Article

A Random Forest Model to Estimating Precipitation Return Periods for Improved Water Management: A Comparative Analysis with Probability Density Functions

Johan Anco-Valdivia , Sebastián Valencia-Félix , Alain Jorge Espinoza Vigil , Guido Anco , Julian Booker * , Julio Juarez-Quispe , Erick Rojas-Chura

Posted Date: 19 November 2024

doi: 10.20944/preprints202411.1477.v1

Keywords: probability distributions; mathematical methods; return period; random forest; algorithms; annual maximum rainfall; artificial intelligence

Article

A Random Forest Model to Estimating Precipitation Return Periods for Improved Water Management: A Comparative Analysis with Probability Density Functions

Johan Anco-Valdivia [1], Sebastián Valencia-Félix [1], Alain Jorge Espinoza Vigil [1], Guido Anco [2], Julian Booker [3,*], Julio Juarez-Quispe [1] and Erick Rojas-Chura [1]

[1] School of Civil Engineering, Universidad Católica de Santa María, Arequipa 04013, Peru
[2] School of Systems and Informatics Engineering, Universidad Nacional Mayor de San Marcos, Lima 15081, Peru
[3] School of Electric, Electronic and Mechanical Engineering, University of Bristol, Bristol, UK
* Correspondence: j.d.booker@bristol.ac.uk

Abstract: Precipitation within specific return periods plays a crucial role in the design of hydraulic infrastructure for water management. Traditional analytical approaches involve collecting annual maximum precipitation data from a station followed by the application of statistical probability distributions, and selecting the best-fit distribution based on goodness-of-fit tests (e.g., Kolmogorov-Smirnov). However, this methodology relies on current data, raising concerns about its suitability for outdated data. This study aims to compare Probability Density Functions (PDFs) with the Random Forest (RF) machine learning algorithm for estimating precipitation at different return periods. Using data from five stations located in various parts of the Arequipa province in Peru, it was evaluated the performance of both methods using the Root Mean Square Error (RMSE) metric. The results show that RF outperforms PDFs in most cases, yielding lower RMSE values for precipitation estimates at return periods of 2, 5, 10, 20, 50, and 100 years for the studied stations.

Keywords: probability distributions; mathematical methods; return period; random forest; algorithms; annual maximum rainfall; artificial intelligence

1. Introduction

Floods are one of the most frequent and destructive types of disasters, impacting communities worldwide. They are responsible for up to 40% of global natural disasters, causing almost half of all natural hazard-related fatalities, and exhibiting a worrying increase [1]. As a result, the necessity for sustainable urban planning that integrates demand-driven hydraulic infrastructure becomes paramount.

In this way, according to Bertilsson, *et al.* [3], climate change and rapid urbanization present significant obstacles to sustainable urban planning, while the proliferation of megacities exacerbates flood risk by increasing impervious surfaces and altering hydrological pathways and increased urbanization and population concentration increase exposure to natural hazards [2]. It is important to consider that, a effective flood risk management necessitates a well-informed and accurate assessment of flood hazards, acknowledging the likelihood of increased extreme flood events due to global warming [1]. Making accurate estimations of intense precipitation is key to improving early warnings and protecting the population, therefore the analysis of data from climate stations is of great relevance, as noted by Padji, *et al.* [4].

Furthermore, this analysis is essential in areas such as planning infrastructure resilient to extreme conditions and efficiently managing water resources. In this sense, an effective approach to improving flood planning is based on the ability to anticipate extremes of rainfall [5].

The theory of statistical extremes shows that the frequency of extreme events is more closely tied to changes in climate variability than to fluctuations in the mean climate state [6]. Additionally, the process of distribution fitting involves matching a statistical distribution to a dataset derived from random processes and is a critical step in capturing the underlying patterns. Therefore, since probability distributions are essential for quantifying uncertainty, selecting the wrong distribution can lead to flawed conclusions [7]. There are numerous studies that highlight the significance of identifying the distribution that best fits the data from a meteorological station, as demonstrated by Mandal and Choudhury [8] that investigated the annual, seasonal and monthly maximum daily rainfall patterns in Sagar Island, situated on the continental shelf of the Bay of Bengal. The study revealed that the normal distribution provided the best fit for annual, post-monsoon and summer seasons.

In Asia, Zhai, *et al.* [10] applied methodologies for analyzing rainfall predictions, focusing on probability distribution models such as Pearson Type III, Pareto-Burr-Feller, generalized extreme value (GEV), and Weibull. These models were calibrated with historical data to predict extreme rainfall. Additionally, they implemented the Mann-Kendall test to detect long-term trends in precipitation. These techniques help improve the prediction of the recurrence of extreme events and assess long-term trends in the frequency and magnitude of rainfall, providing a solid foundation for designing drainage infrastructure and flood management systems. Moreover, Si, *et al.* [11] addressed the issue by constructing regional meteorological stations, enabling a more in-depth study of extreme rainfall, even with short data series. They employed sampling methods based on peaks over threshold and the generalized Pareto distribution to optimize spatial interpolation parameters, improving the accuracy of daily extreme rainfall predictions at regional stations.

As an alternative to the mathematical methods described previously, the integration of Big Data and Artificial Intelligence has enabled Machine Learning (ML) to emerge as a promising tool in weather forecasting, leveraging its strengths in addressing nonlinear complexities and uncovering previously unknown relationships within the Earth’s climate system [12]. In this sense, this method could be useful to determine the return period, which is an essential metric that quantifies the probability of extreme events such as floods and droughts, which can inflict significant harm on society and the environment. This probabilistic concept is widely utilized in hydrological studies and has garnered increased attention due to the necessity of effectively managing complex processes in an evolving environmental landscape [13]. In the realm of water resources design and management, return period analysis serves as a crucial tool for risk assessment and communication [14].

However, access to updated data is not always feasible due to various hindrances, including the scarcity of meteorological stations in the region, inadequate maintenance of measurement equipment, and restrictions on access to up-to-date data owing to privacy policy constraints. For instance, in the United Arab Emirates, Branch, *et al.* [10] highlighted the need to improve predictions of extreme events, particularly in arid regions. They employed high-resolution solutions validated with data from meteorological stations, despite the limitations posed by the scarce information provided.

Therefore, this paper presents an innovative approach for regions facing similar data availability issues, which is based on utilizing the Random Forest (RF) algorithm to estimate precipitation for a specified return period instead of relying on probability density functions (PDFs). For instance, Sun, *et al.* [15] emphasize that, compared to traditional methods, RF is effective in capturing nonlinear relationships between precipitation and predictive variables, such as the Normalized Difference Vegetation Index (NDVI), Land Surface Temperature (LST), and topographical features.

Furthermore, Papacharalampous, *et al.* [16] focus on the application of RF through quantile regression, conducting a large-scale comparison among various algorithms to identify the most effective one. This suggests that RF can significantly enhance the accuracy of precipitation predictions. Finally, Hassan, *et al.* [17] highlight the necessity of integrating models like RF to improve precipitation forecasting by leveraging significant patterns and relevant attributes to optimize model performance. They underscore the importance of developing hybrid classifiers and addressing limitations such as reliance on historical data and regional variability, thereby promoting a more automated and advanced approach to meteorological analysis.

This study is structured as follows: The Data and Study Area section provides an overview of the region and data used. The Methodology section describes the PDFs employed and the RF algorithm, including the comparison metric. The Results section presents precipitation estimates for various return periods using both methods, along with a comparative analysis. The Discussion section elaborates on the differences between the two methods, explaining their performance. Lastly, the Conclusion section summarizes the key findings and suggests avenues for future research.

2. Data and Study Area

Arequipa province, located in southern Peru, is a diverse region nestled in the Andean highlands. Bordering the Pacific Ocean to the West and the Andes mountains to the East. The climate is characterized by mild temperatures and low humidity throughout the year. It is important to recognize that Arequipa is a thriving economic hub, driven by mining, agriculture, tourism and manufacturing. The region is rich in mineral resources, including copper, zinc and gold. Agriculture is another significant activity, with potato, corn and wheat crops. Tourism attracts visitors worldwide with its breathtaking natural landscapes, including the Colca Canyon, Misti Volcano and Andagua Valley.

The preceding description provides an overview of some characteristics of the Arequipa province. However, to perform a comparative analysis of methods, it is necessary to have access to the corresponding data. To this end, the data was obtained from Peruvian National Meteorology and Hydrology Service (SENAMHI) for 5 stations distributed across the Arequipa province, as shown in Figure 1. However, these data sets vary in length for each station and are not updated to the last year, with most stations reaching only up to 2014. The specific data lengths for each station are: La Pampilla (83 years), La Joya (49 years), El Frayle (50 years), Las Salinas (51 years) and Chiguata (49 years). The analysis was conducted considering the following years for each station: 1965 to 2013, and the maximum annual precipitation was determined for each year to initiate the subsequent comparison of methods.

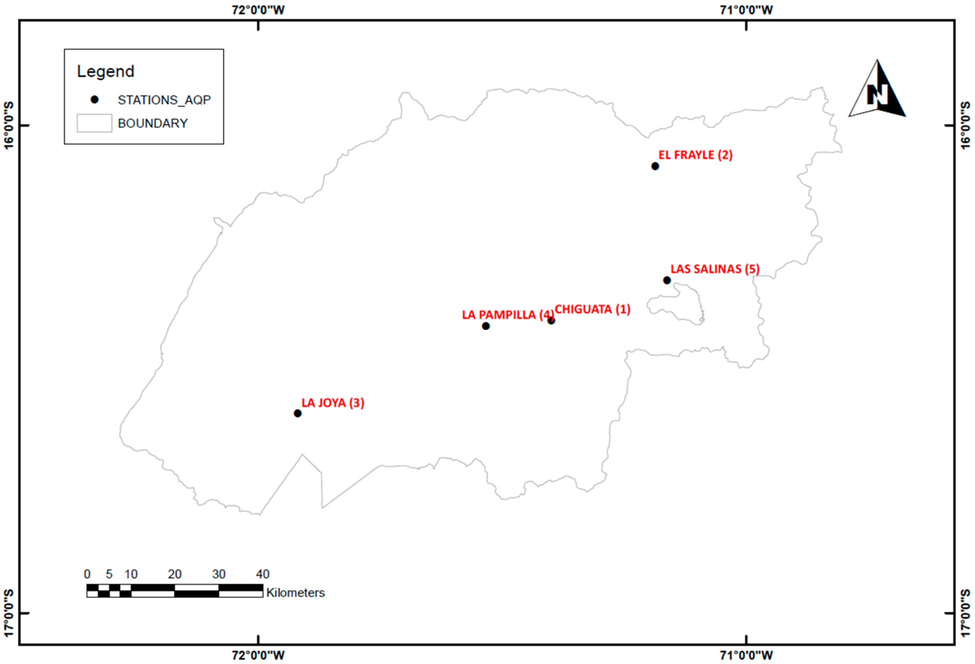

Figure 1. Distribution map of stations in Arequipa province.

3. Methodology

doi:10.20944/preprints202411.1477.v1

When selecting the most appropriate distribution for a specific location, it is crucial to consider the variety of distribution models available. This section outlines the distribution models, the RF algorithm, goodness-of-fit test and the metric used in the study, root mean square error (RMSE).

3.1. Commonly used Probability Distributions

Statistical distributions are essential tools for modeling and analyzing complex phenomena and enable researchers to describe and predict extreme events. In the next lines, the mathematic expressions are found with some applications of these distributions.

3.1.1. Normal

The distribution is a symmetric probability model characterized by its mean and standard deviation parameters. The probability density function (PDF) is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right]$$

for the range of $-\infty < x < \infty$. Annual precipitation and runoff analyses often rely on the Normal distribution [18].

3.1.2. Log-Normal

A continuous probability model suitable for skewed data. Its logarithmic transformation ensures positive values, making it ideal for modeling variables with lower bounds.

$$f(x) = \frac{1}{x\sigma_Y\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma_Y^2}(\ln(x)-\mu_Y)^2\right]$$

where the range of random variable is $x > 0$. The two parameters are expressed as follows:

$$\sigma_Y = \left[\ln\left(1+\frac{\sigma_X^2}{\mu_X^2}\right)\right]^{\frac{1}{2}}$$

$$\mu_Y = \ln(\mu_X) - \frac{1}{2}\sigma_Y^2$$

3.1.3. Pearson Type 3

A versatile, three-parameter model suitable for skewed data, its Gamma distribution-like properties enable the analysis of variables with varying coefficients of variation, making it applicable to precipitation, streamflow and water quality data.

$$f(x) = \frac{1}{|\alpha|\Gamma(\beta)}\left[\left(\frac{x-\xi}{\alpha}\right)^{\beta-1}\right]\exp\left[-\frac{(x-\xi)}{\alpha}\right]$$

Its parameters are:

$$\beta = 4/\gamma^2$$

$$\alpha = \sigma_\gamma/2$$

$$\xi = \mu - 2\sigma/\gamma$$

This distribution is prominently employed in hydrology [5].

3.1.4. Log Pearson Type 3

A flexible probability model that combines logarithmic transformation with the Pearson Type 3 distribution, this combination enables analysis of skewed data with extreme values.

$$f(x) = \frac{1}{|\alpha|x\Gamma(\beta)}\left[\left(\frac{\ln(x)-\xi}{\alpha}\right)^{\beta-1}\right]\exp\left[-\frac{(ln(x)-\xi)}{\alpha}\right]$$

Phien and Ajirajah [19] assessed the suitability of this distribution for modeling flood and maximum rainfall data, as well as its applicability to annual rainfall and streamflow sequences.

3.1.5. Generalized Extreme Value (GEV)

It is a flexible model for extreme events, its cumulative distribution function allows for the analysis of both upper and lower tails, making it suitable for modeling extreme precipitation, flood and drought events.

$$f(x) = \alpha^{-1}\exp[-(1-k)y - \exp(-y)],$$

$$y = -k^{-1}\log\left\{1 - \frac{k(x-\xi)}{\alpha}\right\}, k \neq 0$$

$$y = \frac{x-\xi}{\alpha}, k = 0$$

The GEV distribution is extensively endorsed in European countries for its exceptional ability to accurately model flood data [20].

3.1.6. GEV MIN (L-moments)

A variant of the GEV model, specifically designed for analyzing extreme minima, its inverse cumulative distribution function enables the modeling of low-frequency events, such as droughts and minimum streamflow.

$$f(x) = \exp\left(-[1 + \xi(\frac{x-\mu}{\sigma})^{-\frac{1}{\xi}}\right), \qquad for\ 1 + \xi(\frac{x-\mu}{\sigma}) > 0$$

3.1.7. GEV MAX (kappa specified, L-moments)

A specialized GEV model for extreme maxima, its cumulative distribution function is tailored for analyzing high-frequency events, such as floods, heavy precipitation and maximum streamflow.

$$f(x) = exp\left(-[1 - k(\frac{x-\mu}{\sigma})^{\frac{1}{k}}\right), \qquad for\ 1 - k(\frac{x-\mu}{\sigma}) > 0$$

3.2. Goodness-of-Fit Test

To assess the validity of a specified probability distribution model, a goodness-of-fit test statistics are applied. A plethora of normality tests are available, including Empirical Distribution Function (EDF) tests, which quantify the divergence between empirical and theoretical distributions [21]. Prominent EDF tests include the Kolmogorov-Smirnov (K-S), the Anderson-Darling (A-D) test, and Cramer-Von Mises test [22]. The K-S was applied in this study, due to its simplicity and computational ease of implementation, coupled with the fact that it does not require the estimation of additional parameters.

Furthermore, RMSE was employed to evaluate model performance and identify the best-fitting model.

3.2.1. Kolmogorov-Smirnov (K-S) Test

This test is a non-parametric statistical test used to assess the goodness-of-fit between observed and theoretical distributions. It quantifies the maximum distance between cumulative distribution functions, enabling detection of significant deviations. It is designed to compare the empirical cumulative frequency $S_n(x)$ with the cdf of an assumed theoretical distribution $F_x(x)$. For a sample size n, the values are sorted in a non-decreasing sequence, $X_1 < X_2 < \ldots < X_n$ and the K-S statistic is applied to each data value in the ascending order.

$$S_n(x) = 0;\ if\ X < X_1$$

$$= \frac{k}{n};\ \ if\ X_k \leq X < X_{k+1}$$

$$= 1;\ if\ X > X_n$$

The K-S test statistic is the maximum difference between $S_n(x)$ and $F_x(x)$

$$D_n = \max|F_x(x) - S_n(x)|$$

doi:10.20944/preprints202411.1477.v1

$$P(D_n \leq D_n^{\alpha}) = 1 - \alpha$$

Following the next syntax, the critical value is D_n^{α}, the significance level is α and k is the rank order of the data set.

3.2.2. Root Mean Square Error (RMSE)

The lowest RMSE values signifies the best-fitting model, yielding the standard deviation of the prediction uncertainty. It assesses the difference between actual and estimated values. The RMSE has the following expression:

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(x_i - X)^2}$$

where x_i is the estimated value and X is the actual value.

3.3. Return Period

Probability of occurrence in a given time:

$$P(x \geq x_T) = \frac{1}{T}$$

$$T = \frac{1}{1 - P(x \leq x_T)}$$

Probability of occurrence based on observed data

$$P(i) = \frac{i}{N+1}$$

where i is the position of the observation and n is the number of observations.

3.4. Random Forest

By integrating multiple Decision Trees, RF achieves robust predictions through averaging, exhibiting superiority in handling high-dimensional feature spaces and complex data structures, thereby ensuring reliable performance. Represent a machine learning approach that merges the principles of classification and regression trees with bagging techniques, incorporating an added level of randomness [23]. RF allows for effortless tuning of parameters and requires minimal computational resources [24]. Its effectiveness has been widely established in multiple fields [25], such as Behrens, *et al.* [26] applied it in his analysis of inflation forecasting. It is a supervised learning algorithm that utilizes regression trees as its base learner. Notably, RF can generate multiple trees without pruning, thereby enhancing its predictive capabilities. The training process incorporates two sources of randomness. Firstly, the algorithm employs bootstrap sampling, where each tree is constructed from a random subset of the original data with replacement, allowing for repeated samples. Secondly, RF introduces feature randomness by randomly selecting a subset of candidate predictor variables at each node, and choosing the optimal split based on the most suitable value. The accuracy of the predictions is sensitive to hyperparameter tuning, yet no standardized approach exists for selecting optimal parameters, necessitating a trial-and-error approach [25]. For this study, the sklearn library was utilized to create the RF model.

Supervised Learning

Supervised learning algorithms are designed to derive a function that integrates a group of variables to forecast another variable. The input variables in this function are known as predictor variables, which can also be referred to as independent variables, exogenous variables, covariates, or features. The variable that is being predicted is termed the dependent variable, which may also be called the predictand, response variable, outcome, endogenous variable, target variable, or output.

These algorithms are divided into two main categories based on the nature of the dependent variable: regression and classification. In regression algorithms, the dependent variable is numerical, while in classification algorithms, the dependent variable is categorical [23]. For this study, due to the

doi:10.20944/preprints202411.1477.v1

nature of the case, it is considered a regression problem, attributable to the characteristic of the predictand variable being precipitation for a return period, which is an inherently continuous numerical variable.

4. Results

The principal goal of this study is to determine the differences calculating the precipitation for each return period using the best-fit distribution and the RF algorithm for each station. Knowledge of return periods for extreme events facilitates the evaluation of risk exposure and potential damage from severe weather events, such as floods and intense rainfall, thereby informing policy and decision-making processes [5].

4.1. Best-Fit Distributions for Each Station and Comparison with Random Forest

The traditional approach involves using the annual maximum precipitation corresponding to each station, followed by ordering all this data from highest to lowest to determine the probability of occurrence for each value. This can be graphed to better understand the occurrence probabilities in relation to precipitation levels. At this point, probability density functions (PDFs) come into play, as they are able to fit the data and then evaluate the chosen function for a specific value, which in this case is the return period. However, the choice of the function is not arbitrary; it must undergo a goodness-of-fit test. In this study, the Kolmogorov-Smirnov (K-S) test was selected.

Following this line of thought, the innovative approach incorporates the Random Forest (RF) algorithm, which, like the traditional approach, starts by using the annual maximum precipitation and the probability of occurrence for each value. The difference lies in the fact that RF learns the patterns in the data and can adapt more effectively than traditional statistical functions. It is important to note that the data for this research were handled as follows: one training set (comprising 60% of the data for each station), one test set (comprising 20% of the data for each station), and one validation set (comprising 20% of the data for each station). The 60% allocated for training ensures that RF model has sufficient data to capture complex relationships between variables and adjust its parameters, the allocation of 20% for testing allows for precise evaluation of the trained model’s performance, and finally, the 20% allocated for validation enables an exhaustive evaluation of the model’s hyperparameters and selection of the optimal set. The parameters used for developing the algorithm were: 300 decision trees and a random state of 42. The decision to opt for 300 decision trees was made because a larger number of trees is capable of reducing variance, thus increasing the model's stability. It also provides a reasonable balance between accuracy and computational complexity. Additionally, a random state of 42 was chosen to ensure the reproducibility of the results and to avoid bias in feature selection.

In this way, the model can learn from the respective training set and improve its predictions. The test set was then subjected to the RMSE metric to evaluate accuracy, and the PDFs were similarly evaluated using the same test set that was used for the RF algorithm. The results for each station are shown in Table 1.

Table 1. Statistical results and Best-Fit Distribution for each station with comparison of RMSE for Best-Fit Distribution and Random Forest.

N°	Station Name	Mean	Standard Deviation	Best Fit Statistic Results		RMSE of Best-Fit Distribution	RMSE of RF
				By K-S Test	D max		
1	Chiguata	21.8	11.4	GEV - Min (L-Moments)	0.05815	1.3572	0.9108
2	El Frayle*	23.6	8.9	GEV - Max (k specified, L-Moments)	0.06645	1.1408	1.4561

3	La Joya	1.8	3	Pearson III	0.1000 1	0.3698	0.0779
4	La Pampilla	16.7	18.4	Log Pearson III	0.0417 4	1.204	0.621
5	Las Salinas	19.8	9	GEV - Min (L-Moments)	0.0503	1.7653	1.1432

It should be noted that the El Frayle station has a missing value in its annual maximum precipitation dataset, for which the moving averages method was used to fill in the missing data.

To better illustrate the behavior of the data and how they are distributed, Figures 2–6 show the different graphs of annual maximum precipitation as a function of the probability of occurrence and how the PDFs fit the data in their respective ways.

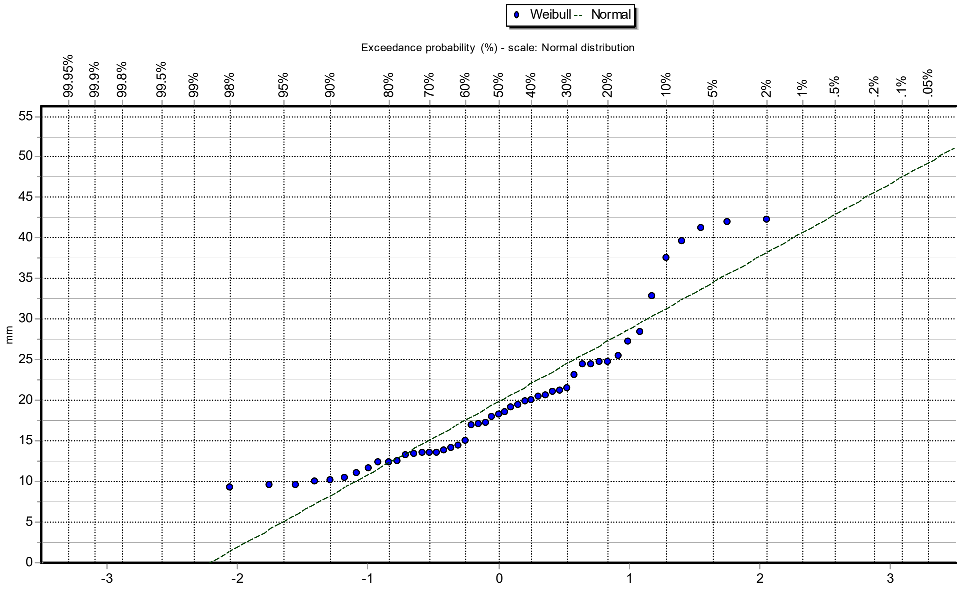

Figure 2. Q-Q plot of Normal distribution at Station: Las Salinas.

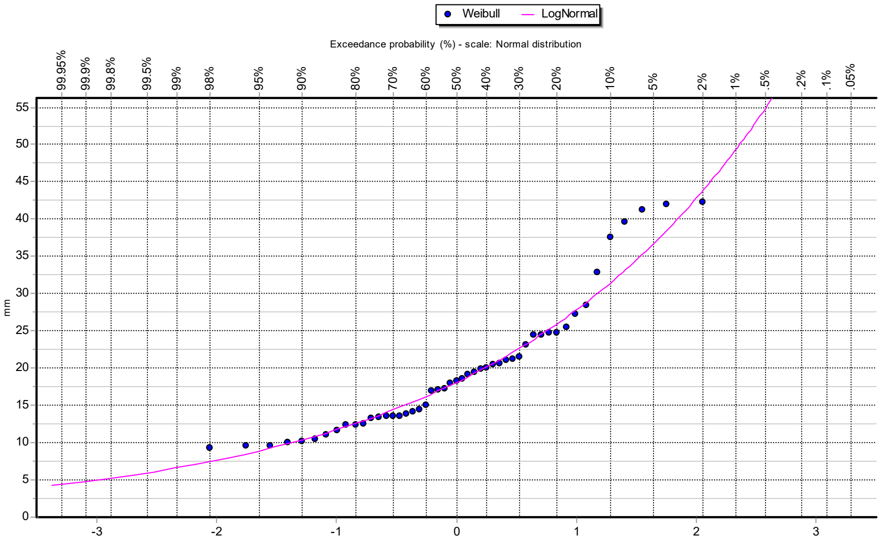

Figure 3. Q-Q plot of Log Normal distribution at Station: Las Salinas.

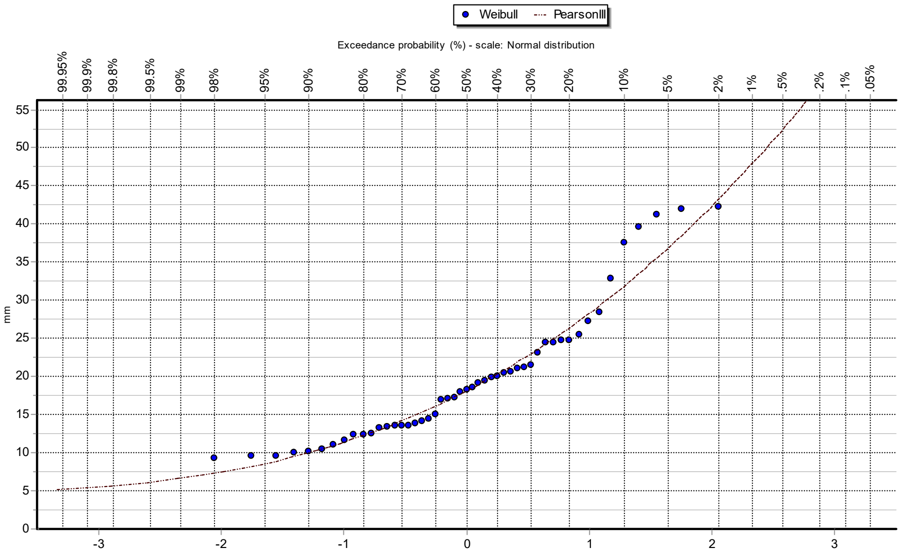

Figure 4. Q-Q plot of Pearson III distribution at Station: Las Salinas.

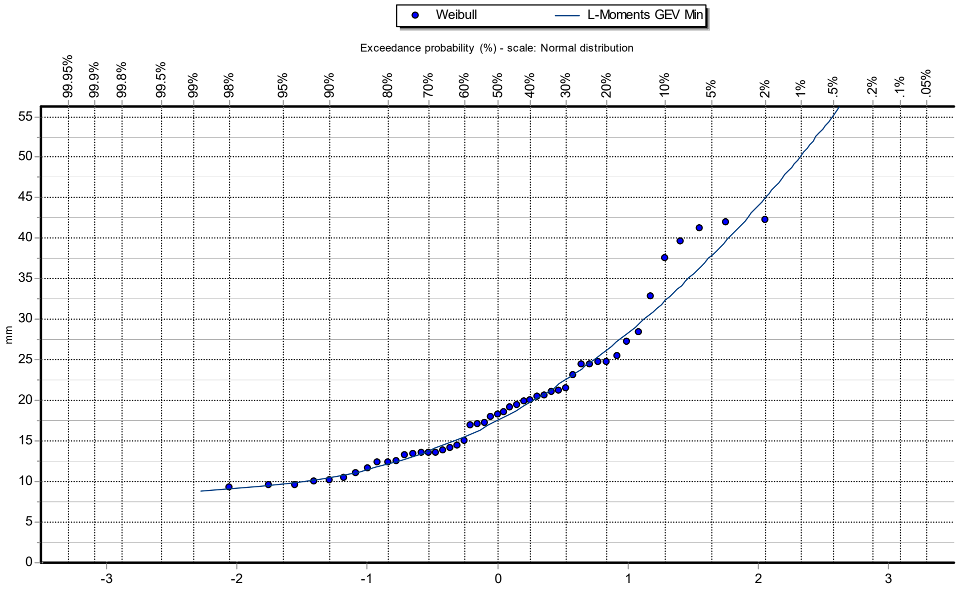

Figure 5. Q-Q plot of GEV Min distribution at Station: Las Salinas.

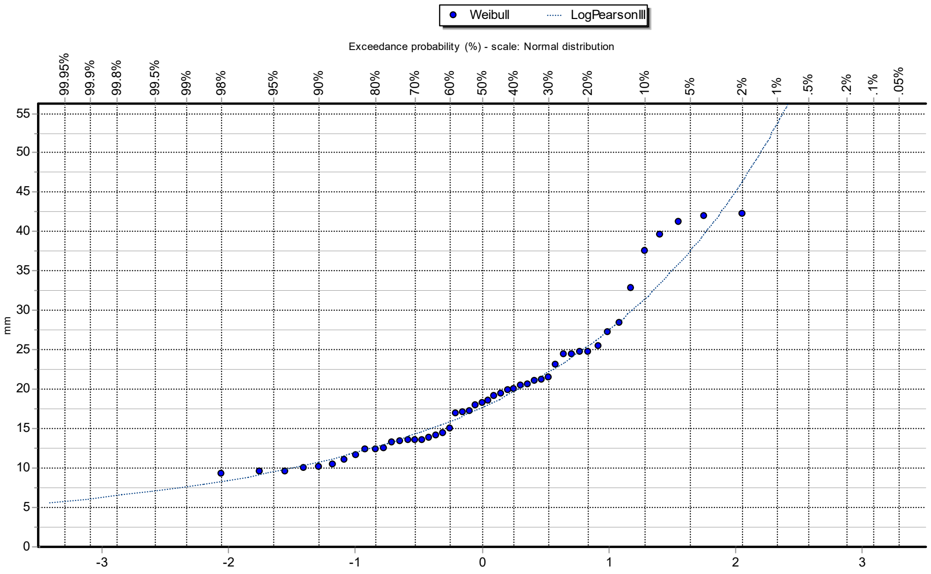

Figure 6. Q-Q plot of Log Pearson III distribution at Station: Las Salinas.

4.2. Estimating the Return Period Using the Best-Fit Distribution and Random Forest

Once the best-fitting PDFs for each station are identified, the precipitation for each selected return period can be determined. Table 2 presents the return periods (2, 5, 10, 20, 50, and 100 years) along with the corresponding results.

Table 2. Return Periods results with PDFs.

N°	Station Name	Return Period using the Best Fit Distribution (mm)					
		2	5	10	20	50	100
1	Chiguata	20.0623	31.205	37.7953	43.5582	50.3423	55.0198
2	El Frayle	21.4111	28.9544	34.7028	40.8594	49.8822	57.5199
3	La Joya	0.50552	2.44163	4.77694	7.544	11.6142	14.9007
4	La Pampilla	12.1171	23.2693	32.7275	43.3751	59.5553	73.5711
5	Las Salinas	17.5087	26.2854	32.2689	37.9196	45.0286	50.1978

Figure 7 shows the distribution of the predictions obtained from the PDFs listed in Table 2, which in this case represent precipitation values, as a function of the return period.

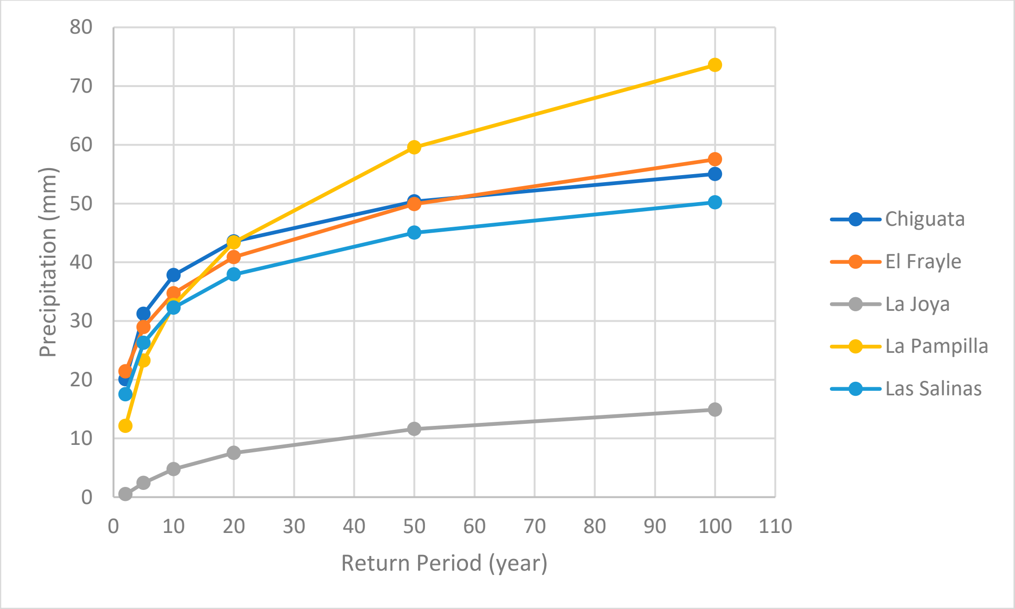

Figure 7. Return Periods for each station using PDF.

On the other hand, the precipitation results for the previously mentioned return periods are displayed in Table 3.

Table 3. Return Periods results with RF.

N°	Station Name	Return Period using Random Forest (mm)					
		2	**5**	**10**	**20**	**50**	**100**
1	Chiguata	20.4393	33.797	39.641	43.304	46.5453	46.5453
2	El Frayle	21.6073	30.0127	33.9537	40.8773	51.587	51.587
3	La Joya	0.7347	3.02	4.3387	6.3653	13.676	13.676
4	La Pampilla	12.2967	23.4273	29.6063	42.4827	91.5047	91.5047
5	Las Salinas	18.3933	24.8417	34.188	41.0773	41.6797	41.6797

Finally, the distribution of the values is shown in Figure 8.

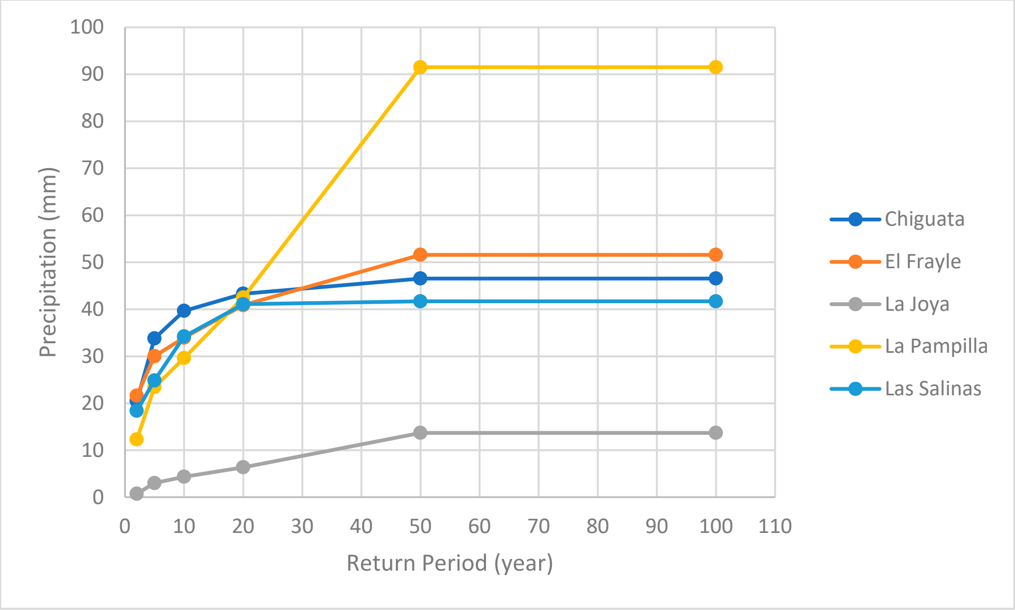

Figure 8. Return Periods for each station using RF.

5. Discussion

Achieving national disaster resilience to hydrological phenomenon is essential to countries such as Peru [27,28], where adequately estimating precipitations is a great challenge.

Geographic location and surrounding environment are fundamental factors in the variation of the rainfall pattern, as is the case of Bangladesh, where the PDFs that fit the best were GEV (36% of the stations), Pearson 3 (26% of the stations) and Log Pearson 3 (26% of stations) [5]. While in the case of Arequipa province, the analysis results showed that, following the traditional approach, 40% of the stations were better fitted with the GEV-MIN (L-Moments) function after performing the Kolmogorov-Smirnov (K-S) test.

However, the scope of new technologies allows innovative approaches to be used, such as the Random Forest algorithm, which is used for predictions in different fields. One of the uses in the field of hydraulics is the prediction of DAM displacement [25], where it was validated, that RF is a powerful tool for monitoring tasks.

In this particular case, to predict the precipitation for a specific return period in Arequipa province, 80% of the stations were better fitted with RF compared to PDFs, as they exhibited lower RMSE values, making it a viable alternative to the traditional approach adopted by much of the world.

The analysis conducted showed a similarity in precipitations for different return periods when using both methods (PDFs and RF), as evidenced in Tables 2 and 3 of the Results section. However, it can be observed that for the 50- and 100-year periods, the values double. This may be due to the lack of variability in the data, since when a predictive model has data with low variability, it cannot fully understand the complexity of the data, resulting in vague predictions for high return periods. On the other hand, PDFs are designed to model extreme values and are better able to capture the variability of the data.

As has been observed in different parts of the world, there are studies analyzing rainfall patterns with probability functions for cities, regions, and even entire countries. However, to date, no such studies exist in Peru.

6. Conclusions

This study aims to provide an innovative approach to the calculation of precipitation for different return periods with a new method to improve accuracy, as well as offer an alternative for regions facing similar cases of data scarcity. Currently, there are cities and regions around the world lacking updated information from their meteorological stations due to various reasons, such as high maintenance costs, low public interest, among others.

Improved precision in precipitation estimates for various return periods has a significant impact on water resource planning and management. Achieving higher accuracy in these estimates enables more effective and efficient planning of water resources including allocation for agriculture, industry, and domestic consumption. This, in turn, reduces flood risk and associated damages by identifying high-risk areas and taking preventive measures. Moreover, optimal water supply allocation is ensured through efficient distribution of available water, meeting the needs of populations and industries. Improved precipitation estimates also yield significant cost savings in infrastructure, maintenance, and operation of water systems. This integrated approach to resource management is crucial for ensuring sustainability and resilience of water systems in the face of increasing climate change.

Finally, a future research line involves analyzing daily, monthly, and annual rainfall patterns in Peruvian regions. Along these lines, there are other machine learning algorithms capable of performing regression, as done in this study (e.g., neural networks). Thus, another research avenue is to explore the calculation using different algorithms and to analyze the differences each one offers.

Author Contributions: Conceptualization, J.A.-V., S.V.-F and G.A.; methodology, J.A.-V., S.V.-F, G.A; software, J.A.-V., S.V.-F and J.J.-Q; validation, J.A.-V., G.A. and S.V.-F.; formal analysis, J.A.-V., G.A and S.V.-F; investigation, J.A.-V., G.A., S.V.-F and J.J.-Q; resources, J.A.-V., A.J.E.V, J.J.-Q and E.R.-C; data curation, J.A.-V. and J.J.-Q; writing—original draft preparation, J.A.-V., S.V.-F and J.B; writing—review and editing, J.A.-V., S.V.-F, A.J.E.V. and J.B; visualization, J.A.-V., J.J.-Q and E.R.-C; supervision, J.A.-V., A.J.E.V and J.B; project administration, J.A.-V., A.J.E.V and J.B; funding acquisition, A.J.E.V., J.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available on request from the corresponding author, J.B. The data are not publicly available due to ethical restrictions.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Aureli, F.; Mignosa, P.; Prost, F.; Dazzi, S. Hydrological and Hydraulic Flood Hazard Modeling in Poorly Gauged Catchments: An Analysis in Northern Italy. *Hydrology* **2021**, *8*, doi:10.3390/hydrology8040149.
2. Gao, M.; Wang, Z.; Yang, H. Review of Urban Flood Resilience: Insights from Scientometric and Systematic Analysis. *Int. J. Environ. Res. Public Health* **2022**, *19*, 8837.
3. Bertilsson, L.; Wiklund, K.; de Moura Tebaldi, I.; Rezende, O.M.; Veról, A.P.; Miguez, M.G. Urban flood resilience – A multi-criteria index to integrate flood resilience into urban planning. *J. Hydrol.* **2019**, *573*, 970-982, doi:https://doi.org/10.1016/j.jhydrol.2018.06.052.
4. Padji, C.; Meukaleuni, C.; Mezoue Adiang, C.; Bongue, D.; Monkam, D. Estimation of return dates and return levels of extreme rainfall in the city of Douala, Cameroon. *Heliyon* **2024**, *10*, doi:10.1016/j.heliyon.2024.e34832.
5. Alam, M.A.; Emura, K.; Farnham, C.; Yuan, J. Best-Fit Probability Distributions and Return Periods for Maximum Monthly Rainfall in Bangladesh. *Climate* **2018**, *6*, 9.
6. Katz, R.W.; Brown, B.G. Extreme events in a changing climate: Variability is more important than averages. *Climatic Change* **1992**, *21*, 289-302, doi:10.1007/BF00139728.
7. Khudri, M.M.; Sadia, F. Determination of the Best Fit Probability Distribution for Annual Extreme Precipitation in Bangladesh. *European Journal of Scientific Research* **2013**, *103*, 391-404.

8. Mandal, S.; Choudhury, B.U. Estimation and prediction of maximum daily rainfall at Sagar Island using best fit probability models. *Theoretical and Applied Climatology* **2015**, *121*, 87-97, doi:10.1007/s00704-014-1212-1.
9. Branch, O.; Schwitalla, T.; Temimi, M.; Fonseca, R.; Nelli, N.; Weston, M.; Milovac, J.; Wulfmeyer, V. Seasonal and diurnal performance of daily forecasts with WRF V3.8.1 over the United Arab Emirates. *Geoscientific Model Dev.* **2021**, *14*, 1615-1637, doi:10.5194/gmd-14-1615-2021.
10. Zhai, W.; Wang, Z.; Feng, Y.; Xue, L.; Ma, Z.; Tian, L.; Sun, H. Developing the Actual Precipitation Probability Distribution Based on the Complete Daily Series. *Sustainability* **2023**, *15*, 13136.
11. Si, L.; Shao, Q.; Zhao, L.; Wei, T.; Hou, J.; Huang, J. Adjusting and refinement of daily extreme precipitation based on high-density weather stations. *J. Nat. Disasters* **2023**, *32*, 145-159, doi:10.13577/j.jnd.2023.0314.
12. Li, H.; Li, M. Modeling of Precipitation Prediction Based on Causal Analysis and Machine Learning. *Atmosphere* **2023**, *14*, 1396.
13. Volpi, E. On return period and probability of failure in hydrology. *Wiley Interdiscip. Rev.: Water* **2019**, *6*, doi:10.1002/WAT2.1340.
14. Alam, M.A.; Farnham, C.; Emura, K. Best-Fit Probability Models for Maximum Monthly Rainfall in Bangladesh Using Gaussian Mixture Distributions. *Geosciences* **2018**, *8*, doi:10.3390/geosciences8040138.
15. Sun, T.; Yan, N.; Zhu, W.; Zhuang, Q. Assessing a machine learning-based downscaling framework for obtaining 1km daily precipitation from GPM data. *Heliyon* **2024**, *10*, doi:10.1016/j.heliyon.2024.e36368.
16. Papacharalampous, G.; Tyralis, H.; Doulamis, N.; Doulamis, A. Uncertainty estimation of machine learning spatial precipitation predictions from satellite data. *Mach. Learn.: Sci. Technol.* **2024**, *5*, doi:10.1088/2632-2153/ad63f3.
17. Hassan, M.M.; Rony, A.; Khan, M.; Hassan, M.; Yasmin, F.; Nag, A.; Zarin, T.; Bairagi, A.; Alshathri, S.; El-Shafai, W. Machine Learning-Based Rainfall Prediction: Unveiling Insights and Forecasting for Improved Preparedness. *IEEE Access* **2024**, *11*, 132196-132222, doi:10.1109/ACCESS.2023.3333876.
18. Markovic, R. Probability Functions of Best Fit to Distribution of Annual Precipitation and Runoff. **1965**.
19. Phien, H.N.; Ajirajah, T.J. Applications of the log Pearson type-3 distribution in hydrology. *J. Hydrol.* **1984**, *73*, 359-372, doi:10.1016/0022-1694(84)90008-8.
20. Salinas, J.L.; Castellarin, A.; Kohnová, S.; Kjeldsen, T.R. Regional parent flood frequency distributions in Europe - Part 2: Climate and scale controls. *Hydrol. Earth Syst. Sci.* **2014**, *18*, 4391-4401, doi:10.5194/hess-18-4391-2014.
21. Dufour, J.-M.; Farhat, A.; Gardiol, L.; Khalaf, L. Simulation-based finite sample normality tests in linear regressions. *The Econometrics Journal* **1998**, *1*, C154-C173.
22. Arshad, M.; M.T, R.; Ahmad, M. Anderson Darling and Modified Anderson Darling Tests for Generalized Pareto Distribution. *Journal of Applied Sciences* **2003**, *3(2)*, doi:10.3923/jas.2003.85.88.
23. Tyralis, H.; Papacharalampous, G.; Langousis, A. A brief review of random forests for water scientists and practitioners and their recent history in water resources. *Water (Switzerland)* **2019**, *11*, doi:10.3390/w11050910.
24. Ahmadi, S.M.; Balahang, S.; Abolfathi, S. Predicting the hydraulic response of critical transport infrastructures during extreme flood events. *Engineering Applications of Artificial Intelligence* **2024**, *133*, 108573, doi:https://doi.org/10.1016/j.engappai.2024.108573.
25. Su, Y.; Weng, K.; Lin, C.; Zheng, Z. An Improved Random Forest Model for the Prediction of Dam Displacement. *IEEE Access* **2021**, *PP*, 1-1, doi:10.1109/ACCESS.2021.3049578.

26. Behrens, C.; Pierdzioch, C.; Risse, M. Testing the optimality of inflation forecasts under flexible loss with random forests. *Economic Modelling* **2018**, *72*, 270-277, doi:https://doi.org/10.1016/j.econmod.2018.02.004.
27. Espinoza Vigil, A.J.; Booker, J.D. Building national disaster resilience: assessment of ENSO-driven disasters in Peru. *International Journal of Disaster Resilience in the Built Environment* **2023**, *14*, 423-433, doi:10.1108/IJDRBE-10-2022-0102.
28. Valencia-Félix, S.; Anco-Valdivia, J.; Espinoza Vigil, A.J.; Hidalgo Valdivia, A.V.; Sanchez-Carigga, C. Review of Green Water Systems for Urban Flood Resilience: Literature and Codes. *Water* **2024**, *16*, doi:10.3390/w16202908.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.