FLOOD SUSCEPTIBILITY THROUGH MACHINE LEARNING AND CENSUS DATA IN THE ITAJAÍ RIVER BASIN, BRAZIL

Manuel Mendoza Colos [1]; Guilherme Palermo Coelho [2]; Saket Pande [3]; Murilo Cesar Lucas [4]

RESUMO - Este estudo avaliou a suscetibilidade a inundações e a exposição demográfica na bacia do rio Itajaí, Brasil, por meio da integração de aprendizado de máquina, dados geoespaciais e setores censitários. Para isso, foi aplicado o algoritmo Random Forest (RF), incorporando 12 preditores topográficos, hidrológicos, climáticos e antropogênicos, incluindo declividade, elevação, uso e cobertura do solo (LULC), distância dos rios, precipitação, NDVI, geologia e índices derivados do relevo. O modelo apresentou um desempenho excepcional, com precisão de 0,98, recall de 0,97, acurácia de 0,98, F1-Score de 0,98 e AUC de 0,99, confirmando sua capacidade de diferenciar áreas suscetíveis e não suscetíveis a inundações. A análise SHAP (SHapley Additive exPlanations) revelou que a declividade foi o preditor mais influente, seguido pelo LULC, elevação e distância dos rios. Além disso, as interações mostraram que as zonas próximas aos leitos dos rios, com declividade baixa e baixa elevação, aumentam a probabilidade de inundações, especialmente quando coincidem com áreas agrícolas ou urbanizadas. O mapa de suscetibilidade revelou uma concentração de zonas de alta e muito alta suscetibilidade em setores costeiros e urbanizados. Além disso, a análise demográfica mostrou que apenas 4,2% da área total concentra 47,4% da população e 46,8% das residências, enquanto as classes alta e muito alta reúnem 65,5% da população e 64,7% das unidades habitacionais. Esses resultados evidenciam a necessidade de fomentar a resiliência nas zonas mais vulneráveis.

Palavras-chave: Inundação, Suscetibilidade, Sensoriamento Remoto, Cidades resilientes.

ABSTRACT - This study assessed flood susceptibility and demographic exposure in the Itajaí River basin, Brazil, through the integration of machine learning, geospatial data, and census sectors. The Random Forest (RF) algorithm was applied, incorporating 12 predictors including topographic, hydrological, climatic, and anthropogenic factors, such as slope, elevation, land use and land cover (LULC), distance to rivers, precipitation, NDVI, geology, and relief-derived indices. The model performed exceptionally, with a precision of 0.98, recall of 0.97, accuracy of 0.98, F1-score of 0.98, and AUC of 0.99, confirming its ability to differentiate between flood-prone and non-flood-prone areas. The SHAP (SHapley Additive exPlanations) analysis revealed that slope was the most influential predictor, followed by LULC, elevation, and distance to rivers. Additionally, interactions showed that areas near rivers, with low slopes and low elevation, increase the likelihood of flooding, particularly when they coincide with agricultural or urbanized areas. The susceptibility map revealed a concentration of high and very high susceptibility areas in coastal and urban sectors. Moreover, the demographic analysis showed that only 4.2% of the total area concentrates 47.4% of the population and 46.8% of the housing units, while the high and very high classes together account for 65.5% of the population and 64.7% of the housing units. These results highlight the need to promote resilience in the most vulnerable areas.

Keywords: Flood, Susceptibility, Remote Sensing, Resilient cities.

[1] Afiliação: Faculdade de Tecnologia, Unicamp, R. Paschoal Marmo, 1888, Jardim Nova Itália, 13.484-332, Limeira – SP, m249533@g.unicamp.br
[2] Afiliação: Faculdade de Tecnologia, Unicamp, R. Paschoal Marmo, 1888, Jardim Nova Itália, 13.484-332, Limeira – SP, gpcoelho@unicamp.br
[3] Afiliação:Department of Civil Engineering and Geosciences, Delft University of Technology, Delft, the Netherlands, S.Pande@tudelft.nl
[4] Afiliação: Faculdade de Tecnologia, Unicamp, R. Paschoal Marmo, 1888, Jardim Nova Itália, 13.484-332, Limeira – SP, mlucas@unicamp.br

INTRODUCTION

Floods represent an increasing threat affecting more than 1.81 billion inhabitants, approximately 23% of the global population. Projections suggest this number could reach 2.3 billion by 2050 (Rentschler, Salhab, and Jafino, 2022). Climate change intensifies extreme rainfall events in the river basins of Brazil by increasing their frequency and intensity (Ballarin et al., 2024). This situation is further exacerbated by uncontrolled urbanization and surface impermeabilization, which increases runoff and hydrological vulnerability, increasing the vulnerability of urban areas due to social and economic factors (Widya et al., 2024).

The identification of flood-prone areas, known as flood susceptibility, is crucial to addressing this risk. Traditional models, such as the Hydrologic Engineering Center's River Analysis System (HEC-RAS) and the Storm Water Management Model (SWMM), although accurate, are costly and difficult to apply due to the need for large amounts of data and parameters (Amiri et al., 2024). In contrast, machine learning approaches, such as Random Forest (RF), emerge as an efficient alternative for handling large geospatial datasets, and are less dependent on physically-based factors, as they allow for rapid updates and analysis of changing conditions (Abdo et al., 2025). However, it is necessary to incorporate explainable methods, such as SHapley Additive exPlanations (SHAP) values, to understand how different conditioning predictors contribute to flood risk, providing a clearer and more precise view of the elements influencing the model's results (Aydin and Iban, 2023).

This study aims to map flood susceptibility in the Itajaí River Basin using the RF model with SHAP-based interpretation to assess demographic exposure. This approach not only identifies the areas most prone to flooding but also highlight how hydrological vulnerability overlaps with demograpic, providing a solid scientific foundation for developing more equitable and locally adapted territorial planning strategies.

MATERIALS AND METHODS

The study followed a structured and sequential methodology (Figure 1), aiming to map and evaluate the demographic exposure to flood susceptibility in the Itajaí River basin using the Random Forest (RF) model. In the first stage, the conditioning predictors were organized and standardized to a common spatial resolution. Additionally, training and testing points were generated and processed from flood susceptibility vector polygons provided by the Brazilian Geological Service through the GeoSGB platform (https://geoportal.sgb.gov.br/geosgb/) . Furthermore, population density data from the 2022 Demographic Census of the Instituto Brasileiro de Geografia e Estatística (IBGE) were processed.

In the second stage, the RF model was trained by splitting the data into 70% for training and 30% for testing. In the third stage, the model's performance was evaluated using metrics widely established in the literature, while the contribution of the explanatory predictors was analyzed using the SHAP method. Finally, in the fourth stage, the population's exposure to flood susceptibility was examined using the census sectors from the 2022 IBGE Census.

To map areas susceptible to flooding, it is essential to inspect those that were affected by previous events. Historical news reports of floods in municipalities such as Blumenau, Itajaí, Ilhota, Gaspar, and Rio do Sul were used. The Itajaí municipality database was also used, available at: (https://arcgis.itajai.sc.gov.br/portal/apps/webappviewer/index.html?id=4e097e762ffc484e84648905f0d75347)

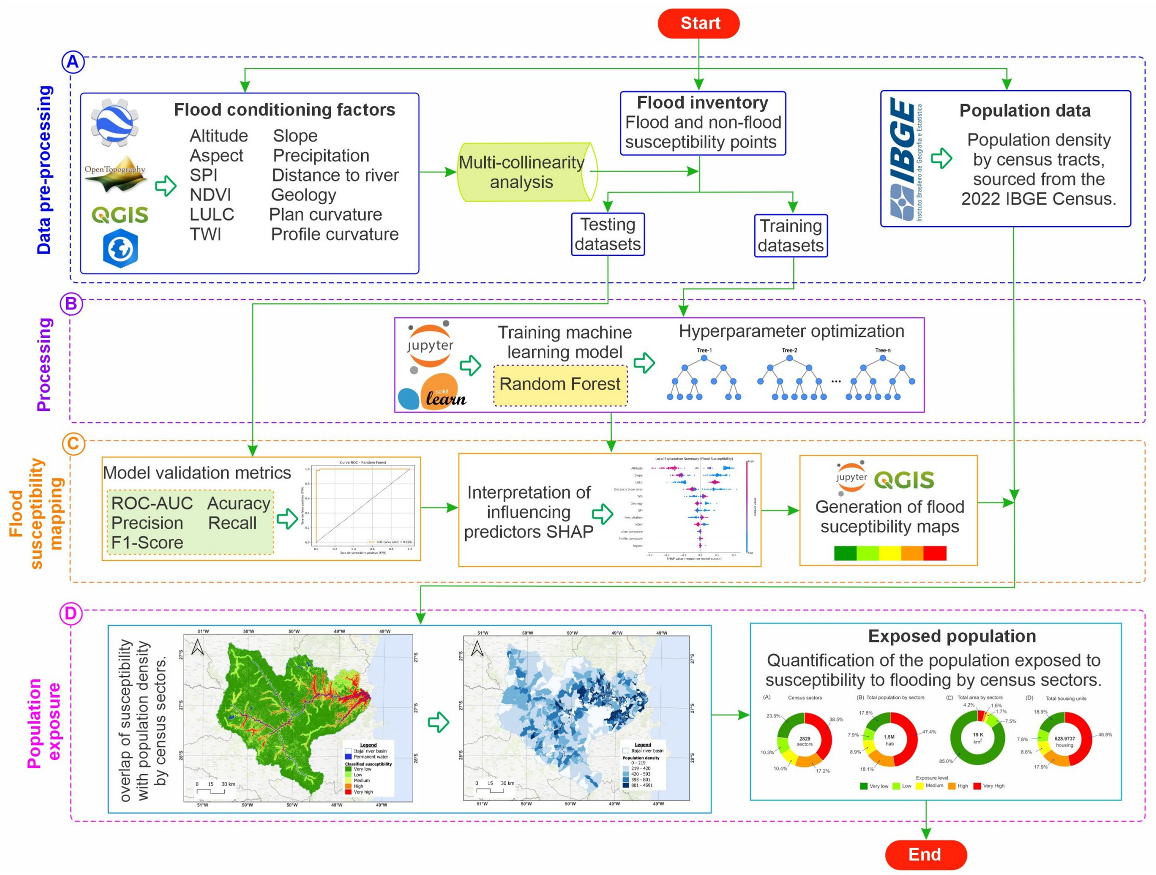

Figure 1: Methodological flow diagram of the study to assess the susceptibility to flooding and the exposure of the population in the Itajaí River Basin.

Study area

The Itajaí River Basin, located in the State of Santa Catarina, Brazil (Figure 2), is one of the regions most affected by floods in the country, due to its flat topography and the presence of alluvial valleys. Covering approximately 15,000 km², it has an estimated population of 1,511,423 inhabitants, with a population density of 94.68 inhab/km² (IBGE, 2022). Land use and occupation are characterized by the predominance of urbanized areas, such as in the municipalities of Blumenau, Itajaí, and Brusque, and agricultural zones destined for the cultivation of corn, soybeans, beans, and vegetables. It still preserves important remnants of the Atlantic Forest, covering about 6,219 km², which represents 40.62% of its total area (Luis et al., 2015). The climate is humid subtropical, with Cfa (hot summers) and Cfb (mild summers) classifications, featuring average annual temperatures between 14°C and 20°C and an average annual precipitation of 1,550 mm, factors that contribute to the recurrence of extreme hydrological events.

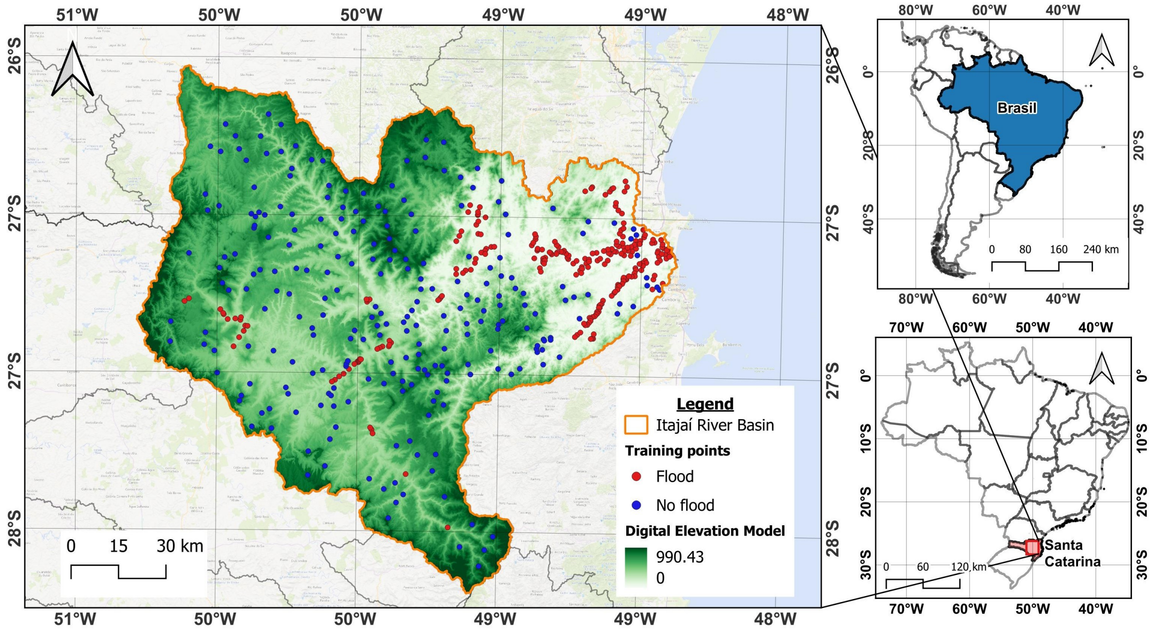

Figure 2: Geographic location of the Itajaí River Basin. The red and blue dots indicate areas susceptible to flooding and areas not susceptible to flooding, respectively. These dots are from the flood susceptibility database, which covers the period from 2012 to 2019 and was obtained from the Brazilian Geological Survey.

Flood Susceptibility

Flood susceptibility represents the likelihood that an area will be impacted by flooding based on its physical and environmental features, allowing for the identification of vulnerable zones before hydrological events (Widya et al., 2024). In this study, 12 predictors were chosen due to their relevance in surface runoff and water accumulation (Waleed and Sajjad, 2025), and all were reprojected to a uniform 30 m resolution for consistency across the dataset (Abdo et al., 2025). The Digital Elevation Model (DEM) was derived from Shuttle Radar Topography Mission (SRTM) data to obtain elevation, slope, and orientation, influencing flow direction and velocity (Bui et al., 2023). Additionally, we calculated plan and profile curvature, the Topographic Wetness Index (TWI) to identify saturation zones, and the Stream Power Index (SPI) to capture the erosive potential in high-energy basins (Amiri et al., 2024).

Hydrological influence was represented by the distance to the river network (Amiri et al., 2024), complemented by accumulated precipitation data from the Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) dataset, NDVI for vegetation density, and lithological and land-use data from ESA WorldCover 2021 to account for soil permeability (Waleed and Sajjad, 2025).

Random Forest (RF)

We selected the RF method for its ability to assess hydrological risks and model nonlinear interactions in complex environments (Bui et al., 2023). RF is robust against overfitting in high-dimensional data and heterogeneous spatial contexts, outperforming traditional statistical approaches in terms of accuracy and spatial stability (Yang et al., 2025). To optimize performance, we performed hyperparameter tuning on the training set using 5-fold cross-validation combined with Grid Search, evaluating 324 parameter combinations. The optimal configuration consisted of 200 trees, no maximum depth limit (None), and the square root of the total number of features as the selection criterion at each split (Abdo et al., 2025).

The model was constructed with a binary encoding of the dependent variable (1 = flood susceptibility; 0 = no flood susceptibility), and the conditioning factors were selected after performing a correlation analysis of the 12 predictors used in the model, as well as calculating the Variance Inflation Factor (VIF) to mitigate multicollinearity (Bui et al., 2023). To improve robustness, the specific adjustment resulted in a minimum of 5 samples required for node splitting and a minimum of 1 sample per leaf node (Aydin and Iban, 2023). This allowed the integration of morphometric, hydrological, climatic, and anthropogenic predictors without making parametric assumptions about the data (Knighton et al., 2020).

Evaluation of the RF Model Performance

The performance of the RF model was evaluated using standard metrics for flood susceptibility (Ahmed et al., 2022). The ROC curve and AUC were analyzed to assess discrimination capability, while accuracy, recall, precision, and F1-score —derived from the confusion matrix— measured the model's ability to distinguish between flood-prone and non-flood-prone areas (Arabameri et al., 2022). These indicators ensure robustness and allow for an objective comparison with previous studies (Ahmed et al., 2022).

To complement this, we implemented SHAP (SHapley Additive exPlanations) (Lundberg et al., 2020), which provides both global and local interpretations by quantifying each predictor's contribution (Waleed and Sajjad, 2025). Summary and dependence plots helped identify the most influential predictors, their directionality, and the interdependencies between topographic and environmental variables, offering insights into the physical mechanisms underlying susceptibility (Hajji et al., 2025).

Demographic Exposure to Susceptibility

The flood susceptibility map was integrated with demographic data to go beyond the physical characterization of susceptibility and advance towards a direct assessment of human exposure (Rentschler, Salhab, and Jafino, 2022). This integration allows us to identify where high susceptibility overlaps with high population concentrations (Waleed and Sajjad, 2025).

Using census sectors as the smallest spatial unit of analysis due to their high resolution and demographic detail (Knighton et al., 2020), the overlay of census data with the susceptibility map helped identify spatial asymmetries, where relatively small areas concentrate disproportionately large populations (Tellman et al., 2021). A susceptibility class was assigned to each census sector based on the predominant (modal) susceptibility value of the pixels within each section (Arabameri et al., 2022).

RESULTS AND DISCUSSION

The Random Forest (RF) model used in this study to evaluate flood susceptibility in the Itajaí watershed showed outstanding performance with a precision of 0.98, a recall of 0.97, and an overall accuracy of 0.98, placing it at an excellent performance (Bui et al., 2023). These metrics align with the results of Oluwadare et al. (2025) in the Sorocaba sub-region, São Paulo, where the RF model achieved F1-Score and AUC metrics between 0.94 and 1.00 (in our case, an F1-Score of 0.98 and an AUC of 0.99).

The SHAP analysis (Figure 3) revealed that slope was the most influential predictor in the model, followed by land use and land cover (LULC), elevation, and distance to river. These findings emphasize the role of topography in flood susceptibility within the basin.

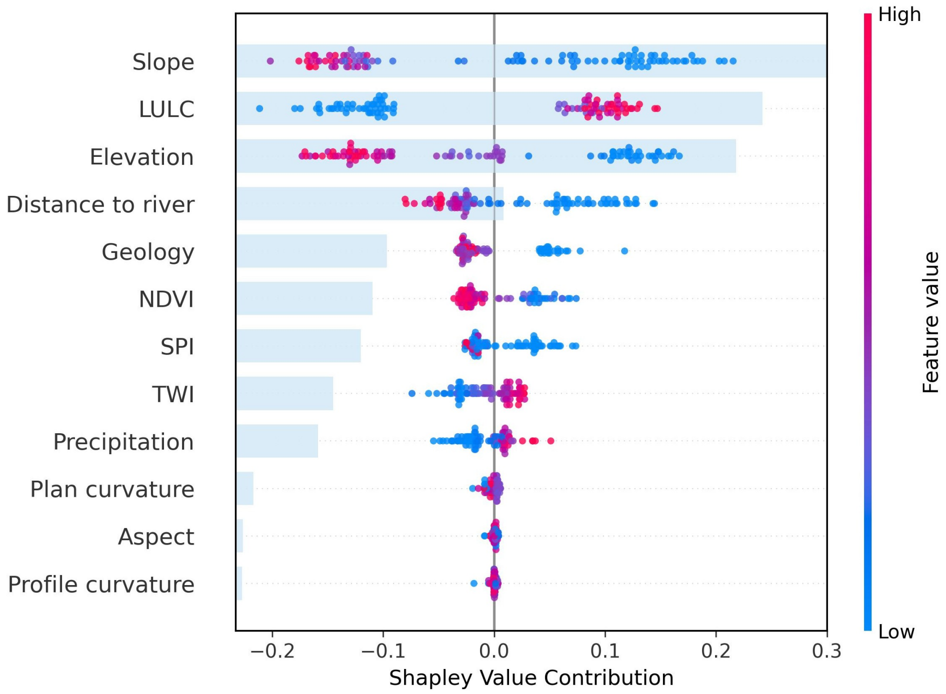

Figure 3: Global SHAP analysis of predictor influence. The horizontal part represents the positive or negative influence of each predictor, while the vertical heatmap bar represents an observation, where the color indicates its value (red: high, blue: low).

Low slope and elevation values, represented by blue dots on the right side of the plot, significantly increased susceptibility. This pattern was reinforced by the influence of areas located closer to the river, supporting the hydrological principle that zones near watercourses are more prone to overflow (Abdo et al., 2025). Moreover, the LULC predictor showed that certain land use classes, represented by red dots on the right side, increased the probability of flooding. Similarly, low values of geology, NDVI, and SPI also contributed to flood occurrence, although with a relatively lower influence.

The SHAP interaction plots reveal how the combination of predictors influences flood susceptibility positively or negatively. In (Figure 4.a), the combination of distances closer to the river (< 50 m), represented on the horizontal axis, and low slopes (< 5%), shown on the vertical color bar, produces positive SHAP values, increasing the prediction of flood susceptibility. This result is consistent with Amiri et al. (2024), who indicate that high-risk areas are associated with minimal slopes and proximity to river channels. A similar pattern is observed in the interaction between LULC and distance to the river (Figure 4.b), where tree cover (class 1) shows negative SHAP values, while croplands (class 3) and built-up areas (class 4), when located closer to the river (< 50 m), contribute positively to flood susceptibility. This transition to positive values in anthropogenic land covers aligns with Debnath et al. (2024), who state that urbanization creates impermeable surfaces that increase runoff and overload drainage systems.

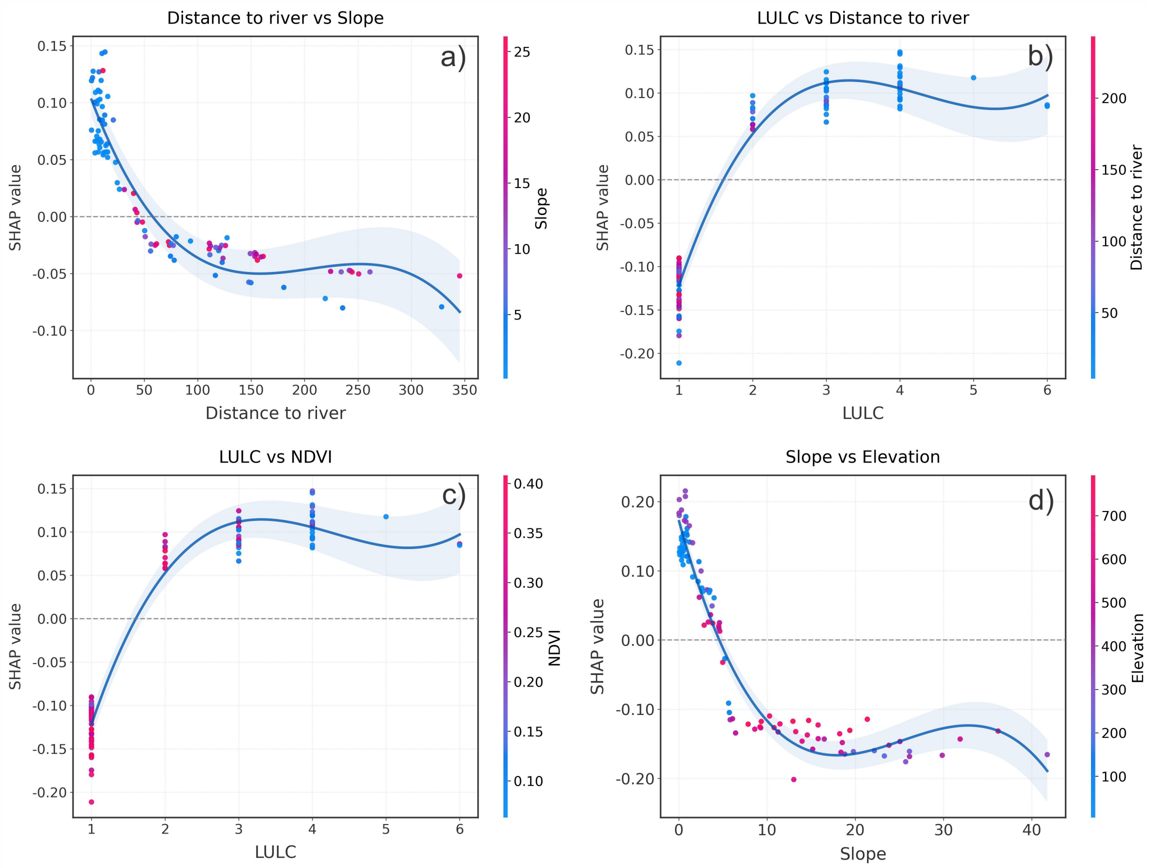

Figure 4: Interaction analysis using SHAP values with a smoothed trend and a 95% confidence interval. The horizontal line at zero represents the division between the positive and negative influence of the combined predictor variables. The LULC classes are: 1 = tree cover, 2 = grasslands, 3 = croplands, 4 = built-up areas, 5 = bare/sparse vegetation, 6 = herbaceous wetlands. a) Distance to river vs Slope, b) LULC vs Distance to river, c) LULC vs NDVI, and d) Slope vs Elevation.

Vegetation conditions also play an important role in explaining the model behavior. In the interaction between LULC and NDVI (Figure 4.c), tree cover (class 1), when combined with high NDVI values, presents negative SHAP values, indicating a lower contribution to flood susceptibility. This result agrees with Hajji et al. (2025), who state that high NDVI acts as a physical mitigation mechanism by reducing the velocity of surface runoff. In contrast, croplands (class 3) show a transition in NDVI values, while urbanized areas (class 4), combined with low NDVI values, present positive SHAP values in the prediction. This behavior aligns with Abdo et al. (2025), who indicate that susceptibility increases in areas with low NDVI due to reduced vegetation density and lower soil infiltration capacity. Finally, the interaction between slope and elevation (Figure 4.d) shows that near-zero slopes combined with low elevations generate positive SHAP values, indicating a direct contribution to flooding and validating the robustness of the model (Amiri et al., 2024).

To generate the susceptibility map, the Natural Breaks (Jenks) classification method was used, categorizing the areas into five susceptibility classes. The results showed a clear heterogeneity, with a high concentration of high and very high susceptibility areas in coastal regions where the largest population is located (Figure 5). Waleed and Sajjad (2025) argue that coastal proximity is a critical factor, as these areas are the most affected during heavy rainfall and storm surges. This highlights the need for adaptation strategies that account for the geographic location of vulnerable areas.

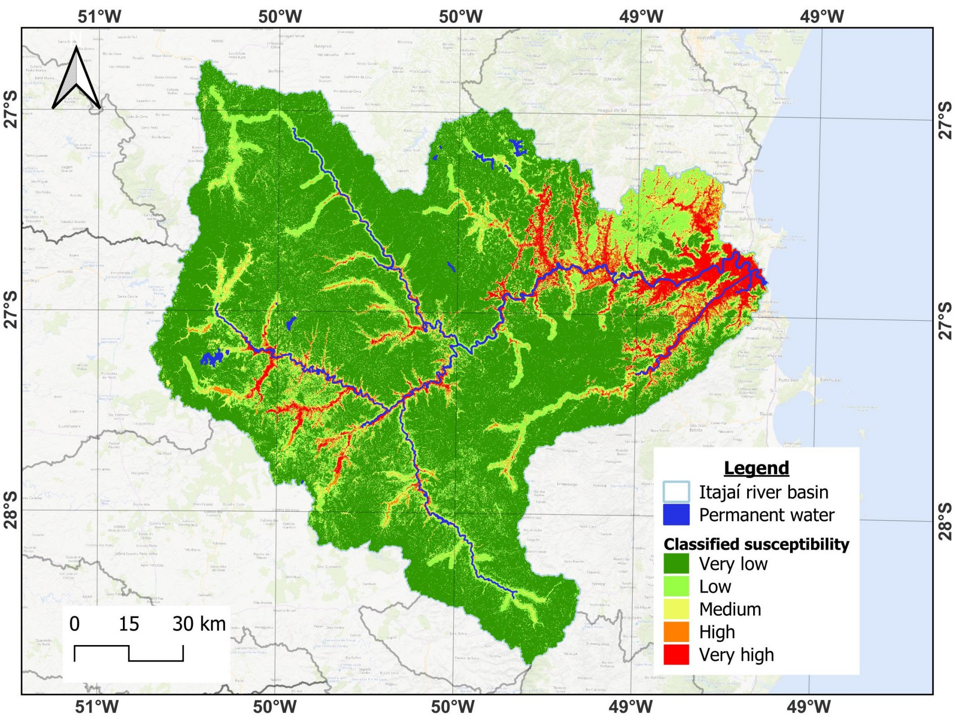

Figure 5: Flood susceptibility map of the Itajaí River Basin classified using natural breaks.

Furthermore, the population exposure assessment showed a marked convergence between the spatial distribution of susceptibility and demographic concentration. When analyzing the distribution across the different exposure levels (Figure 6), it was observed that although areas classified as "Very High" susceptibility represent only 4.2% of the total area (approximately 630 km^2), they concentrate a staggering 47.4% of the total population (approximately 711 thousand inhabitants) and 46.8% of the housing units (approximately 294,360 dwellings). This high concentration of people in flood-prone zones underscores the importance of integrating demographic data into susceptibility assessments, highlighting the extreme vulnerability of these sectors and the pressing need for urgent interventions.

This pattern reflects a significant demographic concentration in vulnerable areas, in line with studies highlighting how urbanized areas are more exposed due to economic growth (i.e., easier access to waterways, greater trade opportunities, and the availability of land at lower prices), which weakens the perception of susceptibility and encourages land occupation in high-risk zones (Arabameri et al., 2022). Furthermore, the "High" and "Very High" exposure levels together represent 65.5% of the population and 64.7% of housing units, indicating structural vulnerability within the state's urbanization. These findings demonstrate the need to improve urban planning and flood resilience, as most socioeconomic infrastructure is located in high-susceptibility sectors.

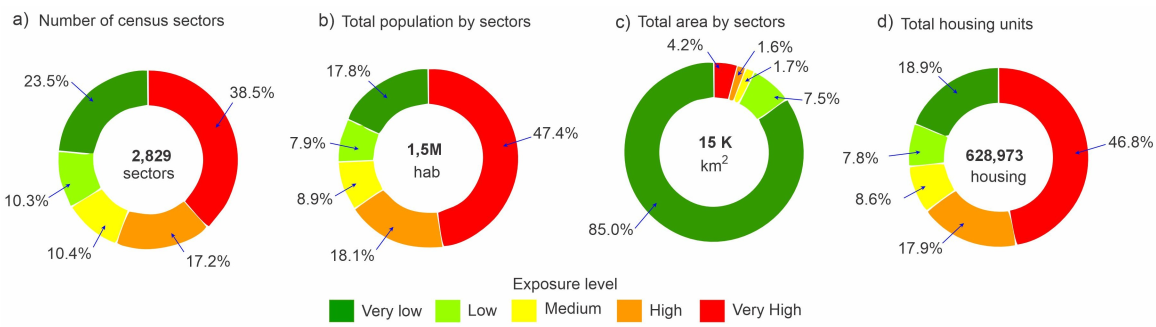

Figure 6: Analysis of exposure to flood susceptibility by census tracts. a) Number of census sectors; b) Total population by census sectors; c) Total area by census sectors; d) Total housing units

CONCLUSION

The results confirm the effectiveness of the RF model in predicting flood susceptibility, highlighting the key influence of slope, land use and land cover (LULC), elevation, and proximity to rivers, which were the main predictors controlling susceptibility. The SHAP analysis identified that the combination of distances close to rivers with low slopes, as well as cropland and built-up areas (LULC classes) located near rivers, increases the probability of flooding. Additionally, there is a clear convergence between areas of high flood susceptibility and high demographic density, with 47.4% of the total population residing in only 4.2% of the total area of the census sectors, demonstrating high demographic vulnerability to extreme events. This trend is also observed in urbanization, where 46.8% of all housing units are located in areas of high flood susceptibility. These findings underscore the need to foster resilience in the most vulnerable areas.

ACKNOWLEDGEMENTS

This study was funded by the São Paulo Research Foundation (FAPESP) (Grant No. 2025/16161-0). The first author acknowledges the National Council for Scientific and Technological Development (CNPq) for awarding a master scholarship. The authors acknowledge financial support from the Fundo de Apoio ao Ensino, à Pesquisa e à Extensão (FAEPEX) of the University of Campinas (UNICAMP) for supporting the visit of Professor Saket Pande (Grant No. 2448/26) and funding the participation of the first author in XVI ENAU and VI SRRU.

REFERENCES

Abdo, Hazem Ghassan et al. (2025). "Machine learning-based assessment of flood susceptibility in the Eastern Mediterranean: a case study of Baniyas River basin". In: *Geomatics, Natural Hazards and Risk* 16.1, p. 2524417.

Ahmed, Ishita Afreen et al. (2022). "Flood susceptibility modeling in the urban watershed of Guwahati using improved metaheuristic-based ensemble machine learning algorithms". In: *Geocarto International* 37.26, pp. 12238–12266.

Amiri, Afshin et al. (2024). "A novel machine learning tool for current and future flood susceptibility mapping by integrating remote sensing and geographic information systems". In: *Journal of Hydrology* 632, p. 130936.

Arabameri, Alireza et al. (2022). "Flood susceptibility mapping using meta-heuristic algorithms". In: *Geomatics, Natural Hazards and Risk* 13.1, pp. 949–974.

Aydin, Halit Enes and Muzaffer Can Iban (2023). "Predicting and analyzing flood susceptibility using boosting-based ensemble machine learning algorithms with SHapley Additive exPlanations". In: *Natural Hazards* 116.3, pp. 2957–2991.

Ballarin, André S et al. (2024). "Frequency rather than intensity drives projected changes of rainfall events in Brazil". In: *Earth's Future* 12.1, e2023EF004053.

Bui, Quynh Duy et al. (2023). "Flood risk mapping and analysis using an integrated framework of machine learning models and analytic hierarchy process". In: *Risk Analysis* 43.7, pp. 1478–1495.

Debnath, Jatan et al. (2024). "Flood susceptibility assessment of the Agartala Urban Watershed, India, using machine learning algorithm". In: *Environmental Monitoring and Assessment* 196.2, p. 110.

Hajji, Sonia et al. (2025). "Enhancing flood prediction through remote sensing, machine learning, and Google Earth Engine". In: *Frontiers in Water* 7, p. 1514047.

Knighton, James et al. (2020). "Predicting flood insurance claims with hydrologic and socioeconomic demographics via machine learning: Exploring the roles of topography, minority populations, and political dissimilarity". In: *Journal of Environmental Management* 272, p. 111051.

Lundberg, Scott M et al. (2020). "From local explanations to global understanding with explainable AI for trees". In: *Nature machine intelligence* 2.1, pp. 56–67.

Oluwadare, Temitope Seun et al. (2025). "Applying Machine Learning Algorithms for Spatial Modeling of Flood Susceptibility Prediction over São Paulo Sub-Region". In: *Land* 14.5, p. 985.

Rentschler, Jun, Melda Salhab, and Bramka Arga Jafino (2022). "Flood exposure and poverty in 188 countries". In: *Nature communications* 13.1, p. 3527.

Tellman, Beth et al. (2021). "Satellite imaging reveals increased proportion of population exposed to floods". In: *Nature* 596.7870, pp. 80–86.

Waleed, Mirza and Muhammad Sajjad (2025). "Advancing flood susceptibility prediction: A comparative assessment and scalability analysis of machine learning algorithms via artificial intelligence in high-risk regions of Pakistan". In: *Journal of Flood Risk Management* 18.1, e13047.

Widya, Liadira Kusuma et al. (2024). "Flood susceptibility mapping of Cheongju, South Korea based on the integration of environmental factors using various machine learning approaches". In: *Journal of Environmental Management* 364, p. 121291.

Yang, Jian et al. (2025). "Enhancing flood susceptibility predictions by using certainty factor in non-flood selection: a case study of Guangdong Province with four tree-based machine learning models". In: *Stochastic Environmental Research and Risk Assessment* 39.7, pp. 3123–3146.