RBRH

*Revista Brasileira de Recursos Hídricos*
*Brazilian Journal of Water Resources*
*Versão On-line ISSN 2318-0331*
*RBRH*, Porto Alegre, v. 28, e9, 2023
*Scientific/Technical Article*

*https://doi.org/10.1590/2318-0331.282320230008*

# Estimation of monthly rainfall missing data in Southwestern Colombia: comparing different methods

## *Estimativa de dados faltantes de precipitação mensal no sudoeste da Colômbia: comparação de diferentes métodos*

**Juan Sebastián Del Castillo-Gómez[1], Teresita Canchala[1], Wilmar Alexander Torres-López[1], Yesid Carvajal-Escobar[1] & Camilo Ocampo-Marulanda[1]**

[1]Escuela de Ingeniería de los Recursos Naturales y del Ambiente, Universidad del Valle, Facultad de Ingeniería, Cali, Valle del Cauca, Colombia
E-mails: juan.s.castillo@correounivalle.edu.co (JSCG), teresita.canchala@correounivalle.edu.co (TC), wilmar.alexander.torres@correounivalle.edu.co (WATL), yesid.carvajal@correounivalle.edu.co (YCE), camilo.ocampo@correounivalle.edu.co (COM)

Received: January 20, 2023 - Revised: March 17, 2023 - Accepted: March 30, 2023

## ABSTRACT

Historical rainfall records are relevant in hydrometeorological studies because they provide information on the spatial features, frequency, and amount of precipitated water in a specific place, therefore, it is essential to make an adequate estimation of missing data. This study evaluated four methods for estimating missing monthly rainfall data at 46-gauge stations in southwestern Colombia covering 1983-2019. The performance of the Normal Ratio (NR), Principal Components Regression (PCR), Principal Least Square Regression (PLSR), and Artificial Neural Networks (ANN) methods were compared using three standardized error metrics: Root Mean Square Error (RMSE), Percent BIAS (PBIAS), and Mean Absolute Error (MAE). The results generally showed a better performance of the nonlinear ANN method. Regarding the linear methods, the best performance was registered by the PLSR, followed by the PCR. The results suggest the applicability of the ANN method in regions with a low density of stations and a high percentage of missing data, such as southwestern Colombia.

**Keywords:** Rainfall time series; Data reconstruction; Estimation methods; Artificial Neural Networks (ANN).

## RESUMO

Os registros históricos de precipitação são relevantes para os estudos hidrometeorológicos porque fornecem informações sobre as características espaciais, frequência e volume de precipitação de água em um local específico, portanto, é essencial realizar uma estimativa adequada dos dados faltantes. Esta pesquisa avaliou quatro métodos para estimativa de dados de precipitação faltantes mensalmente em 46 estações de medição no sudoeste da Colômbia, abrangendo os anos de 1983 à 2019. O desempenho dos métodos de Razão Normal (NR), Regressão de Componentes Principais (PCR), Regressão por Mínimos Quadrados Principais (RMQP) e Redes Neurais Artificiais (RNA) foram comparados usando três métricas padronizadas para os erros: Raiz do Erro Médio Quadrático (REMQ), Percentagem BIAS (PBIAS) e Erro Absoluto Médio (EMA). Os resultados com frequência mostraram um melhor desempenho do método ANN não-linear. Com relação aos métodos lineares, o melhor desempenho foi registrado pelo PLSR, seguido pelo PCR. Os resultados sugerem a aplicabilidade do método ANN em regiões com baixa densidade de estações de medição e alta porcentagem de dados faltantes, como o sudoeste da Colômbia.

**Palavras-chave:** Séries cronológicas de chuvas; Reconstrução de dados; Métodos de comparação; Redes Neurais Artificiais (ANN).

## INTRODUCTION

Proper water management depends mainly on the quantity and quality of hydroclimatological information. Knowledge of the spatial distribution and seasonality of rainfall is relevant for the economy and society of a region (Morales-Acuña et al., 2021). Reliable and complete data is required to take action on actions such as flood and drought risk prediction and assessment, rainfall forecasting, dry spells, desertification, climate variability studies, and water resource planning, among others (Canchala et al., 2019; Kuok et al., 2010; Miró et al., 2018). Knowing the processes involved in the water balance of a region is essential due to the influence of climate change on precipitation worldwide with increasingly variable patterns, according to the projections of the Intergovernmental Panel on Climate Change (2022).

Usually, missing data in rainfall time series have become a common problem in gauge stations (Ramos-Calzado et al., 2008; Torres et al., 2015; Souza & Leal, 2017; Pinheiro et al., 2022) due to measurement instrument failure, observation errors, and outliers. In addition, it is common to find logistical, economic, and accessibility limitations in the field, which make it challenging to establish and maintain networks for measuring hydrometeorological variables (Cruz-Roa & Barrios, 2018).

Addi et al. (2022) and Taghi et al. (2017) highlight the need to evaluate rainfall missing data estimation methods to allow characterizing its spatio-temporal behaviour in detail and climate trend analysis. In this sense, different methods have been developed to estimate missing data (Miró et al., 2017); among these, there are traditional methods based on the estimation of missing data from the nearest neighbouring stations (Auer et al., 2005; Burhanuddin et al., 2017; Cruz-Roa & Barrios, 2018; Silva et al., 2007; Domonkos, 2015; Ramos-Calzado et al., 2008; Taghi et al., 2017), methods based on spatial interpolation (IDW) widely used in hydrology (Armanuos et al., 2020; Cerón et al., 2021a; Silva et al., 2007; Lee & Kang, 2015; Morales Martínez et al., 2019), multiple regression methods (Bárdossy & Pegram, 2011; Moraes Cordeiro & Blanco, 2021; Santos et al., 2021; Francisco, 2015; Teegavarapu, 2012), advanced linear methods such as Iterated Least Squares approach (DeGaetano & Allen, 2002), and of multiple linear regression derived from Empirical Orthogonal Functions (EOF's) with principal component analysis (Shahrokhi et al., 2020).

Likewise, more recently used methods that consider nonlinear relationships, such as artificial neural networks (Canchala et al., 2020c; Canchala et al., 2019; Chiu et al., 2021; Kajornrit et al., 2012; Khalili et al., 2016; Londhe et al., 2015; Ocampo-Marulanda et al., 2021), genetic programming (Ismail et al., 2021; Khalili et al., 2016), among others. The latter are characterized by having higher demands for computational cost and knowledge compared to traditional methods and multiple linear regression.

Southwestern Colombia recurrently presents a lack of information related to climate records due to aspects such as the low density of gauge stations and social and public order problems that make it difficult to access them compared to other regions of Colombia (Canchala et al., 2019). This problem affects hydroclimatological analyzes for water resource management and planning purposes. On the other hand, this region presents a complex climatology, due to the influence of climatic phenomena such as El Niño Southern Oscillation (ENSO), Madden-Julián Waves, Chocó low-level Jet, among others (Canchala et al., 2020a; Cerón et al., 2021b; Puertas & Carvajal, 2008; Rueda & Poveda, 2006; Sedano-Cruz et al., 2013; Torres, 2012), coupled with its proximity to Ecuador and influence of the Intertropical Convergence Zone (ITCZ).

In this sense, the main objective of this research is to evaluate and compare four methods for estimating missing monthly rainfall data at 46-gauge stations in southwestern Colombia with historical data between 1983-2019, using four performance metrics. Hence, this article is arranged as follows: Section 2 describes the study area and data. Section 3 describes the methodology used. Section 4 includes the results and discussion, and finally, section 5 shows the conclusions.

## STUDY AREA AND DATA

### Study area

The study area is in Southwestern Colombia (Department of Nariño). It has an approximate area of 33,268 km2 that occupies a geostrategic position because the Andes Mountain range crosses its limits, the Tropical Pacific Ocean. It registers significant topographic changes in small distances (Canchala et al., 2020b), which provides the regional diversity of reliefs, thermal floors and microclimates (See Figure 1).

Furthermore, it is in the south of the Colombian Biogeographic Chocó; recognized as a biodiversity hotspot that shelters approximately 3% of the world's plant species (Poveda & Mesa, 2000), characterized by registering three important rainfall core that range between 7000 mm. year$^{-1}$ to 9000 mm. year $^{-1}$ (Cerón et al., 2021b). Likewise, it registers three natural regions where 52% is made up of the Pacific region, with high rainfall that ranges between 3000-7000 mm.year$^{-1}$, 40% belongs to the Andean region, where the presence of moors, volcanoes, and relief stand out. Rugged with rainfall that ranges between 1000-2000 mm.year$^{-1}$, and the remaining 8% belongs to the Amazon jungle region where there is high biodiversity of communities and species and rainfall that varies between 3500 mm.year$^{-1}$ – 4500 mm.year$^{-1}$ (Canchala et al., 2019, 2020a; Cerón et al., 2021b).

### Rainfall dataset

Monthly rainfall time series with 37 years of observation (1983-2019) from 46-gauge stations located in southwestern Colombia were used. The data was provided by Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM) of Colombia, and its spatial distribution is shown in Figure 1.

## METHODS

Four methodologies for estimating missing monthly rainfall data in Southwestern Colombia were used: Normal Ratio (NR), Principal Components Regression (PCR), Principal Least Square Regression (PLSR), and Artificial Neural Networks. (ANN). The first three methods are linear, and the last corresponds to a nonlinear methodology. Initially, the rainfall records were organized, and the

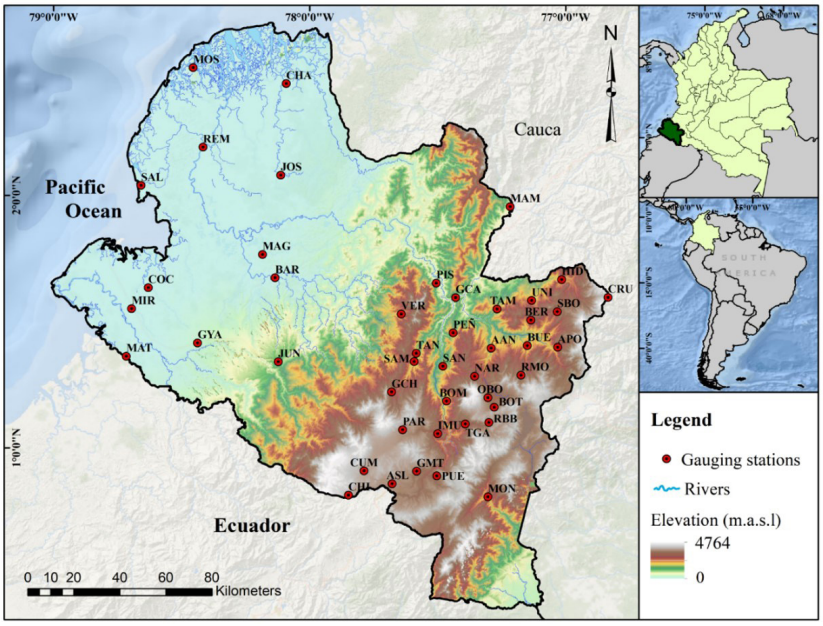


**Figure 1.** Geographic location of the study area and spatial distribution of rain gauge stations.

exploratory and confirmatory data analysis was carried out, where the behaviour of the rainfall was determined through descriptive statistical measures, as well as the percentage of missing data in the 46 study stations. Subsequently, homogeneous rainfall groups were formed using statistical cluster analysis and similarity measures for the three linear methods. Finally, missing data were imputed, and method performance was assessed using four performance metrics: Root Mean Square Error (RMSE), Percent bias (PBIAS), Mean Absolute Error (MAE), and Pearson correlation coefficient (r). The best method for estimating missing data was selected based on these results. In the flowchart of Figure 2, the methodology used to estimate the missing data in the different study stations is registered.

## Pre-processing data

Preliminary analysis of the rainfall time series was performed through exploratory and confirmatory data analysis. This analysis consists of applying graphical and quantitative statistical tools to identify patterns and anomalies in the data and observe the behaviour of rainfall in the selected stations (Zhao et al., 2011; Castro et al., 2012).

Furthermore, the consistency of the data was verified using the Spearman correlation coefficient ($\varrho$), used to quantify the degree of correspondence between the stations and identify anomalous gauge stations. $\varrho$ ranges from -1 to 1, with the maximum (minimum) value being the perfect positive (negative) correlation. The statistical significance of the correlations was determined using the T-student test with a significance level of $\alpha$ =0.05.

For the NR method, the Spearman correlation coefficient was used as a quantitative criterion for selecting neighbouring stations for the estimation of missing data since it is not multiple estimation methods. Stations with values of $\varrho \geq 0.5$ were selected, a criterion used by Cruz-Roa & Barrios (2018), establishing that the stations with correlation coefficients constitute a homogeneous hydrological cluster.

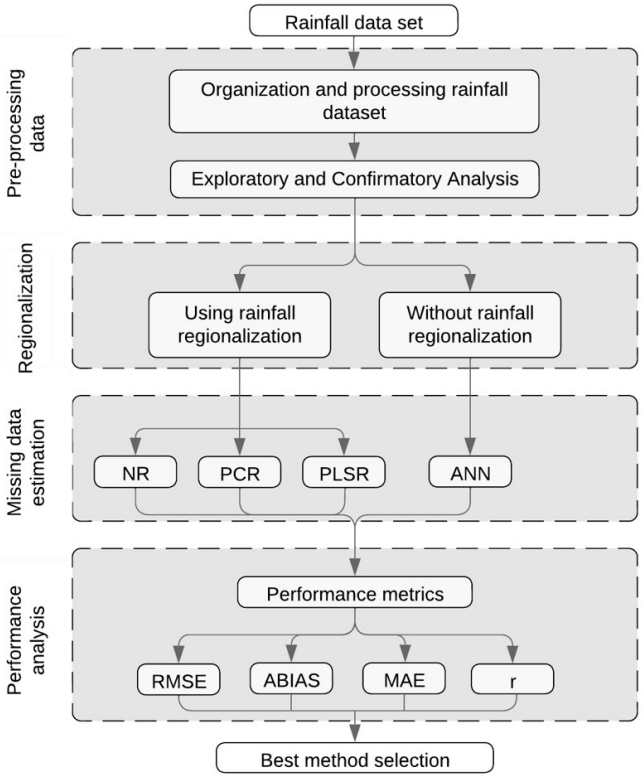


**Figure 2.** Flowchart of the methodology used in the study.

## Regionalization

Defining homogeneous rainfall regions is essential for hydrological applications, particularly for regions with high spatio-temporal variability in rainfall patterns (Zhang et al., 2016). Therefore, cluster analysis (CA) was used to classify stations with similar characteristics and properties.

For clustering purposes to regionalize the monthly rainfall, we use the HCA with square of Euclidean Distance for measure the observations, represented by Equation 1, and Ward's method for the linkage rule. According to Hershey et al. (2010), Darand & Reza (2014), Gois et al. (2020) and Silva (2020) this fusion results in the most distinctive clusters.:

$$d_{xy} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (1)$$

where $x_i$ and $y_i$ are elements of comparison, which in this case represent the rainfall between stations $x$ and $y$ at the i-th instant of time.

Subsequently, hierarchical groups were formed by Ward's method, allowing the grouping of gauge stations with similar behaviours of rainfall. Finally, each region's spatial coherence and rainfall patterns were verified.

This study chose Ward's method because it has been the clustering technique most used in climate regionalization (Fazel et al., 2018; Canchala et al., 2022). Overall, it outperforms other algorithms in terms of segregation, providing relatively dense groups with minimums within-group variance. Ward's method recognizes the minimum variance within groups, joining elements with a minimal sum of squares between them (Hervada-Sala & Jarauta-Bragulat, 2004; Santos et al., 2015).

## Missing data estimation

There are different methodologies for estimating missing data and reconstructing hydroclimatological time series (Adilah & Hannani, 2021; Chiu et al., 2021; Shahrokhi et al., 2020; Canchala et al., 2019; Lai et al., 2019; Morales-Martínez et al., 2019; Cruz-Roa & Barrios, 2018; Miró et al., 2017; Kim & Pachepsky, 2010; Burhanuddin et al., 2017; Zuccolotto, 2012; Silva et al., 2007; Paulhus & Kohler, 1952). Due to the complexity of the study area, four estimation methods for missing data were evaluated, including conventional (linear) and artificial intelligence (nonlinear) methods, which were NR using neighbouring stations, PCR, PLSR, and ANN.

### Normal Ratio (NR)

It is a non-multivariate conventional method initially used by Paulhus & Kohler (1952) and is currently used in numerous hydrological studies (Adilah & Hannani, 2021; Moraes Cordeiro & Blanco, 2021; Burhanuddin et al., 2016; Caldera et al., 2016; Arango et al., 2012; Puertas & Carvajal, 2008). It is used when there are missing data in a specific month for a weather station with neighbouring stations without missing records for the same months as the station of interest and with similar topographical features. For this, the expression of Equation 2 is used:

$$P_x = \frac{1}{n} \sum_{i=1}^{n} \frac{N_x}{N_i} P_i \quad (2)$$

where; $n$ is the number of pluviometric stations with continuous record data close to station x, in which the record will be completed, $P_x$ is the rainfall of station x during the month to be completed, $P_i$ is the rainfall of stations 1 up to n during the month to be completed, $N_x$ is the multiannual monthly rainfall of station x, and $N_i$ the multiannual monthly average rainfall of stations 1 to n.

### Principal Components Regression (PCR)

Principal Component Analysis (PCA) or Empirical Orthogonal Functions (FOEs) is a statistical technique used in climate research as a tool to analyze meteorological series with high spatio-temporal dimensionality and noise (Carvajal-Escobar & Marco, 2005; Taylor et al., 2013). Together with regression, they are multivariate techniques that allow data from multiple monitoring sites to be incorporated into a statistical model while minimizing the effects of non-correlation between the measured variables (Shlens, 2014), which is why it is widely used in hydroclimatology to extract linear relationships between variables in a data set (Lu & Hsieh, 2003), reduce dimensionality and avoid multicollinearity (Cerón et al., 2021a; Ocampo-Marulanda et al., 2022).

Principal Components Regression is a method introduced by Massy (1965) that allows transforming the p original explanatory variables into a new set of k uncorrelated variables, called Principal Components (PCs), with $k < p$. These variables are later used as explanatory variables in a multiple linear regression model to estimate the value of the response variable (Wyatt et al., 2020).

A subset of PCs was selected to improve the quality of the estimation of the model, and the Kaiser-Gutman (K-G) criterion was used to determine the number of these to use. The K-G criterion recommends choosing the first k PCs associated with eigenvalues greater than 1.0. In addition to the above, the empirical criterion of the percentage of accumulated variance was also used, which suggests keeping the PCs that accumulate an explained variance of approximately 80% % (Cuadras, 2007).

The relationship between the set of gauge stations was established using Equation 3, based on the multiple linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + e \quad (3)$$

where the variable Y represents the rainfall of one of the gauge stations, $\beta_0$ is the intercept, $\beta_1,\ldots,\beta_p$ is the partial regression coefficients associated with the p gauge stations $X_j\,(j = 1,2,\ldots,p)$ to be considered predictors, and e represents the random error component associated with the regression model.

The predictors were transformed considering $k < p$ linear combinations to reduce the possible effect of multicollinearity by including monitoring stations that may present a spatial correlation as predictor variables:

$$Z_l = a_{1l} X_1 + a_{2l} X_2 + \ldots + a_{pl} X_p \quad (4)$$

with $l = 1,\ldots,k$, called components, and subsequently adjusting a linear regression model (Equation 5) using them as new predictors:

$$Y = \alpha_0 + \alpha_1 Z_1 + \alpha_2 Z_2 + \ldots + \alpha_k Z_k + e \quad (5)$$

Additionally, if the coefficients $a_{jl}$ (loads or weights) are selected as follows $\sum_{j=1}^{p} a_{jl}a_{jm} = 0, si\, l \neq m$, components $Z_l$ will be orthogonal, therefore avoiding potential problems of multicollinearity. The procedure described above was carried out iteratively, varying the gauge station to be considered as the response variable and using the other available stations as predictor variables.

## Principal Least Square Regression (PLSR)

Partial Least Squares (PLS) regression is a model for multivariate prediction or estimation of data (Andersson, 2009). It was introduced by Wold (1975) and combined features of principal component analysis and multiple regression. Like Principal Component Regression (PCR), it allows the original explanatory variables to be transformed into a new set of independent variables (components) (Wold et al., 2001), with the difference that PCR regression does not consider the response variable to determine the components, since it only maximizes the explained variance of the set of predictor variables to be introduced in the regression model; while PLS regression determines these components taking into account both the original predictor variables and the response variable (it maximizes the covariance between the original explanatory variables and the response variable(s).

PLS regression with a single response variable is called PLS1; if there is more than one response variable, the PLS regression is called the PLS2 regression (Andersson, 2009). In this research, the PLS1 algorithm was used, and the empirical criterion of the percentage was used, which suggests conserving the components that allow accumulating at least 80% of the explained variability of the response variable.

## Artificial Neural Networks (ANN)

Nonlinear Principal Component Analysis (NLPCA) is a method using ANN, proposed by Scholz et al. (2005) and used in the estimation of missing hydroclimatological data by Miró et al. (2017) and Canchala et al. (2019). This method uses an auto-associative neural network approach based on decoding (second phase of NLPCA), better known as inverse NLPCA, which corresponds to a nonlinear generalization of the standard Principal Component Analysis (PCA).

The NLPCA uses a reconstruction function $\Phi_{gen} : y \rightarrow \bar{x}$which is performed by a feed-forward network $\bar{x} = \varphi_{gen}(w, y) = W_4 g(W_3 y)$, where the objective of $\varphi_{gen}$ is to estimate the data set $\bar{x}$ approximated the target data y, minimizing the root mean square error $x - \bar{x}^2$ (See Figure 3). The NLPCA toolbox used in this study is available at Scholz (2023).

The organization structure of the input and output layers is carried out through different types of architectures; the most used model is the Multi-Layer Perceptron (MLP), which has a structure with an input layer, one or several layers hidden and an output layer (Canchala et al., 2020a). Rumelhart developed this model, also called the error propagation model or back-propagation model, which uses the Delta Learning Rule learning method (Demir & Keskin, 2021). For this research, the architecture employed in Canchala et al. (2019) was used.

A learning algorithm was used for preliminary recognition of the data and identification of historical rainfall patterns of the time series associated with different weather phenomena that influence the region. This was done using the back-propagation algorithm, which propagates the error between the actual output data and the estimated output data in the neural network. The above is described by Equation 6:

$$E = \sum_{j=1}^{M} E_j = \sum_{m=1}^{PM} \sum_{j=1}^{n} \left( x_{mj} - \bar{x}_{mj} \right)^2 \quad (6)$$

where E is the total error, M is the number of input data, $E_j$ is the error of the squared difference between the actual data ($x_{mj}$) and the estimated data ($\hat{x}_{mj}$) (Khalili et al., 2016).

For the ANNs, the architecture [46-45-46] was used, and different iterations were evaluated in the neural network (5,000, 8,000 and 10,000). The best results were obtained using

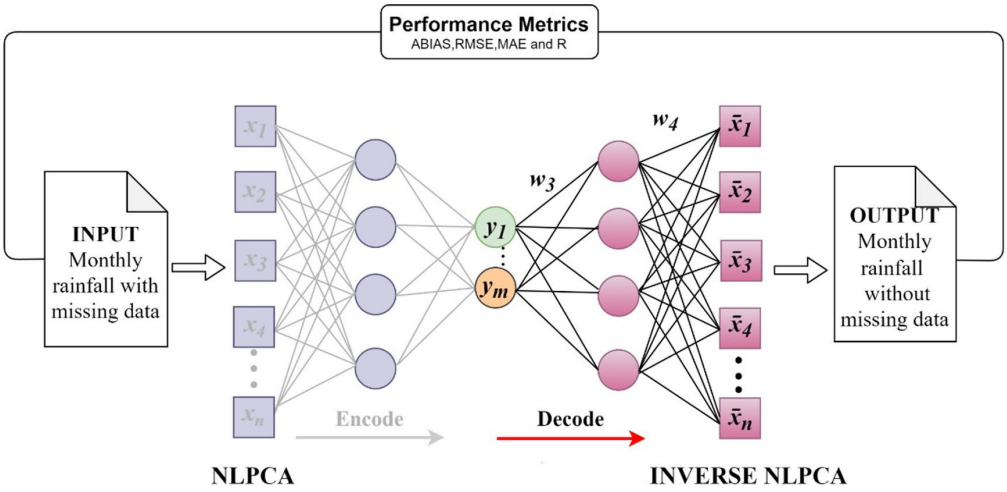


**Figure 3.** Flowchart of the NLPCA Inverse. $x_n$ is the input layer (Monthly rainfall dataset), $y_m$ is the bottleneck layer of the NLPCA model, and $\bar{x}_n$ is the output layer (Monthly rainfall dataset reconstructed).

10,000 iterations; this greater number of iterations allowed better recognition of the database and identification of atypical patterns resulting from climatic phenomena and anomalies that influence the study area.

### Performance metrics

The evaluation of the estimation of missing data of the four (4) methods used was carried out through three (3) standardized performance metrics: RMSE, MAE, and ABIAS, which allowed verifying the adjustment in each of the gauge stations studied. Also, the Pearson correlation coefficient (r) was used to measure the similarity between the observed and estimated data. The metrics used are represented in Table 1.

## RESULTS AND DISCUSSION

### Descriptive statistical analysis

Table 2 shows the descriptive statistics of the gauge stations' monthly rainfall time series for the period 1983-2019.

The studied gauge stations showed altitude differences, varying from 3 m.a.s.l in the plain of the Pacific region to 3120 m.a.s.l in the Andes Mountain range. The fluctuations of rainfall in the territory are mainly due to the altitude differences, in addition to orographic aspects, presence of the Andes mountain range, limits with the Pacific Ocean, proximity to the Amazon basin, and influence of climatic phenomena at different time scales, such as the ENSO phenomenon, the Madden-Julián Oscillation, and ITCZ migration, among others (Cerón et al., 2021a; Puertas & Carvajal, 2008; Rueda & Poveda, 2006; Serna et al., 2018; Torres-Pineda & Pabón-Caicedo, 2017). It is highlighted that the areas with the highest average annual rainfall are located in the Pacific region, with records of up to 8689.39 mm.year$^{-1}$.

In general, the monthly rainfall time series showed high variation with respect to the mean, ranging between 174.11 mm.year$^{1}$ and 1347.6 mm.year$^{1,}$ with variation coefficients that ranged between 9% and 41%, with the greatest variability being recorded in the Pacific region. This region is characterized by not presenting a defined rainfall trend (Guzmán et al., 2014); some areas register high rainfall during the year and, in others, low rainfall, mainly in stations near the Pacific Ocean. These results are consistent with Cerón et al. (2021b), who reported a core of high rainfall in the south of the Colombian Pacific (3000 to 7000 mm.year$^{-1}$), and Canchala et al. (2022) who reported in the Pacific region of Nariño rainfall ranges from 2500 to 8650 mm.year$^{-1}$.

The high variability of rainfall in the study area is mainly due to the movement of the ITCZ, the influence of the Chocó jet stream, and the incidence of the La Niña and El Niño phases of the ENSO, which have been documented by different authors who have studied the hydroclimatology of Colombia and specifically the southwestern Colombian (Canchala et al., 2022; Cerón et al., 2021b; Poveda & Mesa, 1999; Urrea et al., 2019). Furthermore, the asymmetry coefficient was positive (negative) in 80.44% (19.56%) of the stations, indicating that in these stations, most of the rainfall values are higher (lower) than the average.

### Regionalization of monthly rainfall

The regionalization of the monthly rainfall was performed using Ward's method, which showed the conformation of 3 homogeneous regions formed into three groups of 33-, 12- and 1-gauge station, which is consistent with the natural regions of the Andean region (AR), Pacific region (PR) and Amazon region (AMR), respectively, as shown in Figure 4. As well, these results are consistent with the rainfall regionalization performed by Canchala et al. (2022) through the application of the nonlinear technique called Kohonen's self-organized maps, also with Guzmán et al. (2014), using PCA and Jaramillo-Robledo & Chaves-Córdoba (2000) who used statistical clusters.

The three identified homogeneous regions show different patterns of rainfall interannual variability. The gauge stations of the PR are characterized by registering average rainfall between

**Table 1.** Performance metrics used.

| Name | Abbreviation | Equation | Perfect score |
|---|---|---|---|
| Root Mean Square Error | RMSE | $\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Z_i-\hat{Z}_i\right)^2}$ | 0.0 |
| Mean Absolute Error | MAE | $\frac{1}{n}\sum_{i=1}^{n}\left\|Z_i-\hat{Z}_i\right\|$ | 0.0 |
| Absolute BIAS | ABIAS | $\frac{\sum_{i=1}^{n}\left\|Z_i-\hat{Z}_i\right\|}{\sum_{i=1}^{n}z_i}$ | 0.0 |
| Pearson Correlation | r | $\frac{\sum_{i=1}^{n}\left(Z_i-\overline{Z}\right)\left(\hat{Z}_i-\underline{\hat{Z}}\right)}{\sqrt{\sum_{i=1}^{n}\left(Z_i-\overline{Z}\right)^2\sum_{i=1}^{n}\left(\hat{z}_i-\underline{\hat{z}}\right)^2}}$ | 1.0 |

where, $z_i$ is the original observed value of monthly rainfall in month i, $\hat{Z}_i$ is the estimated value of monthly rainfall in month *i*, $\overline{Z}$ is the average of observed rainfall, $\underline{\hat{z}}$ is the average of rainfall estimated, and *n* is the number of observations (months).

**Table 2.** Analysis of descriptive statistics of the monthly rainfall in the gauge stations of Nariño (1983-2019).

| Gauge Station | ID | Region* | Elevation (m.a.s.l) | Rainfall mean (mm.year$^{-1}$) | Rainfall std dev (mm.year$^{-1}$) | Rainfall CV (%) | SC* | Missing data (%) |
|---|---|---|---|---|---|---|---|---|
| A. Nariño | AAN | AR | 1796 | 1191.8 | 270.07 | 23 | 0.64 | 1.13 |
| A. San Luis | ASL | AR | 2961 | 875.4 | 169.40 | 19 | 0.30 | 0.68 |
| Aponte | APO | AR | 1800 | 1541.3 | 414.85 | 27 | 0.61 | 0.68 |
| Barbacoas | BAR | PR | 32 | 6714.9 | 969.90 | 14 | 0.31 | 3.60 |
| Berruecos | BER | AR | 2200 | 1736.1 | 342.65 | 20 | 0.28 | 3.83 |
| Bombona | BOM | AR | 1493 | 1040.6 | 217.91 | 21 | 0.42 | 1.35 |
| Botana | BOT | AR | 2820 | 927.6 | 175.11 | 19 | 0.55 | 3.60 |
| Buesaco | BUE | AR | 220 | 1269.7 | 369.11 | 29 | 0.36 | 1.80 |
| Chiles | CHI | AR | 3100 | 1091.0 | 231.66 | 21 | -0.13 | 1.80 |
| Coco | COC | PR | 20 | 2680.1 | 974.36 | 36 | 0.40 | 3.38 |
| Cumbal | CUM | AR | 392 | 891.5 | 176.40 | 20 | 0.68 | 0.23 |
| El Charco | CHA | PR | 50 | 3604.4 | 609.77 | 17 | 0.21 | 7.66 |
| Guachavéz | GCH | AR | 2834 | 1648.1 | 327.70 | 20 | 0.10 | 0.90 |
| Gualmatán | GMT | AR | 2830 | 942.2 | 296.59 | 31 | 3.00 | 2.03 |
| Guasca | GCA | AR | 500 | 584.3 | 185.89 | 32 | 0.42 | 8.11 |
| Guayacana | GYA | AR | 100 | 6056.7 | 864.41 | 14 | -0.38 | 6.98 |
| Hidromayo | HID | AR | 1820 | 1332.4 | 328.46 | 25 | 0.56 | 1.58 |
| Imués | IMU | AR | 2550 | 999.6 | 229.43 | 23 | 0.14 | 1.13 |
| José Tapaje | JOS | PR | 80 | 4762.9 | 1242.69 | 26 | -0.55 | 2.93 |
| Junín | JUN | PR | 950 | 8689.3 | 1048.47 | 12 | 0.33 | 2.70 |
| La Cruz | CRU | AR | 2248 | 1343.3 | 331.75 | 25 | 0.79 | 0.45 |
| Magüí | MAG | PR | 100 | 4893.1 | 1347.60 | 28 | -0.26 | 6.31 |
| Mamaconde | MAM | AR | 650 | 1325.9 | 392.22 | 30 | 0.12 | 1.80 |
| Mataje | MAT | PR | 100 | 3458.1 | 852.29 | 25 | 0.56 | 9.46 |
| Mira | MIR | PR | 16 | 2988.4 | 699.15 | 23 | 0.16 | 1.80 |
| Monopamba | MON | AMR | 1776 | 3188.3 | 293.71 | 9 | -0.79 | 3.38 |
| Mosquera | MOS | PR | 10 | 3620.3 | 822.46 | 23 | 0.43 | 4.05 |
| Nariño | NAR | AR | 2590 | 1985.7 | 488.99 | 25 | 0.40 | 1.35 |
| Obonuco | OBO | AR | 2710 | 806.3 | 250.73 | 31 | 1.83 | 7.21 |
| Paraíso | PAR | AR | 3120 | 995.6 | 192.79 | 19 | 0.11 | 0.68 |
| Peñol | PEÑ | AR | 1620 | 1100.1 | 239.15 | 22 | 0.34 | 0.23 |
| Pisanda | PIS | AR | 350 | 1253.0 | 299.73 | 24 | 0.95 | 0.00 |
| Puerres | PUE | AR | 2817 | 1025.0 | 174.11 | 17 | 0.77 | 0.23 |
| Remolino | REM | PR | 40 | 2797.2 | 1151.03 | 41 | 0.13 | 11.94 |
| Rio Bobo | RBB | AR | 364 | 1100.0 | 252.51 | 23 | 0.08 | 0.68 |
| Rosal Monte | RMO | AR | 2568 | 1346.4 | 332.48 | 25 | 0.48 | 0.90 |
| Salahonda | SAL | PR | 3 | 4800.9 | 1182.88 | 25 | -0.37 | 3.15 |
| Samaniego | SAM | AR | 1700 | 1346.4 | 480.47 | 33 | -0.19 | 0.00 |
| San Bernardo | SBO | AR | 2190 | 2014.9 | 370.66 | 18 | 0.25 | 1.58 |
| Sande | SND | AR | 840 | 4503.4 | 1309.49 | 29 | -0.72 | 13.06 |
| Sandoná | SAN | AR | 20 | 1142.3 | 364.61 | 32 | 0.20 | 1.58 |
| Taminango | TAM | AR | 1875 | 1687.2 | 329.08 | 20 | 0.49 | 0.45 |
| Tanama | TAN | AR | 1500 | 1351.9 | 252.53 | 19 | 0.14 | 0.45 |
| Tangua | TGA | AR | 2420 | 1004.1 | 232.64 | 23 | 0.08 | 0.45 |
| La Unión | UNI | AR | 1745 | 1982.9 | 441.05 | 22 | -0.03 | 1.13 |
| Vergel | VER | PR | 1770 | 2514.6 | 456.44 | 18 | 0.24 | 3.38 |

PR: Pacific Region, AR: Andean Region, AMR: Amazon Region. *SC: Skewness Coefficient.

2500-8689 mm.year$^{-1}$ (See Figure 4), with a monomodal regime where the highest (lowest) rainfall is recorded from April to July (August to March) (See Figure 4a). On the other hand, the gauge stations located in AR show a variation of average rainfall between 600 to 1500 mm.year$^{-1}$, showing a bimodal cycle (see Figure 4b), where there are two periods of high rainfall corresponding to March-April-May (MAM) and October-November-December (OND). Finally, the MON station was classified in the AMR, with rainfall between 2500-4500 mm.year$^{-1}$, registering a monomodal cycle (See Figure 4c) with maximum rainfall in the months of June-July-August (JJA). These results are consistent with findings reported by Canchala et al. (2019), Canchala et al. (2020a), and Cerón et al.

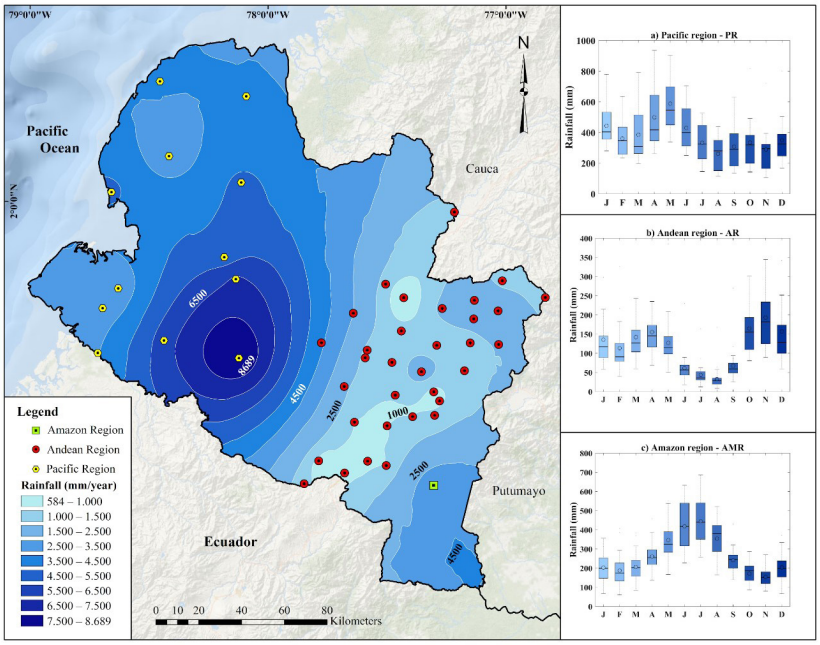


**Figure 4.** Average annual rainfall in Nariño (1983-2019) and rainfall regime for the three identified natural regions a) Pacific Region, b) Andean Region, and c) Amazon Region.

(2021b), who studied the rainfall variability in the department of Nariño associated with ENSO and other climatic anomalies in the period 1983-2016 and identified in their studies the PR and AR, as two homogeneous regions with different rainfall regimes. In contrast, Arango et al. (2012) and Guzmán et al. (2014) identified three regions: PR, AR and AMR in the study area, with different interannual patterns consistent with those identified in this study.

## Gauge stations correlations

According to the correlation criterion for selecting stations ($\varrho \geq 0.5$) described in the methodology for the NR method and according to the results obtained, nine neighbouring stations were selected, homogeneously distributed in the three regions. For the PR, the BAR, MIR, and RBB stations were selected; for the AR, the PIS, RMO, and SAM were selected; and for the AMR, which only has one gauge station (MON), three neighbouring stations of the department of Putumayo CPC, CHP, and TTS (located to the east of the study region) were used. More details about the correlation results for PR, AR and AMR are depicted in Supplementary Materials S1, S2, and S3, respectively.

## Missing data estimation

The missing data estimation results using the NR, PCR, PLSR, and ANN methods in the PR, AR, and AMR regions are shown in Figures 5, 6, and 7, respectively. For purposes of comparative analysis, the results obtained at the gauge station with the highest percentage of missing data for each region are depicted in this study. In this sense, the gauge stations selected for PR, AR and AMR were Mataje (MAT-9.46%), Guasca (GCA-8.11%) and Monopamba (MON-3.38%), respectively. The results of the missing data estimation of all gauge stations are available in Supplementary Material S4.

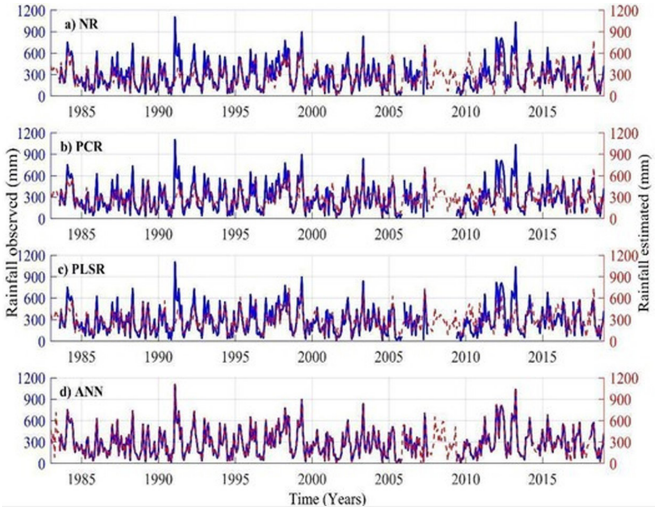


**Figure 5.** Comparison of time series between observed and estimated rainfall for the methods evaluated at the MAT gauge station (PR). The blue line indicates observed rainfall, and the red line indicates the estimated rainfall.

Figure 5 shows the missing data estimation obtained for MAT gauge station. For the NR (Figure 5a), PCR (Figure 5b), and PLSR (Figure 5c) methods, underestimated values were obtained in the periods between the years 1986-1988, 1991-1994, 1997-1999, and 2013-2014, highlighting that in 1998 the observed value was underestimated by more than 60%. On the other hand,

overestimates of 33% (38%) were obtained for the 2010-2011 (2016-2019) period. These periods mostly coincide with events related to typical increases and decreases in rainfall in the study area, mainly associated with the ENSO phenomenon described by Trenberth (2018), which in turn are consistent with the results obtained in the missing data estimation in southwestern Colombia by Canchala et al. (2020a), and Ocampo-Marulanda et al. (2022).

The results of the missing data estimation in PR show that these three methods are sensitive to rainfall variability. In contrast, the missing data estimation using the ANN method shows that the reconstructed time series for MAT gauge station does not record high underestimates and overestimates (Figure 5d); instead, the time series estimated follows the behaviour of the observed values. This result shows that the ANN method has a high capacity to reconstruct historical data despite the high rainfall variability in the region. This shows agreement with the results obtained by Santos et al. (2021), Chiu et al. (2021), Canchala et al. (2019), Khalili et al. (2016), and Miró et al. (2017), who used ANN models to estimate missing data of rainfall time series in areas with high variability and obtained low estimation errors, indicating a high capacity in the reconstruction of time series in areas with climatic complexity.

Figure 6 shows the rainfall missing data estimation of the GCA gauge station of the AR. In general, the time series presented high variations in rainfall, highlighting high rainfall in the years 1985, 2002, 2004, and 2010, where rainfall was greater than 200 mm.month$^{-1}$. Particularly in the year 1985, the NR (Figure 6a), PCA (Figure 6b) and PLSR (Figure 6c) methods registered underestimates of 65%, 70%, and 63%, respectively. In contrast, the ANN method (Figure 6d) registered less underestimation (26%), showing superiority in its performance over the other methods.

Figure 7 shows the rainfall missing data estimation for the MON gauge station of the AMR, where maximum rainfall events are observed in the years 1986, 1988, 1992, 2002, 2011, 2013, and 2015, which were underestimated with the methods PCR (Figure 7b) and PLSR (Figure 7c). For its part, the estimation made by NR (Figure 7a), registered a better fit compared to PCR and PLSR, showing that, for some periods, this method was superior to the two linear methods; however, the estimates made with ANN (Figure 7d) showed high precision and accuracy in the reconstruction of the series. For example, the maximum event recorded in July 1986 was underestimated by 6%, 51%, 53%, and 0% with the NR, PCR, PLSR, and ANN methods, respectively.

Broadly, the best rainfall missing data estimation in the gauge stations of the three regions PR, AR and AMR was obtained with the nonlinear ANN method. This method showed superiority and high capacity for time series reconstruction even when high rainfall variability is recorded, which can reduce the performance of the methods. The results of the linear methods NR, PCR, and PLSR, were not the best; however, it is highlighted that in some periods the estimations depicted high similarity with the observed values; being consistent with the results obtained by Silva et al. (2007) in some regions of Sri Lanka; Pizarro et al. (2009) in the Maulé region of Chile; Cruz-Roa & Barrios (2018) in the Coello river basin, in the department of Tolima, Colombia; Morales-Martínez et al. (2019) in Tabasco, Mexico; Armanuos et al. (2020) in Ethiopia and Adilah & Hannani (2021) in the state of Pahang, Malaysia, in

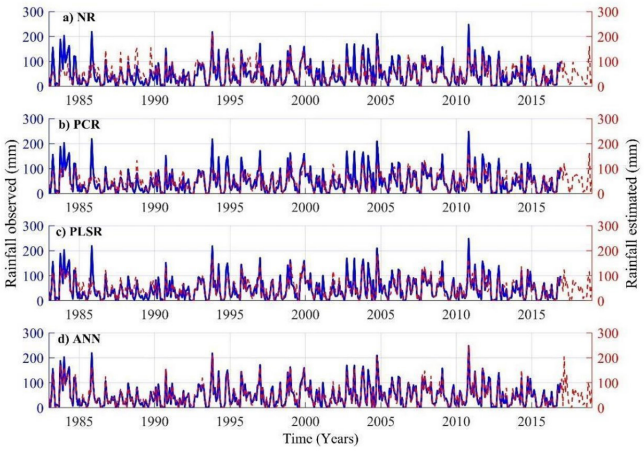


**Figure 6.** Comparison of time series between observed and estimated rainfall for the methods evaluated at the GCA gauge station (AR). The blue line indicates observed rainfall, and the red line indicates the estimated rainfall.

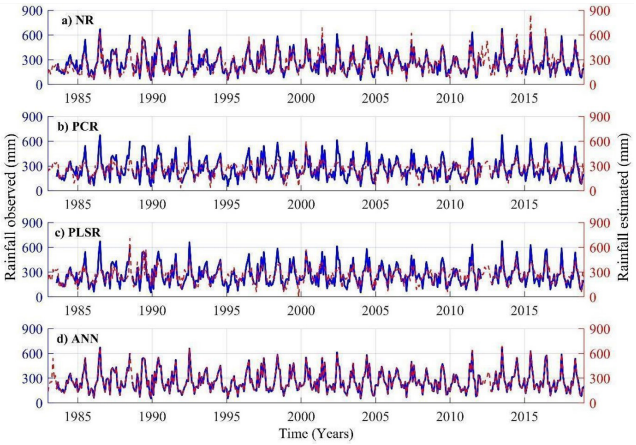


**Figure 7.** Comparison of time series between observed and estimated rainfall for the methods evaluated at the MON station (AMR).

which the NR obtained a good fit in specific periods of the time series evaluated, and in some cases, throughout the study period.

## Performance metrics

The performance metrics described in the methodology section were used to quantify the performance of each of the missing data estimation methods evaluated in this research. The standardized performance metrics are shown in Figure 8, which range between 0.0 and 0.75, where the best (worst) performance is identified with values close to 0.0 (0.75). According to the three-error metrics, RMSE, MAE, and ABIAS, NR was the method with the largest errors, which ranged from 0.24-0.62, 0.23-0.44, and 0.18-0.44, respectively. In contrast, the ANN method depicted superiority

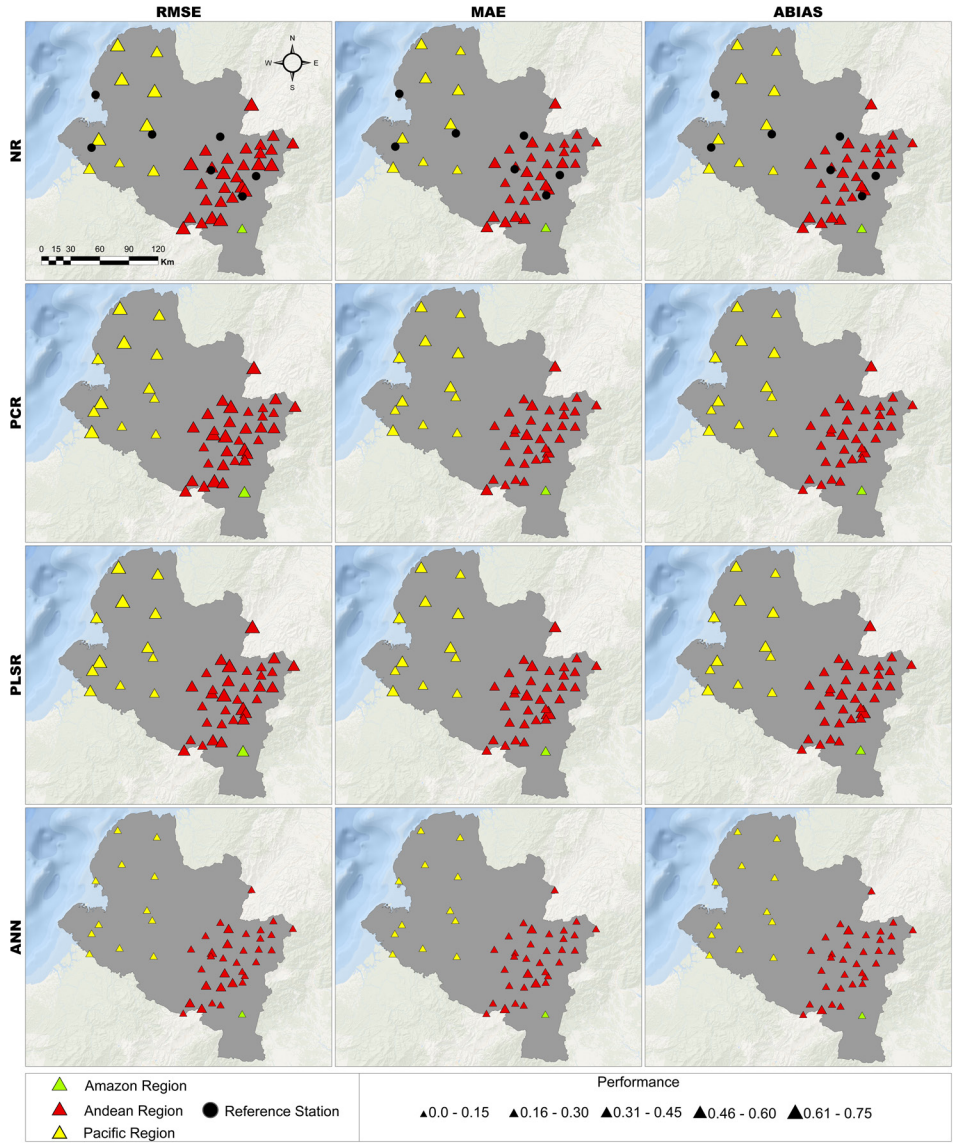


**Figure 8.** Performance metrics between observed and estimated rainfall using NR, PCR, PLSR, and ANN. The reference stations correspond to the neighbouring stations used for the NR method.

in the imputation of missing over the other methods registering errors that ranged between 0.01-0.29 (RMSE), 0.01-0.22 (MAE), and 0.01-0.22 (ABIAS). On the other hand, it was evidenced that the estimated errors using NR, PCR, and PLSR were higher in the PR than in the AR and AMR, while using the ANN method, the lowest errors in their order were registered in the PR (0.01-0.03), AR (0.01-0.29) and AMR (0.02). For the PCR method, the error metrics ranged between 0.25-0.59 (RMSE), 0.18-0.41 (MAE) and 0.18-0.41 (ABIAS) and for the PLSR, the error metrics ranged 0.22-0.54 (RMSE), 0.16-0.48 (MAE) and 0.18-0.41 (ABIAS).

In general, the best method for rainfall missing data estimation in Southwestern Colombia is ANN, due to registering the lowest errors in the three natural regions of Nariño. It is highlighted that the best performance was recorded in the reconstruction of the rainfall time series of the PR, which registers high variability (See Table 2) and requires greater precision in their estimation, considering that it is a region with low-density of gauge stations, and therefore with lack hydroclimatological information. This result contrasts with Canchala et al. (2019). They obtained higher RMSE values in the PR and lower values in the AR of southwestern Colombia, highlighting differences in the study period and the number of gauge stations evaluated. Regarding the three linear methods, the best performance was registered by the PLSR, followed by the PCR and the NR.

Finally, to measure the statistical relationship between the observed and the estimated time series using the four methods already described, the Pearson correlation coefficient was estimated (See Figure 9).

The correlation coefficients confirmed the results obtained through the performance metrics. The Pearson coefficient for the estimation performed by the ANN method ranged between 0.81 - 1.00, showing a high correlation between the estimated and observed values. In contrast, it was observed that the lowest correlations were recorded in the estimates of the NR method, highlighting

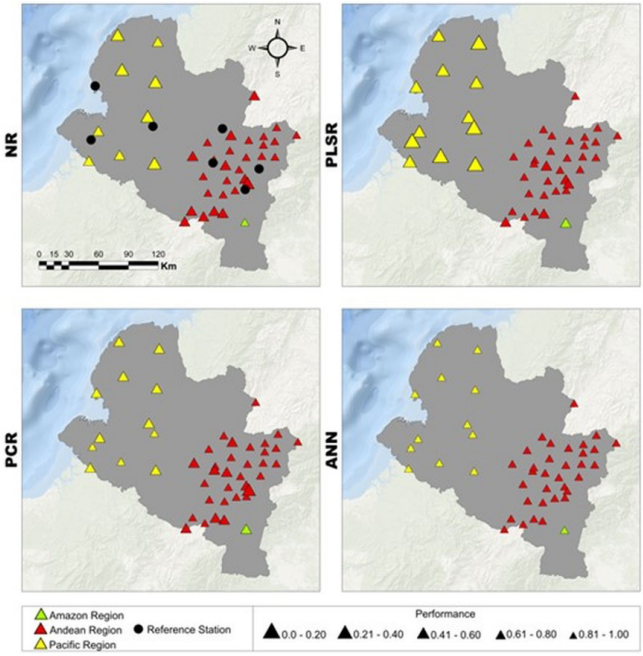


**Figure 9.** Pearson correlations between observed and estimated rainfall using NR, PCR, PLSR, and ANN.

that the lowest coefficients are observed in the PR, where there is a high variability of rainfall (See Table 2). Furthermore, using the three linear methods, the best estimations were obtained in the AR. In contrast, with the nonlinear method (ANN), the best estimates were recorded in the PR, showing a high capacity for recognising the nonlinear relationships that allow overcoming difficulties associated with the high variability, scarce information, and low density of stations in this region.

The results obtained in this study are consistent with Miró et al. (2017), who evaluated ten missing data imputation methods and reported that the best results were obtained with nonlinear methods, highlighting the methods based on the NLPCA approach. Additionally, it is consistent with Canchala et al. (2019) and Ocampo-Marulanda et al. (2021). They used methodologies based on the NLPCA approach to monthly rainfall missing data estimation and extreme rainfall indices, registering low errors in their estimation. On the other hand, this research shows that methods based on artificial neural networks have a higher capacity to fill in missing data than linear inference methods, as shown in the studies made by Kim & Pachepsky, (2010) in Chesapeake Bay, United States; Teegavarapu (2012) in Kentucky, United States; Londhe et al. (2015) in Pune district, Maharashtra, India; Miró et al. (2017) in the Iberian Peninsula, Spain; Demir & Keskin (2021) in Turkey and Khalili et al. (2016) in Mashhad, Iran. Finally, it is highlighted that among the linear methods evaluated (PLSR and PCR), they show good performance and are easy to apply, being very useful when knowledge of artificial intelligence is unavailable.

## CONCLUSIONS

The following conclusions are reached in the study:

The ANN is the best method for the monthly rainfall missing data estimation in the AR, PR, and AMR of Southwestern Colombia due to this nonlinear method shows a high capacity to identify atypical patterns and reconstruct time series without the need to implement auxiliary variables such as altitude, geographical position, among others. This result is relevant for the study area, characterized by recording complex topographical conditions due to the Andes Mountains, the influence of the ENSO climate variability mode that occurs in the Pacific Ocean, the proximity to the Amazon Forest, and other conditions that influence the variability of the rainfall.

Among the linear methods evaluated, PLSR was the method that registered the best performance. However, the results were similar to those obtained by the PCR method. For both methods, the best results were obtained in the AR and AMR, contrary to the PR, which showed the highest errors and low correlation rates. Furthermore, it is highlighted that for the PLSR, a smaller number of CPs were used, which explained a significant percentage of the variance of the data. This aspect represents an advantage over PCR.

Using the NR method, a good performance was obtained in the time series reconstruction, very similar to that obtained in some gauge stations by the PCR and PLSR methods, showing synchrony between the observed and estimated data in extreme periods of low and high rainfall linked to ENSO Phenomenon. This is a positive aspect because it is a conventional method, easy to apply and does not require specific software knowledge; however, it cannot be used in areas without neighbouring stations.

Finally, we found that in areas with a lower density of gauge stations, such as PR (12-gauge stations) and AMR (1 gauge station), the performance of the ANN method was better compared to AR (33 stations), where a lower performance was observed. According to the results obtained in the metrics used, contrary to the NR, PCR, and PLSR methods, which presented better results in AR. Notably, the ANN method recognized the different rainfall patterns in the evaluated time series that are associated with nonlinear behaviour, which improved the quality of the estimates.

## ACKNOWLEDGEMENTS

This work was supported by the Universidad del Valle through the research Projects CI 0521171 and CI 21010, and MINCIENCIAS for funding the research project "Análisis de eventos extremos de precipitación asociados a variabilidad y cambio climático para la implementación de estrategias de adaptación en sistemas productivos agrícolas de Nariño". The second author has received research support from Fondo Nacional de Financiamiento para la Ciencia, la Tecnología y la Innovación Francisco José de Caldas – MINCIENCIAS through the "Convocatoria No 891 de 2020 para el fortalecimiento de vocaciones y formación en CTeI para la reactivación económica en el marco de la postpandemia 2020". The authors thank the Research Group in Water Resources Engineering and Soil – IREHISA Research Group in Applied Statistics - INFERIR, at the Universidad del Valle for their contributions in this research paper. Finally, special thanks to Instituto de Hidrología, Meteorología y Estudios Ambientales - IDEAM for providing the database containing the monthly rainfall in the Department of Nariño.

## REFERENCES

Addi, M., Gyasi-Agyei, Y., Obuobie, E., & Amekudzi, L. K. (2022). Evaluation of imputation techniques for infilling missing daily rainfall records on river basins in Ghana. *Hydrological Sciences Journal*, *67*(4), 613-627. http://dx.doi.org/10.1080/02626667.2022.2030868.

Adilah, N., & Hannani, H. (2021). Comparison of methods to estimate missing rainfall data for short term period at UMP gambang. *IOP Conference Series. Earth and Environmental Science*, *682*(1), 012027. http://dx.doi.org/10.1088/1755-1315/682/1/012027.

Andersson, M. (2009). A comparison of nine PLS1 algorithms. *Journal of Chemometrics*, *23*(10), 518-529. http://dx.doi.org/10.1002/cem.1248.

Arango, C., Dorado, J., Guzmán, D., & Ruiz, J. (2012). Climatología trimestral de Colombia. Instituto de Hidrología. *Meteorología y Estudios Ambientales*, *1*(1), 19.

Armanuos, A. M., Al-Ansari, N., & Yaseen, Z. M. (2020). Cross assessment of twenty-one different methods for missing precipitation data estimation. *Atmosphere*, *11*(4), 1-35. http://dx.doi.org/10.3390/atmos11040389.

Auer, I., Böhm, R., Jurković, A., Orlik, A., Potzmann, R., Schöner, W., Ungersböck, M., Brunetti, M., Nanni, T., Maugeri, M., Briffa, K., Jones, P., Efthymiadis, D., Mestre, O., Moisselin, J. M., Begert, M., Brazdil, R., Bochnicek, O., Cegnar, T., Gajić-Čapka, M., Zaninović, K., Majstorović, Ž., Szalai, S., Szentimrey, T., & Mercalli, L. (2005). A new instrumental precipitation dataset for the greater alpine region for the period 1800-2002. *International Journal of Climatology*, *25*(2), 139-166. http://dx.doi.org/10.1002/joc.1135.

Bárdossy, A., & Pegram, G. (2011). Downscaling precipitation using regional climate models and circulation patterns toward hydrology. *Water Resources Research*, *47*(4), http://dx.doi.org/10.1029/2010WR009689.

Burhanuddin, S. N. Z. A., Deni, S. M., & Ramli, N. M. (2017). Imputation of missing rainfall data using revised normal ratio method. *Advanced Science Letters*, *23*(11), 10981-10985. http://dx.doi.org/10.1166/asl.2017.10203.

Burhanuddin, S. N. Z. A., Mohd Deni, S., & Mohamed Ramli, N. (2016). Revised Normal Rtio Methods for Imputation of Missing Rainfall Data. *Scientific Research Journal*, *13*(1), 84-97.

Caldera, H., Piyathisse, V., & Nandalal, K. (2016). A comparison of methods of estimating missing daily rainfall data. *Engineer: Journal of the Institution of Engineers*, *4*(49), 1-8. http://dx.doi.org/10.4038/engineer.v49i4.7232.

Canchala, T., Alfonso-Morales, W., Carvajal-Escobar, Y., Cerón, W. L., & Caicedo-Bravo, E. (2020a). Monthly rainfall anomalies forecasting for southwestern Colombia using artificial neural networks approaches. *Water*, *12*(9), 2628. http://dx.doi.org/10.3390/w12092628.

Canchala, T., Alfonso-Morales, W., Cerón, W. L., Carvajal-Escobar, Y., & Caicedo-Bravo, E. (2020b). Teleconnections between monthly rainfall variability and large-scale climate indices in Southwestern Colombia. *Water*, *12*(7), 1-20. http://dx.doi.org/10.3390/w12071863.

Canchala, T., Carvajal-Escobar, Y., Alfonso-Morales, W., Loaiza Cerón, W., & Caicedo, E. (2019). Estimation of missing data of monthly rainfall in southwestern Colombia using artificial neural networks. *Data in Brief*, *26*, 104517. http://dx.doi.org/10.1016/j.dib.2019.104517.

Canchala, T., Cerón, W. L., Francés, F., Carvajal-Escobar, Y., Andreoli, R. V., Kayano, M. T., Alfonso-Morales, W., Caicedo-Bravo, E., & Souza, R. A. F. (2020c). Streamflow variability in colombian pacific basins and their teleconnections with climate indices. *Water*, *12*(2), 526. http://dx.doi.org/10.3390/w12020526.

Canchala, T., Ocampo-Marulanda, C., Alfonso-Morales, W., Carvajal-Escobar, Y., Ceron, W., & Caicedo-Bravo, E. (2022). Techniques for monthly rainfall regionalization in southwestern Colombia. *Anais da Academia Brasileira de Ciências*, *94*(4), 48. http://dx.doi.org/10.1590/0001-3765202220201000.

Carvajal-Escobar, Y., & Marco, J. (2005). Caudal Mensual Utilizando Variables. *Ingeniería y Competitividad*, *7*(1), 15.

Castro, L. M., Carvajal-Escobar, Y., & Ávila, Á. J. (2012). Análisis clúster como técnica de análisis exploratorio de registros múltiples en datos meteorológicos. *Ingeniería de Recursos Naturales y Del Ambiente*, *1*(11), 11-20.

Cerón, W. L., Andreoli, R. V., Kayano, M. T., Canchala, T., Carvajal-Escobar, Y., & Souza, R. A. F. (2021a). Comparison of spatial interpolation methods for annual and seasonal rainfall in two hotspots of biodiversity in South America. *Anais da Academia Brasileira de Ciências*, *93*(1), 1-22. http://dx.doi.org/10.1590/0001-3765202120190674.

Cerón, W. L., Kayano, M. T., Andreoli, R. V., Canchala, T., Carvajal-Escobar, Y., & Alfonso-Morales, W. (2021b). Rainfall variability

in southwestern Colombia: changes in ENSO-related features. *Pure and Applied Geophysics*, *178*(3), 1087-1103. http://dx.doi.org/10.1007/s00024-021-02673-7.

Chiu, P. C., Selamat, A., Krejcar, O., Kuok, K. K., Herrera-Viedma, E., & Fenza, G. (2021). Imputation of rainfall data using the sine cosine function fitting neural network. *International Journal of Interactive Multimedia and Artificial Intelligence*, *6*(7), 39-48. http://dx.doi.org/10.9781/ijimai.2021.08.013.

Cruz-Roa, A. F., & Barrios, M. I. (2018). Estimación de datos faltantes de lluvia mensual a través de la asimilación de información satelital y pluviométrica en una cuenca andina tropical. *Idesia*, *36*(3), 107-117. http://dx.doi.org/10.4067/S0718-34292018005001601.

Cuadras, C. M. (2007). *Nuevos metodos de analisis multivariante* (Vol. 20, No. 3). Barcelona: CMC Editions.

Darand, M., & Reza, M. (2014). Regionalization of precipitation regimes in iran using principal component analysis and hierarchical clustering analysis. *Environmental Processes*, *1*(4), 517-532. http://dx.doi.org/10.1007/s40710-014-0039-1.

DeGaetano, A. T., & Allen, R. J. (2002). Trends in twentieth-century temperature extremes across the United States. *Journal of Climate*, *15*(22), 3188-3205. http://dx.doi.org/10.1175/1520-0442(2002)015<3188:TITCTE>2.0.CO;2.

Demir, C., & Keskin, S. (2021). Artificial neural network approach for nonlinear principal components analysis. *International Journal of Current Research*, *13*(1), 15987-15992.

Domonkos, P. (2015). Homogenization of precipitation time series with ACMANT. *Theoretical and Applied Climatology*, *122*(1–2), 303-314. http://dx.doi.org/10.1007/s00704-014-1298-5.

Fazel, N., Berndtsson, R., Bertacchi, C., Madani, K., & Kløve, B. (2018). Regionalization of precipitation characteristics in Iran's Lake Urmia basin. *Theoretical and Applied Climatology*, *132*(1-2), 363-373. http://dx.doi.org/10.1007/s00704-017-2090-0.

Francisco, C.-A. D. (2015). Estimación simultánea de datos hidrológicos anuales faltantes en múltiples sitios. *Ingeniería, Investigación y Tecnología*, *16*(2), 295-306. http://dx.doi.org/10.1016/j.riit.2015.03.013.

Gois, G., Oliveira-Júnior, J. F., Silva, C. A., Serafini, B., De Bodas, P. M., & Sousa, A. H. (2020). Statistical normality and homogeneity of a 71-year rainfall dataset for the state of Rio de Janeiro-Brazil. *Theoretical and Applied Climatology*, *141*(3-4), 1573-1591. http://dx.doi.org/10.1007/s00704-020-03270-9.

Guzmán, D., Ruíz, J., & Cadena, M. (2014). Regionalización de Colombia según la estacionalidad de la precipitación media mensual, através de Componentes Principales (ACP). Instituto de Hidrología. *Meteorología y Estudios Ambientales*, *1*, 1-54.

Hershey, R., Mizell, S., & Earman, S. (2010). Chemical and physical characteristics of springs discharging from regional low systems of the carbonate-rock province of the Great Basin, western United States. *Hydrogeology Journal*, *18*(4), 1007-1026. http://dx.doi.org/10.1007/s10040-009-0571-7.

Hervada-Sala, C., & Jarauta-Bragulat, E. (2004). A program to perform Ward's clustering method on several regionalized variables. *Computers & Geosciences*, *30*(8), 881-886. http://dx.doi.org/10.1016/j.cageo.2004.07.003.

Intergovernmental Panel on Climate Change – IPCC. (2022). *Climate change 2022: impacts, adaptation and vulnerability - summary for policymakers*. Geneva: IPCC.

Ismail, A. R., Aziz, N. A., Ralib, A. M., Abidin, N. Z., & Bashath, S. S. (2021). A particle swarm optimization levy flight algorithm for imputation of missing creatinine dataset. *International Journal of Advances in Intelligent Informatics*, *7*(2), 225-236. http://dx.doi.org/10.26555/ijain.v7i2.677.

Jaramillo-Robledo, Á., & Chaves-Córdoba, B. (2000). Distribución de la precipitación en Colombia analizada mediante conglomeración estadística. *Cenicafé*, *51*(2), 102-113.

Kajornrit, J., Wong, K. W., & Fung, C. C. (2012). Estimation of missing precipitation records using modular artificial neural networks. In: *Proceedings of the 19th international conference on Neural Information Processing* (pp. 52-59). Berlin: Springer.

Khalili, N., Khodashenas, S. R., Davary, K., Baygi, M. M., & Karimaldini, F. (2016). Prediction of rainfall using artificial neural networks for synoptic station of Mashhad: a case study. *Arabian Journal of Geosciences*, *9*(13), 624. http://dx.doi.org/10.1007/s12517-016-2633-1.

Kim, J. W., & Pachepsky, Y. A. (2010). Reconstructing missing daily precipitation data using regression trees and artificial neural networks for SWAT streamflow simulation. *Journal of Hydrology (Amsterdam)*, *394*(3–4), 305-314. http://dx.doi.org/10.1016/j.jhydrol.2010.09.005.

Kuok, K. K., Harun, S., & Shamsuddin, S. M. (2010). Particle swarm optimization feedforward neural network for modeling runoff. *International Journal of Environmental Science and Technology*, *7*(1), 67-78. http://dx.doi.org/10.1007/BF03326118.

Lai, W. Y., Kuok, K. K., Gato-Trinidad, S., & Derrick, K. X. L. (2019). A study on sequential K-nearest neighbor (SKNN) imputation for treating missing rainfall data. *International Journal of Advanced Trends in Computer Science and Engineering*, *8*(3), 363-368. http://dx.doi.org/10.30534/ijatcse/2019/05832019.

Lee, H., & Kang, K. (2015). Interpolation of missing precipitation data using Kernel estimations for hydrologic modeling. *Advances in Meteorology*, *2015*, 1-12. http://dx.doi.org/10.1155/2015/935868.

Londhe, S., Dixit, P., Shah, S., & Narkhede, S. (2015). Infilling of missing daily rainfall records using artificial neural network. *ISH Journal of Hydraulic Engineering*, *21*(3), 255-264. http://dx.doi.org/10.1080/09715010.2015.1016126.

Lu, B., & Hsieh, W. W. (2003). Simplified nonlinear principal component analysis. In *Proceedings of the International Joint Conference on Neural Networks* (pp. 759-763). New York: IEEE. http://dx.doi.org/10.1109/IJCNN.2003.1223477.

Massy, W. (1965). Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, *60*(309), 234-256. http://dx.doi.org/10.1080/01621459.1965.10480787.

Miró, J. J., Caselles, V., & Estrela, M. J. (2017). Multiple imputation of rainfall missing data in the Iberian Mediterranean context. *Atmospheric Research*, *197*, 313-330. http://dx.doi.org/10.1016/j.atmosres.2017.07.016.

Miró, J. J., Estrela, M. J., Caselles, V., & Gómez, I. (2018). Spatial and temporal rainfall changes in the Júcar and Segura basins (1955-2016): fine-scale trends. *International Journal of Climatology*, *38*(13), 4699-4722. http://dx.doi.org/10.1002/joc.5689.

Moraes Cordeiro, A. L., & Blanco, C. J. C. (2021). Assessment of satellite products for filling rainfall data gaps in the Amazon region. *Natural Resource Modeling*, *34*(2), http://dx.doi.org/10.1111/nrm.12298.

Morales Martínez, J. L., Horta-Rangel, F. A., Segovia-Domínguez, I., Robles Morua, A., & Hernández, J. H. (2019). Analysis of a new spatial interpolation weighting method to estimate missing data applied to rainfall records. *Atmósfera*, *32*(3), 237-259. http://dx.doi.org/10.20937/ATM.2019.32.03.06.

Morales-Acuña, E., Linero-Cueto, J. R., & Canales, F. A. (2021). Assessment of precipitation variability and trends based on satellite estimations for a heterogeneous Colombian region. *Hydrology*, *8*(3), 1-20. http://dx.doi.org/10.3390/hydrology8030128.

Ocampo-Marulanda, C., Cerón, W. L., Avila-Diaz, A., Canchala, T., Alfonso-Morales, W., Kayano, M. T., & Torres, R. R. (2021). Missing data estimation in extreme rainfall indices for the Metropolitan area of Cali - Colombia: an approach based on artificial neural networks. *Data in Brief*, *39*, 107592. http://dx.doi.org/10.1016/j.dib.2021.107592.

Ocampo-Marulanda, C., Fernández-Álvarez, C., Cerón, W. L., Canchala, T., Carvajal-Escobar, Y., & Alfonso-Morales, W. (2022). A spatiotemporal assessment of the high-resolution CHIRPS rainfall dataset in southwestern Colombia using combined principal component analysis. *Ain Shams Engineering Journal*, *13*(5), 101739. http://dx.doi.org/10.1016/j.asej.2022.101739.

Paulhus, J., & Kohler, M. (1952). Monthly weather review. *Monthly Weather Review*, *80*(8), 129-133. http://dx.doi.org/10.1175/1520-0493(1952)080<0129:IOMPR>2.0.CO;2.

Pinheiro, E., Cavalcante, C. J., Silva, P., & Oliveira Júnior, F. (2022). MODWT-ANN hybrid models for daily precipitation estimates with time-delayed entries in Amazon region. *Environmental Monitoring and Assessment*, *194*(296), 296. http://dx.doi.org/10.1007/s10661-022-09939-0.

Pizarro, R., Ausensi, P., Aravena, D., Sangüesa, C., León, L., & Balocchi, F. (2009). Evaluación de métodos hidrológicos para la completación de datos faltantes de precipitación en estaciones de la región del Maule, Chile. *Aqua-LAC*, *1*(2), 172-185. http://dx.doi.org/10.29104/phi-aqualac/2009-v1-2-07.

Poveda, G., & Mesa, O. J. (1999). La corriente de chorro superficial del oeste ("del Chocó") y otras dos corrientes de chorro en Colombia: climatología y variabilidad durante las fases del ENSO. *Revista Academia Colombiana de Ciencias de la Tierra*, *23*(89), 517-528.

Poveda, G., & Mesa, O. J. (2000). On the existence of Lloró (the rainiest locality on earth): enhanced ocean-land-atmosphere interaction by a low-level jet. *Geophysical Research Letters*, *27*(11), 1675-1678. http://dx.doi.org/10.1029/1999GL006091.

Puertas, O., & Carvajal, Y. (2008). Incidencia de El Niño-Oscilación del Sur en la precipitación y la temperatura del aire en Colombia, utilizando el Climate Explorer. *Revista Científica Ingeniería y Desarrollo*, *23*, 104-118.

Ramos-Calzado, P., Gómez-Camacho, J., Pérez-Bernal, F., & Pita-López, M. F. (2008). A novel approach to precipitation series completion in climatological datasets: application to Andalusia. *International Journal of Climatology*, *28*(11), 1525-1534. http://dx.doi.org/10.1002/joc.1657.

Rueda, O. A., & Poveda, G. (2006). Variabilidad espacial y temporal del chorro del Chocó y su efecto en la hidroclimatología del Pacífico Colombiano. *Meteorología Colombiana*, *10*, 132-145.

Santos, E. B., Lucio, P. S., & Silva, C. M. S. (2015). Precipitation regionalization of the Brazilian Amazon. *Atmospheric Science Letters*, *16*(3), 185-192. http://dx.doi.org/10.1002/asl2.535.

Santos, E. P., Dias, R. L. S., Maciel, I. P., Kolling Neto, A., & Silva, D. D. (2021). Estimation of missing hydrological data in monthly rainfall series using meteorological satellite data. *Environmental Earth Sciences*, *80*(3), 1-9. http://dx.doi.org/10.1007/s12665-021-09409-9.

Scholz, M. (2023). *Nonlinear PCA*. Retrieved in 2023, January 20, from http://www.nlpca.org/matlab.html

Scholz, M., Kaplan, F., Guy, C. L., Kopka, J., & Selbig, J. (2005). Nonlinear PCA: a missing data approach. *Bioinformatics*, *21*(20), 3887-3895. http://dx.doi.org/10.1093/bioinformatics/bti634.

Sedano-Cruz, K., Carvajal-Escoar, Y., & Ávila, A. (2013). Análisis de aspectos que incrementan el riesgo de inundaciones en Colombia. *Luna Azul*, *1*(37), 219-238.

Serna, L. M., Arias, P. A., & Vieira, S. C. (2018). Las corrientes superficiales de chorro del Chocó y el Caribe durante los eventos de El Niño y El Niño Modoki. *Revista de la Academia Colombiana de Ciencias Exactas, Físicas y Naturales*, *42*(165), 410. http://dx.doi.org/10.18257/raccefyn.705.

Shahrokhi, Z., Sohrabi, M. R., & Nik, S. M. (2020). The application of artificial intelligence system and regression methods based on

the spectrophotometric method for fast simultaneous determination of naphazoline and antazoline in ophthalmic formulation. *Optik*, *203*, 164010. http://dx.doi.org/10.1016/j.ijleo.2019.164010.

Shlens, J. (2014). A tutorial on principal component analysis. *arXiv*, *1404.1100*, 1-12.

Silva, M. (2020). Rainfall extremes and drought in Northeast Brazil and its relationship with El Niño–Southern Oscillation. *International Journal of Climatology*, *41*(S1), 1-25. http://dx.doi.org/10.1002/joc.6835.

Silva, R. P., Dayawansa, N. D. K., & Ratnasiri, M. D. (2007). A comparison of methods used in estimating missing rainfall data. *Journal of Agricultural Sciences*, *3*(2), 101. http://dx.doi.org/10.4038/jas.v3i2.8107.

Souza, C., & Leal, M. F. (2017). Analise comparativa de dados de precipitação gerados pelo "Climate Prediction Center – CPC" versus dados observados para o Sul do Brasil. *Revista Brasileira de Geografia Física*, *10*(4), 1180-1198. http://dx.doi.org/10.26848/rbgf.v10.4.p1180-1198.

Taghi, S., Rezazadeh-Joudi, A., & Kusiak, A. (2017). Assessment of different methods for estimation of missing data in precipitation studies. *Nordic Hydrology*, *48*(4), 1032-1044. http://dx.doi.org/10.2166/nh.2016.364.

Taylor, M., Losch, M., Wenzel, M., & Jens, S. (2013). On the sensitivity of field reconstruction and prediction using empirical orthogonal functions derived from Gappy data. *Journal of Climate*, *26*(22), 9194-9205. http://dx.doi.org/10.1175/JCLI-D-13-00089.1.

Teegavarapu, R. S. V. (2012). Spatial interpolation using nonlinear mathematical programming models for estimation of missing precipitation records. *Hydrological Sciences Journal*, *57*(3), 383-406. http://dx.doi.org/10.1080/02626667.2012.665994.

Torres, C. E. (2012). *Efecto de las ondas Madden-Julian en la precipitación sobre algunas regiones del territorio colombiano* (Maestría thesis). Universidad Nacional de Colombia, Bogotá D.C.

Torres, C., Coll, R., Oliveira, J., Gois, G., & Sarmento, A. (2015). Avaliação das Estimativas de Precipitação do Produto 3B43-TRMM do Estado do Amazonas. *Floresta e Ambiente*, *22*(3), 279-286. http://dx.doi.org/10.1590/2179-8087.112114.

Torres-Pineda, C. E., & Pabón-Caicedo, J. D. (2017). Variabilidad intraestacional de la precipitación en Colombia y su relación con la oscilación de Madden-Julian. *Revista de la Academia Colombiana de Ciencias Exactas, Físicas y Naturales*, *41*(158), 79. http://dx.doi.org/10.18257/raccefyn.380.

Trenberth, K. (2018). *Climate Data NINO SST INDICES (NINO 1+2, 3, 3.4, 4; ONI AND TNI). NCAR. Climate Data Guide*. Retrieved in 2023, January 20, from https://climatedataguide.ucar.edu/climate-data/nino-sst-indices-nino-12-3-34-4-oni-and-tni

Urrea, V., Ochoa, A., & Mesa, O. (2019). Seasonality of Rainfall in Colombia. *Water Resources Research*, *55*(5), 4149-4162. http://dx.doi.org/10.1029/2018WR023316.

Wold, H. (1975). Soft Modelling by Latent Variables: The Nonlinear Iterative Partial Least Squares (NIPALS) Approach. *Journal of Applied Probability*, *12*(S1), 117-142. http://dx.doi.org/10.1017/S0021900200047604.

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, *58*(2), 109-130. http://dx.doi.org/10.1016/S0169-7439(01)00155-1.

Wyatt, B. M., Ochsner, T. E., Krueger, E. S., & Jones, E. T. (2020). In-situ soil moisture data improve seasonal streamflow forecast accuracy in rainfall-dominated watersheds. *Journal of Hydrology (Amsterdam)*, *590*(August), 125404. http://dx.doi.org/10.1016/j.jhydrol.2020.125404.

Zhang, Y., Moges, S., & Block, P. (2016). Optimal cluster analysis for objective regionalization of seasonal precipitation in regions of high spatial-temporal variability: application to Western Ethiopia. *Journal of Climate*, *29*(10), 3697-3717. http://dx.doi.org/10.1175/JCLI-D-15-0582.1.

Zhao, J., Chevalier, F., Pietriga, E., & Balakrishnan, R. (2011). Exploratory analysis of time-series with chronolenses. *IEEE Transactions on Visualization and Computer Graphics*, *17*(12), 2422-2431. http://dx.doi.org/10.1109/TVCG.2011.195.

Zuccolotto, P. (2012). Principal component analysis with interval imputed missing values. *AStA. Advances in Statistical Analysis*, *96*(1), 1-23. http://dx.doi.org/10.1007/s10182-011-0164-3.

## Authors contributions

Juan Sebastián Del Castillo-Gómez: Conceptualization, data curation, formal analysis, methodology, software, writing – original draft.

Teresita Canchala: Conceptualization, data curation, formal analysis, investigation, methodology, software, supervision, writing – review & editing.

Wilmar Alexander Torres-López: Conceptualization, formal analysis, methodology, validation.

Yesid Carvajal-Escobar: Conceptualization, formal analysis, methodology, supervision.

Camilo Ocampo-Marulanda: Formal analysis, methodology, writing – review & editing.

**Editor-in-Chief:** Adilson Pinheiro

**Associated Editor:** Carlos Henrique Ribeiro Lima

## SUPPLEMENTARY MATERIAL

Supplementary material accompanies this paper.

**S1**. Spearman correlation of monthly rainfall between 12-gauge stations in the PR for 1983-2019.

**S2**. Spearman correlation of monthly rainfall between 33-gauge stations in the AR for 1983-2019.

**S3**. Spearman correlation of monthly rainfall between 8-gauge stations in the AMR and Putumayo for 1983-2019.

This material is available as part of the online article from https://doi.org/10.1590/2318-0331.282320230008