water

MDPI

Article

Filling Gaps in Daily Precipitation Series Using Regression and Machine Learning in Inter-Andean Watersheds

Marcelo Portuguez-Maurtua [1,2,3,*], José Luis Arumi [2,4], Octavio Lagos [2,4], Alejandra Stehr [5] and Nestor Montalvo Arquiñigo [3]

1 Doctoral Program in Water Resources and Energy for Agriculture, Universidad de Concepcion, Av. Vicente Mendez 595, Chillan 3812120, Chile
2 CRHIAM Water Research Center, Universidad de Concepcion, Victoria 1295, Concepcion 4070386, Chile; jarumi@udec.cl (J.L.A.); octaviolagos@udec.cl (O.L.)
3 Water Resources Department, College of Agricultural Engineering, Universidad Nacional Agraria La Molina, Av. La Molina s/n, Lima 15024, Peru; nmontalvo@lamolina.edu.pe
4 Water Resources Department, College of Agriculture Engineering, Universidad de Concepción, Av. Vicente Mendez 595, Chillan 3812120, Chile
5 Centro de Ciencias Ambientales EULA-Chile, Departamento de Sistemas Acuáticos, Facultad de Ciencias Ambientales, Universidad de Concepción, Concepcion 4070386, Chile; astehr@udec.cl
* Correspondence: mportuguez@lamolina.edu.pe; Tel.: +51-1-949-377-610

Abstract: As precipitation is a fundamental component of the global hydrological cycle that governs water resource distribution, the understanding of its temporal and spatial behavior is of great interest, and exact estimates of it are crucial in multiple lines of research. Meteorological data provide input for hydroclimatic models and predictions, which generally lack complete series. Many studies have addressed techniques to fill gaps in precipitation series at annual and monthly scales, but few have provided results at a daily scale due to the complexity of orographic characteristics and in some cases the non-linearity of precipitation. The objective of this study was to assess different methods of filling gaps in daily precipitation data using regression model (RM) and machine learning (ML) techniques. RM included linear regression (LRM) and multiple regression (MRM) algorithms, while ML included multiple regression algorithms (ML-MRM), K-nearest neighbors (ML-KNN), gradient boosting trees (ML-GBT), and random forest (ML-RF). This study covered the Malas, Omas, and Cañete River (MOC) watersheds, which are located on the Pacific Slope of central Peru, and a nineteen-year period of records (2001–2019). To assess model performance, different statistical metrics were applied. The results showed that the optimized machine learning (OML) models presented the least variability in estimation errors and the best approximation of the actual data from the study zone. In addition, this investigation shows that ML interprets and analyzes non-linear relationships between rain gauges at a daily scale and can be used as an efficient method of filling gaps in daily precipitation series.

Keywords: precipitation gap filling; regression; machine learning; standard normal homogeneity test; K-nearest neighbors; gradient boosting tree; random forest

Citation: Portuguez-Maurtua, M.; Arumi, J.L.; Lagos, O.; Stehr, A.; Montalvo Arquiñigo, N. Filling Gaps in Daily Precipitation Series Using Regression and Machine Learning in Inter-Andean Watersheds. *Water* **2022**, *14*, 1799. https://doi.org/10.3390/w14111799

Academic Editors: Zheng Duan and Scott Curtis

Received: 26 March 2022
Accepted: 27 May 2022
Published: 2 June 2022

1. Introduction

Precipitation is a fundamental component of the hydrological cycle that governs water resource distribution [1]. The hydrological cycle of a given region is directly related to its topography, geology, physical mechanisms, and climate; precipitation is the most important phenomenon [2,3]. Precipitation, due to its high spatiotemporal variability and the large number of interconnected variables involved, is one of the most difficult atmospheric variables to characterize, estimate and forecast [4], especially on a daily scale, due to its high spatial and temporal variability [5]. The understanding of the temporal and spatial behavior of precipitation is of great interest, especially in studies on climatic risks [6]. In addition, exact estimates of precipitation are fundamental in multiple lines of research, as they serve as the basis for statistical models and analysis [7,8].

Peru is composed of three slopes, one that drains into the Pacific Ocean (Pacific Slope), another that drains into the Atlantic Ocean (Atlantic Slope), and a third that drains into Lake Titicaca (Titicaca Slope) [9]. The Peruvian Pacific Slope is located in tropical latitudes, and precipitation there is mainly influenced by orographic conditions, the ocean, and the atmosphere [10,11]. The spatial distribution of rainfall stations in Peru is heterogenous. In addition, precipitation series are frequently incomplete, which complicates the hydrological or climatological characterization of a given place [12].

Recent studies on the Peruvian Pacific Slope and coast have allowed it to be classified into homogenous regions [11], which has aided the understanding of spatial and seasonable precipitation variability patterns [9,13]. There are numerous methods of precipitation series gap filling, including least squares regression, predictive mean matching, nearest neighbor techniques, decision tree techniques, gradient boosting, and artificial neural networks [12,14–17]. In addition, geostatistical methods such as ordinary kriging tend to overestimate the number of rainy days and underestimate their magnitudes, and a negative correlation is even found in several reports between nearby stations [18–20]. In addition, the authors Huang et al. [21] and Gorshenin et al. [22] have evaluated the k-nearest-neighbor algorithm, together with machine learning models, such as multilayer perceptron (MLP), support vector machine (SVM) and random forest (RF), with promising results. A study in Germany used machine learning (ML) techniques, analyzing non-linear relationships between spatially distributed rain gauges [12]. In addition, in a recent study conducted by Bellido-Jiménez et al. [23] to fill possible gaps in precipitation datasets, in semi-arid regions of Andalusia, several machine learning models (MLP, SVM and RF) were tested, showing good results using neighboring data with MLP.

However, studies on precipitation series gap filling have mainly addressed annual and monthly scales [24–26]. Similarly, there are other studies that have addressed regional-scale development techniques, merging estimates based on quantile mapping, spatial interpolation, machine learning, and multi-strategy fusion [27,28], with few investigations focused on a daily scale, due to the complexity of orographic characteristics and in some cases the non-linearity of precipitation series between neighboring stations [8,29,30]. The objective of this study was to fill gaps in daily precipitation series through comparative analysis of regression model (RM) and ML techniques. RM included linear (LRM) and multiple regression models (MRM). For ML, multiple regression models (ML-MRM), K-nearest neighbors (ML-KNN), gradient boosting trees (ML-GBT), and random forest (ML-RF) were used. In addition, an optimization process, optimized machine learning (OML), was used with the multiple regression (OML-MRM), K-nearest neighbors (OML-KNN), gradient boosting tree (OML-GBT), and random forest models (OML-RF), for a network of 17 rainfall stations located in the Malas, Omas, and Cañete River (MOC) watersheds. We assessed the efficiency of the results obtained from each model using statistical metrics. However, in order to guarantee reliable results using raw rainfall data, it is an essential requirement to perform the quality control process, such as the homogenization of the daily rainfall series, which allowed the detection of observation and measurement errors, which are problems that occur in a rainfall observation network [31]. In addition, it is important to identify homogeneous zones through the regionalization process, using up to three methods as a means of verifying the results.

The aim of this study was to demonstrate that ML techniques can interpret and analyze non-linear relationships between rain gauges at a daily scale and can be used as an efficient method of filling gaps in daily precipitation series. The results of the gap-filled precipitation series can be used in future investigations to evaluate the performance of the of the daily precipitation data obtained from satellite sensors based on a hydrological model and evaluate its performance based on time series of discharges measured at hydrometric stations. Finally, the results of this study showed that the ML models presented better approximations to the actual data than the RM models.

The structure of the paper is organized as follows. Section 2 shows the information about the locations, the dataset, the theoretical background of the different machine learning

(ML) models evaluated, the preprocessing algorithms and evaluation metrics, in addition to the quality control of the dataset by homogenization and regionalization. Then, in Sections 3 and 4, the results are reported and discussed, respectively. Finally, Section 5 describes the conclusions reached in this work.

2. Materials and Methods

2.1. Study Area

The study area comprised the Mala, Omas, and Cañete River (MOC) watersheds (Figure 1), located in the central part of the Peruvian Pacific Slope and coast; its total area is 9496 km^2 (2250, 1167 and 6079 km^2, MOC basins, respectively). The area is characterized by a significant latitudinal gradient that goes from 0 to 6500 masl; above 2500 masl is the wet watershed area [32]. The rivers flow from east to west from the Andes to the Pacific Ocean, with bare, steep slopes that favor significant swelling, floods, and erosion during heavy rainfall episodes [9]. In addition, in normal conditions, the region is influenced by the South Pacific High, in combination with the Humboldt current that produces dry, stable conditions with moist air trapped below the inversion layer at about 1000 masl, and it presents major seasonal and interannual precipitation variability [9,11,13].

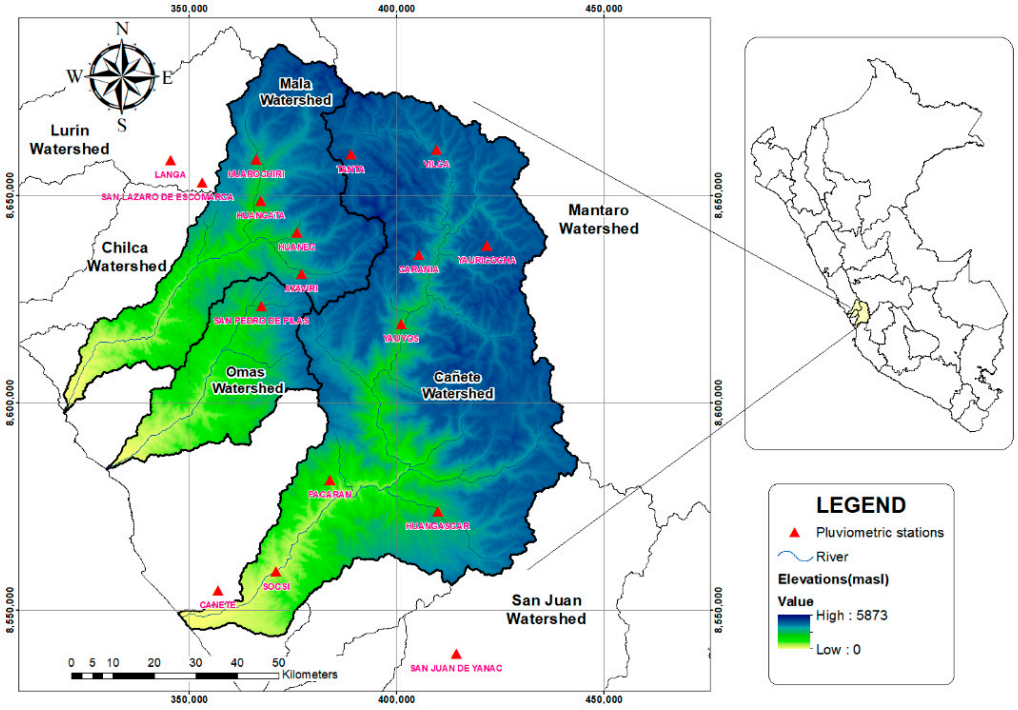


Figure 1. Elevation of the study area, main rivers, selected watershed boundaries, and location of rainfall stations.

2.2. Methods

The methodology has four stages; Figure 2 shows the methodological diagram. The first stage is the collection of available daily precipitation information from within and near the study area. The second stage is the exploratory analysis and homogenization of rainfall data. The third stage is the regionalization process, which includes the use of the Ward, K-means, and regional vector analysis methods (RVM). Finally, the fourth stage consists of the filling of gaps in daily precipitation series using the RM and ML methods. In addition, the performance of each model was evaluated using metrics.

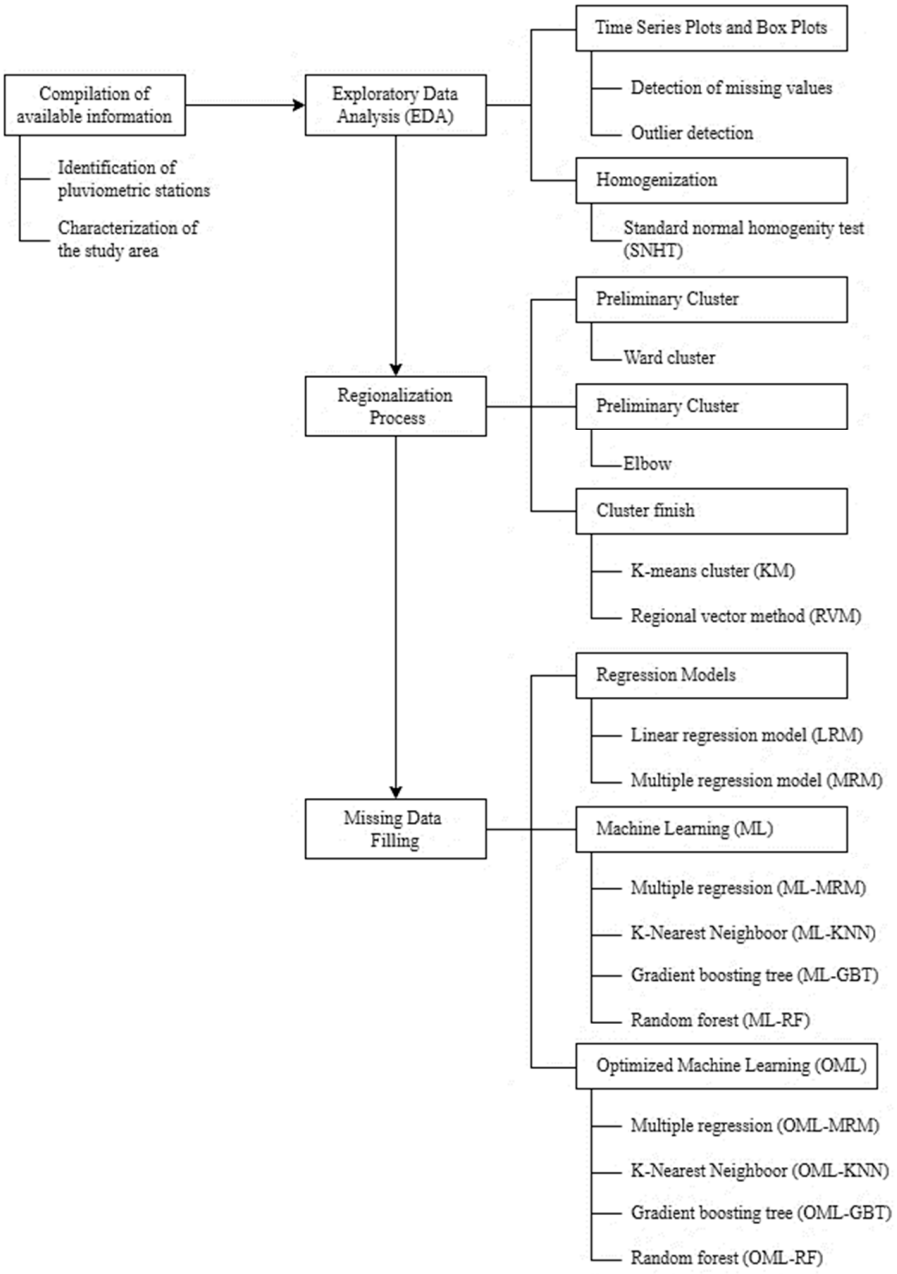


Figure 2. Methodological diagram for daily precipitation series gap filling.

2.2.1. Collection of Available Information

A total of 17 rainfall stations were selected, some with records since 1965, others since 1980, etc., all of which had periods with irregular records. The stations are part of the network managed by the National Meteorology and Hydrology Service of Peru (SENAMHI, https://www.senamhi.gob.pe/mapas/mapa-estaciones/mapadepesta1.php accessed on: 5 March 2020). In addition, stations located outside the study region and those inactive during the selected period were discarded. Similarly, there are rainfall stations with more than 10% missing (empty) data relative to the total length of the analyzed series. Figure 1 shows the spatial locations of the rainfall stations. In addition, Table 1 shows the geographic coordinates, quantity of observed data, and quantity of missing data.

Table 1. Rainfall stations of MOC watersheds, 2001–2019 period.

ID	Stations	Coordinates		Altitude	Observed Data		Missing Data	
		Latitude	Longitude	(masl)	No of Data	(%)	No of Data	(%)
1	Ayaviri	−12.38	−76.13	3228	6881	99.2	58	0.8
2	Cañete	−13.07	−76.32	158	3830	55.2	3109	44.8
3	Carania	−12.34	−75.87	3875	6939	100	0	0
4	Huancata	−12.22	−76.22	2700	6939	100	0	0
5	Huangascar	−12.9	−75.83	2533	6908	99.6	31	0.4
6	Huañec	−12.29	−76.14	3205	6939	100	0	0
7	Huarochiri	−12.13	−76.23	3154	6787	97.8	152	2.2
8	Langa	−12.13	−76.42	2863	6484	93.4	455	6.6
9	Pacaran	−12.83	−76.07	700	5132	74	1807	26
10	San Juan de Yanac	−13.21	−75.79	2550	6482	93.4	457	6.6
11	San Lazaro de Escomarca	−12.18	−76.35	3758	6486	93.5	453	6.5
12	San Pedro de Pilas	−12.45	−76.22	2600	6909	99.6	30	0.4
13	Socsi	−13.03	−76.19	500	4687	67.5	2252	32.5
14	Tanta	−12.12	−76.02	4323	6819	98.3	120	1.7
15	Vilca	−12.11	−75.83	3864	6297	90.7	642	9.3
16	Yauricocha	−12.32	−75.72	4675	6818	98.3	121	1.7
17	Yauyos	−12.49	−75.91	2327	6878	99.1	61	0.9

2.2.2. Exploratory Data Analysis (EDA)

It is an essential requirement to guarantee reliable results using raw rainfall data, the application of quality control procedures, by means of graphs and the homogenization of time series, allowing the detection of observation and measurement errors, supported in recommended by Estévez et al. [31]. This process was carried out in two phases: first, a time series graph and boxplot, which allowed the identification of missing values and outliers; this process is performed in Python. Second, in order to determine inconsistencies at the stations, which could stem from a change in instrument location, variations in the conditions at the measurement site, or an observer change, the data were analyzed using the standard normal homogeneity test (SNHT), as described by [33–35].

The SNHT was developed by [36] and modified by [37,38]; it uses Y to denote the candidate series and Y_i to denote a specific value (for example, cumulative annual precipitation or mean annual temperature) in the year (or other unit of time) i. In addition, X_j denotes one of the surrounding reference sites (jth of a total of k) and X_{ji} a specific value from that site. The following equations were used to detect the relative non-homogeneities (traditionally used in precipitation studies):

$$Q_i = \frac{Y_i}{\left\{ \frac{\left[\sum_{j=1}^{k} \rho_j^2 X_{ji} \frac{\hat{Y}}{X_j} \right]}{\sum_{j=1}^{k} \rho_j^2} \right\}} \quad (1)$$

and

$$Q_i = Y_i - \left\{ \frac{\sum_{j=1}^{k} \rho_j^2 \left[X_{ji} - \hat{X}_j + \hat{Y} \right]}{\sum_{j=1}^{k} \rho_j^2} \right\} \quad (2)$$

where Q_i is the ratio in Equation (1) and the difference in Equation (2) in a specific year i; $\hat{Y}$ represents the multi-annual mean of the candidate time series; and ρ_j is the correlation coefficient between the test variable Y and the reference variable X_j [36,38,39]. This method is implemented in the Climatol package for R language [34]. Climatol has three normalization methods: division by mean values, subtraction of means, and complete standardization; here, we opted for subtraction of means, as the minimum precipitation values can be zero [34,40,41]. On a preliminary basis, Climatol was run for a monthly time

step, identifying breaks; based on these breaks, Climatol was run again for a daily time step. The results show graphs of absolute maximum autocorrelation (ACmx), SNHT, root mean square error (RMSE), and percentage of original data (POD).

2.2.3. Regionalization Process

This section describes the regionalization process, which was performed using three methods. In the first method, Ward's hierarchical clustering analysis was applied. This method is also known as "minimum variance" grouping, where Ward's objective function of the [42] algorithm minimizes the sum of squared deviations of the attribute vectors from the centroid of their respective groups; instead of merging samples or clusters as a function of distance, it starts by assigning "zero variance" to all clusters. This method was applied to ascertain the preliminary clustering of the stations [43]. This process was carried out by programming in R language.

In the second method, non-hierarchical K-means clustering (KM) was applied, which is a statistical technique designed to assign objects to a fixed number of clusters according to a set of specified variables [11,44]. It consists of obtaining a partition that minimizes intraclass inertia. This is achieved locally (it depends on the initial points) using the Euclidian distance between individuals and the moving centers used for aggregation. The KM algorithm is an iterative procedure in which the attribute vectors move from one group to another to minimize the value of the objective function, F, defined in Equation (3).

$$F = \sum_{k=1}^{k} \sum_{j=1}^{m} \sum_{i=1}^{N_k} d^2\left(y_{ij}^k - y_{\bullet j}^k\right) \tag{3}$$

In Equation (3), k indicates the number of groups, N_k represents the number of attribute vectors in group k; ${y^k}_{ij}$ denotes the rescaled value of attribute j in attribute vector i assigned to group k; and ${y^k}_{\bullet j}$ is the mean value of attribute j for group k (Equation (4)) [43,45].

$$y_{\bullet j}^k = \frac{\sum_{i=1}^{N_k} y_{ij}^k}{N_K} \tag{4}$$

However, one of the problems encountered when applying the KM method lies in choosing the number of clusters. Although there is no single criterion for choosing the number of clusters, here we used the elbow method, implementing it by programming in R language.

Finally, the regional vector method (RVM), described by [10,11,44], was the third to be applied, in order to corroborate the previously obtained results. It consists of creating a fictitious station (vector) with average values from all stations in the zone. This method is aimed at the homogenization and completion–extension of precipitation data [46,47] and is based on the creation of an "average value" "vector" station. This concept refers to the calculation of a weighted average of rainfall anomalies for each station, overcoming the effects of stations with extreme and low rainfall values and problems associated with the weight of the rainiest stations relative to the least rainy ones.

This method applies the least squares method to find annual regional rainfall indices Z_i and extended mean precipitation P_j, which is achieved by minimizing the expression [10,11,45]:

$$S = \sum_{I=1}^{N} \sum_{J=1}^{M} \left(\frac{P_{ij}}{P_j} - Z_i\right)^2 \tag{5}$$

where i is the index of the year; j is the index of the station; N is the number of years; M is the number of stations; P_{ij} is annual precipitation at station j in year i; P_j is mean precipitation extended to a period of N years; and, finally, Z_i is the regional rainfall index of year i. This process was carried out using the Hydracces program [48].

2.2.4. Gap-Filling Model

In this stage of the study, the results from the regionalization process were used. The RM and ML techniques were applied for each homogenous region. The daily precipitation series were graphed for each homogenous region, allowing the dates with missing data to be identified. In addition, the intensity of the relationships between stations was analyzed using Pearson coefficient correlations [29,30].

To apply the LRM and MRM techniques, in both cases, target stations (Y) and variables to predict were identified. Predictor stations (X) were identified for LRM and multiple predictor stations (X_m) for MRM. LRM is a computing procedure based on the alternate least squares algorithm (ALS) [49]. It has two steps: first estimating the relationship between predictors and missing values and then using the trend equation to fill the gaps [50], in accordance with Equation (6):

$$P_i(t) = a + b * P_i(t) \quad (6)$$

The values of a and b can be estimated using Equations (7) and (8), respectively.

$$a = \overline{y} - b\overline{x} \quad (7)$$

$$b = \frac{\sum_{i=1}^{n} xy - \frac{\sum_{i=1}^{n} x \sum_{i=1}^{n} y}{n}}{\sum_{i=1}^{n} x^2 - \frac{\left(\sum_{i=1}^{n} x\right) x^2}{n}} \quad (8)$$

where $\overline{y}$ and $\overline{x}$ are mean values of the data series of the reference and similarity stations, respectively [50,51].

Meanwhile, MRM is a statistical technique that consists of finding a linear relationship between a dependent variable and more than one independent variable. It can be represented using the following equation:

$$Y_i = a + b_1X_1 + b_2X_2 + \ldots + b_mX_m + C \quad (9)$$

where Y_i is the dependent variable; X_1, X_2, … X_m are the independent variables; a is the intersection; b_1, b_2, … b_m are the multiple regression coefficients, estimated using the method of least squares; and C is the error term [50,51]. ML is a scientific discipline in the artificial intelligence field that creates systems that learn automatically [8,14]. For gap filling using this technique, the data available at each station were divided randomly to generate a training dataset (train) and test dataset (test) in proportions of 75% and 25%, respectively [8]. The algorithms implemented were MRM, K-nearest neighbors (KNN), gradient boosting trees (GBT), and random forest (RF). In addition, an optimization process was carried out, generating OML-MRM, OML-KNN, OML-GBT, and OML-RF models. These algorithms were implemented using the Python programming language. KNN is a non-parametric method that can be used for both classification and regression.

The result is calculated based on the weighting of a number of nearest neighbors in the attribute space based on a distance function; the most common is Euclidian distance for continuous data [8]. GBT is a method in which multiple decision trees are iteratively fit to the data, and each tree is based on the previous tree to reduce losses and improve performance. It is based on the boosting principle, that is, on the creation of a set of weak learners to improve prediction precision [8,52]. This method has three advantages: first, it does not require the application of a direct physical model; second, it serves as a computationally feasible method of capturing complex non-linear interactions between variables and a response [52,53]; and finally, it presents almost no overfitting problems, which is an important advantage, as many models over- or underestimate results [14,52,53]. RF was proposed by [54]. It is a semi-unsupervised non-parametric algorithm in the decision tree family that consists of a set of uncorrelated trees to produce predictions for classification and regression tasks [55].

2.2.5. Bayesian Optimization

One of the critical aspects of machine learning models' efficiency is hyperparameter selection. It is very important to establish the correct values; performance can change drastically from excellent to very poor. A common practice in the scientific community uses a trial and error technique, where different values, ranging from tens to thousands of possibilities, are evaluated [23,31]. Therefore, efficiently setting the hyperparameter space is essential, because if the hyperparameter space is ample, the algorithm wastes significant time in non-promising configurations (apart from being very slow). On the other hand, when the hyperparameter space is small, an accurate hyperparameter configuration set may be missing, even though it is fast [23,31].

Bayesian optimization was used to estimate the hyperparameters due to its great popularity in machine learning models and its good performance in optimization [56,57]. The procedure consists of four steps, as described by [23]: (1) define the hyperparameter space; (2) the algorithm considers previous evaluations to choose the next set of values to be evaluated (acquisition function); (3) to assess the new hyperparameter configuration using an objective function; and (4) if the optimization process has not finished yet, it goes to the second point. In this work, this algorithm was implemented using Python.

2.2.6. Evaluation Metrics

To assess the efficiency of the developed models, coefficient of determination (R^2), root mean square error (RMSE), Nash–Sutcliffe coefficient (NSE) and percentage bias (PBIAS) were used [8,51,58]. All of them are mathematically expressed as Equations (10)–(13), respectively:

$$R^2 = \frac{\left[\sum_{t=1}^{n}\left(P_{obs} - \overline{P}_{obs}\right)\left(P_{ped} - \overline{P}_{pred}\right)\right]^2}{\sum_{t=1}^{n}\left(P_{obs} - \overline{P}_{obs}\right)^2 \sum_{t=1}^{n}\left(P_{pred} - \overline{P}_{pred}\right)^2} \quad (10)$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(P_{obs,i} - P_{pred,i}\right)^2}{n}} \quad (11)$$

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{n}\left(P_{pred,i} - P_{obs,i}\right)^2}{\sum_{i=1}^{n}\left(P_{pred,i} - \overline{P}_{pred,i}\right)^2} \quad (12)$$

$$\mathrm{PB} = \frac{\sum_{i=1}^{n}\left(P_{obs,i} - P_{pred,i}\right) \times 100}{\sum_{i=1}^{n} P_{pred,i}} \quad (13)$$

where n represents the number of prediction days, P_{obs} corresponds to the measured value for a specific day, P_{pred} is the predicted value, i represents measurement on a specific day, and $\overline{P}_{pred}$ correspond to the average measured and predicted values, respectively.

3. Results

3.1. Analysis of Missing Data, Outliers, and Homogenization

In Figure 3, the bar graph shows the quantity of unavailable precipitation data by station; there are three stations with more than 10% missing data (Cañete, Socsi, and Pacaran), while the remaining stations present less than 10% missing data. The Cañete, Socsi and Pacaran rainfall stations are located in the lower part of the basin, which is characterized by being dry almost all year round (less than 20 mm/year).

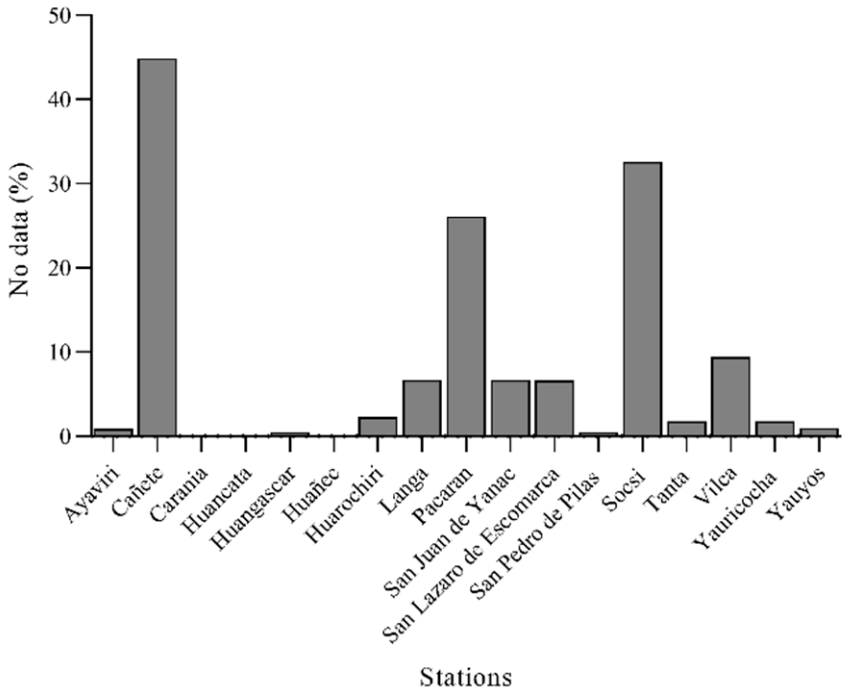


Figure 3. Missing daily precipitation data: quantity of unavailable daily precipitation data as a percentage by station.

Figure 4a shows the boxplots for daily precipitation series of each station; these contain a large number of scattered values, which initially could be considered outliers. However, it should be taken into account that daily precipitation shows high temporal and spatial variability patterns. Figure 4b shows the boxplot at a monthly scale, showing smaller dispersions, probably lower outliers, reflecting less spatial and temporal variability.

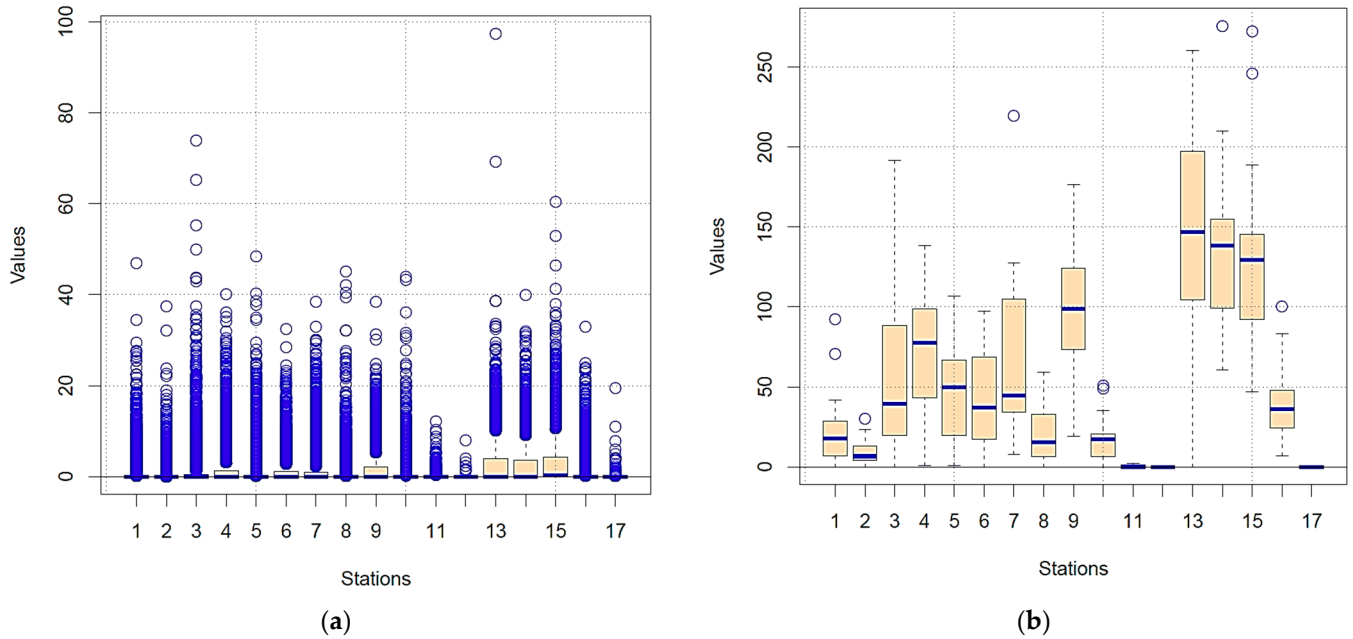


Figure 4. Exploratory analysis of outliers: (**a**) Daily series and (**b**) Monthly series.

Figure 5a shows the results of Pearson coefficient correlations; high spatial and temporal variability on a daily scale are observed, complicating the detection of homogeneities. The high variability of daily records compared to that of monthly or annual values makes it very difficult to directly apply methods for identifying inhomogeneities at the daily scale; in accordance with the recommendations of [34], the homogenization process was performed at a monthly scale, at which it is possible to detect cutoffs or breakpoints. Once the breakpoints were identified, the homogenization process was carried out on a daily scale using the Climatol package in R (https://cran.r-project.org/web/packages/climatol/index.html accessed on: 5 May 2020) [34,59]. The results in Figure 5a,b show the correlation between the original normalized series and the reference series obtained based on the other stations. The reference series was constructed based on the average value of the nearest stations, which is weighted by the inverse of the distance from the analysis station [34,39,41]. The daily-scale correlation results present a maximum value of 0.40 and a minimum below zero (Figure 5a); the monthly-scale results reach values close to 1.0 (Figure 5b). This analysis was carried out for all the stations.

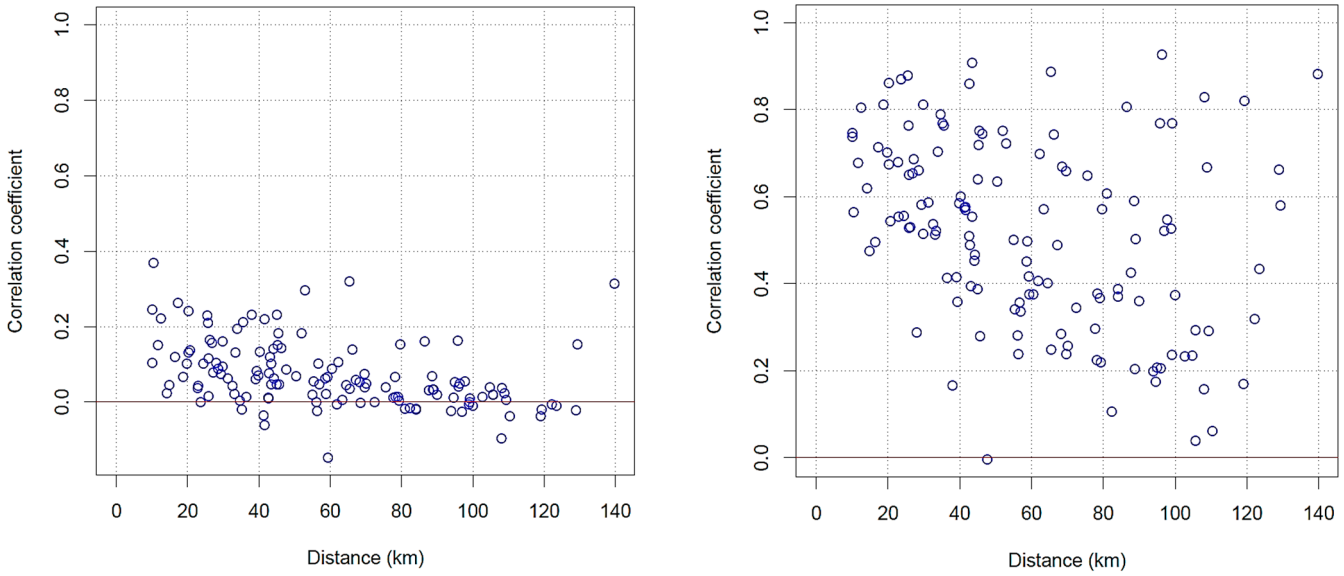


Figure 5. Correlogram between stations: (**a**) Daily precipitation series and (**b**) Monthly precipitation series.

Climatol provided overall absolute maximum autocorrelation (ACmx), SNHT, root mean square error (RMSE), and percentage of original data (POD) results. The ACmx values are not significant until the third quartile of the series (0.34); the values are below 60% autocorrelation, which indicates that the series are non-seasonal (Figure 6a). The series present anomalies in SNHT values between the original and homogenized series; the values range from 9.10 to 80.90, with the exception of the Cañete station, which reaches a maximum of 228, creating a rather wide variation spectrum (Figure 6b). RMSE presents high variation, with a minimum value of 1.26 and a maximum of 4.79 (Figure 6c). Finally, POD, which compares the original and homogenized data series, presents high values, meaning that the original data available are of good quality (Figure 6d). In addition, results of the analysis of homogeneity by station were obtained (Table 2). The Cañete, Socsi, and Pacaran stations presented ACmx values above 0.60, SNHT values above 90.0, and POD values above 10%. Only RMSE presented low values.

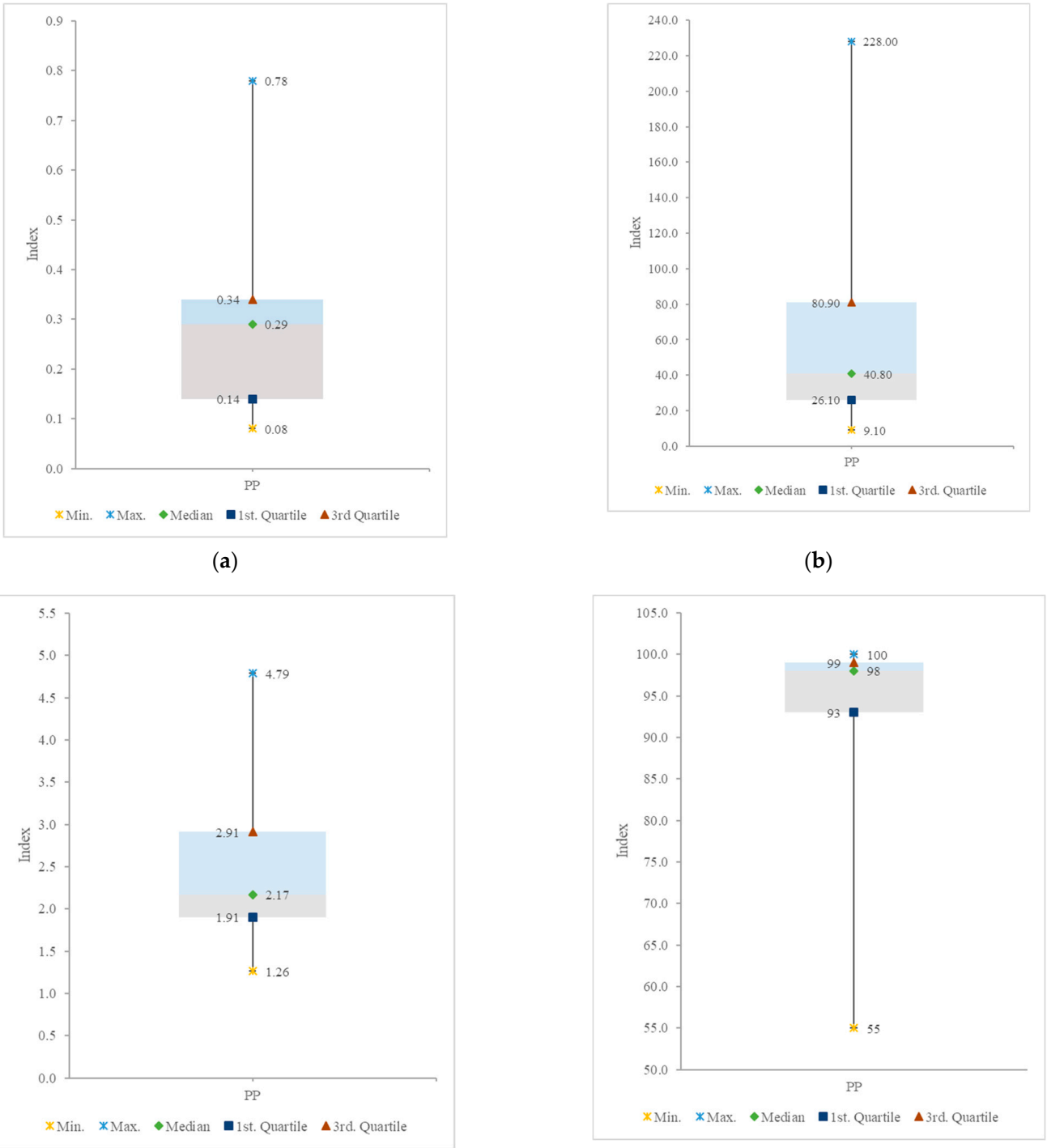


Figure 6. Homogeneity analysis statistics: (**a**) Station maximum absolute autocorrelation (ACmx), (**b**) Standard normal homogeneity test (SNHT), (**c**) Root mean squared error (RMSE) and (**d**) Percentage of original data (POD).

Table 2. Homogeneity analysis statistics for each station.

Stations	ACmx	SNHT	RMSE	POD
Ayaviri	0.19	47.4	2.9	99
Cañete	0.65	228.0	1.3	55
Carania	0.20	26.1	2.6	100
Huancata	0.33	95.0	2.4	100
Huangascar	0.14	35.9	1.9	99
Huañec	0.29	68.7	2.2	100
Huarochiri	0.13	55.0	2.7	97
Langa	0.08	80.9	2.0	93
Pacaran	0.73	166.1	1.3	73
San Juan de Yanac	0.10	21.5	1.5	93
San Lazaro de Escomarca	0.32	20.9	3.4	93
San Pedro de Pilas	0.15	13.4	2.0	99
Socsi	0.78	40.8	1.4	67
Tanta	0.34	155.6	4.8	98
Vilca	0.34	30.5	3.9	90
Yauricocha	0.36	38.6	4.6	98
Yauyos	0.08	9.1	1.9	99

3.2. Regionalization Analysis

The results provided by ward showed four groups of regions for the 17 stations (Figure 7). This process, implemented based on R code, allowed the initial clustering to be ascertained. Clustering analysis with KM is a method that creates the most heterogenous clusters possible; that is, the objects in the k-clusters must be as similar as possible to those that belong to their cluster and completely unlike the objects in other clusters [11]. A fundamental point in the application of KM is to ascertain the optimum number of clusters. There are many criteria for choosing the optimal number of clusters, however; for this study, the elbow method (EM) was used due to its extensive application in diverse hydrological studies with good results. The optimal cluster or region value is shown in Figure 8. According to EM analysis, the optimal number of regions was four. In addition, KM was used to define the stations belonging to each homogenous region. Table 3 details the number and names of the stations in each region.

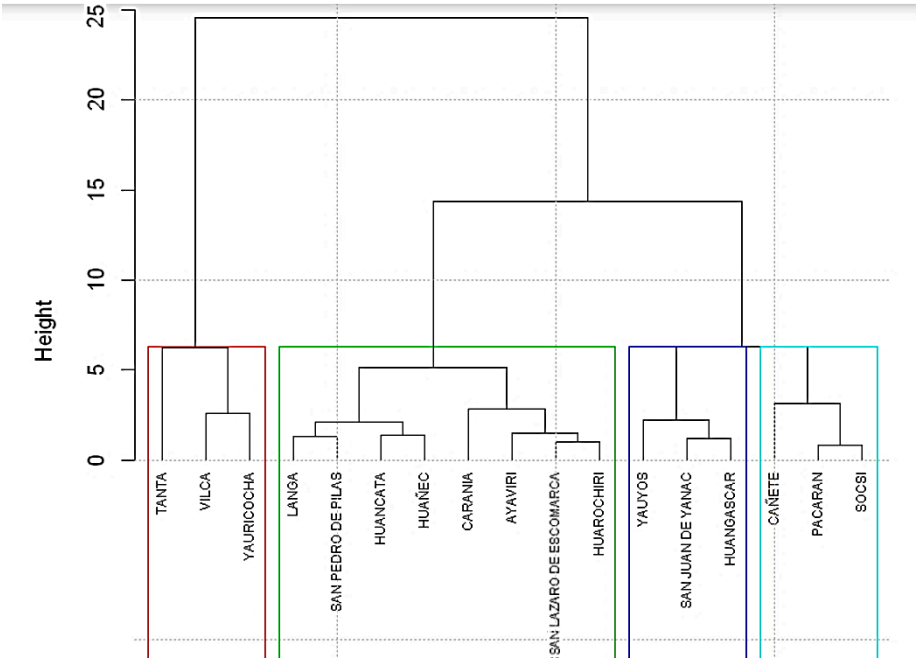


Figure 7. Ward method clustering: Dendrogram (2001–2019 period).

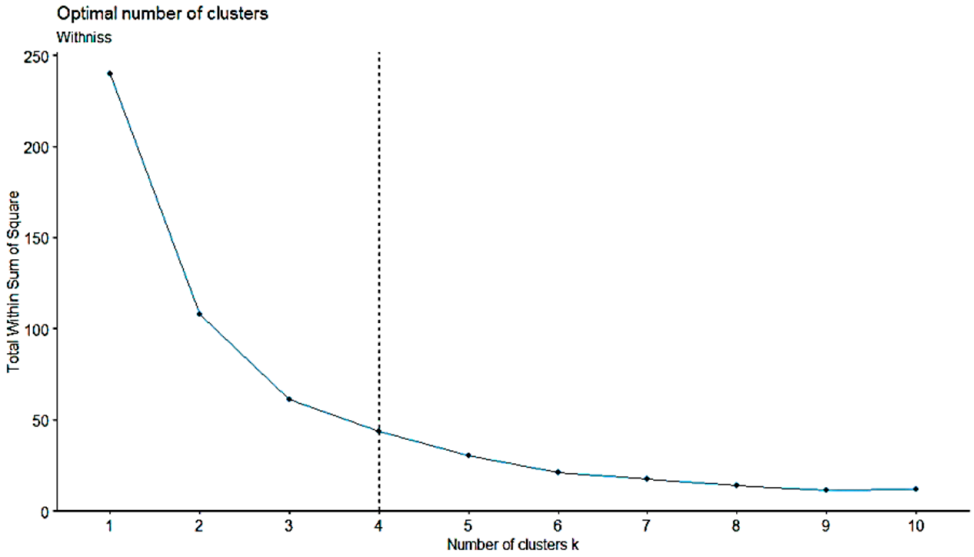


Figure 8. Optimal number of clusters according to the elbow method.

Table 3. Annual regional vector indices—Region 1.

Station	Time (Years)	Standard Deviation	Station/Vector Correlation
Langa	16	0.252	0.882
San Lazaro de Escomarca	16	0.263	0.664
Ayaviri	17	0.116	0.904
Huancata	19	0.341	0.863
Huañec	19	0.187	0.751
Huarochiri	14	0.159	0.851
Carania	19	0.191	0.679

The results obtained with ward and KM indicate that precipitation during the evaluated period was not similar at every station throughout the watersheds. The application of the ward and KM methods was performed using code written in R, and for EM, the code written in Python.

Finally, the RVM method was applied to validate the results obtained based on the described models. The Hydraccess program was used to apply RVM (https://hybam.obsmip.fr/es/hydraccess-3 accessed on: 5 May 2020). The results show clustering of stations with similar behaviors in terms of interannual precipitation variation, taking the standard deviation and correlation coefficient/vector as indicators. The regions are considered homogenous if the values of the standard deviation (SD) are lower than 0.4 and the correlation coefficient/vector values are above 0.7 [11]. The final results show the clustering of rainfall stations into homogenous regions.

The RVM method was used to obtain three final clusters that, in accordance with their statistics and analysis of the results, included the stations that are shown in Table 3 and Tables S1 and S2 (Supplementary Material), and Figure 9, and Figures S1 and S2 (Supplementary Material). It was not possible to analyze cluster 3, as its stations presented a high percentage of missing data.

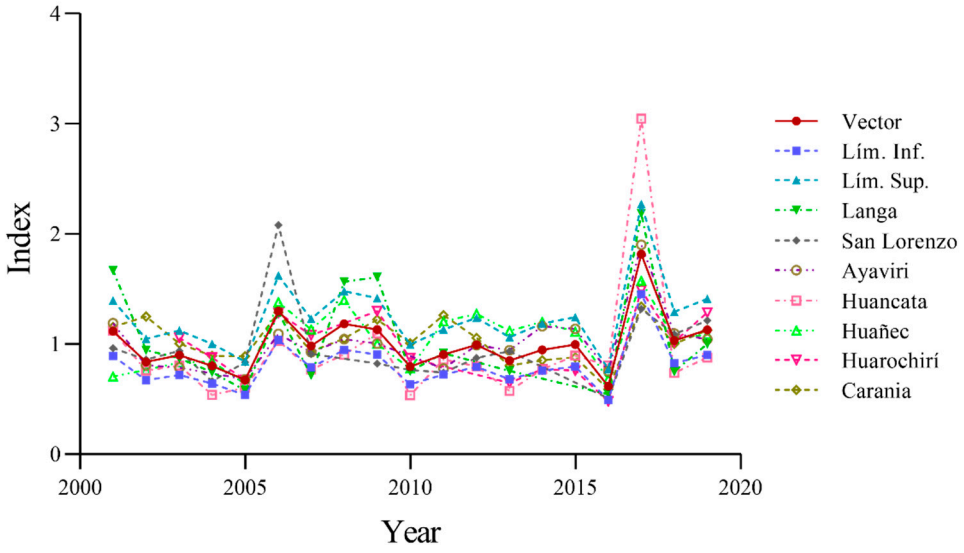


Figure 9. Annual indices of the regional vector and stations in Region 1.

The results obtained with the ward and KM methods are consistent in the number of homogenous regions. However, the number of stations in Regions 1 and 2 presented a slight discrepancy between the results obtained with ward and KM; therefore, the results obtained with RVM were used for verification, showing accord between the KM and RVM results. Table S3 (Supplementary Material) shows the final results of the homogenous region clustering. In addition, Figure 10 shows the regionalization of rain gauge stations based on the KM and RVM results.

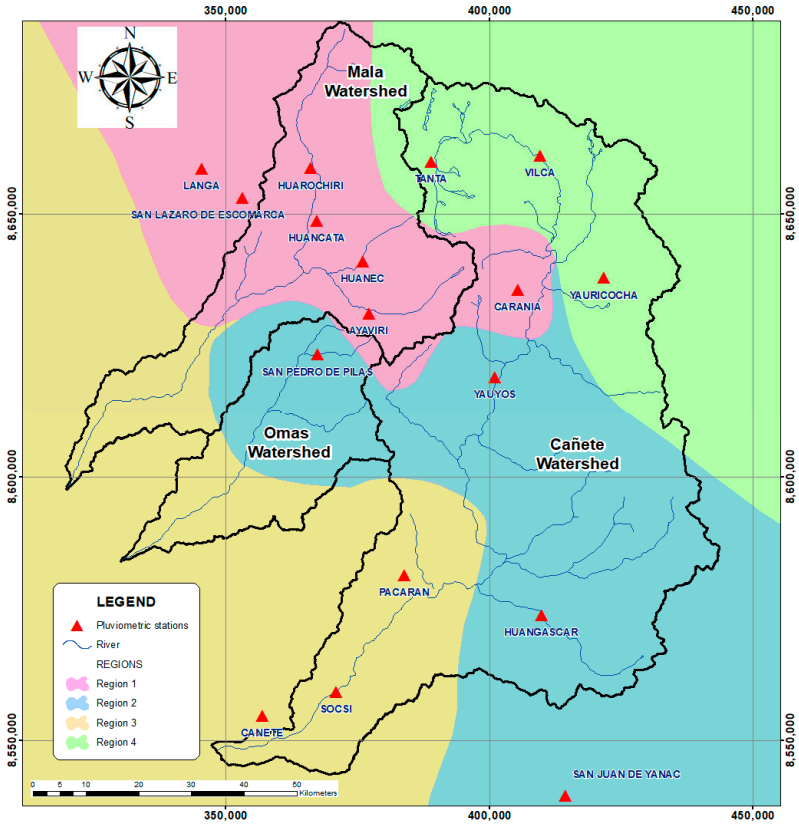


Figure 10. Regionalization of the stations according to the KM and MVR methods.

3.3. Analysis of the Series Gap-Filling Process

In Table 4 and Tables S4 and S5 (Supplementary Material), the correlation values corresponding to the stations clustered by homogenous region are shown. The correlations were below 0.60 and above 0.38; below 0.58 and above 0.32, and below 0.45 and above 0.37 in Regions 1, 2, and 4, respectively. These coefficients are considered acceptable given the dry conditions, with more than 90% of the rain gauge records close to zero throughout the year due to the hydroclimatic conditions, with any value greater than zero causing high variability [11]. Therefore, this analysis allowed the level of representation using Pearson coefficient correlations within a region to be highlighted.

Table 4. Correlation coefficient—Region 1.

	Ayaviri	Carania	Huancata	Huañec	Huarochiri	San Lazaro de Escomar-ca	Langa
Ayaviri	1						
Carania	0.48	1					
Huancata	0.60	0.47	1				
Huañec	0.51	0.43	0.49	1			
Huarochiri	0.56	0.55	0.58	0.46	1		
San Lazaro de Escomarca	0.45	0.40	0.43	0.39	0.44	1	
Langa	0.45	0.38	0.46	0.38	0.46	0.55	1

In the application of RM for filling missing precipitation data, the models were generated based on the homogenous regions. For the LRM algorithm, the Ayaviri station was designated as variable Y and the Huancata station was designated as variable X based on the greater Pearson coefficient correlations value. The other stations with missing data were selected in a similar manner. Table 5 shows the Y and X variables for each region. Meanwhile, for MRM, the procedure was similar to that of the previous case; the Ayaviri station was identified as the Y variable and all the remaining stations (Huancata, Langa, San Lazaro de Escomarca, Huañec, Huarochiri, and Carania) were identified as X_m (see Table 5).

Table 5. Identification of target stations (Y) and predictor stations by homogenous region.

Regions	Target Station (Y)	Predictor Station (X)	Multiple Predictor Stations (X_m)
Region 1	Ayaviri	Huancata	Huancata, Langa, San Lazaro de Escomarca, Huañec, Huarochiri, Carania
	Huarochiri	Huancata	Huancata, Langa, San Lazaro de Escomarca, Ayaviri, Huañec, Carania
	San Lazaro de Escomarca	Langa	Langa, Ayaviri, Huancata, Huañec, Huarochiri, Carania
	Langa	San Lazaro de Escomarca	San Lazaro de Escomarca, Ayaviri, Huancata, Huañec, Huarochiri, Carania
Region 2	San Pedro de Pilas	Huangascar	Huangascar, San Juan de Yanac, Yauyos
	Huangascar	San Pedro de Pilas	San Pedro de Pilas, Yauyos, San Juan de Yanac
	Yauyos	San Pedro de Pilas	San Pedro de Pilas, Huangascar, San Juan de Yanac
	San Juan de Yanac	San Pedro de Pilas	San Pedro de Pilas, Huangascar, Yauyos
Region 4	Tanta	Vilca	Vilca and Yauricocha
	Yauricocha	Vilca	Vilca and Tanta
	Vilca	Yauricocha	Yauricocha and Tanta

For the application of the different gap-filling algorithms, this process was carried out independently in each homogenous region. The Y stations with missing data in each homogenous region were identified, as were the X_m stations corresponding to each target station. The Y and X_m stations for each homogenous region are shown in Table 5.

For the filling of missing precipitation data with ML, the analysis was also carried out independently in each homogenous region. First, the available data were divided, with one portion for training and another for testing (75% and 25% respectively); this division was performed randomly. Then, the ML-MRM, ML-KNN, ML-GBT and ML-RF were selected along with their respective parameters (see Table 6). In addition, considering that many models contain parameters that cannot learn from training data, it was necessary to carry out an optimization process. To this end, it was important to ascertain the hyperparameter values using the Bayesian Optimization method. The results of a model can depend largely on the values taken by its hyperparameters; however, it cannot be known beforehand what values are suitable. The most common means of finding optimal values is testing different possibilities; in this study, optimization processes were carried out for the OML-MRM, OML-KNN, OML-GBT, and OML-RF algorithms (Table 7).

Table 6. Parameter and hyperparameter values for the ML algorithms.

Algorithm	Parameters [Values]	Hyperparameters [Values]
Multiple Regression	alpha [1] solver [‘auto’] modelo[Ridge]	alpha [logspace(–5, 5, 500)] solver [‘auto’] modelo[Ridge]
K-nearest neighbors	n_neighbors [5] leaf_size [30] algoritm [‘auto’] modelo[KNeighborsRegressor]	n_neighbours [linspace(1, 100, 500] leaf_size [1, 3] algoritm [‘auto’] modelo[KNeighborsRegressor]
Gradient boosting tree	n_estimators [100] max_feature [‘none’] max_depth [3] subsample [1] modelo[GradientBoostingRegressor]	n_estimators [50, 100, 1000, 2000] max_feature [‘auto’, 3, 5, 7] max_depth [‘None’, 3, 5, 10, 20] subsample [0.5, 0.7, 1] modelo[GradientBoostingRegressor]
Random forest	n_estimators [100] max_feature [‘auto’] max_depth [‘None’] modelo[RandomForestRegressor]	n_estimators [50, 100, 1000, 2000] max_feature [‘auto’, 3, 5, 7] max_depth [‘None’, 3, 5, 10, 20] modelo[RandomForestRegressor]

Based on the Y and X_m variables, ML was first applied for default parameter values using the ML-MRM, ML-KNN, ML-GBT, and ML-RF models. It was also applied using parameters called hyperparameters, generating the OML-MRM, OML-KNN, OML-GBT, and OML-RF models. This process allowed the model parameters to be optimized. The parameter and hyperparameter values used in the algorithms created in Python are shown in Table 6.

Table 6 describes the parameter and hyperparameter values used in each algorithm in ML. It is observed that only one parameter value was assigned when using the default algorithm. However, for the algorithm optimization process, a wide range of values was defined, and using the Bayesian optimization method, the optimal hyperparameters were estimated.

3.4. Assessment of Model Performance

To assess the performance of the models, different statistical metrics—R^2, RMSE, NSE and PBIAS—were used for both datasets (training and test). The obtained results are presented in Table 7 and Tables S6 and S7 (Supplementary Material). These statistics were calculated for the 2001–2019 period; periods with missing data were not considered.

Many linear models, among them LRM, contain parameters that cannot learn from training data, making it necessary for the modeler to establish them. In addition, to establish the predictive capacity of ML, which consists of testing how close its predictions are to the actual values of the response variable, a set of observations is needed, with its corresponding response variables, but that the model has not "seen", that is, which have not participated in its initial fitting. Finally, to assess the performance of models by comparing predicted and actual precipitation values, the use of statistical metrics is important.

Table 7. Model efficiency according to fit statistics—Region 1.

Stations	Samples	Statistics	LRM	MRM	Machine Learning				Optimized Machine Learning			
					MRM	KNN	GBT	RF	MRM	KNN	GBT	RF
Ayaviri	Train	R^2	0.36	0.49	0.48	0.57	0.64	0.89	0.48	0.49	0.59	0.59
	Train	RMSE	3.15	2.81	2.87	2.61	2.39	1.36	2.87	2.89	2.55	2.58
	Train	NSE	0.36	0.49	0.48	0.57	0.64	0.88	0.48	0.47	0.59	0.58
	Train	PBIAS	0.00	0.00	0.00	3.92	0.00	−1.73	0.00	0.45	0.00	0.38
	Test	R^2			0.52	0.38	0.49	0.45	0.52	0.48	0.71	0.70
	Test	RMSE			2.62	3.03	2.75	2.86	2.62	2.83	2.05	2.14
	Test	NSE			0.52	0.36	0.47	0.43	0.52	0.44	0.71	0.68
	Test	PBIAS			0.00	0.67	−8.46	−10.65	0.00	21.64	0.00	1.01
Huarochiri	Train	R^2	0.34	0.49	0.49	0.60	0.65	0.92	0.49	0.51	0.60	0.61
	Train	RMSE	3.12	2.74	2.80	2.47	2.32	1.19	2.80	2.76	2.49	2.48
	Train	NSE	0.34	0.49	0.49	0.60	0.65	0.91	0.49	0.50	0.60	0.60
	Train	PBIAS	0.00	0.00	0.00	6.48	0.00	−1.39	0.00	4.38	0.00	0.47
	Test	R^2			0.52	0.41	0.51	0.49	0.53	0.53	0.73	0.73
	Test	RMSE			2.54	2.83	2.58	2.64	2.51	2.57	1.93	1.96
	Test	NSE			0.52	0.40	0.50	0.48	0.53	0.51	0.72	0.71
	Test	PBIAS			−1.72	7.12	−5.63	−9.30	0.00	18.36	0.00	0.95
San Lazaro de Esco-marca	Train	R^2	0.30	0.38	0.38	0.49	0.65	0.90	0.38	0.41	0.54	0.45
	Train	RMSE	3.44	3.22	3.16	2.87	2.42	1.41	3.17	3.11	2.75	3.03
	Train	NSE	0.30	0.38	0.38	0.49	0.64	0.88	0.38	0.40	0.53	0.43
	Train	PBIAS	0.00	0.00	0.00	7.01	0.00	−1.96	0.00	10.88	0.00	0.14
	Test	R^2			0.42	0.28	0.34	0.37	0.41	0.43	0.73	0.56
	Test	RMSE			3.33	3.73	3.55	3.46	3.34	3.35	2.31	2.98
	Test	NSE			0.42	0.27	0.33	0.37	0.41	0.41	0.72	0.53
	Test	PBIAS			0.00	16.25	9.41	1.78	0.00	14.86	0.00	−0.05
Langa	Train	R^2	0.30	0.39	0.40	0.53	0.68	0.93	0.40	0.45	0.59	0.60
	Train	RMSE	1.98	1.85	1.85	1.64	1.37	0.71	1.85	1.80	1.55	1.55
	Train	NSE	0.30	0.39	0.40	0.53	0.67	0.91	0.40	0.43	0.58	0.58
	Train	PBIAS	0.00	0.00	0.00	5.79	0.00	−3.09	0.00	10.61	0.00	0.60
	Test	R^2			0.36	0.24	0.32	0.31	0.37	0.36	0.70	0.70
	Test	RMSE			1.85	2.09	1.94	1.98	1.83	1.87	1.28	1.33
	Test	NSE			0.35	0.17	0.28	0.26	0.37	0.34	0.69	0.67
	Test	PBIAS			−4.32	0.76	−7.22	−16.36	0.00	17.93	0.00	1.23

The R^2 values for the dataset (training and test) present a correlation between the *Y* and *X* variables in each model. For the Ayaviri station, which belongs to homogenous region 1 (see Table 7), the ML-RF model gives the best R^2 value (0.89) for the training data; however, for the test dataset, this value is reduced by nearly half (R^2 = 0.45). For optimized ML, the training and test R2 values are close to each other, and in some cases, the R^2 values are better for the test datasets than the training datasets. RMSE is a measure of the variance of residuals, which allows the magnitude of deviation of simulated values from observed values to be quantified; the LRM model presents the greatest RMSE (3.15) for the Ayaviri station. It was also observed that the test dataset generally presents a lower RMSE, particularly with the optimized ML models (OML-GBT and OML-RF).

The NSE is a tool that measures the predictive capacity of a model, which can take values between $-\infty$ and 1.0, with 1.0 being the optimal value. Values between 0.0 and 1.0 are generally seen as acceptable performance levels, while values equal to or less than 0.0 indicate that the mean of the observed values is a better predictor than the simulated value, indicating inadequate performance [60]. In accordance with the results shown in Table 7, values for the Ayaviri station are between 0.36 and 0.88 for both datasets (test and training). However, the ML models present values very close to 1.0 (ML-RF, NSE = 0.88) for the training dataset, indicating an acceptable level of performance.

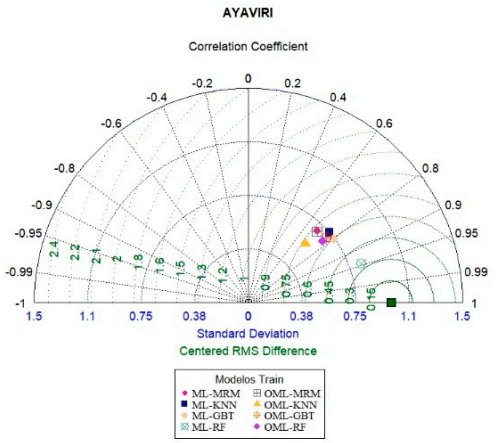
PBIAS measures the tendency of simulated data to be larger or smaller than their observed counterparts; its optimal value is 0. Positive values indicate a model with an underestimation bias and negative values indicate an overestimation bias. For the Ayaviri station, the OML-KNN presents high underestimation (PBIAS = 21.64), while the ML-RF model presents high overestimation (PBIAS = −10.65). However, the LRM, MRM, ML-MRM, OML-MRM, and OML-GBT models present an optimal PBIAS value for both datasets (training and test).

4. Discussion

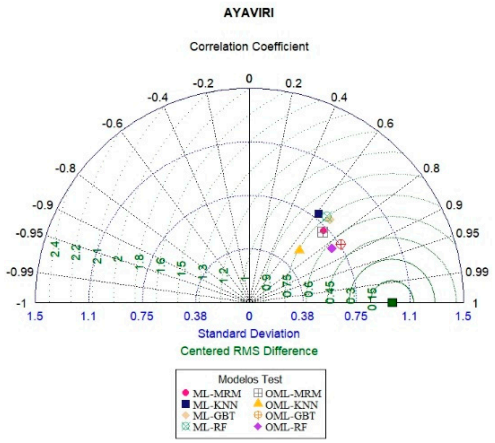
Based on the results obtained in the exploratory analysis, the Cañete, Socsi and Pacaran stations presented large quantities of missing values (over 10%); they also failed the homogeneity test. Therefore, the initial number of stations (17) was reduced to 14. In this study, the RM and ML methods were used to fill gaps in daily precipitation series at stations located in the MOC watersheds on the Peruvian Pacific Slope and coast. The procedure was carried out in three stages: collection of information on daily precipitation series, exploratory analysis, and homogenization. Therefore, it is essential to implement quality control procedures for raw rainfall data to ensure their reliability for use. In addition, preliminary ward cluster analysis, followed by KM and RVM analysis, through which three homogenous regions that concisely represent the relationship between precipitation variability and altitude were identified; and, finally, RM and ML were applied as a method of filling gaps in precipitation series.

RM and ML are customizable and easy-to-implement techniques that seek the best performance for a given problem among numerous algorithms. ML analyses with hyperparameter values (OML-MRM, OML-KNN, OML-GBT, and OML-RF) presented the best data recovery performance, demonstrating that ML models can extract additional information from data that by nature present noisy characteristics due to their high spatial and temporal variability [8,34,39]. In general terms, the decision tree methods (OML-GBT and OML-RF) perform the regression task well; however, some variations are observed that demonstrate that not all ML algorithms are equal in datasets that are superficially similar and can vary widely in terms of their prediction power. This also underlines the variation in the mechanisms of ML algorithms, even though all of them are capable of extracting information from non-linear and noisy datasets.

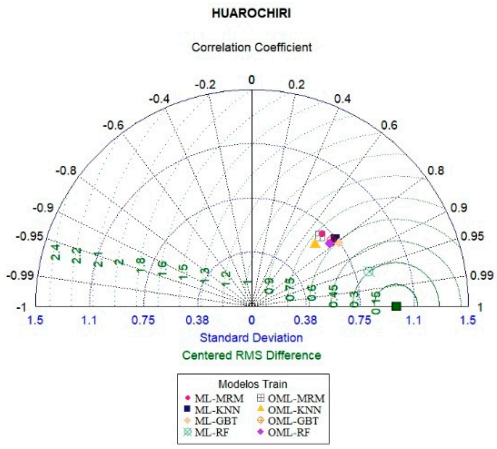
Figure 11a shows the Taylor diagram for the Ayaviri station for the training dataset; the ML-RF model presents the best results, with prediction precipitation the most consistent with observed precipitation (R^2 = 0.89, RMSE = 1.36, NSE = 0.88, and PBIAS = −1.73). Figure 11b shows the Taylor diagram for the Ayaviri station for the test dataset; the OML-GBT and OML-RF present the best results (R^2 = 0.71, RMSE = 2.05, NSE = 0.71, and PBIAS = 0.00 and R^2 = 0.70, RMSE = 2.14, NSE = 0.68 y PBIAS = 1.01, respectively). The analyses of the other stations (Huarochiri, San Lazaro de Escomarca, and Langa), are shown in Figure 11c–h; all these stations are located in homogenous region 1, and the values of the results obtained for them are similar to those of the Ayaviri station. Likewise, it is observed that in terms of the statistical metrics for the training and test datasets, the optimized ML models present the best results, particularly the OML-GBT and OML-RF models. The results of the analysis of the statistical metrics are shown in the figures. For the Ayaviri station, the OML-RF model presents a slight underestimation, while the results of the OML-GBT model are more efficient. Finally, in regions 2 and 4, Figures S3 and S4 respectively (Supplementary Material), the OML-GBT and OML-RF present the best results in terms of statistical metrics.



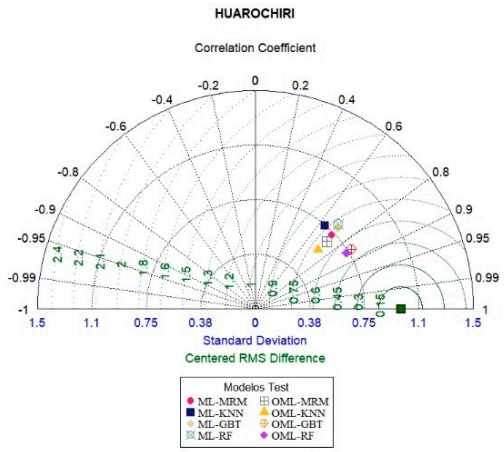
(**a**)



(**b**)



(**c**)



(**d**)

Figure 11. *Cont.*

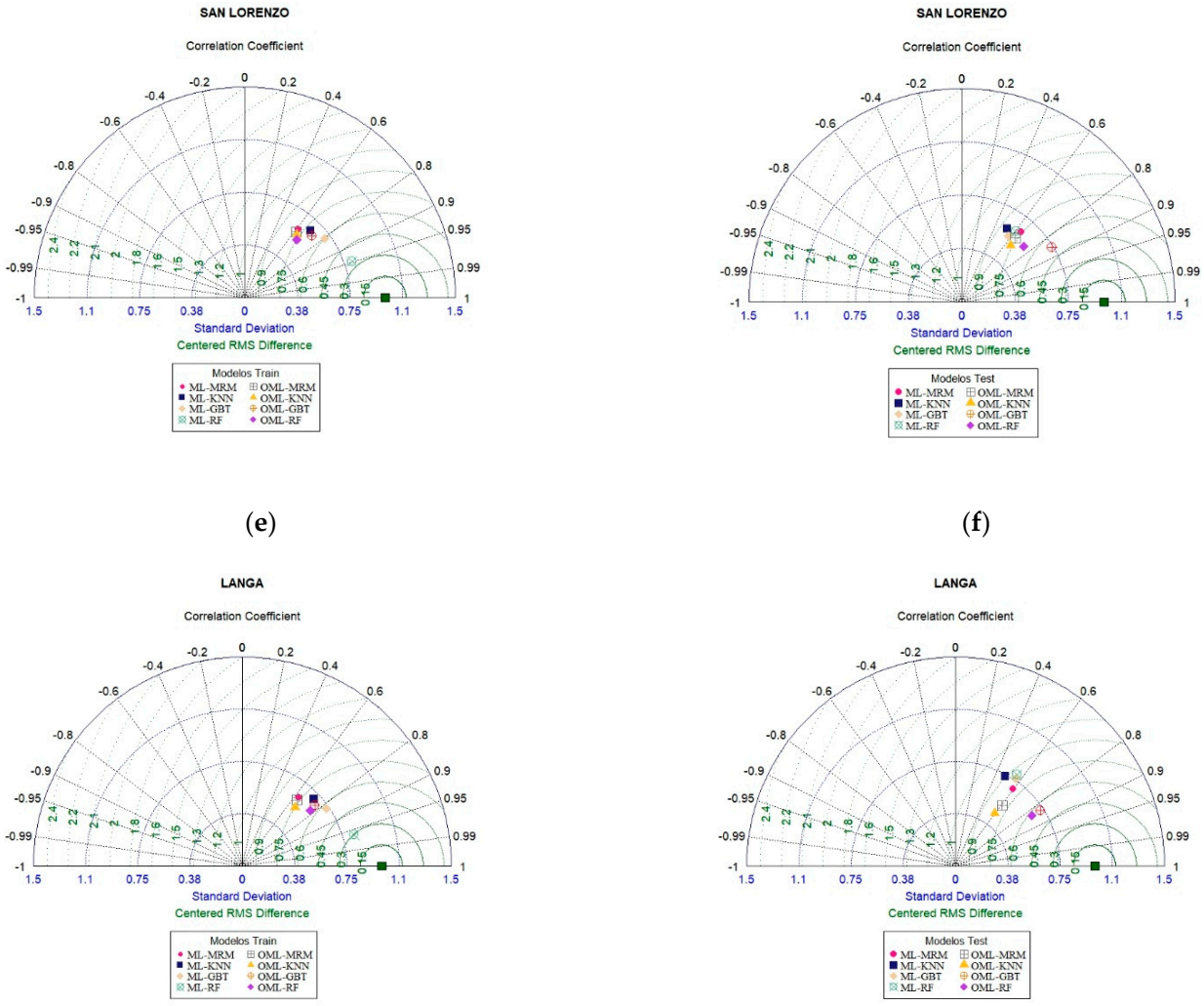


Figure 11. Taylor diagrams that show a statistical comparison (normalized standard deviation and correlation coefficient) of observed precipitation and modeled precipitation based on precipitation datasets (training and test) for four stations: (**a**) Ayaviri (training), (**b**) Ayaviri (test), (**c**) Huarochiri (training), (**d**) Huarochiri (test), (**e**) San Lazaro de Escomarca (training), (**f**) San Lazaro de Escomarca (test), (**g**) Langa (training), and (**h**) Langa (test).

5. Conclusions

This study has demonstrated the performance advantages of ML techniques for filling gaps in daily precipitation series as well as the potential of ML models in the optimization process using hyperparameter values for training (75%) and test datasets (25%), based on the efficiencies of the statistical metrics. However, it is important to note that a quality control raw rainfall data and regionalization process are necessary, which allows homogenous regions to be identified. Precipitation along the Peruvian Pacific Slope is highly influenced by El Niño, with marked positive asymmetry of strong events, and La Niña, with non-Gaussian distribution of precipitation data, which limits to a certain extent the linear analysis approach [9]. Finally, the results obtained in this study showed that the OML-GBT and OML-RF models presented the least variability in estimation errors and the best approximation to the actual data, efficiently interpreting the spatiotemporal variability of precipitation, as demonstrated by the analyzed statistical metrics.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/w14111799/s1, Figure S1: Annual indices of the regional vector and stations in Region 2; Figure S2: Annual indices of the regional vector and stations in Region 4; Figure S3: Taylor diagrams that show a statistical comparison (normalized standard deviation and correlation coefficient) of observed precipitation and modeled precipitation based on precipitation datasets (training and test) for four stations: (a) San Pedro de Pilas (training), (b) San Pedro de Pilas (test), (c) Huangascar (training), (d) Huangascar (test), (e) Yayos (training), (f) Yayos (test), (g) San Juan de Yanac (training), and (h) San Juan de Yanac (test); Figure S4: Taylor diagrams that show a statistical comparison (normalized standard deviation and correlation coefficient) of observed precipitation and modeled precipitation based on precipitation datasets (training and test) for four stations: (a) Tanta (training), (b) Tanta (test), (c) Yauricocha (training), (d) Yauricocha (test), (e) Vilca (training), (f) Vilca (test); Table S1: Annual regional vector indices—Region 2; Table S2: Annual regional vector indices—Region 4; Table S3: K-means clustering (2001–2019 period); Table S4: Correlation coefficient—Region 2; Table S5: Correlation coefficient—Region 4; Table S6: Model efficiency according to fit statistics—Region 2; Table S7: Model efficiency according to fit statistics—Region 4.

Author Contributions: Conceptualization, M.P.-M. and J.L.A.; methodology, M.P.-M.; software, M.P.-M.; validation, M.P.-M. and J.L.A.; formal analysis, M.P.-M.; investigation, M.P.-M., J.L.A., O.L., A.S. and N.M.A.; writing—original draft preparation, M.P.-M.; writing—review and editing, M.P.-M., J.L.A., O.L., A.S. and N.M.A.; visualization, M.P.-M. and N.M.A.; supervision, J.L.A. and A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the CRHIAM Water Research Center: ANID/FONDAP/15130015.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available at https://www.mdpi.com/article (accessed on 25 March 2022), time series, algorithms and other.

Acknowledgments: CRHIAM Water Research Center, Project ANID/FONDAP/15130015, Universidad Nacional Agraria La Molina, Eliana Contreras López and Ricardo León Ochoa.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Li, D.; Christakos, G.; Ding, X.; Wu, J. Adequacy of TRMM satellite rainfall data in driving the SWAT modeling of Tiaoxi catchment (Taihu lake basin, China). *J. Hydrol.* **2018**, *556*, 1139–1152. [CrossRef]
2. Santos, L.O.F.D.; Querino, C.A.S.; Querino, J.K.A.D.S.; Pedreira Junior, A.L.; Moura, A.R.D.M.; Machado, N.G.; Biudes, M.S. Validation of rainfall data estimated by GPM satellite on Southern Amazon region. *Rev. Ambiente Água* **2019**, *14*. [CrossRef]
3. Zambrano-Bigiarini, M.; Nauditt, A.; Birkel, C.; Verbist, K.; Ribbe, L. Temporal and spatial evaluation of satellite-based rainfall estimates across the complex topographical and climatic gradients of Chile. *Hydrol. Earth Syst. Sci.* **2017**, *21*, 1295. [CrossRef]
4. Jiang, L.; Wu, J. *Hybrid PSO and GA for Neural Network Evolutionary in Monthly Rainfall Forecasting*; Springer: Berlin/Heidelberg, Germany, 2013.
5. Cramer, S.; Kampouridis, M.; Freitas, A.A.; Alexandridis, A.K. An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives. *Expert Syst. Appl.* **2017**, *85*, 169–181. [CrossRef]
6. Chen, F.; Gao, Y.; Wang, Y.; Qin, F.; Li, X. Downscaling satellite-derived daily precipitation products with an integrated framework. *Int. J. Climatol.* **2019**, *39*, 1287–1304. Available online: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/joc.5879 (accessed on 25 March 2022). [CrossRef]
7. Bai, P.; Liu, X. Evaluation of five satellite-based precipitation products in two gauge-scarce basins on the Tibetan Plateau. *Remote Sens.* **2018**, *10*, 1316. [CrossRef]
8. Chivers, B.D.; Wallbank, J.; Cole, S.J.; Sebek, O.; Stanley, S.; Fry, M.; Leontidis, G. Imputation of missing sub-hourly precipitation data in a large sensor network: A machine learning approach. *J. Hydrol.* **2020**, *588*, 125126. [CrossRef]
9. Lavado Casimiro, W.S.; Ronchail, J.; Labat, D.; Espinoza, J.C.; Guyot, J.L. Basin-scale analysis of rainfall and runoff in Perú (1969–2004): Pacific, Titicaca and Amazonas drainages. *Hydrol. Sci. J.* **2012**, *57*, 625–642. [CrossRef]

10. Espinoza Villar, J.C.; Ronchail, J.; Guyot, J.L.; Cochonneau, G.; Naziano, F.; Lavado, W.; De Oliveira, E.; Pombosa, R.; Vauchel, P. Spatio-temporal rainfall variability in the Amazon basin countries (Brazil, Peru, Bolivia, Colombia, and Ecuador). *Int. J. Climatol.* **2009**, *29*, 1574–1594. Available online: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/joc.1791 (accessed on 25 March 2022). [CrossRef]
11. Rau, P.; Bourrel, L.; Labat, D.; Melo, P.; Dewitte, B.; Frappart, F.; Lavado, W.; Felipe, O. Regionalization of rainfall over the Peruvian Pacific slope and coast. *Int. J. Climatol.* **2017**, *37*, 143–158. [CrossRef]
12. Körner, P.; Kronenberg, R.; Genzel, S.; Bernhofer, C. Introducing Gradient Boosting as a universal gap filling tool for meteorological time series. *Meteorol. Z.* **2018**, *27*, 369–376. [CrossRef]
13. Lavado Casimiro, W.; Espinoza, J.C. Impactos de El Niño y La Niña en las lluvias del Perú (1965–2007). *Rev. Bras. De Meteorol.* **2014**, *29*, 171–182. [CrossRef]
14. Bertsimas, D.; Pawlowski, C.; Zhuo, Y.D. From Predictive Methods to Missing Data Imputation: An Optimization Approach. *J. Mach. Learn. Res.* **2017**, *18*, 7133–7171. [CrossRef]
15. Teegavarapu, R.S.V.; Chandramouli, V. Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. *J. Hydrol.* **2005**, *312*, 191–206. [CrossRef]
16. Barrios, A.; Trincado, G.; Garreaud, R. Alternative approaches for estimating missing climate data: Application to monthly precipitation records in South-Central Chile. *For. Ecosyst.* **2018**, *5*, 28. [CrossRef]
17. Xia, Y.; Fabian, P.; Stohl, A. Winterhalter, Forest climatology: Estimation of missing values for Bavaria, Germany. *Agric. For. Meteorol.* **1999**, *96*, 131–144. [CrossRef]
18. Bostan, P.A.; Heuvelink, G.B.M.; Akyurek, S.Z. Comparison of regression and kriging techniques for mapping the average annual precipitation of Turkey. *Int. J. Appl. Earth Obs. Geoinf.* **2012**, *19*, 115–126. [CrossRef]
19. Mair, A.; Fares, A. Comparison of Rainfall Interpolation Methods in a Mountainous Region of a Tropical Island. *J. Hydrol. Eng.* **2011**, *16*, 371–383. [CrossRef]
20. Simolo, C.; Brunetti, M.; Maugeri, M.; Nanni, T. Improving estimation of missing values in daily precipitation series by a probability density function-preserving approach. *Int. J. Climatol.* **2010**, *30*, 1564–1576. [CrossRef]
21. Huang, M.; Lin, R.; Huang, S.; Xing, T. A novel approach for precipitation forecast via improved K-nearest neighbor algorithm. *Adv. Eng. Inform.* **2017**, *33*, 89–95. [CrossRef]
22. Gorshenin, A.; Lebedeva, M.; Lukina, S.; Yakovleva, A. Application of Machine Learning Algorithms to Handle Missing Values in Precipitation Data. In *Distributed Computer and Communication Networks*; Lecture Notes in Computer Science; Vishnevskiy, V., Samouylov, K., Kozyrev, D., Eds.; Springer International Publishing: Cham, Switzerland, 2019; p. 11965. [CrossRef]
23. Bellido-Jiménez, J.A.; Gualda, J.E.; García-Marín, A.P. Assessing Machine Learning Models for Gap Filling Daily Rainfall Series in a Semiarid Region of Spain. *Atmosphere* **2021**, *12*, 1158. [CrossRef]
24. Devi, U.; Shekhar, M.S.; Singh, G.P.; Rao, N.N.; Bhatt, U.S. Methodological application of quantile mapping to generate precipitation data over Northwest Himalaya. *Int. J. Climatol.* **2019**, *39*, 3160–3170. [CrossRef]
25. Estévez, J.; Bellido-Jiménez, J.A.; Liu, X.; García-Marín, A.P. Monthly Precipitation Forecasts Using Wavelet Neural Networks Models in a Semiarid Environment. *Water* **2020**, *12*, 1909. [CrossRef]
26. Sattari, M.T.; Rezazadeh-Joudi, A.; Kusiak, A. Assessment of different methods for estimation of missing data in precipitation studies. *Hydrol. Res.* **2016**, *48*, 1032–1044. [CrossRef]
27. Tang, G.; Clark, M.P.; Newman, A.J.; Wood, A.W.; Papalexiou, S.M.; Vionnet, V.; Whitfield, P.H. SCDNA: A serially complete precipitation and temperature dataset for North America from 1979 to 2018. *Earth Syst. Sci. Data* **2020**, *12*, 2381–2409. [CrossRef]
28. Tang, G.; Clark, M.P.; Papalexiou, S.M. SC-Earth: A Station-Based Serially Complete Earth Dataset from 1950 to 2019. *J. Clim.* **2021**, *34*, 6493–6511. [CrossRef]
29. Carrera-Villacrés, D.V.; Guevara-García, P.V.; Tamayo-Bacacela, L.C.; Balarezo-Aguilar, A.L.; Narváez-Rivera, C.A.; Morocho-López, D.R. Relleno de series anuales de datos meteorológicos mediante métodos estadísticos en la zona costera e interandina del Ecuador, y cálculo de la precipitación media. *Idesia* **2016**, *34*, 81–90. [CrossRef]
30. Luna Romero, A.E.; Lavado Casimiro, W.S. Evaluación de métodos hidrológicos para la completación de datos faltantes de precipitación en estaciones de la cuenta Jetepeque, Perú. *Rev. Tecnológica-ESPOL* **2015**, *28*, 42–52.
31. Estévez, J.; Gavilán, P.; Giráldez, J.V. Guidelines on validation procedures for meteorological data from automatic weather stations. *J. Hydrol.* **2011**, *402*, 144–154. [CrossRef]
32. Portuguez Maurtua, D.M. Aplicación de la Geoestadística a Modelos Hidrológicos en la cuenca del río Cañete. Master's Thesis, Universidad Nacional Agraria La Molina, Lima, Peru, 2017.
33. Guevara Ochoa, C.; Briceño, N.; Zimmermann, E.D.; Vives, L.S.; Blanco, M.; Cazenave, G.; Ares, M.G. Relleno de series de precipitación diaria para largos periodos de tiempo en zonas de llanura: Caso de estudio cuenca superior del arroyo del Azul. *Geoacta* **2017**, *42*, 38–60. Available online: http://www.scielo.org.ar/scielo.php?script=sci_arttext&pid=S1852-77442017000100004&lng=es (accessed on 5 March 2020).
34. Guijarro, J. Homogenization of climatic series with Climatol. *Rep. Técnico State Meteorol. Agency (AEMET)* **2018**, *3*, 1–20. Available online: https://www.climatol.eu/homog_climatol-en.pdf (accessed on 5 March 2020).
35. Toreti, A.; Kuglitsch, F.G.; Xoplaki, E.; Della-Marta, P.; Aguilar, E.; Prohom, M.; Luterbacher, J. A note on the use of the standard normal homogeneity test (SNHT) to detect inhomogeneities in climatic time series. *Int. J. Climatol.* **2011**, *31*, 630–632. [CrossRef]
36. Alexandersson, H. A homogeneity test applied to precipitation data. *J. Climatol.* **1986**, *6*, 661–675. [CrossRef]

37. Alexandersson, H.; Moberg, A. Homogenization of swedish temperature data. *Part I: Homogeneity test for linear trends. Int. J. Climatol.* **1997**, *17*, 25–34.
38. Moberg, A.; Alexandersson, H. Homogenization of swedish temperature data. *Part ii: Homogenized gridded air temperature compared with a subset of global gridded air temperature since 1861. Int. J. Climatol.* **1997**, *17*, 35–54.
39. Pandzic, K.; Kobold, M.; Oskorus, D.; Biondic, B.; Biondic, R.; Bonacci, O.; Likso, T.; Curic, O. Standard normal homogeneity test as a tool to detect change points in climate-related river discharge variation: Case study of the Kupa River Basin. *Hydrol. Sci. J.* **2020**, *65*, 227–241. [CrossRef]
40. Ahmad, N.H.; Deni, S.M. Homogeneity test on daily rainfall series for Malaysia. *Mat. Malays. J. Ind. Appl. Math.* **2013**, *29*, 141–150.
41. Marcolini, G.; Bellin, A.; Chiogna, G. Performance of the Standard Normal Homogeneity Test for the homogenization of mean seasonal snow depth time series. *Int. J. Climatol.* **2017**, *37*, 1267–1277. [CrossRef]
42. Ward, J.H. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.* **1963**, *58*, 236–244. [CrossRef]
43. Yashwant, S.; Sananse, S. Comparisons of Different Methods of Cluster Analysis with Application to Rainfall Data. *Int. J. Innov. Res. Sci.* **2015**, *4*, 10861. [CrossRef]
44. Luna Vera, J.A.; Domínguez Mora, R.T. Un método para el análisis de frecuencia regional de lluvias máximas diarias: Aplicación en los Andes bolivianos. *Ingeniare Rev. Chil. De Ing.* **2013**, *21*, 111–124. [CrossRef]
45. Ilbay, M.L.; Barragán, R.Z.; Lavado-Casimiro, W. Regionalization of precipitation, its aggressiveness and concentration in the Guayas river basin, Ecuador. *La Granja* **2019**, *30*, 57. [CrossRef]
46. Hiez, G. L'homogénéité des données pluviométriques. *Cah. ORSTOM Série Hydrol.* **1977**, *14*, 29–173.
47. Brunet-Moret, Y. Homogénéisation des précipitations. *Bur. Cent. Hydrol. De L'orstom Á Paris* **1979**, *16*, 147–170.
48. Vauchel, P. Hydraccess: Progiciel de gestion et d'exploitation de bases de données hydrologiques. In HYDROMED: Séminaire International les Petits Barrages Dans le Monde Méditerranéen: Recueil des Résumés. In Proceedings of the Les Petits Barrages dans le Monde Méditerranéen: Séminaire International, Tunis, North Africa, 28–31 May 2001.
49. Wang, J.-H.; Hopke, P.K.; Hancewicz, T.M.; Zhang, S.L. Application of modified alternating least squares regression to spectroscopic image analysis. *Anal. Chim. Acta* **2003**, *476*, 93–109. [CrossRef]
50. Bárdossy, A.; Pegram, G. Infilling missing precipitation records—A comparison of a new copula-based method with other techniques. *J. Hydrol.* **2014**, *519*, 1162–1170. [CrossRef]
51. Khosravi, G.; Nafarzadegan, A.R.; Nohegar, A.; Fathizadeh, H.; Malekian, A. A modified distance-weighted approach for filling annual precipitation gaps: Application to different climates of Iran. *Theor. Appl. Climatol.* **2015**, *119*, 33–42. [CrossRef]
52. Natekin, A.; Knoll, A. Gradient boosting machines, a tutorial. *Front. Neurorobotics* **2013**, *7*, 21. [CrossRef]
53. Ma, L.; Zhang, G.; Lu, E. Using the Gradient Boosting Decision Tree to Improve the Delineation of Hourly Rain Areas during the Summer from Advanced Himawari Imager Data. *J. Hydrometeorol.* **2018**, *19*, 761–776. [CrossRef]
54. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
55. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning*; Springer: Berlin/Heidelberg, Germany, 2013.
56. Bellido-Jiménez, J.A.; Estévez, J.; García-Marín, A.P. New machine learning approaches to improve reference evapotranspiration estimates using intra-daily temperature-based variables in a semi-arid region of Spain. *Agric. Water Manag.* **2021**, *245*, 106558. [CrossRef]
57. Bellido-Jiménez, J.A.; Estévez, J.; García-Marín, A.P. Assessing Neural Network Approaches for Solar Radiation Estimates Using Limited Climatic Data in the Mediterranean Sea. *Environ. Sci. Proc.* **2021**, *4*, 19.
58. Gómez Guerrero, J.S.; Aguayo Arias, M.I. Evaluación de desempeño de métodos de relleno de datos pluviométricos en dos zonas morfoestructurales del Centro Sur de Chile. *Investig. Geográficas* **2019**, *99*, 1–16. [CrossRef]
59. Guijarro, J.A.; Guijarro, M.J. Package 'Climatol'. 2019. Available online: https://doi.org/5.gwdg.de/pub/misc/cran/web/packages/climatol/climatol.pdf (accessed on 5 March 2020).
60. Moriasi, D.N.; Arnold, J.G.; Van Liew, M.W.; Bingner, R.L.; Harmel, R.D.; Veith, T.L. Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations. *Trans. ASABE* **2007**, *50*, 885–900. [CrossRef]