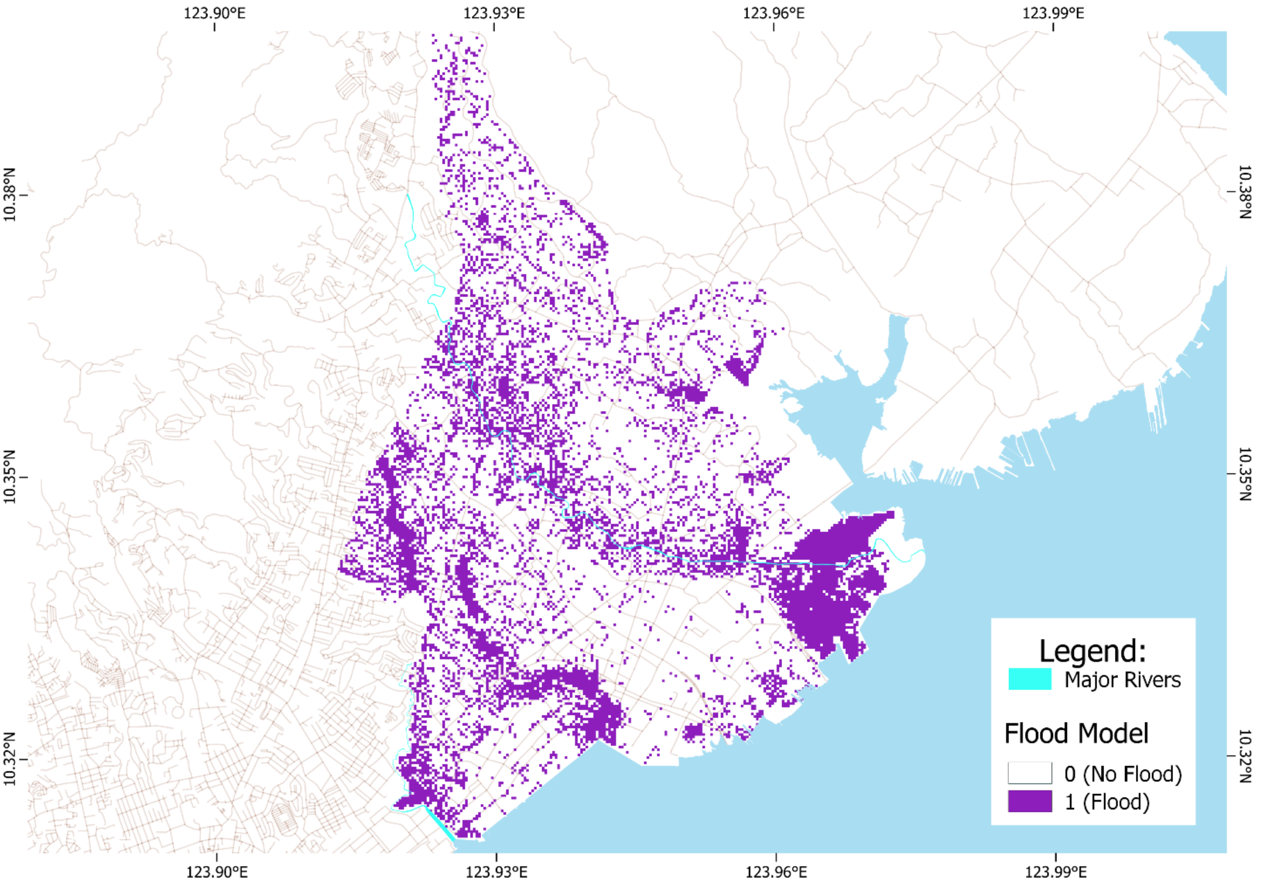


Figure 6. Predicted Urban Flood-prone Areas

3.3 Accuracy

The final flood prediction model's performance is truly impressive, with its accuracy of approximately 80.57% reflecting a high degree of correctness in its classifications. However, what truly distinguishes this model is its remarkable AUC ROC value of 0.8667659. AUC ROC is a critical metric that gauges the model's ability to differentiate between positive and negative cases across varying probability thresholds. An AUC ROC score above 0.5 signifies the model's capacity to make accurate distinctions, and a score approaching 1 indicates exceptional discrimination. In this case, the AUC ROC score of 0.8667659 emphasizes the model's outstanding ability to distinguish between flood and non-flood events, making it exceptionally well-suited for applications where precision is paramount.

The accompanying 95% confidence interval for accuracy, ranging from 0.7744 to 0.8344, bolsters the credibility of this accuracy estimate, providing a confidence range for the true accuracy. This interval suggests a high degree of precision in the model's performance evaluation. Importantly, the model significantly outperforms the No Information Rate (NIR) with a p-value of less than 2e-16, affirming its effectiveness in flood prediction by a substantial margin. Cohen's Kappa coefficient, a measure of agreement beyond chance, stands at 0.6029, indicating a notable level of consensus in the model's predictions, further underlining its reliability.

The model's excellence extends to its sensitivity of 0.7387 and specificity of 0.8590, highlighting its ability to accurately identify both flood and non-flood instances. This balanced performance is a testament to the model's versatility and reliability across diverse flood risk management scenarios. Positive Predictive Value (Pos Pred Value) at 0.8063 emphasizes the model's reliability in correctly predicting flood events when it makes positive predictions, while Negative Predictive Value (Neg Pred Value) at 0.8053 underscores its precision in predicting the absence of floods when it makes negative predictions.

Considering the prevalence of flood events in the dataset, estimated at 0.4429, the model's Detection Rate of 0.3271 demonstrates its capability to capture a substantial portion of actual flood occurrences. Detection Prevalence, at 0.4057, highlights the model's frequency of flood predictions, indicating its practical utility in identifying areas at risk of flooding.

Balanced Accuracy, a comprehensive measure accounting for sensitivity and specificity while accommodating class imbalances, stands at 0.7988, reaffirming the model's ability to provide a balanced and reliable assessment of its overall performance.

In sum, our final flood prediction model's outstanding AUC ROC score, combined with its high accuracy and other robust performance metrics, positions it as an exceptional tool for dependable flood forecasting. These results underscore its potential to make significant contributions to flood risk assessment, early warning systems, and decision-making processes in flood management and disaster preparedness. Further refinements and validations are likely to reveal even greater potential for this model in real-world flood forecasting applications.

4. CONCLUSION

This paper gives a comprehensive analysis of the potential of probability-based modeling to help anticipate urban floods in Mandaue City, Philippines using probability density function (PDF) and GIS analysis. In this study we used the random forest machine learning method to detect flood-prone areas in the city. We obtained positive results for our model runs, with an AUC ROC of over 0.75 and a model accuracy of 0.8057. The results of this study seek to inform stakeholders in the city in order to create specific strategies for urban environments that could be used to reduce flooding risks and more importantly, offers a framework for other cities to apply probability-based modeling to generate tailor-made strategies for flood resilience.

ACKNOWLEDGEMENT

The Smart City Solutions to Urban Flooding Program was used to conduct the study. We thank the Department of Science and Technology (DOST) for their financial assistance. The project's execution was managed by the University of the Philippines Cebu and the DOST Philippine Council for Industry, Energy, and Emerging Technology Research and Development.

REFERENCES

Aldiansyah, S., & Wardani, F. 2023. Evaluation of flood susceptibility prediction based on a resampling method using machine learning. Journal of Water and Climate Change. https://doi.org/10.2166/wcc.2023.494

Breiman, L. 2001. Random Forests. Machine Learning, 45(1), 5–32.

Chan, F. K. S., Griffiths, J. A., Higgitt, D., Xu, S., Zhu, F., Tang, Y. T., & Thorne, C. R. 2018. ‘Sponge city’ in China – a breakthrough of planning and flood risk management in the urban context. Land Use Policy, 76, 772–778.

De Silva, M. M. G. T., & Kawasaki, A. 2020. A local-scale analysis to understand differences in socioeconomic factors affecting economic loss due to floods among different communities. International Journal of Disaster Risk Reduction, 47, 101526.

Esfandiari, M., Jabari, S., McGrath, H., & Coleman, D. 2020. Flood mapping using random forest and identifying the essential conditioning factors; a case study in Fredericton, New Brunswick, Canada. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, V-3-2020, 609–615. https://doi.org/10.5194/isprs-annals-V-3-2020-609-2020

Ho, T. K. 1995. Random Decision Forests. In Proceedings of the 3rd International Conference on Document Analysis and Recognition, 278–282. IEEE.

Lee, S., Kim, J. C., Jung, H. S., Lee, M. J., & Lee, S. 2017. Spatial prediction of flood susceptibility using random-forest and boosted-tree models in Seoul metropolitan city, Korea. Geomatics, Natural Hazards and Risk, 8(2), 1185–1203. https://doi.org/10.1080/19475705.2017.1308971

Masood, M., & Takeuchi, K. 2012. Assessment of flood hazard, vulnerability and risk of mid-eastern Dhaka using DEM and 1D hydrodynamic model. Natural Hazards, 61, 757–770.

Mosavi, A., Golshan, M., Janizadeh, S., Choubin, B., Melesse, A. M., & Dineva, A. A. 2022. Ensemble models of GLM, FDA, MARS, and RF for flood and erosion susceptibility mapping: a priority assessment of sub-basins. Geocarto International, 37(9), 2541–2560.

Rahmati, O., et al. 2019. Urban flood hazard modeling using self-organizing map neural network. Water, 11(11), 2370. https://doi.org/10.3390/w11112370

Santos, G. D. 2021. 2020 Tropical Cyclones in the Philippines: A Review. Tropical Cyclone Research and Review, 10(3), 191-199.

Şen, Z. 2018. Flood Modeling, Prediction and Mitigation. Springer International Publishing, Cham, Switzerland.

Shabani, F., Kumar, L., & Ahmadi, M. 2018. Assessing accuracy methods of species distribution models: AUC, specificity, sensitivity and the true skill statistic. Global Journal of Human Social Science, 18(1), 6–18.

Ul Hassan, M., Hayat, O., & Noreen, Z. 2019. Selecting the best probability distribution for at-site flood frequency analysis; a study of Torne River. SN Applied Sciences, 1, 1629. https://doi.org/10.1007/s42452-019-1584-z

Van, E. T., & Schwarz, A. 2020. Plastic debris in rivers. Wiley Interdisciplinary Reviews: Water, 7(1), e1398.

Wang, Y., Hong, H., Chen, W., Li, S., Panahi, M., Khosravi, K., Shirzadi, A., Shahabi, H., Panahi, S., & Costache, R. 2019. Flood susceptibility mapping in Dingnan County (China) using adaptive neuro-fuzzy inference system with biogeography based optimization and imperialistic competitive algorithm. Journal of Environmental Management, 247, 712–729.

Wang, Y., Li, Z., Tang, Z., & Zeng, G. 2011. A GIS-based spatial multi-criteria approach for flood risk assessment in the Dongting Lake Region, Hunan, Central China. Water Resources Management, 25, 3465–3484.