ESTIMATING EXTREME RAINFALL EQUATION PARAMETER IN SOUTHEAST BRAZIL USING MACHINE LEARNING

Alex Tavares Silva 1
Jader Lugon Junior 2
Wagner Rambaldi Telles 3

ABSTRACT

Purpose: This article explores the use of advanced machine learning techniques, including Random Forests and Deep Learning, to predict parameters of the intense rainfall equation.

Methods: The study applies deep neural networks and random forests to predict parameters of the intense rainfall equation. Random Forests method is employed to handle the heterogeneity of data, while Deep Learning captures non-linear relationships. The application takes place in the state of Rio de Janeiro, with a focus on predicting parameters for specific municipalities using available data from the ANA (Brazilian Water Agency).

Results and Conclusion: The neural network demonstrates accuracy in learning these parameters, with discrepancies attributed to differences in historical data periods. Despite limitations, the neural network shows promise in predictions, while the Random Forest closely aligns with the results of the neural network and the Gumbel method. The algorithms perform less accurately in regions with limited training data, emphasizing the need for additional variables to enhance prediction accuracy.

Research implications: The most significant implication of this research is the potential improvement in intense rainfall forecast, using advanced machine learning techniques such as Random Forests and Deep Learning. Society will benefit with it resulting in better systems, in specific municipalities within the state of Rio de Janeiro, for early warning, risk management, and urban planning.

Keywords: IDF Curves, Random Forests, Deep Learning, Gumbel.

PARÂMETRO ESTIMANDO EQUAÇÃO DE CHUVA EXTREMA NO SUDESTE BRASILEIRO USANDO APRENDIZAGEM DE MÁQUINA

RESUMO

Objetivo: Este artigo explora o uso de técnicas avançadas de aprendizado de máquina, incluindo Florestas Aleatórias e Aprendizado Profundo, para prever parâmetros da equação de chuvas intensas.


Resultados e Conclusão: A rede neural demonstra precisão na aprendizagem desses parâmetros, com discrepâncias atribuídas a diferenças nos períodos de dados históricos. Apesar das limitações, a rede neural mostra promessa em previsões, enquanto a Floresta Aleatória se alinha de perto com os resultados da rede neural e do

1 Instituto Federal Fluminense (IFF), Itaperuna, Rio de Janeiro, Brasil. E-mail: altasilva@gmail.com
Orcid: https://orcid.org/0000-0001-7829-4203
2 Instituto Federal Fluminense (IFF), Macaé, Rio de Janeiro, Brasil. E-mail:jljunior@iff.edu.br
Orcid: https://orcid.org/0000-0001-8030-0713
3 Universidade Federal Fluminense (UFF), Santo Antônio de Pádua, Rio de Janeiro, Brasil.
E-mail:wtelles@id.uff.br Orcid: https://orcid.org/0000-0002-6032-3405
método Gumbel. Os algoritmos têm desempenho menos preciso em regiões com dados de treinamento limitados, destacando a necessidade de variáveis adicionais para aprimorar a precisão das previsões.

**Implicações da investigação:** Uma implicação significativa da pesquisa é a melhoria potencial na previsão de eventos climáticos extremos, como chuvas intensas, utilizando técnicas avançadas de aprendizado de máquina, como Random Forests e Deep Learning. Isso pode ter implicações práticas relevantes para sistemas de alerta precoce, gestão de riscos e planejamento urbano em municípios específicos do estado do Rio de Janeiro.

**Originalidade/Valor:** Uma implicação mais significativa desta pesquisa é o potencial aprimoramento na previsão de chuvas intensas, por meio do uso de técnicas avançadas de aprendizado de máquina, como Florestas Aleatórias e Aprendizado Profundo. A sociedade se beneficiará com isso, resultando em sistemas melhores, especialmente em municípios específicos do estado do Rio de Janeiro, para alertas precoces, gestão de riscos e planejamento urbano.

**Palavras-chave:** Curvas IDF, Random Forests, Deep Learning, Gumbel.

RGSA adota a Licença de Atribuição CC BY do Creative Commons (https://creativecommons.org/licenses/by/4.0/).

1 INTRODUCTION

Hydrological planning is fundamental for the development of countries, especially for nations that depend on agriculture, making it necessary to future problems. According to Darji et al. (2015), accurate forecasting is very important for agriculture dependent countries such as India. Rainfall prediction is important for analyzing the crop productivity, and also to improve the management of water resources.

Sales et al. (2021) conducted a hydrological study using the WRF software for rainfall forecast for the Paraiba do Sul watershed in Rio de Janeiro, Brazil. This method can be applied as the design criterion for heavy rain conditions in engineering projects such as dams, dikes and bridges or for assessing extreme environmental events.

With regard to rainfall, better planning requires advanced knowledge of what will happen. Obtaining past data is essential for attempting any form of prediction. As per Burian and Durrans (2000), the main problem with continuous hydrologic simulation is the management of these massive data sets and the need for long-term meteorological records, in the case precipitation includes.

Rainfall prediction models offer insights into the impact of various climatological variables on precipitation levels. Lately, Deep Learning has facilitated the autonomous labeling of data, enabling the creation of data-driven models for time series datasets (Aswin, 2018).

Several Machine Learning models are applied to various environmental issues, which predict rainfall. This is a challenging task due to the high volatility and complex nature of
atmospheric data. It is considered the most challenging task because its depends on various factors (HUSSAIN, ZOREMSANGA, 2021; KALA, VAIYANATHAN, 2022).

Predicting parameters of the rainfall equation in regions with scarcity of data can directly help in agriculture, since it is possible to predict periods with higher and lower rainfall. Hudnurkar & Rayavaru (2022) state that many researchers have used machine learning algorithms such as k-nearest neighbor (k-NN), support vector machine (SVM), artificial neural network (ANN), decision tree (DT) and random forest (RF) for the purpose of rainfall classification.

The advance of computer power turned Artificial Intelligence research one of the main topic of interest in scientific and engineering applications. It currently encompasses a wide variety of subfields, generally extending to specific tasks. Artificial intelligence is relevant for virtually any intellectual task (RUSSEL, NORVING, 2013). Machine Learning is a subfield of artificial intelligence that focuses on developing algorithms and models that enable computers to learn and improve from past experiences. Instead of explicitly programming rules or instructions, Machine Learning systems use data to train models and with them one can test different scenarios to help the decision making process. They can be used for predicting outcomes, classifying information, identifying patterns, and more. There are different types of Machine Learning algorithms, including supervised, unsupervised, and reinforcement learning.

At its core, a neural network is a computational model inspired by the way the brain performs specific tasks or functions of interest (HAYKIN, 2008). Deep Neural Networks, on the other hand, are a specific technique within Machine Learning that draws inspiration from the structure of the human brain. They consist of layers of interconnected artificial neurons that process information hierarchically. Deep Neural Networks are called "deep" because they have multiple hidden layers between the input layer and the output layer. They are used for complex tasks such as image recognition, natural language processing, automatic translation, and more.

Deep Neural Networks, or Deep Learning, have been particularly effective in handling complex data and extracting important features for decision-making. They are the foundation of many recent advancements in areas such as computer vision and natural language processing. Deep Neural Networks are often referred to as "deep networks" or "deep learning" due to their ability to handle tasks that involve many layers of information processing.

Random Forest (RF) is a technique used in modelling predictions and behavior analysis and is built on decision trees. Lee et al. (2022) compared several models of Machine Learning to estimate of rainfall erosivity factor in Italy and Switzerland and concluded the results RF model had the highest performance.

Deep Neural Networks, or Deep Learning, have been particularly effective in handling complex data and extracting important features for decision-making. They are the foundation of many recent advancements in areas such as computer vision and natural language processing.
The goal of this paper is to present the study conducted in the Southeast region of Brazil, using two techniques: Deep Learning (DP) and Random Forests (RF). Both techniques were used to predict the parameters used in the equation for extreme rainfall, also known as Intensity, Duration and Frequency (IDF) equation.

2 METHODS

2.1 STUDY SITE

The research took place in the state of Rio de Janeiro, in the southeastern region of Brazil, and involved most than four hundred locations. In that state, there were 16,055,174 people according to the Instituto Brasileiro de Geografia (IBGE) in 2022. The state has an area of 43,750 Km², and its population density is 366.97 inhabitants per square kilometer. IBGE is a government agency in Brazil. There were 447 pluviometric stations in the state for training and practicing the neural network. Its parameters were estimated for the following cities: Angra dos Reis, Itaperuna, Resende, and Volta Redonda. These municipalities were chosen due to differences in their altitudes, respectively (130m, 110m, 440m, 360m). The Figure (1) shows the study area.

Figure 1
Region of study (the authors, 2024).
Figure (2) displays the Terrain Digital Elevation Model of the study area. It is evident that the municipalities of Resende and Volta Redonda have higher elevations compared to Itaperuna. Therefore, Volta Redonda features a mixed topography.

2.2 EXTREME RAINFALL EQUATION

Oliveira (2019) conducted a survey of all intensity-duration-frequency relationships of extreme rainfall published in Brazil estimating parameters from 5209 rainfall pluviometric stations containing more than 15 years of observations.

For this study, satellite image data from historical series were used, where the ANA (National Water Agency) stations are located.

The calculated parameters are dimensionless parameters of the general rainfall equation (TUCCI et al., 2015; OLIVEIRA, 2019), according to equation (1).

**Figure 2**

*Digital Elevation Model (the authors, 2024)*
\[ i = \frac{KTR^a}{(t+b)^c} \]  \hspace{1cm} (1)

where:

\( i \) = standard deviation (mm);
\( K, a, b, c \) = dimensionless parameters that depend on the region (mm);
\( TR \) = recurrence time (mm);
\( t \) = duration (min).

\subsection*{2.3 GUMBEL DISTRIBUTION AND KOLMOGOROV-SMIRNOV}

The Gumbel distribution is widely utilized in the field of risk analysis, hydrological engineering, extreme value analysis, and other applications involving the modeling of rare or extreme events. It provides a valuable statistical tool for estimating the probability of occurrence of extreme events and quantifying risks associated with these events. The Gumbel distribution is considered the most widely used extreme value distribution in the frequency analysis of hydrological variables and offers advantages over others by not requiring reference to probability tables. Instead, it only requires the calculation of the mean and standard deviation of annual maximum daily precipitation values (PEREIRA, DUARTE, SARMENTO, 2017).

In order to verify whether the gumbel distribution behaves correctly and is consistent with annual maximum precipitation values, the fitting of data to the statistical distribution needs to be evaluated in terms of quality (SILVA et al., 2021). Among the goodness-of-fit tests available in the literature, this study adopts the Kolmogorov-Smirnov goodness-of-fit test, as it is widely used in this type of analysis.

The fitting to the Gumbel distribution was performed according to Equation (2) (Choi, Choi, 1999), which utilizes the statistical parameters of mean and standard deviation of annual maximum precipitation values to determine the maximum height of precipitation for a 1-day duration corresponding to different return periods (Pereira, Duarte, Sarmento, 2017).

\[ x = \bar{x} - s \left( 0.45 + 0.7977ln\left( \frac{ln\left( \frac{Tr}{Tr-1} \right)}{ln\left( \frac{Tr}{Tr-1} \right)} \right) \right) \]  \hspace{1cm} (2)

\( x \) is the adjusted maximum precipitation (mm);
\( \bar{x} \) is the mean of collected maximum precipitation values (mm);
\( s \) is the standard deviation of annual maximum precipitations (mm);
\( Tr \) is the return period (years).
The Kolmogorov-Smirnov goodness-of-fit test examined values for significance levels of 1%.

For further information on gumbel and the statistical method for rainfall validation smirnov-kolmogorov, see (SILVA et al., 2021).

2.4 DEEP NEURAL NETWORK

According HAYKIN (1999) neural networks have applications in a wide range of fields, including modeling, time series analysis, pattern recognition, signal processing, and control, thanks to a fundamental feature: the ability to learn from input data, whether supervised or unsupervised.

A deep neural network refers to a type of artificial neural network (ANN) that has multiple hidden layers between the input and output layers. These hidden layers enable the network to learn and represent complex patterns and features in data, making it suitable for various machine learning tasks, including image and speech recognition, natural language processing, and more. Deep neural networks are a fundamental component of DP, a subfield of machine learning that focuses on training and using neural networks with many hidden layers to solve intricate problems. In this study, the neural network structure used is as shown in the Figure (3).

The neural network was developed using six layers. All the layers used the activation function Rectified Linear Unit (ReLU), except the last one, which used the Linear activation function. The first layer used four neurons, the second, third, and fourth used seven neurons. The fifth layer used five neurons, and the last layer used four neurons.

It was also configured using the Adam optimizer. The Adam optimizer is a frequently used optimization algorithm in neural network training. It combines the benefits of the RMSprop (Root Mean Square Propagation) optimizer and Stochastic Gradient Descent (SGD). The Adam optimizer adaptively adjusts the learning rates for the network's weights, making it effective in achieving fast convergence during training. It is widely used due to its solid performance across a variety of DP problems.

The MSE loss is a function that measures the average squared difference between the predicted outputs from the neural network and the actual labels in the training data. It's often used as a loss function in regression problems, where the goal is to predict numerical values, such as prices, scores, etc. The lower the value of the MSE loss, the better the neural network
is fitting the training data. The training process involves minimizing this loss so that the network makes more accurate predictions.

The network was trained using parameters obtained through the Luus-Jaakola optimization method after the analysis of satellite imagery in the Rainfall Estimates from Rain Gauge and Satellite Observations (CHIRPS) project. Since 1999, scientists from the United States Geological Survey (USGS) and the Climate Hazard Center (CHC), supported by funding from the United States Agency for International Development (USAID), the National Aeronautics and Space Administration (NASA), and the National Oceanic and Atmospheric Administration (NOAA), have developed techniques to produce rainfall maps, particularly in areas where surface data is scarce.

Activation functions play a fundamental role in neural networks, as they are responsible for introducing non-linearity into the outputs of the network's layers. They determine how neurons respond to weighted inputs, allowing networks to learn to represent complex relationships in data.

The ReLU activation function is one of the most widely used activation functions in neural networks. It is defined as Equation(3).

\[ \text{ReLU}(x) = \max(0, x) \quad (3) \]

In DP, epochs represent the number of times the entire training dataset is passed through the neural network. Each epoch consists of a full cycle of training, during which the neural network adjusts its weights based on the training data. The goal is to train the network for multiple epochs so that it learns patterns in the data and improves its performance. However, it's essential to strike a balance as training for too many epochs can lead to overfitting, where the network becomes overly specialized on the training data and doesn't generalize well to unseen data. In the neural network used in this work, the number of epochs was set to 1000.
These are essential concepts in DP and are commonly used in configuring and training neural networks to solve a variety of tasks, from regression to classification. Each one plays a crucial role in network training and performance.

2.5 RANDOM FORESTS

The RF is a machine learning algorithm known for its effectiveness and versatility in regression and classification problems (LIAW, WIENER, 2002). In the case of regression, a Random Forest model is trained to predict a numerical value as output. The decision trees in the Random Forest are adjusted to estimate a continuous value as output, and the final output of the Random Forest is obtained as an average (or median) of the outputs from all individual trees.

The algorithm used is Random Forest, a commonly used machine learning algorithm that combines the output of multiple decision trees to achieve a single result. As stated by Hijiao et al. (2022), the basic idea of the RF model is to extract repeatedly and randomly K samples from the training dataset with replacement according to the bootstrap resampling method.

The Random Forest is a powerful and versatile technique that can be applied to a wide range of machine learning problems, including regression, depending on the type of output one aims to predict. One of the main advantages of the random forest is its ability to handle high-
dimensional data and address overfitting, a common issue in many machine learning models (BREIMAN, 2001).

Furthermore, the random forest is also capable of handling imbalanced data, making it a popular choice in classification problems where the output classes have unequal distributions (CHEN, LIAW, BREIMAN, 2004). As the Random Forest is a machine learning technique that allows an algorithm to train on data and make predictions based on patterns identified in the training data, it is considered a form of artificial intelligence.

Another way to obtain parameters is by interpolating them through a powerful technique to estimate values in locations without data samples, using the weighted proximity of known data to fill in the gaps.

2.6 TECHNICAL INFORMATION

For this work, the Python programming language and the Keras library were used. Python is a high-level programming language widely used in data science, machine learning, web development, automation, and many other fields. Python is known for its clear and readable syntax, making it a popular choice for programmers of all skill levels. Python's extensive standard library and active community contribute additional packages and modules, making it a versatile language for a wide range of applications.

Keras, on the other hand, is an open-source DP library written in Python. It is designed to be simple and user-friendly, allowing developers to build neural networks quickly and efficiently. Keras is highly modular and can run on various back-end frameworks such as TensorFlow, Theano, and CNTK. Its high-level interface simplifies the creation of neural networks, and it supports a wide range of neural network types, making it a preferred choice for rapid prototyping of DP models. It enjoys extensive documentation and a vibrant user community, making it a powerful tool for various machine learning tasks.

QGIS is an essential tool for professionals working with geospatial data. As open-source software, it offers an accessible and powerful solution for visualizing, analyzing, and editing geographic information. With support for various data formats and a wide range of analysis tools, QGIS allows users to perform a variety of tasks, from creating simple maps to complex research and planning projects. Its user-friendly interface and an active community of users and developers contribute to making it a popular and versatile choice in the world of Geographic Information Systems.
3 RESULTS AND DISCUSSIONS

Predicting parameters of the intense rainfall equation is not an easy task. Using DP and RF to assist in predicting them is an effective and reliable approach.

The training and testing phase of data is of paramount importance in the development and evaluation of neural networks and, indeed, in many types of machine learning models. During the training phase, the neural network learns to map patterns in input data to desired outputs by adjusting its internal weights and parameters to minimize the loss function. Subsequently, the network is evaluated on an independent test dataset to assess its ability to generalize what it learned during training to unseen data. This evaluation helps identify overfitting, where the network fits too closely to the training data and fails to generalize effectively. Additionally, it allows for the fine-tuning of hyper parameters and the validation of the model's performance against task-specific metrics, ensuring that it meets project requirements and can perform effectively in real-world challenges.

Figure 3

Training and validation of parameters.

The training and validation graphs are illustrated in Figure (4). It can be observed that after approximately 400 epochs, the data was adjusted to prevent overfitting and underfitting. In Table (1), the parameters obtained by the neural network are presented, while in Table (2), the results from the RF are provided. Despite the parameters of the intense rainfall equation not
being everlasting, as they are based on past precipitations, it is the most up-to-date literature available.

### Tabela 1

*Parameters obtained through the neural network prediction*

<table>
<thead>
<tr>
<th>City</th>
<th>K</th>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angra dos Reis</td>
<td>887.4977</td>
<td>0.1516</td>
<td>11.8754</td>
<td>0.9118</td>
</tr>
<tr>
<td>Itaperuna</td>
<td>886.6353</td>
<td>0.1821</td>
<td>11.8918</td>
<td>0.8554</td>
</tr>
<tr>
<td>Resende</td>
<td>896.2754</td>
<td>0.1927</td>
<td>11.9300</td>
<td>0.8311</td>
</tr>
<tr>
<td>Volta Redonda</td>
<td>898.9662</td>
<td>0.2105</td>
<td>11.9500</td>
<td>0.7965</td>
</tr>
</tbody>
</table>

### Tabela 2

*Parameters obtained through the Random Forests prediction.*

<table>
<thead>
<tr>
<th>City</th>
<th>K</th>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angra dos Reis</td>
<td>776.2983</td>
<td>0.2301</td>
<td>12.5007</td>
<td>0.7897</td>
</tr>
<tr>
<td>Itaperuna</td>
<td>726.8961</td>
<td>0.2666</td>
<td>12.9263</td>
<td>0.7999</td>
</tr>
<tr>
<td>Resende</td>
<td>883.7685</td>
<td>0.2464</td>
<td>12.6846</td>
<td>0.7934</td>
</tr>
<tr>
<td>Volta Redonda</td>
<td>920.0268</td>
<td>0.2583</td>
<td>12.7291</td>
<td>0.7956</td>
</tr>
</tbody>
</table>

In Figures (5 - 8), the parameters estimated by the deep artificial neural network and those found by the RF were applied to Equation (1). For the purpose of comparison, parameters were calculated using the traditional method (Gumbel) and analyzed through IDF curves in the municipalities. The IDF (Intensity-Duration-Frequency) curves provide valuable insights into the relationship between rainfall intensity, duration, and frequency, allowing for a comprehensive comparison of the Gumbel method's results in both locations. This comparative analysis in municipalities with extensive temporal data aims to enhance our understanding of how different methods, such as Gumbel, perform in estimating and characterizing extreme rainfall events in these specific regions. Such comparisons contribute to the robustness and reliability of hydrological assessments and design considerations in the context of water resource management and infrastructure planning.
Figura 4
Analysis of IDF curves in the municipality of Itaperuna and of machine learning algorithms.

![IDF Curves Itaperuna - TR 5 years](image)

Figura 5
Analysis of IDF curves in the municipality of Volta Redonda and of machine learning algorithms.

![IDF Curves Volta Redonda - TR 5 years](image)
Comparing the results obtained from machine learning techniques with the Gumbel method is of significant importance in various applications, particularly in the fields of hydrology and risk analysis. The Gumbel method is widely employed for modeling rare events, such as climatic extremes, and estimating the probability of their occurrence. This traditional approach has been grounded in established statistical principles over time.

**Figura 6**

*Analysis of IDF curves in the municipality of Angra dos Reis and of machine learning algorithms.*

The introduction of machine learning techniques, such as RF or Neural Networks, into hydrological modeling provides an alternative and more flexible perspective. Comparing the results of these approaches with the Gumbel method allows for the assessment of the algorithms' capacity to handle complex patterns in the data and determines if they efficiently capture extreme events.
Figura 7

Analysis of IDF curves in the municipality of Angra dos Reis and of machine learning algorithms.

This comparison not only shows the effectiveness of machine learning techniques in specific contexts but also highlights situations where these innovative approaches offer additional benefits, such as increased accuracy under certain conditions or the ability to handle more complex data. Ultimately, this comparative analysis contributes to the ongoing advancement in the selection and application of more robust and accurate methods in modeling extreme events and managing associated risks.

4 CONCLUSIONS

In this study, a deep neural network and RF were applied to predict parameters of the intense rainfall equation. The neural network demonstrated the ability to learn these parameters with good accuracy. Discrepancies between the parameters from the literature and those obtained through the neural network may arise due to differences in the time periods of historical series used in the literature and the period covered by satellite imagery. Although, the literature data is considered the most up-to-date for comparison purposes.
This justification is supported by the calibration of the network during the training and validation phases, indicating that the network did not exhibit overfitting or underfitting. In summary, the neural network yielded promising results, demonstrating its proficiency in learning from input data and accurately predicting the output parameters.

Nevertheless, adding more data for training the neural network, such as altitude, type of vegetation, and other variables, can help improve the accuracy of predictions, so it is important to highlight that the neural network shown in this paper makes predictions with only a few input variables. For this reason, it performs well in predictions, considering this limitation.

Random Forest predicted parameters close to those of the Neural Network and those obtained by the Gumbel method. This suggests that the Random Forest algorithm performed comparably to both the Neural Network and the traditional Gumbel method in predicting parameters. The ability of Random Forest to approximate results from different methodologies underscores its versatility and effectiveness in handling diverse datasets and modeling scenarios. Such consistency in predicting parameters across different algorithms enhances confidence in the reliability of the Random Forest approach, showcasing its applicability in a range of predictive tasks within the context of the given study or analysis.

In regions with limited training data for the networks, such as Angra dos Reis, Resende, and Volta Redonda, the algorithms did not estimate intensities as accurately. There were also periods when ANA did not determine the amount of rainfall at these stations, which can lead to a less precise curve.

The techniques employed were able to accurately predict the forecasted values for rainfall lasting 24 hours. However, shorter-duration rains were not well estimated and should be the subject of further studies.

ACKNOWLEDGEMENTS

The authors acknowledge the financial support provided by CAPES, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Finance Code 001), CNPq, Conselho Nacional de Desenvolvimento Científico e Tecnológico, and FAPERJ, Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (Proc. E-26/210.220/2021).

REFERENCES

Estimating Extreme Rainfall Equation Parameter in Southeast Brazil Using Machine Learning


