International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor (2018): 7.426

Homogenization of Monthly Rainfall Data Series in Lerma-Toluca Watershed with Climatol

Ruy Ponce-Cruz[1], Lamine Diakite[2], Alejandro I. Monterroso-Rivas[3], Ronald E. Ontiveros-Capurata[4], Guillermo Crespo-Pichardo[5]

[1, 2, 3]Post Graduate, Student Agricultural Engineering and Integral Use of Water. Autonomous University of Chapingo. 56230. Chapingo, State of México

[4]Mexican Institute of Water Technology. Jiutepec, Morelos, México

[5]Postgraduate College Campus Montecillos- Agrometeorological Station. 56230. Montecillo, State of México

Abstract: ***The lack ofavailability of complete and reliable climatological data series often represents the main difficulty in carrying out hydrological studies. The causes can be diverse, such as human and instrumental errors, false and incomplete records, and the use of outdated equipment at some weather stations, which give rise to the appearance of inhomogeneities unrepresentative of the climatic reality. This study was carried out in the Lerma-Toluca Watershed, Mexico, using145weather stations with monthly 24-hour maximum precipitation data from 1990 to 2015. The homogenization and estimation of the missing data were conducted with the Climatol version 3.1.1 package for the statistical R application (Free software). The statistics considered were: Absolute maximum of autocorrelation of anomalies per station (ACmx), Standard normal homogeneity test (SNHT), Root mean square error (RMSE) and Percentage of original data (POD). The results obtained suggest considering an adequate percentage of original data greater than 60% in the case of this study in order to reduce the RMSE values and keep the stations with the most complete data. The Climatol package was very versatile and practical in the homogenization and estimation of missing precipitation data.***

Keywords: climatic series, maximum precipitation, free software

1. Introduction

The lack of availability of homogeneous and complete climatological data series often represents the main difficulty in carrying out studies on rainfall erosion in watersheds using the Universal Soil Loss Equation. The causes can be diverse, such as coding and instrumental errors, inconsistency in the records, and outdated equipment at some weather stations, which give rise to the appearance of inhomogeneities unrepresentative of climatic variability [1–3].

According to Conrad & Pollack (1962), a series is homogeneous when its variability is due solely to climatic causes, otherwise it is necessary to detect and correct errors, a process known as homogenization [4].

In the case of incomplete series and the presence of missing data, the use of gap-filling techniques is required.

There is an extensive literature dedicated to the different homogenization methods that can be used given the needs[2, 3].

Homogenization and filling in of missing data becomes a laborious and tedious procedure for large volumes of data; it is therefore necessary to use existing applications in the form of free software packages [4, 5].

The main objective of this study is the homogenization of the monthly 24-hour maximum precipitation data and the filling in of the missing data in order to estimate the rainfallerosion in the Lerma-Toluca Watershed, Mexico, as shown in Figure 1, using the Revised version of the Universal Soil Loss Equation (RUSLE-2). Twenty-six monthly precipitation data series (from 1990 to 2015) from145 weather stations were used, with the information provided by the National Meteorological Service of Mexico (SMN).

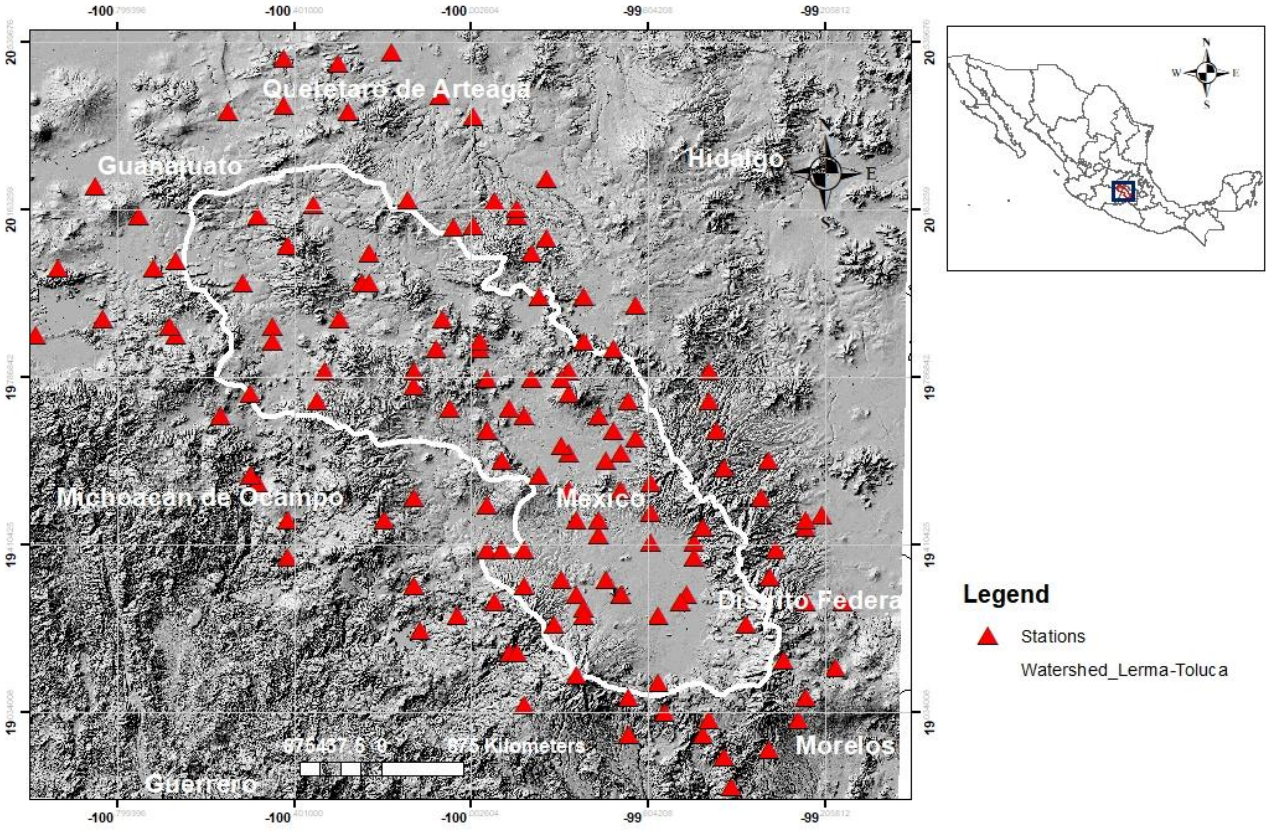

Figure 1: Lerma-Toluca watershed, Mexico

Due to the large volume of data (3770 records), we used the R-language Climatol 3.1.1statistical package[6], which allows the automation of the complete process of homogenization and filling in of missing data, and which has been tested in various parts of the world with daily, monthly and annual precipitation and maximum-minimum temperature data[1, 4, 7].

The methodology used by Climatol is based on the Standard Normal Homogeneity Test (SNHT) proposed by Alexandersson & Moberg (1997).

The results obtained indicate that more than 90% of the data after homogenization are consistent and reliable.

2. Materials and Methods

Study Area

The study area corresponds to the Lerma-Toluca Watershed, located within Lerma-Santiago Hydrological Region No. 12. It extends from the center of the State of Mexico to the northwest of the states of Querétaro and Michoacán. The watershed is located between coordinates 19° 05' and 20° 05' N and 99° 25' and 100° 15' W(Fig. 1). It has an elongated shape with a northwest-southeast orientation and a length of 133 km. It is bordered by the Valley of Mexico and the Panuco River basins to the north and the Balsas Riverto the south. The database is made up ofmonthly 24-hour maximum precipitation series (P_{24})from 145weather stations.

Data preparation

The monthly series were debugged by eliminating those stations that had no records for the 26 years covered by the study. The input files consist of two parts: the first corresponds to the station identification and location data; the second part is the file with monthly data for each station [6]. These input files were converted into "MS-DOS with return to the next line" format to be recognized byClimatol.

Description of the mathematical model used by Climatol

In Climatol, the Standard Normal Homogeneity Test (SNHT) is based on the procedure proposed by Alexandersson (1986) and modified by Alexandersson &Moberg (1997). This test is based on the hypothesis that the rainfall amounts in a candidate station for the test are proportional to some regional averages. This proportionality relationship is expressed in terms of the*Q* ratio between the normalized precipitation values of the candidate station and those of a regional time series defined as a weighted average of several neighboring reference stations. Therefore, the $Q_i$ratio in a specific year is calculated as shown in Equation 1 [8]:

$$Q_i = \frac{A_i}{B_i}, \quad i = 1, \ldots, n \qquad (1)$$

Where:
*i*is the time series
$A_i$is the precipitation function of the candidate station that is calculated using Equation 2[9]:

$$A_i = \frac{P_i}{\bar{P}} \qquad (2)$$

$$B_i = \left\{\left[\sum_{j=1}^{k} \rho_j^2 X_{ji} \bar{Y} / \bar{X}_j\right] / \sum_{j=1}^{k} \rho_j^2\right\} \qquad (3)$$

Where:
Pi is the mean precipitation of the candidate station
*P*is the mean precipitation for the time series
$X_{ij}$is the precipitation at the reference station
$\bar{Y}$and$\bar{X}_j$aremean values of the candidate station and mean values of the neighboring reference sites
$\rho_j$is the weight factor for the reference station j-th. It is defined as the squared correlation coefficient between the candidate series and the j-th reference series[8].
k is the number of reference sites used in a given year of the *time* series, which varies byyear.
To apply the test, the $Q_i$values must be normalized [8–10] using Equation 4:

$$z_i = \frac{Q_i - \bar{Q}}{\sigma_Q} \quad (4)$$

z_i is the normalized precipitation series with mean μ =0and standard deviation σ = 1.

In the null hypothesis test (H_0), the series is homogeneous, i.e. the $z_i$values have a normal distribution N (0,1). In the alternative hypothesis (H_1),the test is inhomogeneous and contains breakpoints in the series. The breakpoints identified in the time series are called *T(k)*series expressed in Equation 5andbroken downin Equation 6 and 7[9, 10].

$$T(k) = k\bar{z}_1^2 + (n-k)\bar{z}_2^2, \; k = 1 \ldots, n \quad (5)$$

$$\bar{z_1} = \frac{1}{k}\sum_{i=1}^{k} z_i \quad (6)$$

$$\bar{z_2} = \frac{1}{(n-k)}\sum_{i=k+1}^{k} z_i \quad (7)$$

Where:
T(k) is the time series
$\bar{z_1}$is theestimated average level of standardizedproportions before a possible change or trend.
$\bar{z_2}$is the estimated average level of proportions after a possible change or trend.
The test statistic denoted in Equation 8 [8, 11] is:

$$T_{max} = \max_{1 \le k \le 0}\{T(k)\} \quad (8)$$

Once the time series is estimated, the SNHT is applied. As in the previous case, the null hypothesis is rejected if the value of the statistic does not reach the level of significance for the size of the sample considered.

Homogenization process with Climatol
Based on the methodology described above, *Climatol* detects anomalous values, which consists of identifying values that do not correspond to a given range. The range of anomalous values for a series was estimated according to its mean and an amplitude of two standard deviations (std=2) given the specific characteristics of the precipitation variable[6, 12]. *Climatol*performs two procedures by default: 1) detection of changes in the mean in blocks (*snht1*), using as the maximum SNHT threshold *snht1*= 25; and 2) detection of changes in the whole series (*snht2*), setting as a maximum threshold *snht2* = 50. The maximum values that exceed the defined thresholds indicate the period where the series changes its normal behavior. The values following the datum in which a shiftwas detected in the mean are stored as a new series and the missing data are completed by the inverse distance method expressed in Equation 9[13]. Once the values are completed, the SNHT is reapplied to the whole series[6].

$$S_r = \frac{\sum_{i=1}^{n} m_i \left(\frac{1}{1+\frac{d_i^2}{wd^2}}\right)}{\sum_{i=1}^{n} \left(\frac{1}{1+\frac{d_i^2}{wd^2}}\right)} \quad (9)$$

Where:
$S_r$is the estimated monthly record for the reference series, $m_i$are the records of all available stations, *d* is the distance in km from the analysis station and the recording station considered and *wd* is the weighting distance in km. Forall cases, a value of*wd= C(0,50,100)*was assigned, considering that there are very little data per month. The anomalous data and missing records were replaced by the value estimated by the correlation, in order to obtain complete series for the analysis periods[7].

3. Results and Discussion

Exploratory analysis of the monthly 24-hour maximum precipitation series (P_{24}).
The exploratory analysis of the precipitation series data (Figure 2) shows a low correlation, with a correlation coefficient less than 50% between stations located up to a distance of 150 km.

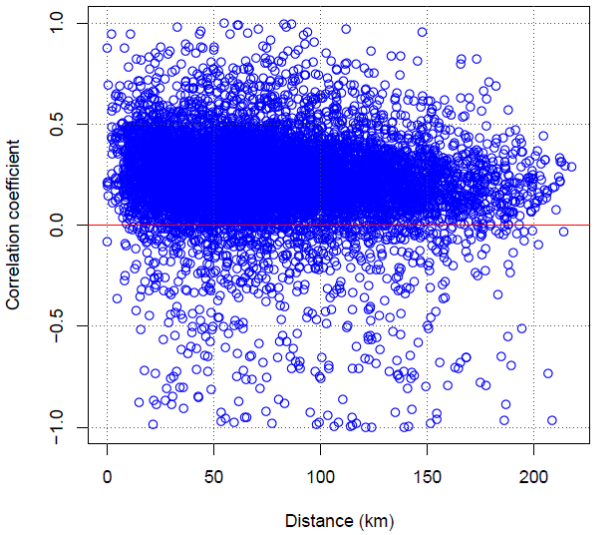

Figure 2: Correlation coefficient distance between stations and their elevation

An example of anomaly detection with Climatol is shown in Figure 3. The anomalies of each station are shown in the form of vertical lines pointing to the date of occurrence. When the maximum value of the shift test in the mean exceeds the established threshold, the position where the series will be cut is marked with a vertical dotted line, establishing that there are different means and series, labeled in its upper part with the test value, as shown in the figure; around the years 1993 and 2014 important shifts are detected (maximum value of the statistic std= 9 and 13). The lower part of the Figure shows in a green line the distance in kilometers to the nearest datum in each time step at logarithmic scale.

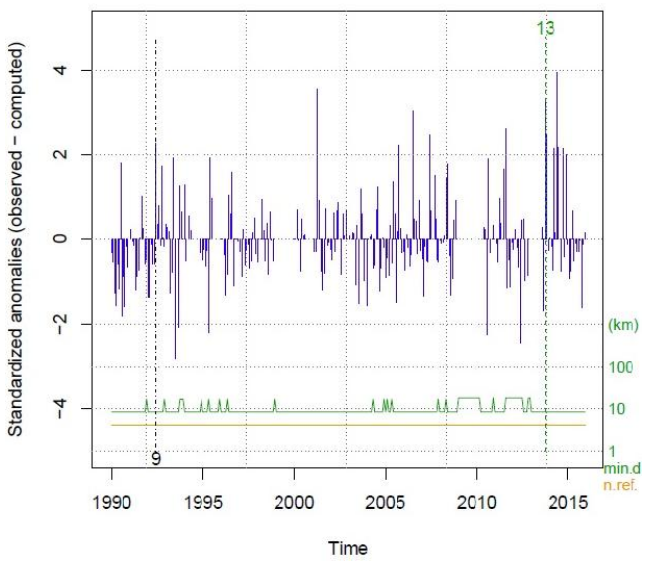

Figure 3: Anomaly detection

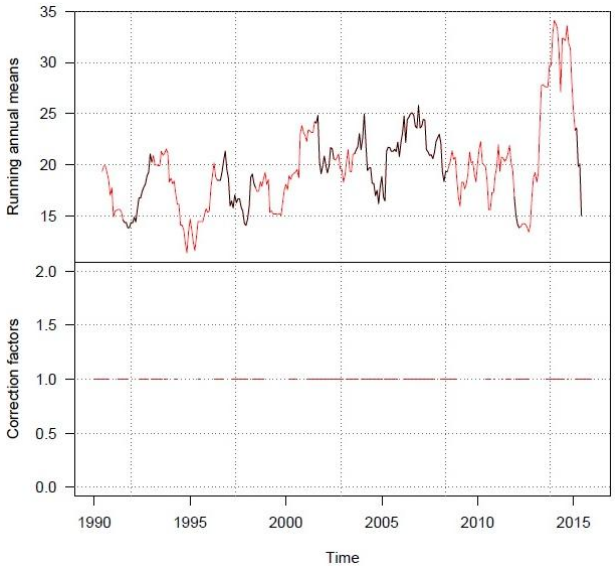

Figure 4: Fitting of series and application of corrections

Once the anomalies in the series have been identified, the fit and application of errors (filling in missing data) are undertaken as shown in Figure 4, which indicates the filled-in missing data with dark line segments.

The homogenization process consisting of the exploratory analysis, anomaly detection, fitting and application of corrections to the precipitation series is done station by station until completing the 145 stations proposed in this study, as in the case of Figures 2, 3 and 4 corresponding to Calvario Station 61(Tlalpan).

To analyze the homogeneity index (SNHT) of allthe weather stations,we represent it geospatially with inverse distance weighting (IDW) and generated a map that shows the different areas of the index. Figure 5 indicates that the areas with the greatest inhomogeneities are located outside the watershedbecause in those areas many stations were not considered and the neighboring stations had many missing data. This result is congruent with Figure 6, where most of the stations are in a homogeneity test (SNHT) range of less than 15, with an RMSE between 5 and 15.

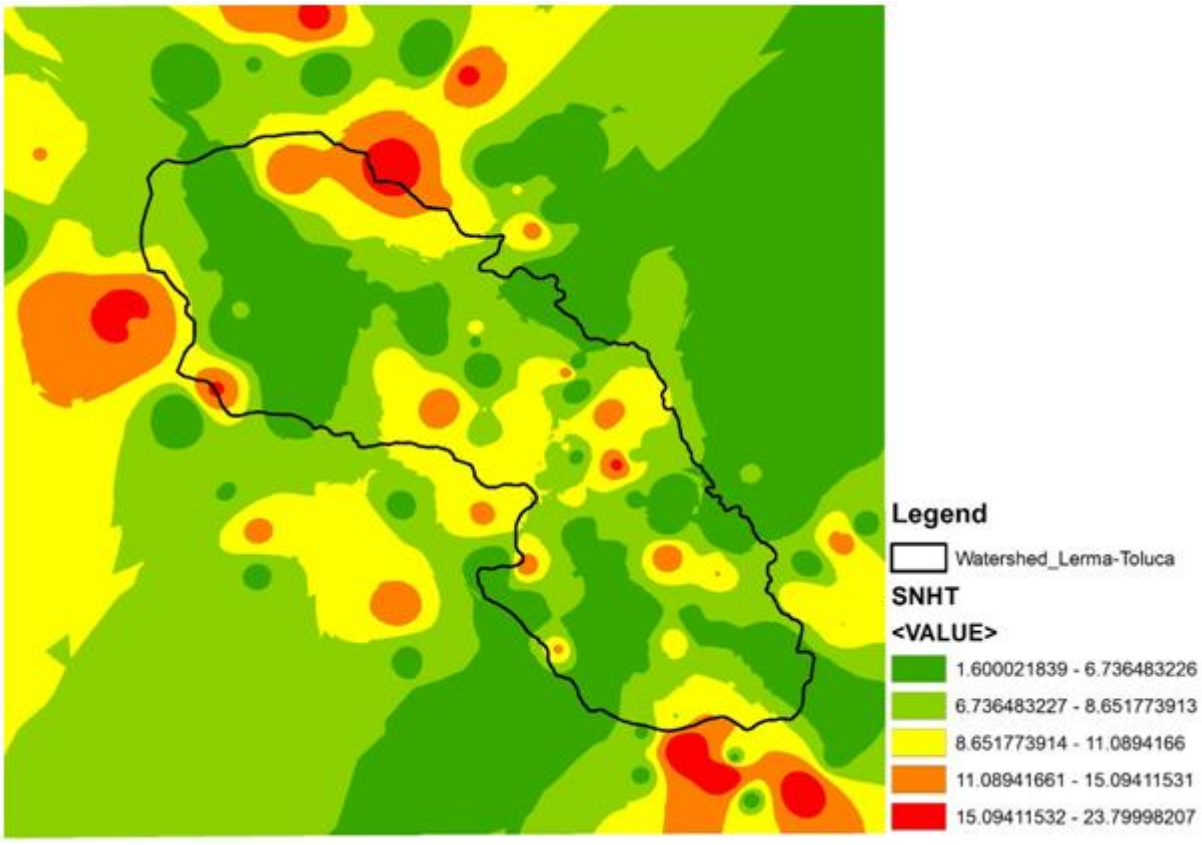

Figure 5: IDW interpolation of SNHT indices

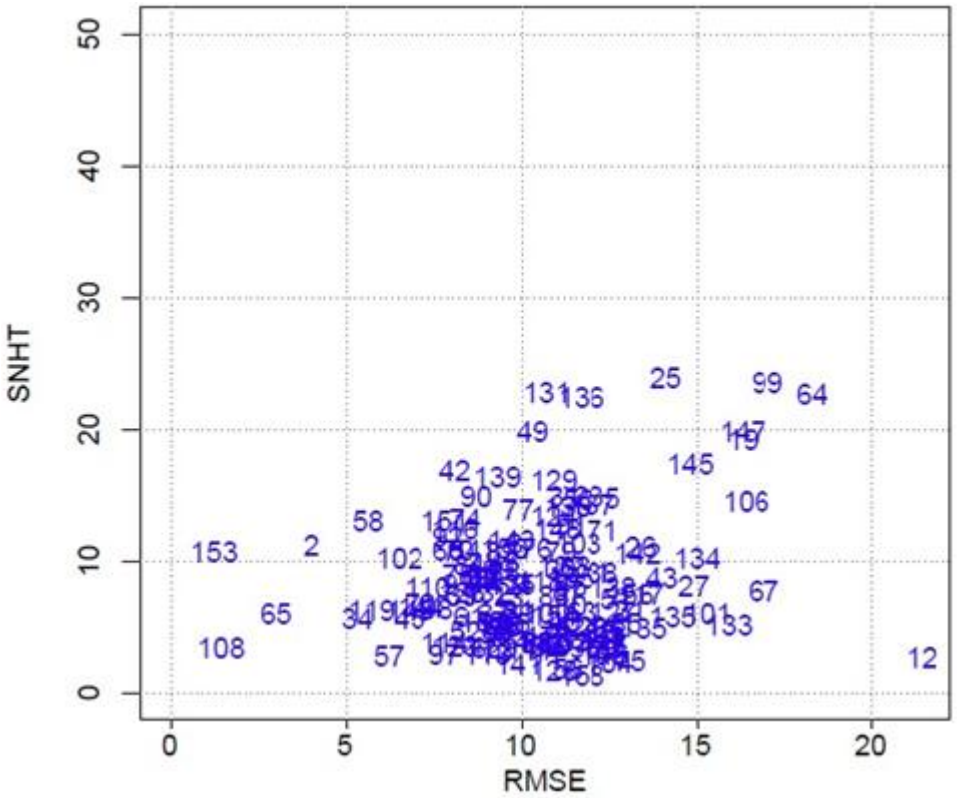

Figure 6: RMSE versusSNHT graph

4. Conclusions

The monthly 24-hour maximum precipitation series of 145 stations, located in the Lerma-Toluca River Watershed, Mexico during the period 1995-2015, were subjected to a process of homogenization and filling in of missing data, through the use of the Climatol 3.1.1 R package. Several tests were performed to fit the predetermined software values to two standard deviations (std=2), the threshold of the homogeneity test to snht1= 25, snht2= 50 and the weighted distance to wd= C(0,50,100).

The exploratory analysis showed little correlation between neighboring stations at very long distances (>150 km) but had a better fit at distances<100 km. Another important factor is the correlation of elevation and distance between stations; a large number of stations with similar elevation and distance favors reliable results. The statistics generated by Climatolallow analyzing the quality of the results. The statistics obtained (ACmx, SNHT, RMSE, POD)in the result indicate 60% original data which based on the geostatistical analysis ensures homogeneity with RMSE less than 15, which is acceptable for the purpose of subsequent studies.

The Climatol package was very versatile and practical in the homogenization and estimation of missing precipitation data. In the case of Mexico, there is no literature concerning the Climatol software, which is becoming popular in the world due to its effectiveness.

References

[1] Hernández García EM., García Valero JA., Palenzuela Cruz JE., Belda Esplugues F (2012). Ejercicio de homogeneización y relleno de series diarias de temperatura máxima, mediante el uso de Climatol. *VIII Congr Int la Asoc Española Climatol*:409–419.

[2] Cristina A., Amílcar C (2009). Homogenization of Climate Data : Review and New Perspectives Using Geostatistics. 291–305.

[3] Aguilar E., Auer I., Brunet M., Peterson TC., Wieringa J (2003). Guidance on metadata and homogenization. *Wmo Td* 1186(October):53.

[4] Acquaotta F., Fratianni S (2014). The Importance of the Quality & Reliabiluty of the Historical Time Series for the Study of Climate Change. *Rev Bras Climatol* 14(14):20–38.

[5] Ribeiro S., Caineta J., Costa AC (2016). Review and discussion of homogenisation methods for climate data. *Phys Chem Earth* 94:167–179.

[6] Guijarro JA (2018). Homogeneización de series climáticas con Climatol Versión 3.1.1. 1:22.

[7] Gaona G., Quentin E., Labus J (2013). Homogeneidad y variabilidad espacial de series meteorológicas del área del proyecto Ciudad del Conomimiento-Yachay. *Av en Ciencias e Ing* 5(2). doi:10.18272/aci.v5i2.138.

[8] Alexandersson H., Moberg A (1997). Homogenization of Swedish Temperature Data. Part I: Homogeneity Test for Linear Trends. *Int J Climatol* 17(1):25–34.

[9] González-Rouco JF., Jiménez JL., Quesada V., Valero F (2001). Quality control and homogeneity of precipitation data in the southwest of Europe. *J Clim* 14(5):964–978.

[10] Marcolini G., Bellin A., Chiogna G (2017). Performance of the Standard Normal Homogeneity Test for the homogenization of mean seasonal snow depth time series. *Int J Climatol* 37(January):1267–1277.

[11] Ahmad NH., Deni SM (2013). Homogeneity Test on Daily Rainfall Series for Malaysia. *Matematika* 29(1):141–150.

[12] Luna MY., Guijarro JA., López JA (2012). A monthly precipitation database for Spain (1851–2008): reconstruction, homogeneity and trends. *Adv Sci Res* 8:1–4.
[13] Wei TC., McGuinness JL (1973). Reciprocal distance squared method, a computer technique for estimating areal precipitation. *ARS-NC*. Available at: http://agris.fao.org/agris-search/search.do?recordID=US201400108710 [Accessed February 4, 2019].

Author Profile

Ruy Ponce-Cruz, Postgraduate Student in AgriculturalEngineering and Integral Use of Water. AutonomousUniversity of Chapingo. 56230. Chapingo, State of México.

Lamine Diakite, Ph.D. and Professor at theAutonomousUniversity of Chapingo. CEPRAE, México and Consultant in RemoteSensing and GIS.

Alejandro I. Monterroso- Rivas, Professor - Full Time Researcher, Division of ForestSciences. AutonomousUniversity of Chapingo, México.

Ronald E. Ontiveros-Capurata, Professor–CONACYT-MexicanInstitute of WaterTechnology. Jiutepec, Morelos, México. Field of specialty: GIS, remoteperception and spatialanalysis.

Guillermo Crespo-Pichardo, Researcher at the Postgraduate College, in charge of the Agrometeorological Station.