Predictive Model Prototype for Flooding in Caxias do Sul

Eduardo Eberhardt Pereira[1], **Daniel Luis Notari**[1]

[1]Universidade de Caxias do Sul (UCS)
Caxias do Sul, Brazil

eepereira@ucs.br, dlnotari@ucs.br

Abstract. *Heavy rainfall events that cause floods are natural and unavoidable, yet result in devastating damage to society and, therefore, must be studied to mitigate their impacts. This study proposes a predictive artificial intelligence model capable of classifying the occurrence of flood events using machine learning techniques, specifically, MLP neural networks. This research will be conducted in the city of Caxias do Sul, Brazil, using meteorological and hydrological data collected by Instituto Nacional de Meteorologia (INMET). The performance of the model will be evaluated using metrics such as accuracy and confusion matrix to explore recall and precision, and the explainability will be investigated using the Shapley Additive Explanations (SHAP) method.*

1. Introduction

The Intergovernmental Panel on Climate Change (IPCC) is an United Nations (UN) entity created to assess the impacts of climate change in the world. In its sixth assessment report [IPCC 2023], the group informed that the average global temperature is 1.1°C higher than it was in the pre-industrial era. Of these, 1.07°C was caused by anthropogenic actions.

In that regard, climate changes go beyond the growth of the average temperature. The same document by [IPCC 2023] states with high confidence that human intervention is contributing to increasing the frequency of more extreme events. Further evidence is given by [World Weather Attribution 2024], which used computer models to quantify human influence in extreme events. Of the sixteen flood events analyzed by the group, fifteen were a result of heavy rainfall caused by anthropogenic actions.

In 2004, approximately a quarter of all registered natural disasters worldwide were flood events [Munich Re Group 2004]. In Brazil, between 2017 and 2022, 55.5 billion reais were spent on damage repair due to heavy rainfall, and 14.8 million people were displaced or killed because of it [CNM 2022]. During the months from September 2023 and May 2024, the Rio Grande do Sul state, in Brazil, suffered several casualties due to three different flood events that broke historical records in terms of rainfall and urban flooding [Marengo et al. 2024].

Floods are one of the most common, fatal, and destructive extreme events [Razavi et al. 2020, Marengo et al. 2024]. They can be defined as the temporary cover of land by water outside of its normal confinement, invading terrains not normally occupied by it [FLOODsite 2005, de Sene and Moreira 2012]. In order to prevent damage from these disasters, efficient disaster management is essential. The process involves several phases, including preparation, response, and recovery [Sharma et al. 2021]. Early warning flood systems enter the preparation phase by guaranteeing that the public defense

is not caught off guard, and prevention mechanisms are well established by the time the event takes place [Sharma et al. 2021].

Early warning systems can be implemented using artificial intelligence and machine learning techniques [Sharma et al. 2021]. This project proposes the development of a machine learning model for early warning of flood events in Caxias do Sul, Brazil, using meteorological and hydrological data from flood events during the period from September 2023 and May 2024.

2. Related Works

Several studies surrounding flood prediction have been conducted over time in different locations, but the methodology is not consistent between these works. [Mosavi et al. 2018] reviewed more than 600 articles about the theme, and concluded that most of the methodologies focus on predicting the rainfall-runoff conversion using different hydrological models' data as input.

However, discussions have been raised on whether these data are reliable or not. [Tucci 2004] and [Ukhurebor et al. 2020] defend that the synthetic data generated by those hydrological models are a decent source to bypass the complexities of installing a monitoring system, but these data will never be as reliable as the observational data collected by sensors.

Considering this issue, [Geron 2019] defends that there's no reason to work with the most complex and robust Artificial Intelligence (AI) models if the data that is training them is inherently problematic. This position agrees with [Seo and Breidenbach 2002] view that the majority of the time spent on Data Science solutions is dedicated to cleaning the data.

Even so, [Klemeš 1983] and [Todini 1988] argue that the absence of a hydrological model in flood prediction makes no sense, and compare it to trying to learn hydrology without considering the water. Those affirmations were regarding regressive models of the time, but can be extended to the AI mentality, considering that this worry is amplified by the nature of the black box models, in which explanation is often disregarded [Russel and Norvig 2022].

This issue could easily be solved by always collecting the data with sensors. However, as pointed by [Mosavi et al. 2018], most works in the topic fight with the lack of reliable and consistent data. Sensors and measures used in complex hydrological models are very difficult to implement, causing very few cases of this sensing to be found in the literature.

Therefore, the literature shows the duality between trying to use hydrological models and data for prediction, guaranteeing explicability, while avoiding using synthetic data and dealing with the scarcity of it. This leads to this study's goal of finding an intermediate, hybrid approach to predict floods, by using simple hydrological models with Multi-Layer Perceptron (MLP) to embrace both hydrological reasoning and data-driven learning.

3. Methodology

This study follows the Sample, Explore, Modify, Model, and Assess (SEMMA) methodology for artificial intelligence development. This method divides the process into sampling and exploring the data, modifying it to fit the application, training the model, and

testing how well it generalizes the data [Sharda et al. 2019]. Since SEMMA is an iterative method, if the results of the test step aren't as good as expected, the process can be restarted from any given step in order to improve the model.

The model will be trained using meteorological data collected by INMET[1], in the city's airport Hugo Cartergiani. These data will be the input for the model. Hydrological models will be integrated as input to the model as well, such as the rational method for rainfall-runoff conversion and the Antecedent Precipitation Index (API) method for keeping the temporal dimension that common neural networks can't store [Viessman and Lewis 2003].

The data will be labeled using alerts from the state's civil defense agency[2]. The civil defense uses the terminology "warning" for possible events and "alert" for events that are taking place. Therefore, only alerts were taken into account for labeling.

Those labels will later be shifted backward in variable time steps, creating a lag that allows the model to learn if the current conditions caused floods in the future. Tests will be made using different numbers of time steps, which will go from one to four days in the future, allowing a study of the efficiency of the prediction in all these forecast horizons.

The model will be a MLP neural network, trained supervised with the data discussed. The output will consist of a single neuron in which the output represents the probability of a flood occurring. That way, if the output is closer to zero, it is classified as no flood, but if is closer to one, a flood is being predicted. Since INMET collects the rain and isolation data daily, the prediction time slot is in days.

The dataset will be split into training and testing sets. After training the model, accuracy and confusion matrix will be used as metrics to determine the model's capacity to predict a flood. Training and testing the model fit, respectively, the model and assess phases of SEMMA methodology. Since the methodology is iterative, if the assessment concludes that the model does not predict flooding well, the process can be restarted at any point, so different approaches can be tested to achieve good enough accuracy [Sharda et al. 2019].

Explainability will be explored using SHAP method for testing how each feature contributes to the model's decision. That way, some hydrological explanations can also be assessed.

4. Partial Results

This study will be conducted as part of the second half of a bachelor's final project during the second semester of 2025. The proposed approach is inspired by the current state of the art. As a novel contribution, this work introduces a classification system for detecting the presence or absence of flooding by balancing hydrological models and AI prediction. While most existing studies focus on predicting basin runoff using synthetic data for training, this study explores whether a machine learning model can directly determine whether flooding will occur in the city under given conditions.

Considering the data is collected daily, this study works with daily predictions, being a temporal limitation of the model, since prediction within hours is arguably more useful and precise considering the problem. Further improvement could be made in the

[1]Available at: https://bdmep.inmet.gov.br

[2]Available at: https://www.defesacivil.rs.gov.br/avisos-e-alertas

future by using hourly data. The spatial range is another limitation of the model, which predicts the flood to occur somewhere in the region where the data was collected, without precision. Further efforts to collect data in different regions of the city would make it possible to provide better precision on where the events would take place.

References

CNM (2022). Prejuízos causados pelas chuvas no brasil entre 2017 e 2022 ultrapassam 55,5 bilhões, revela cnm.

de Sene, E. and Moreira, J. C. (2012). *Geografia geral e do Brasil: espaço geográfico e globalização*, volume 1. Editora Scipione, São Paulo.

FLOODsite (2005). Language of risk.

Geron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, Sebastopol, 2 edition.

IPCC (2023). Climate change 2023: Synthesis report.

Klemeš, V. (1983). Conceptualization and scale in hydrology. *Journal of Hydrology*, 65:1–23.

Marengo, J. A., Dolif, G., Cuartas, A., Camarinha, P., Gonçalves, D., Luiz, R., Silva, L., Alvava, R. C. S., Seluchii, M. E., Moraes, O. L., Soares, W. R., and Nobre, C. A. (2024). O maior desastre climático do brasil: chuvas e inundações no estado do rio grande do sul em abril-maio 2024. *Estudos Avançados*.

Mosavi, A., Ozturk, P., and wing Chau, K. (2018). Flood prediction using machine learning models: Literature review. *Water*, 10(11):1536.

Munich Re Group (2004). Topics geo annual review: Natural catastrophes 2004.

Razavi, S., Gober, P., Maier, H. R., Brouwer, R., and Wheater, H. (2020). In *Anthropocene flooding: challenges for science and society*, volume 34, pages 1996–2000. Hydrological Processes.

Russel, S. J. and Norvig, P. (2022). *Inteligência Artificial: Uma Abordagem Moderna*. gen LTC, 4 edition.

Seo, D.-J. and Breidenbach, J. P. (2002). Real-time correction of spatially nonuniform bias in radar rainfall data using rain gauge measurements. *Journal of Hydrometeorology*, 3(2):93–111.

Sharda, R., Delen, D., and Turban, E. (2019). *Business intelligence e análise de dados para gestão do negócio*. Bookman, 4 edition.

Sharma, K., Anand, D., Sabharwal, M., Tiwari, P. K., Cheikhrouhou, O., and Frikha, T. (2021). A disaster management framework using internet of things-based interconnected devices. *Mathematical Problems in Engineering*, 2021(2629):1–21.

Todini, E. (1988). Rainfall-runoff modeling – past, present and future. *Journal of Hydrology*, 100:341–352.

Tucci, C. E. M. (2004). In *Hidrologia: Ciência e Aplicação*. Editora da UFRGS, Porto Alegre, 4 edition.

Ukhurebor, K. E., Azi, S. O., Aigbe, U. O., Onyancha, R. B., and Emegha, J. O. (2020). Analyzing the uncertainties between reanalysis meteorological data and ground measured meteorological data. *Measurement*, 165:108110.

Viessman, W. J. and Lewis, G. L. (2003). *Introduction to Hydrology*. Prentice Hall, Upper Saddle River, New Jersey, 4 edition.

World Weather Attribution (2024). When risks become reality: Extreme weather in 2024.