Christian Laurio
Theo Guilbert-Hattermann

Flood Risks in India

A little bit of context...

Floods have always been a relevant topic in India, especially due to the **"Indian monsoon"** phenomenon. This climatic event is a major wind system that seasonally reverses its direction. In India, the wind blows from the *northeast* during cooler months and reverses direction to blow from the *southwest* during the warmest months of the year. At the same time, in early summer, the lands and the Himalaya mountains heat up, so that hot air climbs up. Consequently, when the southwest wind meets this warm air, it creates large amounts of rainfall in the country during June and July.

Many factors contribute to flooding, but experts say climate change caused by global warming makes extreme rainfall more likely. India has experienced increasingly extreme weather in recent years, where unrelenting rains came just weeks after an extreme heat most of north India. For instance, in June 2022, 32 of its 35 districts have been affected by the tremendous floods, killing 45 people and displacing more than 4.7 million over one week.

Project Goal

This project aims to analyze and construe floods in India over the years, in order to identify the relevant features and then build a predictive flood model. We have chosen three different datasets from the Kaggle website: *kerala.csv, rainfall in india 1901-2015.csv and flood_risk_dataset_india.csv.* They basically contain the monthly records of rainfall from 1900-2018 from 36 subdivisions of India, with some meteorological, geographical and socio-economic features. Thanks to these information, the goal was to extract some insights for policy makers and apply flood control measures at the end. The final predictive model must indicate if floods will occur or not, use the help of machine learning methods (Random Forests Model, KNN Classifier and Support Vector Machine).

Is the "Indian monsoon" phenomenon visible on data?

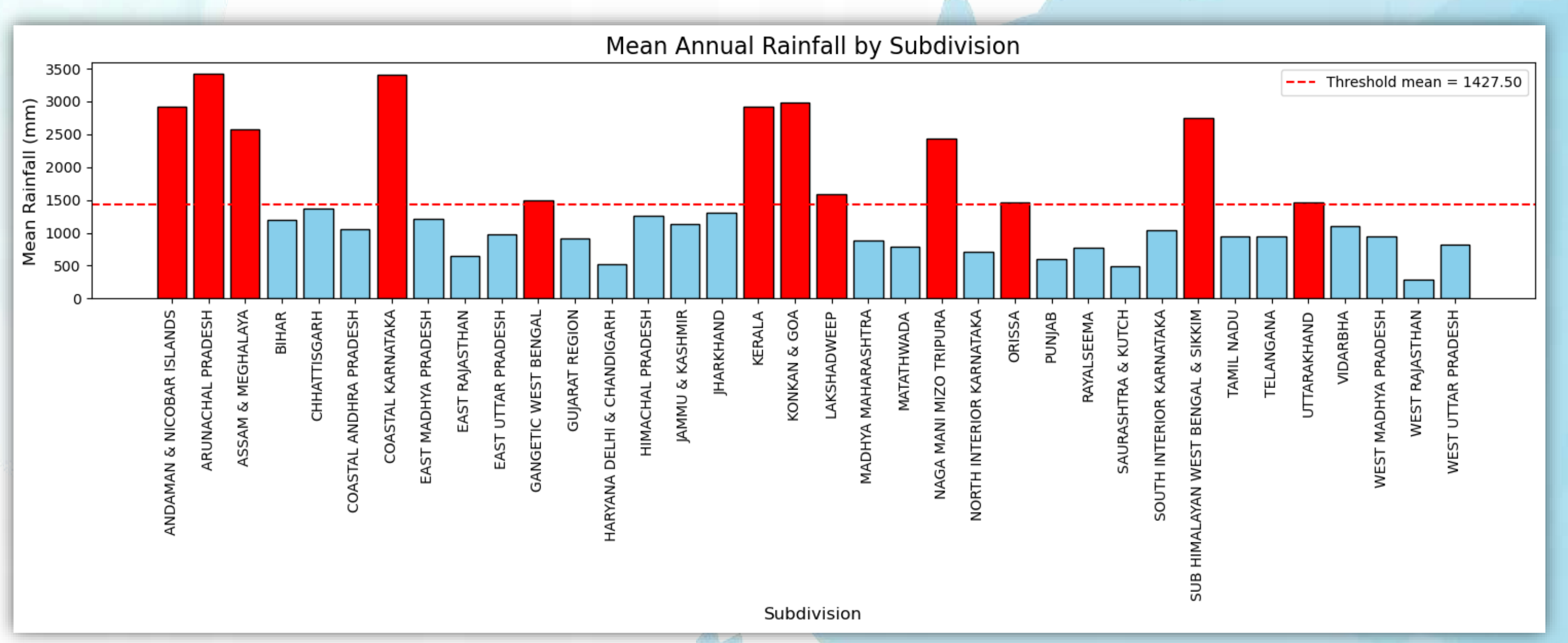


Figure 1: Mean Annual Rainfalls by Subdivision of India

At the beginning, we wanted to see if we could easily identify the areas the most hit by floods in India. That is the reason why we first organized the data by subdivision and then plotted the mean of annual rainfall for each of them.

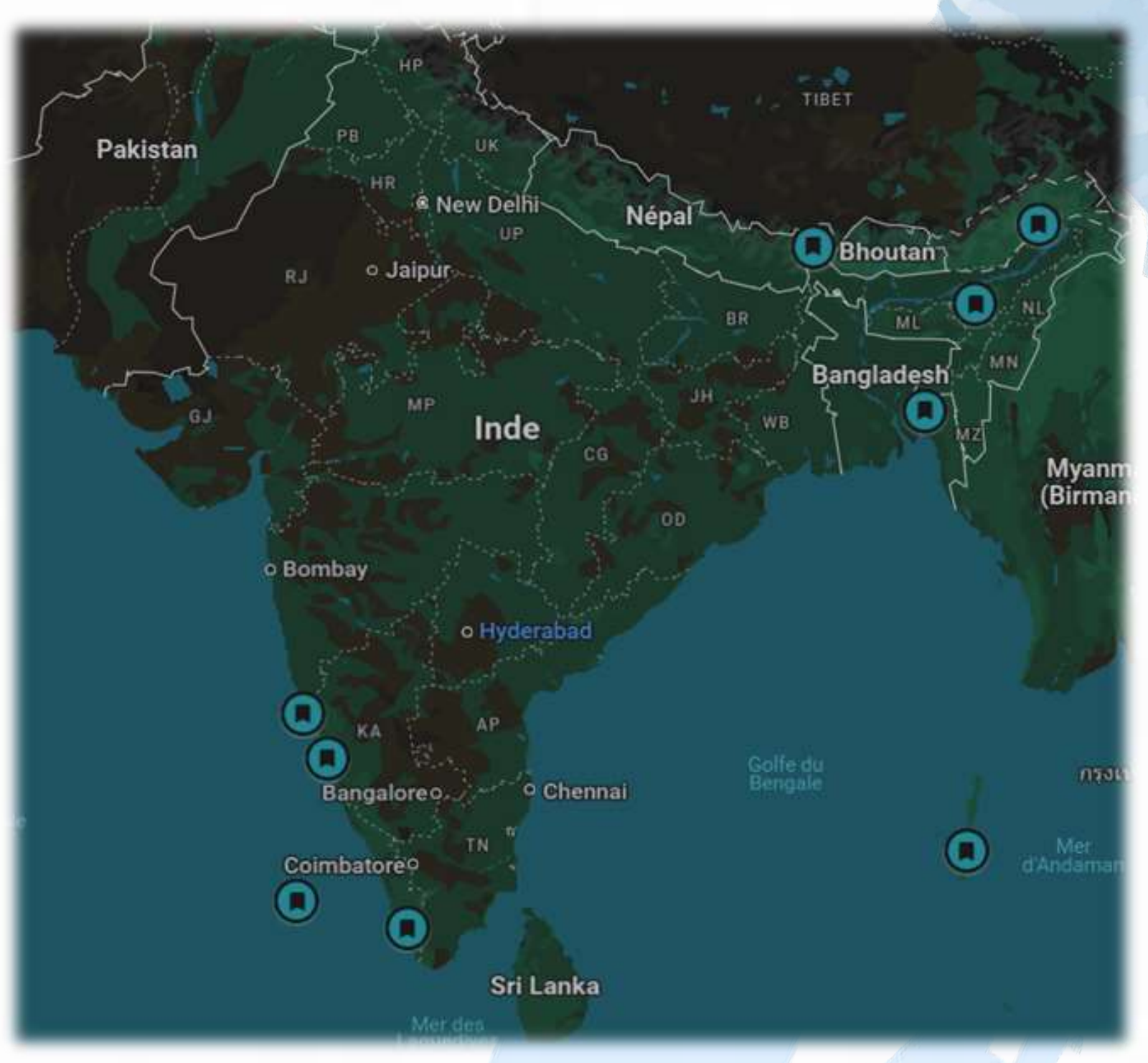


Figure 3: Locations of the areas with the Mean Annual Rainfall above the mean of the whole country

After pinned the locations of the critical subdivisions on a map, we can observe that they are located on the diagonal South-West, North-East of the country. This confirms what we expected with the "Indian monsoon" phenomenon.

Relevant features about floods

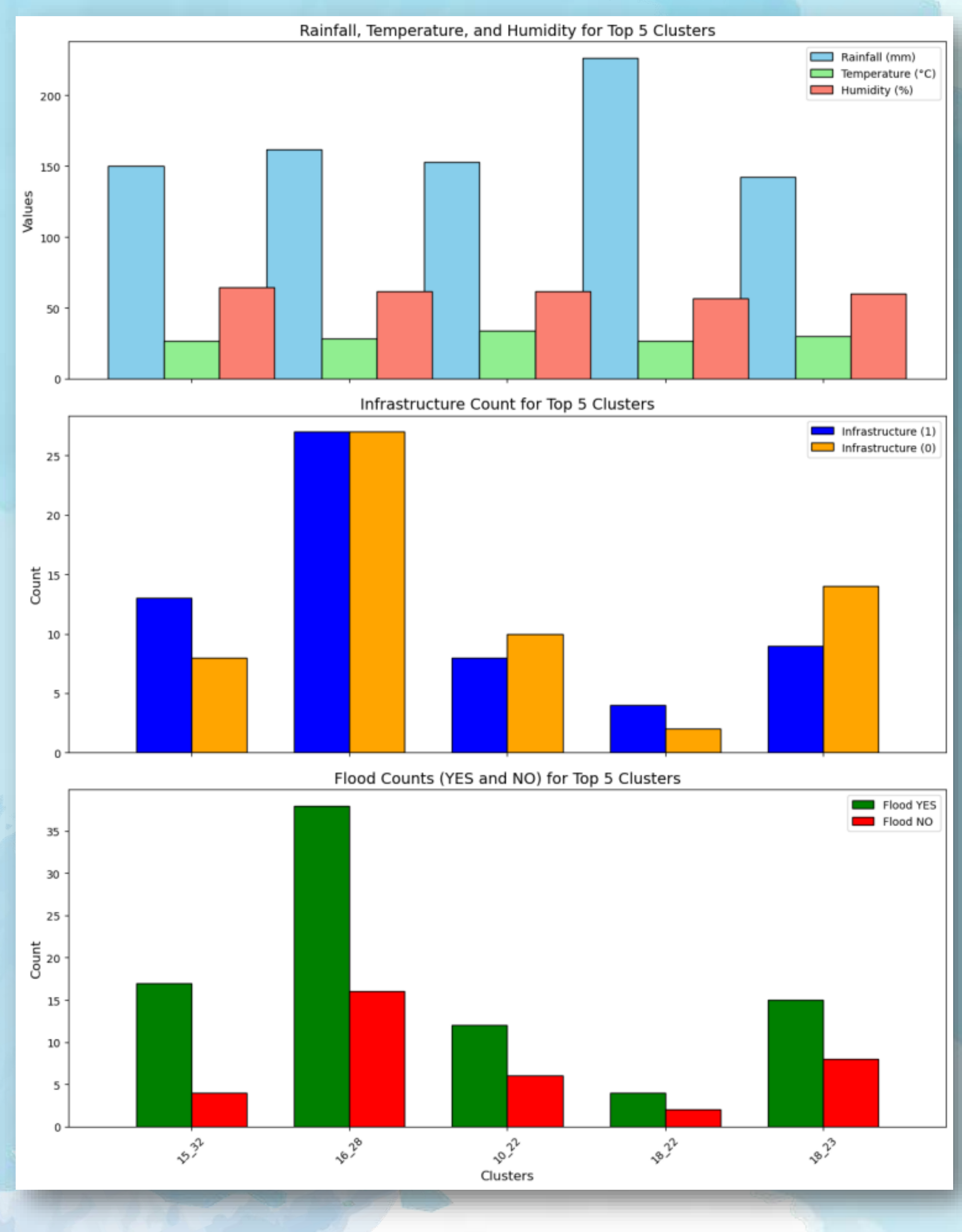


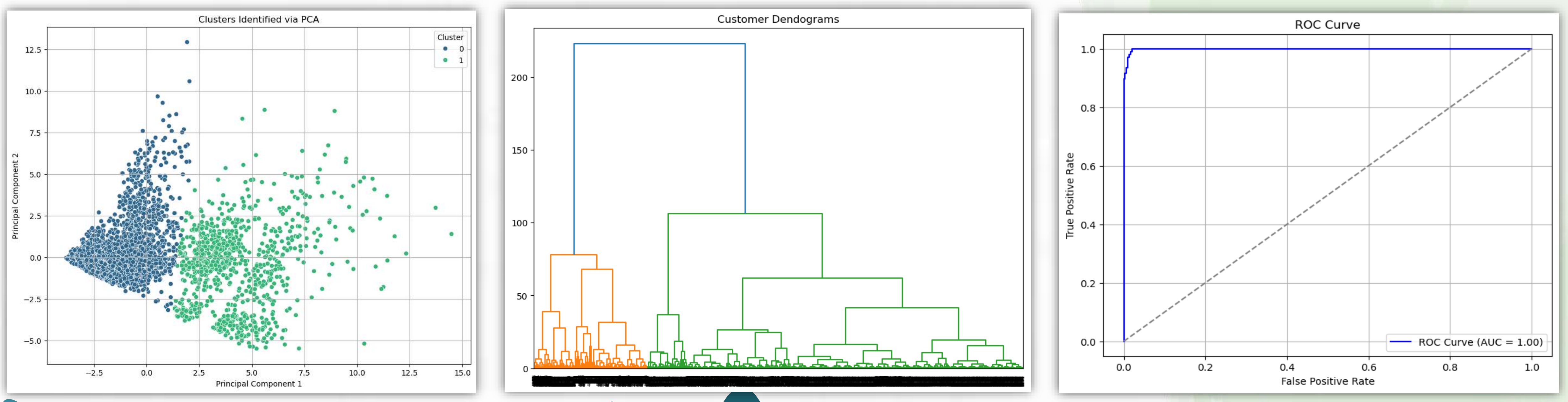
Figure 2: Mean Rainfall, Temperature, Humidity, Flooding Infrastructures and Floods Occurred for the five clusters with the highest floods rate

We wanted to identify the relevant features that could increase the chances to cause floods. For that, we first created clusters based on the latitude (2°) and the longitude (3°), and then counted the floods occurred in each cluster. Afterwards, we sorted them and we only kept the five best clusters where the flood success rate was the highest.

We made the following conclusions. First and foremost, floods are the most important when the temperature and the humidity are high (around 30°C and 60% of humidity), so in summer (Indian monsoon). In addition, the flood control infrastructures are not enough to prevent floods, since they can often occurred even if the areas are equipped with these installations. Finally, rainfalls are not the only criteria to take into account for measuring floods. Indeed, we also compared the areas with the highest and lowest mean rainfall by year. It appeared that rainfalls were two times lower for the second group, but floods occurred as much as the first group.

What the models could learn from the monthly rainfall index in India?

With monthly rainfall index in India from 1901 to 2015, we used the dataset to predict the probability of flooding in different location and years. After data cleaning and pre-processing, we checked if using principal component analysis (PCA) **(Figure A)** and agglomerative hierarchical clustering **(Figure B)** to find a clustering among the dataset of whether there's a flooding or non flooding occurred. We also trained a model using the same dataset using random forest model to predict whether it would flood or not based on certain threshold of the annual monthly rainfall indices in India. Result showed a good prediction model based on its ROC curve **(Figure C).**



Figures (A) plot of PCA, **(B)** AHC diagram, using the rainfall index in India from 1901 to 2015 and **(C)** the ROC curve of the random forests model developed with the same dataset.

In a nutshell...

To conclude, it is difficult to establish a direct link between only some features. Instead, we should combine all these information and add other parameters in order to build a predictive model that could predict more or less precisely floods. Predicting floods is therefore a hard task, where simulations can be very useful in this type of use case.