XXIV ENMC
Encontro Nacional de Modelagem Computacional
XII ECTM
Encontro de Ciência e Tecnologia de Materiais

FROM RAINFALL DATA TO A TWO-DIMENSIONAL DATA-SPACE SEPARATION FOR FLOOD OCCURRENCE

Wagner da Silva Billa[1] - wagner.billa@gmail.com
Leonardo Bacelar Lima Santos[1] - santoslbl@gmail.com
Rogério Galante Negri[2] - rogerio.negri@unesp.br
[1]National Institute for Space Research (INPE), São José dos Campos, São Paulo, Brazil
[2]São Paulo State University (UNESP), São Jose dos Campos, São Paulo, Brazil

Abstract. *Extreme rainfall events are the most important triggers for natural disasters in several countries worldwide. Therefore, rainfall monitoring is a crucial component in early warning systems. In this paper, we use actual flooding data from the city of São Paulo (Brazil) to check if simple linear classifiers using historical accumulated precipitation can predict such events. We use a database for flood events, with geographical location, starting and ending timestamps associated with each flood event's impact. We propose the use of instant maximum precipitation as a time-reference for non-flood events. A linear function on a 2D rainfall data-space (the accumulated precipitation in 2 hours and 96 hours) separates flood and non-flood events. For the studied scenario with a set of 87 flood/non-flood events, associated with a selected rain gauge readings, we've got 11 misclassified events, giving us an accuracy of 87.36%.*

Keywords: *rainfall episodes, floods, flood events, non-flood events, linear classifier*

1. INTRODUCTION

In the last few decades, Extreme Weather Events (EWE) have been happening more frequently and with greater intensity. Heading causes of such events are gas emissions and the greenhouse effect, which increases the planet's temperature and triggers other harmful effects such as floods, mass movements, prolonged droughts, heat waves, typhoons and tornadoes (Akhtar, 2020).

In Brazil, the EWE became more frequent from the second half of the twentieth century. Recently, events like the Catarina hurricane caused floods and landslides as well as several deaths and significant economic losses for the southern region of the country (Marengo, 2009; Stevaux et al., 2009).

A common consequence of EWE are the flooding events, which impacts on several cities around the world, mainly in tropical countries (Dewan, 2013; Ceola et al., 2014). According to the Brazilian Atlas of Natural Disasters (Atlas, 2012), floods correspond to 50% of the occurrences of socio-environmental disasters in the country recorded in recent years and are related to expressive economic impacts. Regarding the city of São Paulo, in 2013, the losses reached R$ 336 million. When its consequences are analyzed at a national level, this amount exceeds R$ 762 million (Haddad & Teixeira, 2015). Therefore, it is clear that monitoring and

predicting such events beyond saving lives are essential to mitigate the impacts of these disasters on the economy.

In the literature, the most common approach for predicting flood events is through the creation of numerical models, often complex and with various types of input data, both qualitative and quantitative (Tehrany et al., 2013; Khosravi et al., 2018; Chapi et al., 2017; Chen et al., 2019). Many of these models attempt to calculate rainfall and runoff ratios using diverse techniques and approaches such as genetic algorithms (Bui et al., 2019), classification and regression trees (Wang et al., 2015; Choubin et al., 2019), random forest (Xie et al., 2017), support vector machines (Degiorgis et al., 2012), fuzzy logic (Darabi et al., 2019), parametric based modeling techniques (Balica et al., 2012) and artificial neural networks (Toth et al., 2000; Shafizadeh et al., 2018).

Nonetheless, more straightforward approaches, such as analyzing historical flooding events combined with rainfall data to distinguish flooding events, are exciting proposals. In this sense, using such information, it is possible to derive a simple yet more abstract approach than the current complex computational models.

In the context above presented discussions, we perform a case study using actual flooding data in São Paulo (Brazil). Our main objective is to verify if a two-dimensional classification based on accumulated rainfall data is enough to distinguish flood and non-flood events.

2. METHODOLOGY

Figure 1 depicts the methodology framework of this study. Initially, the process starts with a data acquisition stage (Section 2.1) followed by an input data analysis (Section 2.2). Posterior, we delimited the study area for the experiment and analysis (Section 2.3). Lastly, we analyze and discuss the achieved results (Section 3).

2.1 DATA ACQUISITION

In this study's scope, it is considered a flood event database composed of traffic interruption reports in the city of São Paulo, Brazil, regarding the period from January 1st, 2015 to December 29th, 2016. Accordingly (Soriano et al., 2016), the pointed period was the last drought crisis in the state of São Paulo, affecting its water supply. On the other hand, several flooding episodes were also observed.

Such database contains $1,928$ entries described in terms of three attributes: (i) exact or "near reference" geographic coordinates of the flood event; (ii) initial timestamp, when the flood started to obstruct the local traffic; and (iii) final timestamp, when the traffic returns to its usual condition. All information was collected and provided by the Climate Emergency Management Center - CGE (CGE, 2021).

Beyond identify flood events, it was also necessary to consider non-flood events. For this purpose, GIDES project (Pimentel et al., 2020) was used as a second data source. In summary, the methodology of the GIDES project tracks all rainfall episodes near rain gauges. Two timestamps define the rainfall concept for GIDES for each event: the initial and the final timestamps. Between these instants may occur periods without precipitation, but it lasts no longer than 24 hours. Moreover, before the initial timestamp and after the final timestamp, there must be a period equal or greater than 24 hours without precipitation. Furthermore, GIDES also registers the timestamp of the highest precipitation value in each rain. From such information, it is possible to check the reports from CGE and consider those that were not pointed as flood

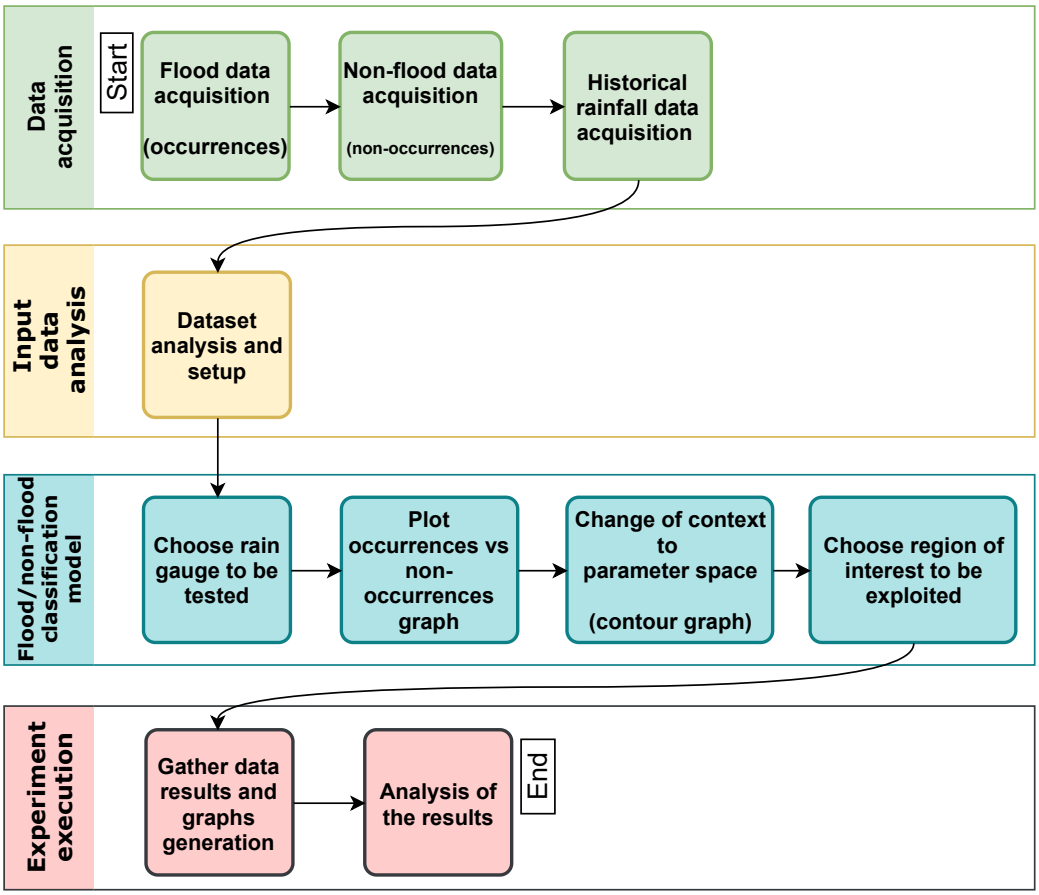

Figure 1- Flowchart for the proposed methodology, following a waterfall model.

events and use them as non-occurrence data. In other words, it is considered the rainfalls that do not cause flood events.

Once gathered data from occurrence and non-occurrence flood events, it is recovered the respective accumulated rainfall registries. For this purpose, it is adopted the rainfall database provided by the Brazilian National Center for Monitoring and Early Warning of Natural Disasters (CEMADEN) (Bernardes et al., 2019).

Figure 2 represents the spatial distribution of reports from CGE (red circles) and rain gauges (blue triangles) in the context of the city of São Paulo. The highlighted rain gauge **#355030862A** is considered as reference for the experiments of Section 3.

2.2 INPUT DATA ANALYSIS

As discussed in Section 2.1, information regarding flood and non-flood events, and their respective rainfall registries, were obtained from different sources. This information is stored into a relational database schema for ease of manipulation and convenience, before being used in an *input data analysis* stage.

Such a process aimed at assigning historical rainfall values to flood events. For this purpose, the values measured by the nearest rain gauge were considered in the event timestamp. A database sanitization analysis was carried out to find anomalous data and non-trusted

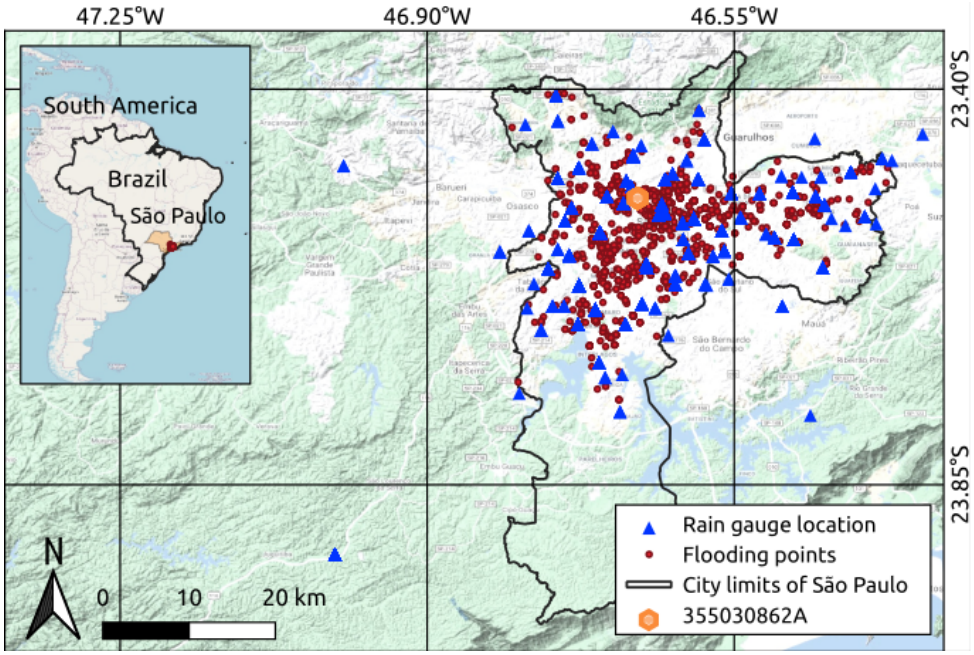

Figure 2- Study area context and spatial distribution of flood events and rain gauges.

information. It was ensured that the rain gauge associated with each event was the closest one and that its distance was less than 2000 meters, because readings farther than this are considered not reliable. All event's timestamps were converted to UTC, since CEMADEN's rain gauge web services for retrieving historical rainfall values relies on this timezone. If that was not handled, we could get wrong values since the city of São Paulo uses the GMT-3 timezone (a three hour difference from the UTC timezone). It was also verified if the accumulated 120 hours reading was greater than zero, necessary condition for the reading to be considered valid.

Considering these initial inconsistencies, some other possible problems were also checked:

- Rain gauge out of order (down or clogged);
- Rain gauge not calibrated;
- Rain gauge internal clock out of synchrony;
- Despite the rain gauge location relatively close to a flood-affected area, the precipitation values were much smaller on it (i.e., localized rain);
- Inaccurate geographic coordinates or "near reference" of flood location.

The first three possibilities represents problems that come from a purely mechanical device: the rain gauge. Those are known issues that might be mitigated or suppressed by periodical maintenance. The fourth problem is assigned to an environmental cause that might happen, where the flooded area is much more affected than the rain gauge location, due to a localized rain. The last one comes from the fact that a flood event represents an affected area instead of a single point location. This is a spatial location problem. Since our occurrences are based on punctual references, maybe the registered location is not as representative as the entire area, which should be treated as a region instead of a point in the analysis.

We could have investigated some other kind of problems but this was not the main objective of the study. Nevertheless, it is important to know that they do exist and can impact the output results and reliability.

2.3 FLOOD AND NON-FLOOD CLASSIFICATION MODEL

In the scope of this work, we propose using a simple linear model to distinguish between flood and non-flood events. For such a purpose, firstly, the focus relies on verifying a two-dimensional representation that allows such separation (i.e., perform a classification).

Formally, let consider that a pair of values regarding the accumulated rainfall over 2 and 96 hours are sufficient and necessary attributes to describe and distinguish flood and non-flood events. Expressing such attributes in the axis of a Cartesian representation, the flood occurrence and non-occurrence take place into this space, whose separation may be given by:

$$f(\mathbf{x}) = (x_2, x_{96})^T \cdot (a, -1) + b \tag{1}$$

where x_2 and x_{96} represents the accumulated rainfall attributes at 2 and 96 hours, respectively; a and b are parameters in such linear model. Accordingly to this model, for a given observation $\mathbf{x} = (x_2, x_{96})$, $f(\mathbf{x}) \geq 0$ implies that $\mathbf{x}$ comprehends a flood event; otherwise, $f(\mathbf{x}) < 0$ implies a non-flood event.

Additionally, from *a prior* knowledge, we may affirm that flood and non-flood events are separated by a decreasing-shaped $f(\cdot)$ model. Such behavior is plausible since flood events are observed for high values in x_2 and/or x_{96}. Reversely, non-flood usually has both small x_2 and x_{96}. Consequently, the parameter a should be a negative value. With respect to b, its value should be positive since x_2 and x_{96} are both greater or equal to zero.

Adopting the model of Equation 1 and assuming that exists a set of observations $\mathbf{x}_i$, for $i = 1, \ldots, m$, where it is also known if such observation represents a flood or non-flood event, several parameters configuration may be assessed according to its suitability in providing a linear separability between flood and non-flood events. The mentioned assessment is conveniently expressed regarding misclassification rates observed on separate flood and non-flood events for a given parameter configuration, where low rates stand for better separations.

3. EXPERIMENTAL RESULTS AND DISCUSSION

Face with the rationale presented in Section 2, an experiment was carried out in order to perform assessment of flood and non-flood detection using linear models in a simple two-dimensional data-space. For this purpose, it was considered an original data set composed by flood and non-flood events inside the influence of one specific rain gauge, denoted by **#355030862A** (see Figure 2).

In Figure 3 is depicted a scatter plot of flood and non-flood events, in terms of 2 and 96 hours of accumulated precipitation, observed in the radii of influence of rain gauge **#355030862A**. As previously discussed (Section 2.3), the separation between such events is achieved by a descending linear model (i.e., negative a).

Considering such a linear model (Equation 1) and the data-space defined in Section 1, the contour plot of Figure 4 express the misclassification rates (or error rates) assigned to each parameter configuration. Such representation highlights a diagonal zone where the parameter configuration delivers low error rates. The parameters assigned to this region play a linear separation with only 11 misclassifications, in an universe of 87 events, representing an accuracy

of 87.36%. A total of 51 a and b solution pairs appear as red diamonds in this area, as shown in Figure 4. The arithmetic mean and the standard deviation for these parameters, regarding the 51 low-error solutions, are equivalent to -7.8235 and 1.5205 for a, and 167.549 and 23.6679 for b, respectively.

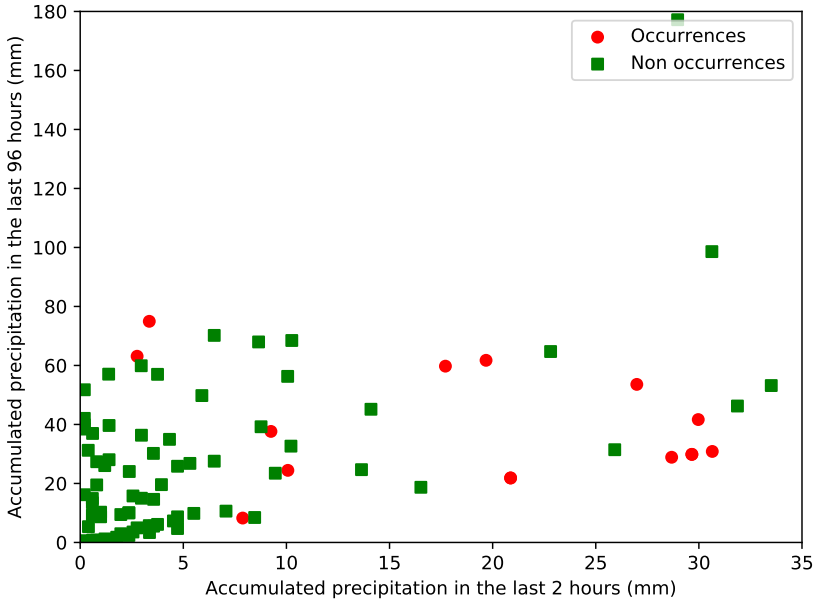

Figure 3- Flood and non-flood occurrences in the neighborhood of rain gauge **#355030862A** in terms of accumulated precipitation values in the past 2 and 96 hours.

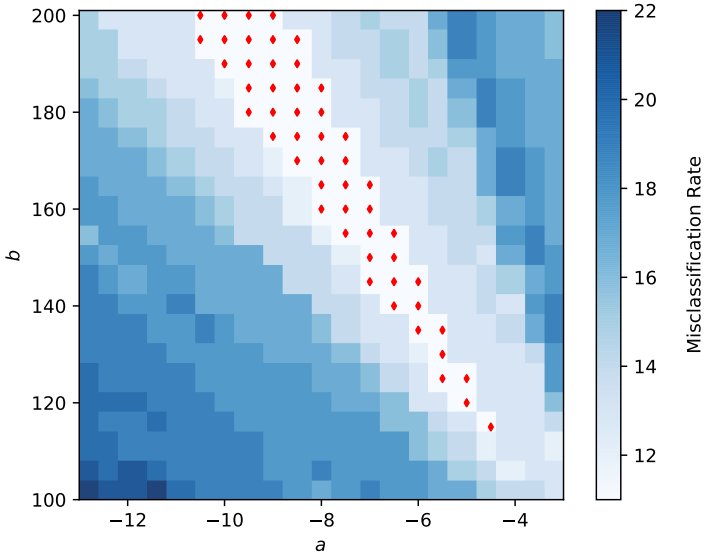

Figure 4- Contour plot of misclassification rate as function of a and b parameters. Red dots are the solution points with the lowest misclassification rate found.

Interesting to note that when looking to the contour plot graphic (Figure 4) we see the visual representation of our mathematical assessment in a clearer and more intuitive way. We see that

the lower misclassification rate appears in a continuous area which we should explore better. In our case, since we used discrete values with defined steps for parameters a and b, we found out 51 solution points in this area. Apparently, either of them represents an optimal solution choice for our linear model because we will reach the lowest misclassification rate (11 events), but since our empirical analysis depends on data and its quality, we shall remember that we have a data driven process herein. This means that the data may not represent the full range of analyzed events variability and a bias may be associated with the solution points.

To reduce the effect of data bias, perhaps the choice of the closest solution point to the means of a and b should be the most suitable value to be used in the linear model. This point is $(-8; 170)$ and it is one of thousands of possibilities in the parameter space presented in Figure 4. Using Equation 1 defined in Section 2.3, the linear model classifier can then be expressed as defined in Equation 2.

$$f(\mathbf{x}) = (x_2, x_{96})^T \cdot (-8, -1) + 170 \tag{2}$$

4. CONCLUSIONS

When plotting the occurrence and non-occurrence points in a two-dimensional graph, using 2 hours and 96 hours precipitation values in the axis, it was impossible to separate both classes without any misclassification point. The minimum misclassification rate was 11, meaning that the classification will still fail in 11 points of our universe of 87 (occurrences and non-occurrences), achieving an accuracy upper-bounded by 87.36%.

Some inconsistencies and other possible problems were described in Section 2.2 to justify this result (rain gauge clogged or down, not calibrated, out of synch, etc.) but, in another way, it can just means that two dimensions of historical rainfall values are insufficient for the correct classification of such events. Maybe adding a third dimension (an intermediate rainfall measure, for instance) could add additional information that makes the classification correct also in current misclassified points. This is a scenario that will be investigated using data from more than 80 rain gauges still available for this analysis. It will be used several Machine Learning algorithms (Decision Trees, Random Forest, Support Vector Machine, K-Nearest Neighbours, Artificial Neural Network and others) to handle such assessments, ease the conclusions, check the hypothesis and search for patterns.

Since this study involved an empirical analysis based on data driven information, a bias may be associated with the solutions found. To reduce its effect the choice of the closest solution point to the means of a and b should be the used in configuration of the linear model used for the classification of new predicted events.

Acknowledgements

This document is the result of the research project funded by FAPESP Grant Numbers 2015/50122-0, 2018/01033-3, 2018/06205-7 and 2021/01305-6; CNPq Grant Number 420338/2018-7 and DFG-IRTG Grant Number 1740/2.

REFERENCES

Rais Akhtar (2020). *Extreme Weather Events and Human Health*. Springer.

Centro Universitário de Estudos (2012). Atlas brasileiro de desastres naturais 1991 a 2010: volume brasil. *Florianópolis: CEPED, UFSC*.

SF Balica, Ioana Popescu, Lindsay Beevers, and Nigel G Wright (2013). Parametric and physically based modelling techniques for flood risk and vulnerability assessment: a comparison. *Environmental modelling & software*, 41:84–92.

Tiago Bernardes, Regina Reani, Rodrigo Conceição, Rafael Luis, Cristina Lourenço, Rogerio Carneiro, Maria Medeiros, and Gustavo Silva (2019). Flood and landslide events database for the municipalities monitored by brazilian center for monitoring and early warnings of natural disasters–cemaden. In *Geophysical Research Abstracts*, volume 21, pages 1–1.

Dieu Tien Bui, Paraskevas Tsangaratos, Phuong-Thao Thi Ngo, Tien Dat Pham, and Binh Thai Pham (2019). Flash flood susceptibility modeling using an optimized fuzzy rule based feature selection technique and tree based ensemble methods. *Science of the total environment*, 668:1038–1054.

Serena Ceola, Francesco Laio, and Alberto Montanari (2014). Satellite nighttime lights reveal increasing human exposure to floods worldwide. *Geophysical Research Letters*, 41(20):7184–7190.

CGE (2021). Centro de Gerenciamento de Emergências Climáticas, São Paulo, Brazil. `https://www.cgesp.org/v3/`

Kamran Chapi, Vijay P Singh, Ataollah Shirzadi, Himan Shahabi, Dieu Tien Bui, Binh Thai Pham, and Khabat Khosravi (2017). A novel hybrid artificial intelligence approach for flood susceptibility assessment. *Environmental modelling & software*, 95:229–245.

Wei Chen, Haoyuan Hong, Shaojun Li, Himan Shahabi, Yi Wang, Xiaojing Wang, and Baharin Bin Ahmad (2019). Flood susceptibility modelling using novel hybrid approach of reduced-error pruning trees with bagging and random subspace ensembles. *Journal of Hydrology*, 575:864–873.

Bahram Choubin, Ehsan Moradi, Mohammad Golshan, Jan Adamowski, Farzaneh Sajedi-Hosseini, and Amir Mosavi (2019). An ensemble prediction of flood susceptibility using multivariate discriminant analysis, classification and regression trees, and support vector machines. *Science of the Total Environment*, 651:2087–2096.

Hamid Darabi, Bahram Choubin, Omid Rahmati, Ali Torabi Haghighi, Biswajeet Pradhan, and Bjørn Kløve (2019). Urban flood risk mapping using the garp and quest models: A comparative study of machine learning techniques. *Journal of hydrology*, 569:142–154.

Massimiliano Degiorgis, Giorgio Gnecco, Silvia Gorni, Giorgio Roth, Marcello Sanguineti, and Angela Celeste Taramasso (2012). Classifiers for the detection of flood-prone areas using remote sensed elevation data. *Journal of hydrology*, 470:302–315.

Ashraf Dewan (2013). *Floods in a megacity: geospatial techniques in assessing hazards, risk and vulnerability*. Springer.

Eduardo Amaral Haddad and Eliane Teixeira (2015). Economic impacts of natural disasters in megacities: The case of floods in são paulo, brazil. *Habitat International*, 45:106–113.

Khabat Khosravi, Binh Thai Pham, Kamran Chapi, Ataollah Shirzadi, Himan Shahabi, Inge Revhaug, Indra Prakash, and Dieu Tien Bui (2018). A comparative assessment of decision trees algorithms for flash flood susceptibility modeling at haraz watershed, northern iran. *Science of the Total Environment*, 627:744–755.

José A MARENGO (2009). Mudanças climáticas, condições meteorológicas extremas e eventos climáticos no brasil. *Fundação Brasileira para o Desenvolvimento Sustentável (FBDS). Mudanças climáticas e eventos extremos no Brasil. Disponível em: http://www. fbds. org. br/fbds/IMG/pdf/doc-504. pdf. Acesso em*, 24.

Jorge Pimentel, Thiago Dutra, Rafael Silva Ribeiro, Pedro Augusto dos Santos Pfaltzgraff, Maria Emília Radomski Brenny, Dario Peixoto, Diogo Rodrigues da Silva, Hideyuki Iwanami, and Tomohiro Nishimura (2020). Risk assessment and hazard mapping technique in the project for strengthening national strategy of integrated natural disaster risk management. *International Journal of Erosion Control Engineering*, 13(1):35–47.

Hossein Shafizadeh-Moghadam, Roozbeh Valavi, Himan Shahabi, Kamran Chapi, and Ataollah Shirzadi (2018). Novel forecasting approaches using combination of machine learning and statistical models for flood susceptibility mapping. *Journal of environmental management*, 217:1–11.

Érico Soriano, Luciana de Resende Londe, Leandro Torres Di Gregorio, Marcos Pellegrini Coutinho, and Leonardo Bacellar Lima Santos (2016). Water crisis in são paulo evaluated under the disaster's point of view. *Ambiente & Sociedade*, 19(1):21–42.

Jose Candido Stevaux, Edgardo M Latrubesse, Maria Lucia de P Hermann, and Samia Aquino (2009). Floods in urban areas of brazil. *Developments in Earth Surface Processes*, 13:245–266.

Mahyat Shafapour Tehrany, Biswajeet Pradhan, and Mustafa Neamah Jebur (2013). Spatial prediction of flood susceptible areas using rule based decision tree (dt) and a novel ensemble bivariate and multivariate statistical models in gis. *Journal of Hydrology*, 504:69–79.

E Toth, A Brath, and A Montanari (2000). Comparison of short-term rainfall prediction models for real-time flood forecasting. *Journal of hydrology*, 239(1-4):132–147.

Zhaoli Wang, Chengguang Lai, Xiaohong Chen, Bing Yang, Shiwei Zhao, and Xiaoyan Bai (2015). Flood hazard risk assessment model based on random forest. *Journal of Hydrology*, 527:1130–1141.

Jiaqiang Xie, Hao Chen, Zhenliang Liao, Xianyong Gu, Dajian Zhu, and Jin Zhang (2017). An integrated assessment of urban flooding mitigation strategies for robust decision making. *Environmental Modelling & Software*, 95:143–155.