An Acad Bras Cienc (2023) 95(2): e20210737 DOI 10.1590/0001-3765202320210737
Anais da Academia Brasileira de Ciências | *Annals of the Brazilian Academy of Sciences*
Printed ISSN 0001-3765 I Online ISSN 1678-2690
www.scielo.br/aabc | www.fb.com/aabcjournal

GEOSCIENCES

Imputation of precipitation data in northeast Brazil

DANIELE T. RODRIGUES, WEBER A. GONÇALVES,
CLÁUDIO MOISÉS S. E SILVA, MARIA HELENA C. SPYRIDES & PAULO SÉRGIO LÚCIO

Abstract: This article evaluates four statistical methods of multiple imputation to fill in the missing data of daily precipitation in Northeast Brazil (NEB). We used a daily database collected by 94 rain gauges distributed in NEB from January 1, 1986 to December 31, 2015. The methods were: random sampling from the observed values; predictive mean matching, Bayesian linear regression; and bootstrap expectation maximization algorithm (BootEm). To compare these methods, missing data from the original series were initially excluded. The next step was to create three scenarios for each method, in which 10%, 20% and 30% of the data were removed at random. The BootEM method presented the best statistical results. With the average bias between the complete series and the imputed series values ranging between -0.91 and 1.30 mm/day. The values of the Pearson correlation ranging between 0.96, 0.91 and 0.86 respectively for 10%, 20% and 30% missing data. We conclude that this is an adequate method for the reconstruction of historical precipitation data in NEB.

Key words: Bootstrap, missing data, multiple imputation, semiarid.

INTRODUCTION

A very common occurrence in the databases of scientific studies is the occurrence of missing data. In the area of climate sciences, these failures are associated with occasional interruptions of automatic weather stations, malfunction of measuring instruments, reorganization of station networks and typing errors, among others. Since in the climate sciences most research depends on observational data from climatic seasons, it is challenging to analyze climatic phenomena. A difficulty factor of these studies is the quality of the data series used, especially with daily scale due to the frequent existence of flaws that make an investigation of a specific location or application of certain indexes unfeasible. The studies, even if presented in reports of the Intergovernmental Panel on Climate Change (IPCC), are limited to carrying out the analyses trying to extract as much information as possible for decision making.

According to Jahan et al. (2018), missing values in precipitation series can make efficient results impossible to obtain in hydrological, agricultural and climatological studies. According to Aybar et al. (2019), missing values pose a major problem to building grid datasets. The imputation of these values is crucial to minimize the lack of homogeneity in grid datasets (Yanto & Rajagopalan 2017). In addition,

most statistical methods assume the absence of missing data, so they can only be applied using complete datasets Honaker et al. (2011).

In this context, several authors have emphasized the lack of quality of historical series of meteorological data, such as Teegavarapu & Chandramouli (2005), Haylock et al. (2006) and Vincent & Yumiko (2006). The biggest obstacle to quantifying climate change is the scarcity of high-quality data, since it is not possible to avoid data failures in historical series. Thus, it is necessary to use alternative methods to treat the data in the best possible way in order to extract information that helps professionals to understand the climate.

One possibility (not advisable) is to exclude periods and/or variables with missing data from the analyses or to ignore the problem. However, this disregards information that may be relevant for the analysis and can induce bias in the final result, especially if the missing data gap is very large (Rubin 1996, Nunes et al. 2009). In addition, the occurrence of missing data makes it difficult to use some statistical methods, which require a complete database for application, such as the analysis of extreme values of a given meteorological variable. To work around this problem, several researchers (Teegavarapu & Chandramouli 2014, Dikbas 2017, Rodrigues et al. 2019) have used statistical methods called data imputation. These methods propose to replace the missing data with plausible estimated values.

A large number of methods exist in the literature on imputing lost data, which can be applied in local or global analyses (Armina et al. 2017). Some of these studies are focused on the analysis of climate series. In this sense, Teegavarapu & Chandramouli (2014) evaluated imputation methods based on optimal proximity, k-nearest neighbor classification and the k-means clustering method, using precipitation data recorded by 13 stations in Kentucky and 41 stations in Florida, USA. Semiromi & Koch (2019) compared two model-based methods (singular spectrum analysis and multichannel spectrum analysis) to reconstruct groundwater level time series observed at 25 piezometric stations in Ardabil Plain, northwest Iran. Jahan et al. (2018) compared eight imputation methods to fill gaps in precipitation series recorded at 27 stations in Bangladesh.

In addition to the methods mentioned above, there are those associated with Bayesian and BootEM imputation, which have been used in different scientific fields. For example, the multiple imputation method by Bayesian linear regression has been used to impute missing data in different datasets, such as: wine dataset, glass identification dataset, comprehensive concrete strength dataset, Indian liver patient dataset and seed dataset (Jadhav et al. 2019). From another perspective, Kaplan & Yavuz (2019) evaluated multiple imputation by Bayesian linear regression using simulated data and also real data from a triennial international survey conducted evaluate education systems worldwide.

With respect to the quality of imputation of meteorological data, Chen et al. (2019) reported the efficiency of the bootstrap expectation maximization algorithm to impute data on rainfall in the Daning River Basin in China (the Xining precipitation dataset, from 1998 to 2008). The authors designed six missing data rate scenarios, 10%, 20%, 30%, 40%, 50% and 60%, and concluded that when compared with other imputation methods, the bootstrap expectation maximization algorithm performed better to fill gaps, irrespective of the percentage of missing data. Izzo et al. (2020) applied the same algorithm to fill gaps in the temperature series, with gaps up to 20%, and rainfall series, with gaps up to 25%, from meteorological stations in the Dominican Republic.

Although these methods have presented promising results, there are few initiatives addressing them in hydroclimatological studies in Brazil and other countries in South America. In general, the methods used in studies in this region are those associated with the nearest neighbor approach, the arithmetic mean with neighboring stations, and the multiple imputation by predictive means matching (Michot et al. 2019) in the Amazon basin. However, recently the bootstrap expectation maximization algorithm has been used to fill data gaps in a small number (8) of meteorological stations with 14% failure during the period from 1980 to 2013 (Marinho et al. 2020), concentrated in the coastal region of Northeast Brazil.

Xavier et al. (2016) generated series through different methods of separation and interpolation of data on evapotranspiration, extreme temperatures and wind speed, organized in a regular grid of the region. Regarding information on wind speed in Brazil, Gilliland & Keim (2018) organized a database composed from different sources (averages of weather stations and re-analyses of global models) to study the variability and trends in wind speed in Brazilian territory during the period from 1980 to 2014. However, the methods adopted in these studies in Brazil have focused on spatial interpolation, which satisfies the need for complete data but at the same time generates information where there are no series. Hence, there is no way to identify whether these methods actually allow suitable inferences about data in regions with low density of original data. Our motivation is to impute data in existing series, collected in situ, that are not available in a given period due to operational or other reasons.

Determining the appropriate analytical approach for imputing missing data is a very sensitive issue, since the use of inappropriate techniques can lead to mistaken conclusions. The imputation of the missing data can be done in two ways, single and multiple. In single imputation, a unique value is estimated for each missing value. This is insufficient to assess the variability of the estimated values for imputation (Rubin 1996, Zhou et al. 2001, Little & Rubin 2002). To overcome this deficiency, the multiple imputation (MI) method was developed by Rubin (1987). In the MI, instead of one, a set of values is estimated for each missing value, and then the estimates are grouped using, for example, the average to obtain a valid inference that reflects data variability (Yozgatligil et al. 2013). Junninen et al. (2004) compared single and multiple imputation methods for imputing missing data in Belfast and Helsinki. The datasets used consisted of NOx, NO2, O3, PM10, SO2 and CO concentrations, all on a time-scale of one per hour (hourly averages), together with four meteorological parameters: wind speed, wind direction, temperature and relative humidity. The authors concluded that, in general, the results obtained by the multiple imputation methods performed better than those obtained by the single imputation methods.

However, only last decade has MI become practicable, due to computational advancement and implementation of its methods in specialized software. MI has become the most suitable class of methods for dealing with missing data (Mcknight et al. 2007). This has led to several focused on different MI methods for filling in database gaps in the area of climate sciences (Lo Presti et al. 2010, Yendra et al. 2013, Yozgatligil et al. 2013, Dikbas 2017).

In this context, in the present study we evaluate four statistical methods of multiple imputation: random sampling from the observed values (Sampling), predictive mean matching (PMM), Bayesian linear regression (Norm) and bootstrap expectation maximization algorithm (BootEM), based on the database about daily precipitation (mm) was used from 94 meteorological stations distributed in Northeast Brazil (NEB). NEB is a region characterized by high interannual spatial variability, with urban

coastal regions having a humid tropical climate and total precipitation above 2,000 mm/year, while the semiarid interior region in some places has rainfall below 300 mm/year (Oliveira et al. 2013). This region is highly vulnerable to periodic droughts (Moura & Shukla 1981, Marengo et al. 2017), a situation that is predicted to worsen with climate change (Torres et al. 2012, Torres & Marengo 2014), leading to a greater number of extreme precipitation events, mainly along the coast (Rodrigues et al. 2020, 2021).

MATERIALS AND METHODS

Database and study region

We used the weather station data from the Meteorological Database for Teaching and Research (BDMEP) of Brazil's National Institute of Meteorology (INMET), available at www.inmet.gov.br. The observations refer to the daily precipitation in millimeters (mm) observed by pluviometers distributed in 94 meteorological stations scattered in NEB. Each series has 30 years, between January 1, 1986 and December 31, 2015. Figure 1 shows the location of NEB and the spatial distribution of the stations.

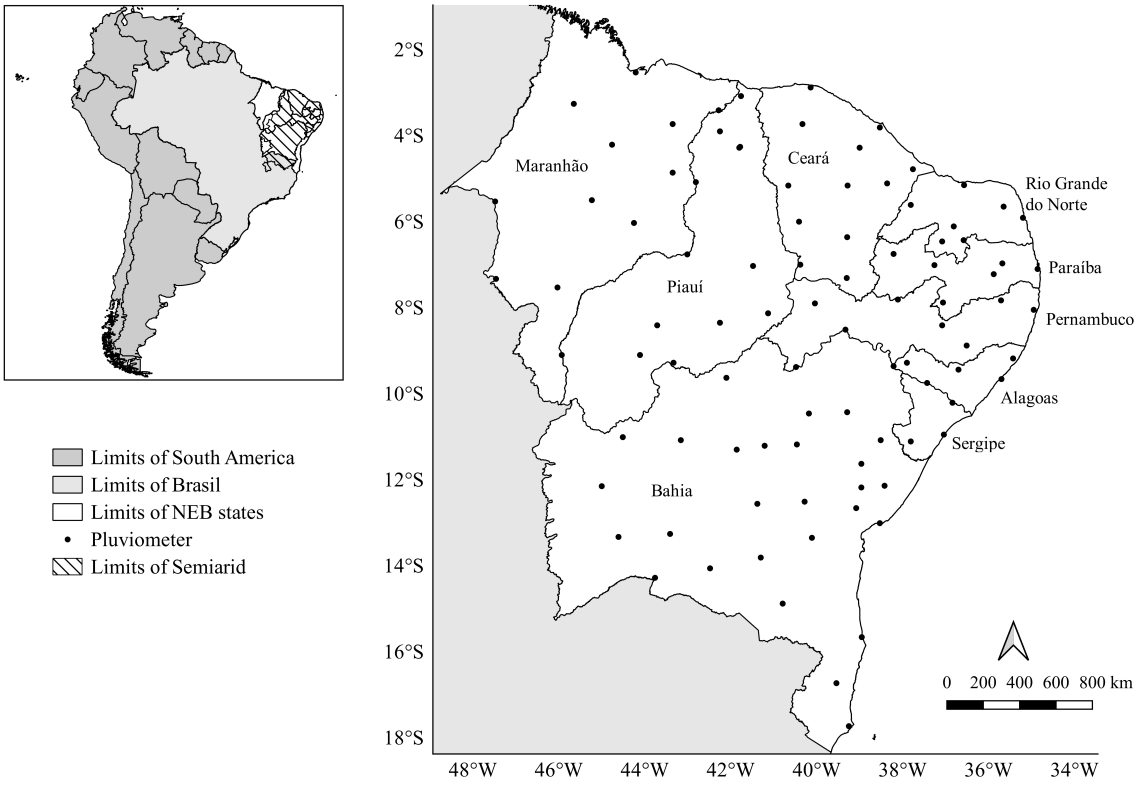

Figure 1. Location of the Brazilian Northeast and the weather stations, represented by circles on the map.

NEB is located between 1°S and 18°S and 34.5°W and 48.5°W. It has an area of approximately 1,558,196 km2, with an official population of 53,078,137 inhabitants according to the most recent census by the Brazilian Institute of Geography and Statistics (IBGE - Instituto Brasileiro de Geografia e

Estatística 2010). NEB basically has three climate types, with average annual rainfall varying from 300 - 2,000 mm/year (Alvares et al. 2013, Oliveira et al. 2017, Rodrigues et al. 2019): a moist coastal climate, prevailing along the coast from the states of Bahia to Rio Grande do Norte; tropical climate in areas of the states of Bahia, Ceará, Maranhão and Piauí; and semiarid climate throughout the hinterlands (Alvares et al. 2013). According to the Northeast Development Superintendency, about 64.65% of the area of NEB is classified as semiarid (SUDENE 2007).

Multiple imputation

Proposed by Rubin (1987), MI basically involves three stages. The first starts with a database of observed and missing values. Through the MI method, each missing valued is imputed m times, obtaining m complete databases, as shown in Figure 2. The method used to impute the m values is the same, but the values differ. The magnitude of these differences reflects the uncertainties about the imputed values. In addition, the observed values must remain the same.

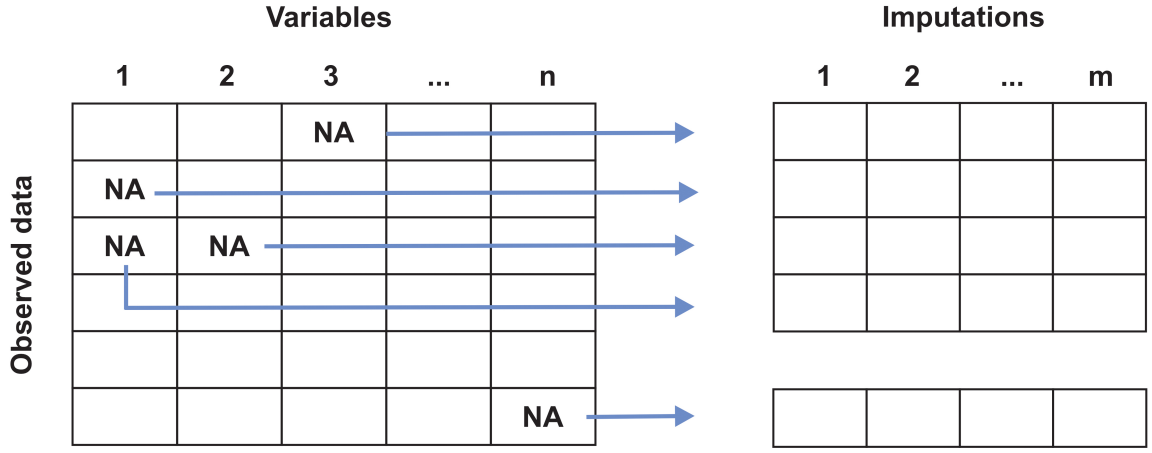

Figure 2. Dataset with m imputations for each unavailable value (NA: Not Available). Source: Adapted from Rubin (1987).

The second stage consists of estimating the parameter (s) of interest for each set of imputed data, through the application of standard analysis methods for complete data. In the third stage, the results obtained can be combined using the rules proposed by Rubin (1987). These are widely reported in the literature (Schafer & Graham 2002, Little & Rubin 2014) and involve multiple imputation, irrespective of the method used to make the imputation.

The purpose is that from each analysis the estimates for the parameter of interest X are obtained, that is, X_j for $j = 1, 2, ..., m$. According to Schafer (1997), X can be any scalar measure to be estimated, such as mean, correlation coefficient, regression coefficient or odds ratio. In this study, the combined estimate was the average of the individual estimates, denoted by Equation (1).

$$\overline{X} = \frac{1}{m}\sum_{j=1}^{m}\hat{X_j} \tag{1}$$

Then the total variance is the combination between the variance within and between the imputations (Equation 2).

$$T = \overline{U} + 1 + \frac{1}{m}B \tag{2}$$

where:

$$\overline{U} = \frac{1}{m}\sum_{j=1}^{m} U_j \quad e \quad B = \frac{1}{m-1}\sum_{j=1}^{m}(\hat{X_j} - \overline{X})^2 \tag{3}$$

In the literature, there are review articles and tutorials that assist researchers in the treatment of lost data. Rubin (1996), Schafer (1997), Schafer & Graham (2002) and Little & Rubin (2019) present a wide view of the existing methods for imputing missing values. Other works, such as Horton & Lipsitz (2001), Acock (2005) and Buhi et al. (2008), have compared some of the techniques available in the main statistical packages.

MI methods

The MI methods used in this study were: random sampling from the observed values (Sampling); predictive mean matching (PMM); Bayesian linear regression (Norm); and bootstrap expectation maximization algorithm (BootEM). Among these, Sampling is considered the simplest. It randomly chooses a value already observed to replace the missing value. PMM is a variant of linear regression (Li et al. 1991, Schafer 1997, Di Zio & Guarnera 2009), in which predicted values are observed and missing. The missing data are replaced by the observed data in which the predicted values for observed and missing are closest. The Norm method is similar to the PMM, since it involves linear regression analysis. However, the imputations are made according to the values predicted for missing values. To perform MI using the methods Sampling, PMM and Norm, the R language multivariate imputation by chained equations package (Buuren & Oudshoorn 2011) was used.

The BootEM method uses the bootstrap resampling technique via the expectation maximization (EM) algorithm. This method imputes the missing values of a time series by estimates generated by the EM algorithms implemented in several bootstrap samples of the original data (Rubin 1994, Honaker et al. 2011). Bootstrap resampling uses existing data as a pseudo-population and selects new samples with replace. For example, if $X_1, ..., X_n$, are independently and identically distributed from an unknown distribution F, this distribution is estimated by $\hat{F}_n(x)$, which is the empirical distribution F_n defined in Equation (4), where $I(X)$ is the indicator function of the set X.

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x) \tag{4}$$

If the pseudo-population is incomplete (with missing data), each bootstrap resampling has a high chance of also being incomplete, making bootstrap estimates inefficient. To improve bootstrap estimates, the expectation maximization algorithm is used (Takahashi 2017). The first stage, the EM algorithm assumes a certain probability distribution. Using the mean and variance-covariance values of this distribution, the expected value of model likelihood is calculated, the likelihood is maximized, model parameters are estimated that maximize these expected values, and then the distribution is updated. The expectation and the maximization stages are repeated until the values converge (Schafer

1997). To perform MI using the methods BootEM, the R language Amelia II: A Program for Missing Data package (Honaker et al. 2011) was used.

Procedures for comparing MI methods

To compare the MI methods, the first procedure was to delete the missing data from the original database. With the complete database (with no missing data), the next step was to create three complete scenarios, in which 10%, 20% and 30% were randomly removed for each station in the database. Once this was done, multiple imputation methods were applied for each scenario, one at a time, thus obtaining 12 imputed databases, four for each scenario (Figure 3).

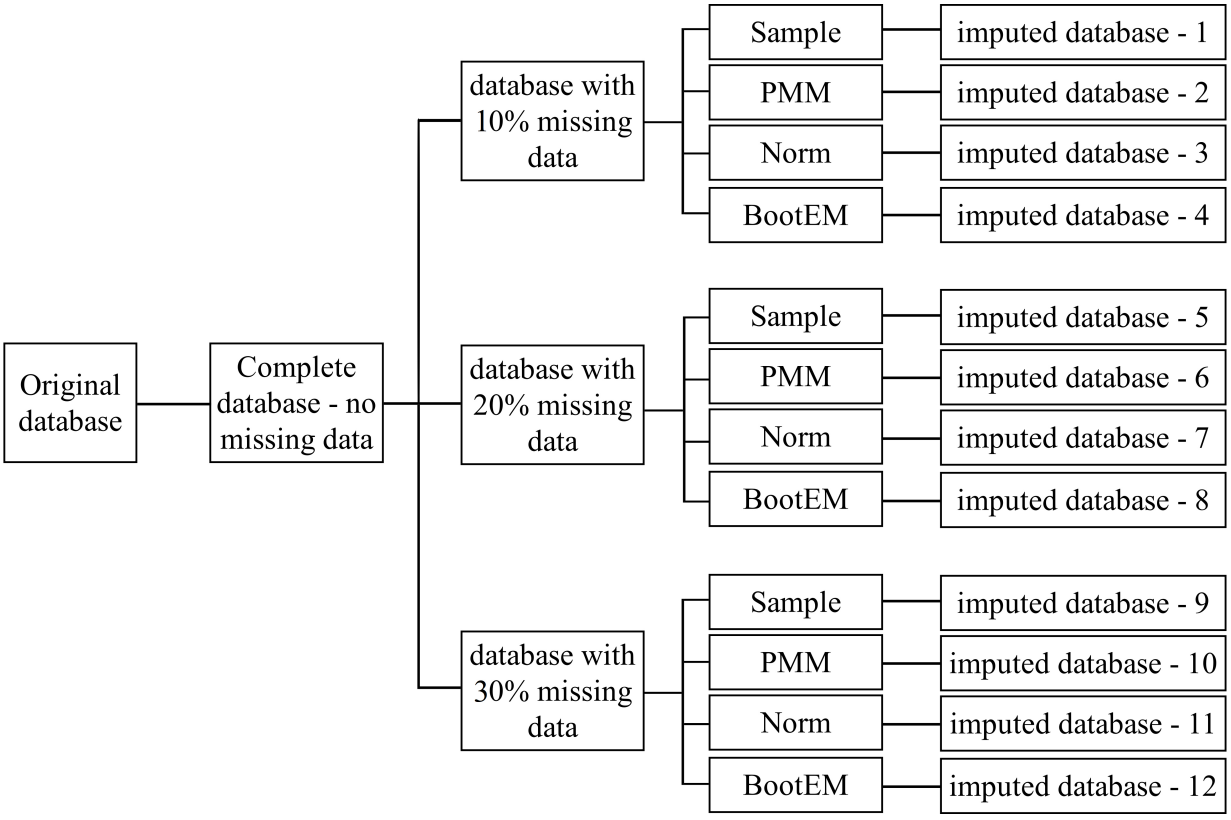

Figure 3. Illustration of the steps to generate the missing and imputed databases.

The comparison between the complete database and each database with the imputed values was carried out by applying the following statistical measures: bias, standard deviation (SD), mean square error (MSE) and Pearson's correlation coefficient (ρ) These are given, respectively, by Equations (5), (6), (7) and (8).

$$Bias = X_i - Y_i \tag{5}$$

$$SD = \hat{\sigma} = \sqrt{\frac{\sum_i^n (Y_i - \overline{Y})^2}{n-1}} \tag{6}$$

$$MSE = \frac{\sum_i^n (X_i - Y_i)^2}{n-1} \tag{7}$$

$$\rho(X,Y) = \frac{\sum_i^n (X_i - \overline{X})(y_i - \overline{Y})}{\sqrt{\sum_i^n (X_i - \overline{X})^2 \sum_i^n (Y_i - \overline{Y})^2}} \tag{8}$$

where X_i is the daily precipitation value of the complete data series, Y_i is the daily precipitation value of the imputed data series, $\overline{X}$ is the average daily precipitation of the complete data series, $\overline{X}$is the average daily precipitation of the imputed data series, and n is the total days of the series.

The probability density function (PDF) was also used. Denoted by $f_X(x)$ a the FDP describes the behavior, in polygon form, of the frequency distribution of a random variable. The probability of the random variable being less than a given value of interest, x, is calculated using the cumulative distribution function (CDF), represented by Equation (9), $F_X(x)$.

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(x)dx \tag{9}$$

The CDF of a continuous random variable is a non descending function, and the expressions are validated: $F_X(-\infty) = 0$ and $F_X(+\infty) = 1$ (Sheather & Jones 1991, Silverman 1998, Venables & Ripley 2002, Scott 2015). Conversely, the corresponding FDP can be obtained by differentiating $F_X(x)$, The FDP, $f_X(x)$, is represented by Equation (10).

$$f_X(x) = \frac{dF_X(x)}{dx} \tag{10}$$

The seasonal analyses were performed quarterly: December, January and February (DJF); March, April and May (MAM); June, July and August (JJA); September, October and November (SON). All analyzes were performed in R language (RC Team 2020).

RESULTS AND DISCUSSION

Statistical analysis of imputation methods

Figure 4 shows the curve of the probability density function of the data in the three different scenarios (10, 20 and 30% of missing data). The distribution is one-tailed with greater probabilities of values close to zero. The distribution is more to the left for the 30% scenario, however, the type of distribution remains the same. Table I shows that for all scenarios and all methods, the bias is low, with a maximum of 0.0072 and a minimum of 0.0001 (values in modulus). This means that, on average, the difference between the actual and the imputed database is close to zero. Furthermore, the values of standard deviation and mean square error increase with increasing percentage of missing data. However, the measures have lower values when the BootEM method is applied, respectively, 0.0256/0.0006; 0.0382/0.0014 and 0.0470/0.0022 for the 10%, 20% and 30% missing data scenarios, respectively.

The quality of the MI, based on the correlation coefficients, varies according to the weather station, method applied and the percentage of missing data (Figure 5). The same results were identified by Teegavarapu & Chandramouli (2014), who evaluated six imputation methods to fill gaps in six different rainfall station series: three time series between 1971-2001, for rainfall in Kentucky, USA, and three time series in the 1994-1999 interval for Florida, USA. In NEB, the correlation between the complete and imputed database decreases as the percentage of missing data increases. As shown in Figure 5,

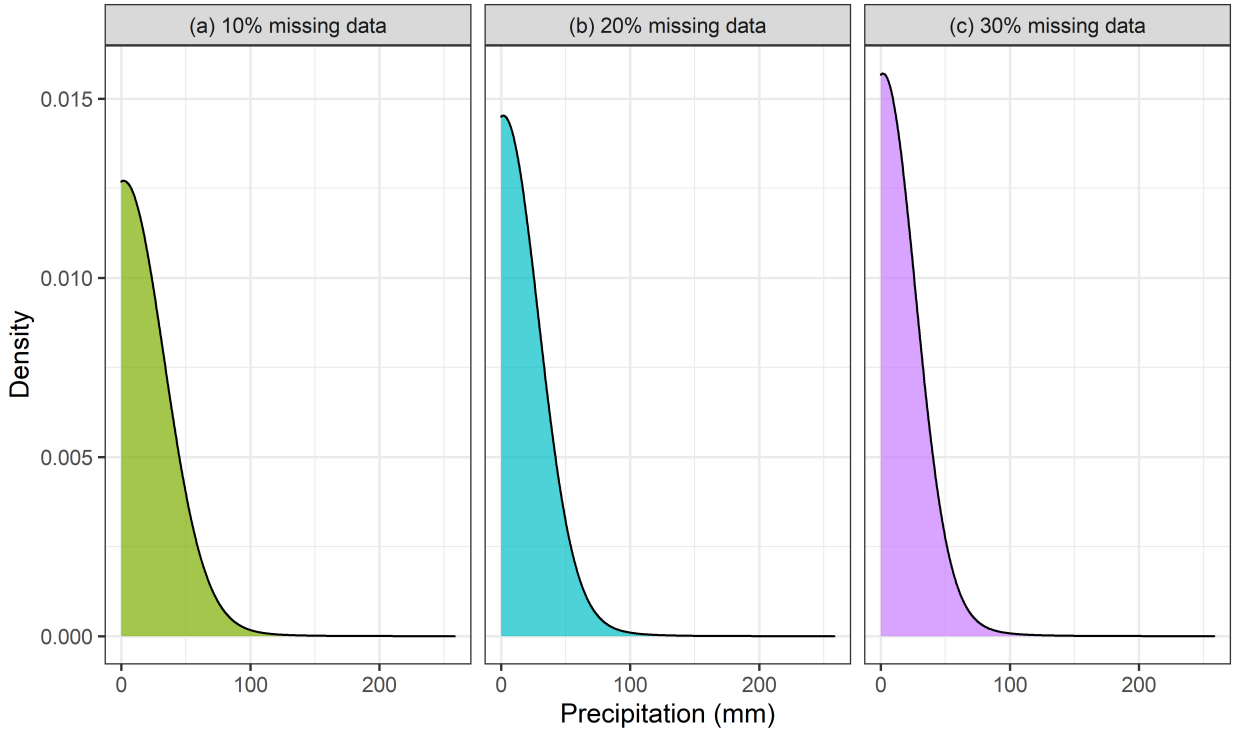

Figure 4. Density of precipitation values randomly removed from the original database (with no missing data), for scenario, in which (a)10% missing data, (b) 20% missing data and (c) 30% missing data.

Table I. Bias*, SD and EQM* measurements of the four imputation methods by scenario.

Scenario	Method	Bias*	SD	EQM*
10%	Amostragem	0,0036	0,0338	0,0011
	Norma	-0,0014	0,0323	0,0010
	PMM	-0,0024	0,0360	0,0013
	BootEM	0,0014	0,0256	0,0006
20%	Sampling	0.0018	0.0517	0.0026
	Norm	-0.0010	0.0478	0.0023
	PMM	-0.0054	0.0404	0.0016
	BootEM	-0.0023	0.0382	0.0014
30%	Sampling	0.0001	0.0715	0.0051
	Norm	0.0033	0.0639	0.0040
	PMM	-0.0072	0.0686	0.0047
	BootEM	0.0062	0.0470	0.0022

*Average obtained from the measurements made for the 94 weather stations.

the BootEM method has the best correlations for all scenarios, and the difference between the MI methods within the same scenario increases as the percentage of missing data increases (Figure 5).

Figure 5. Scatter plot with values of Pearson's correlation coefficients for the four multiple imputation methods, by scenario: (a) 10% missing data; (b) 20% missing data; and (c) 30% missing data.

The information shown in Figure 6 corroborates the statements made previously that the correlation coefficient decreases when the scenario gets worse. However, with 10% missing data, the correlations are always above 0.85, the level of data failures typically found in series pertaining to NEB according to previous climatological analyses Oliveira et al. (2017). There are also different correlation coefficients in NEB within the same scenario. Semiromi & Koch (2019) also observed variations between the quality of 25 imputed time series of groundwater levels in the Ardabil Plain region of northwestern

Iran. Michot et al. (2019) stated that the quality of the imputation methods may be related to the pluviometric density and precipitation regime of the region. Therefore, it is extremely important to consider these factors in studies to evaluate imputed series, as shown by Teegavarapu & Chandramouli (2014), Jahan et al. (2018) and Aybar et al. (2019).

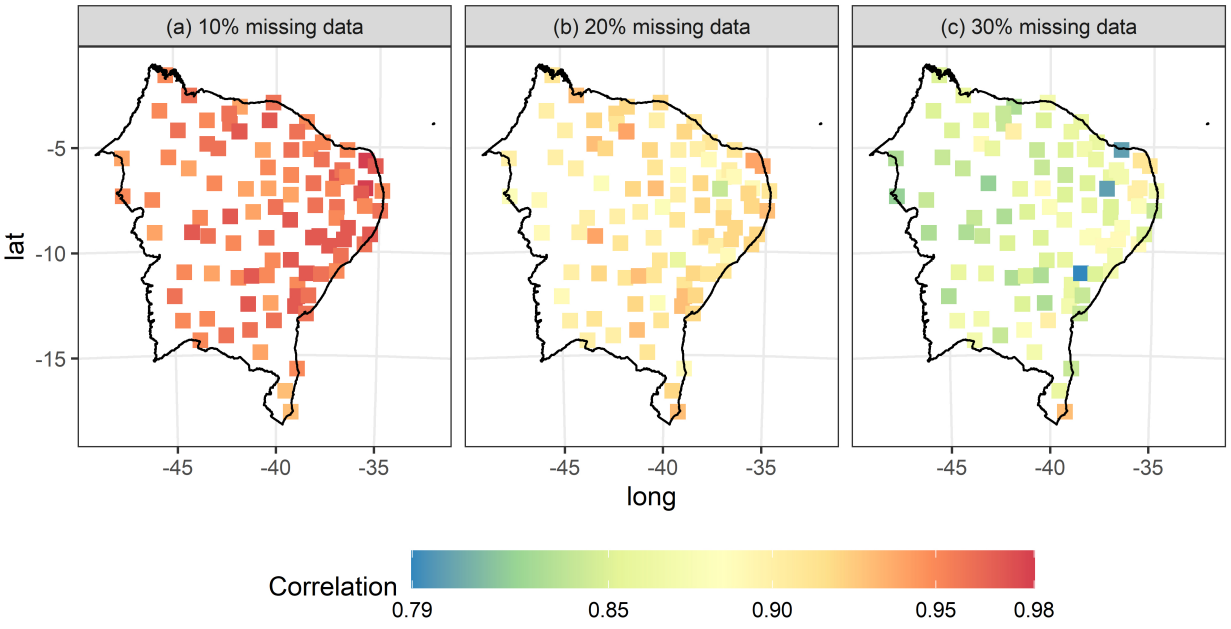

Figure 6. Spatial distribution of Pearson's correlation coefficients for the BootEM method for scenario: (a) 10% missing data; (b) 20% missing data; and (c) 30% missing data.

The information in Figure 6 corroborates the statements made previously that the correlation coefficient decreases when the scenario gets worse. There are again different correlation coefficients in NEB within the same scenario. The first scenario, Figure 6a, has the highest correlation coefficients, ranging from 0.93 to 0.98. Part of the east coast, northern Ceará, Maranhão and central Bahia have better coefficients. In the second scenario, Figure 6b, in general, the north and east coast have the best coefficients, while the lowest correlation coefficients belongs to the central area of Pernambuco and the east of Bahia. In this scenario the correlation coefficients range from 0.85 to 0.95. The worst case, Figure 6c, is the one that obtained the greatest variability of the correlation coefficients, 0.8 to 0.93. In this, the west of NEB stands out, with lower correlation coefficients, and east of Rio Grande do Norte with the highest coefficients, close to 0.93. Also, for the south of Bahia the correlation coefficients are close to 0.9 for the three scenarios under study.

To complement the comparison between the original and imputed data, we present in Figure 7 the curves of PDF from the complete and imputed data (Figure 7a) over NEB and over two regions with distinct precipitation regimes, São Luís (Figure 7b) and Natal (Figure 7c). The distribution of complete data in the NEB presents a trimodal pattern with the main maximum around 55 mm, followed by relative peaks at 120 mm and 200 mm, respectively, consistent with previous analyzes of climatological

studies on extreme values in the NEB (Oliveira et al. 2017, Rodrigues et al. 2020, Costa et al. 2020). The imputed data show the same trimodal pattern, however, the higher frequency values are shifted to relatively higher values while the lower values decrease, in agreement with Figure 4, in which there was a higher frequency of data close to zero as the number of missing data increases.

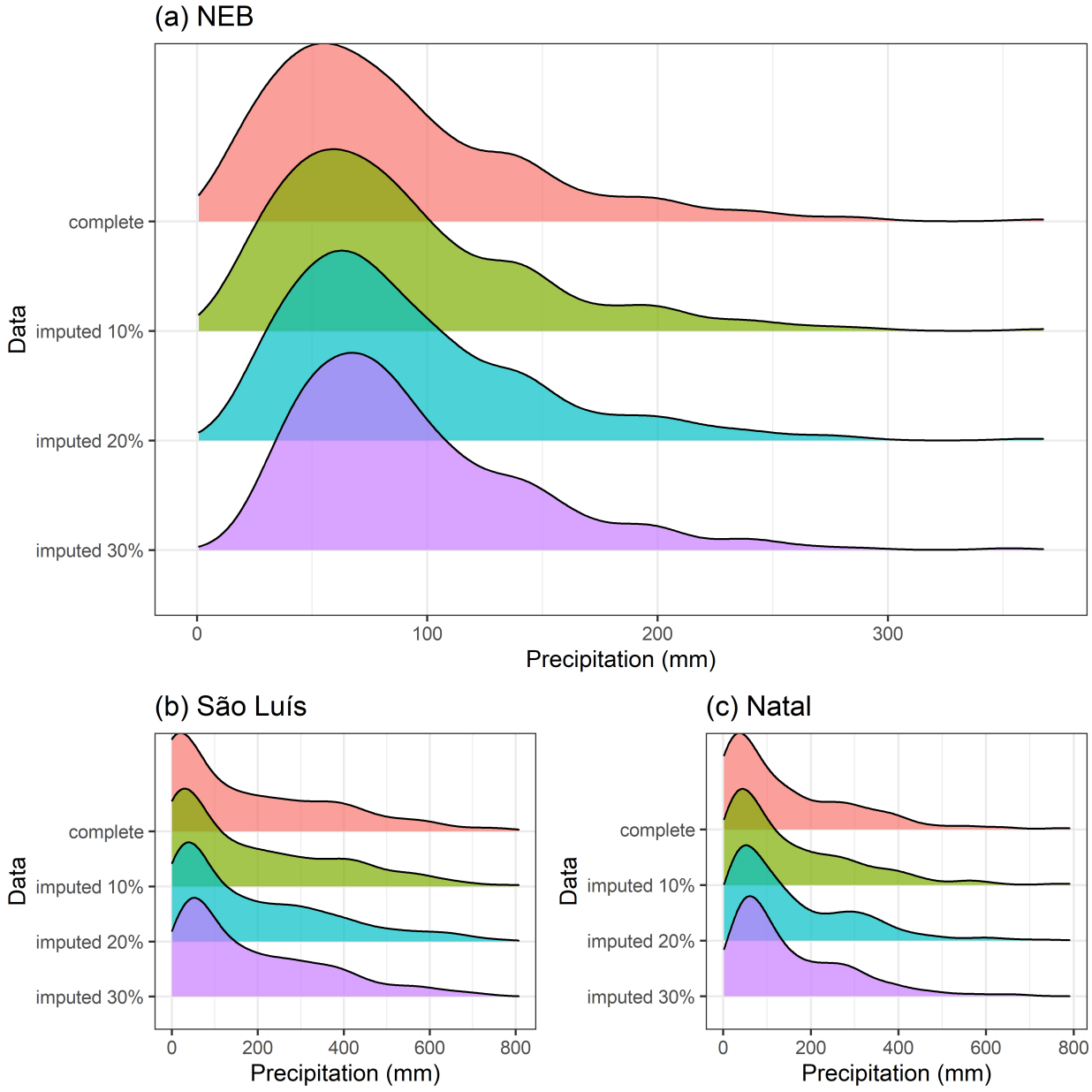

Figure 7. Density of precipitation values from complete data and imputed data for 10%, 20% and 30% missing data in (a) NEB; (b) São Luís (c) Natal.

Regarding the distribution of precipitation values in the two regions (Figures 7b-c), we verified that the imputed data preserve the distribution (with two well-defined maximums) of the complete data. At the same time, it indicates values around 400 mm in the second maximum for São Luis,

while in Natal this maximum is around 300 mm. The relative maxima in São Luis are explained by the proximity to the Amazon basin, the role of the ZCIT and the influence of squall lines that form on the northern coast of NEB (Oliveira et al. 2017, Rodrigues et al. 2020). On the other hand, extreme precipitation values in Natal are influenced by ZCIT and by the action of easterly wave disturbances (EWD), which are common in the months of May to July on the east coast of the NEB (Oliveira et al. 2013, Gomes et al. 2019). The occurrence of extreme precipitation events has also been reported in previous NEB studies for both cities analyzed and the accumulated averages are consistent with other datasets independent of those analyzed here, for example, data estimated via satellites (Rodrigues et al. 2020) or through data collected in automatic surface stations (Rodrigues et al. 2021).

Seasonal and spatial analysis of precipitation

Considering the average value of NEB as a whole (Figure 8), it appears that the rains present well-marked annual seasonality, with periods of drought, having minimum monthly values below 30 mm and maximum values above 200 mm. The month of January 2004 had the largest accumulated rainfall (365 mm), associated with a persistent High Level Cyclonic Vortex, which lasted for more than 10 days that month (Seluchi 2004, Oliveira et al. 2013, Costa et al. 2020). On the other hand, the years between 2012 and 2015 suffered from severe drought (de Medeiros et al. 2020), with maximum monthly rainfall always below 150 mm.

In general, the seasonal behavior of precipitation in NEB, between the complete data series and that imputed by the BootEM method, is similar over time for the three scenarios (Figure 8). When the percentage of missing data increases, the difference between the series also increases, especially in the extreme precipitation values.

Precipitation in NEB has marked temporal variability (Figure 9), with the highest accumulation of precipitation concentrated in the MAM quarter (Figure 9). The climatological average in this quarter is around 150 mm, but it is important to note that the outliers (e.g., monthly rainfall above 350 mm in DFJ) do not occur in the precipitation period with the highest average, as also observed in previous studies (Palharini & Vila 2017, Rodrigues et al. 2019, Gomes et al. 2019). This is basically due to the nature of the meteorological systems that operate during the drier months. They are more convective in nature, with typical cases of easterly wave disturbances or mesoscale convective systems in JJA or cases of High Level Cyclonic Vortexes or the persistence of the South Atlantic Convergence Zone in DJF, as described by Oliveira et al. (2017).

On the other hand, precipitation in the rainiest period is closely linked to the effect of the Intertropical Convergence Zone, which is the main large-scale system operating in the region (Kousky 1979, Moura & Shukla 1981, Bombardi et al. 2018). The positive aspect is that in both scenarios, the imputation method is able to faithfully capture the seasonality of precipitation observed in the NEB, although in the drier months the boxplot of the imputed precipitation showed slightly higher medians compared to the original data.

The data imputed by the BootEM method representing the seasonal period in NEB, in the three studied scenarios, are shown in Figure 9a-c. Apparently, the imputed values overestimate the values of the complete data series. This occurs especially in the JJA and SON periods, when the lowest precipitation accumulated in NEB happens (Rodrigues et al. 2019). A study by Eischeid et al. (2000) to

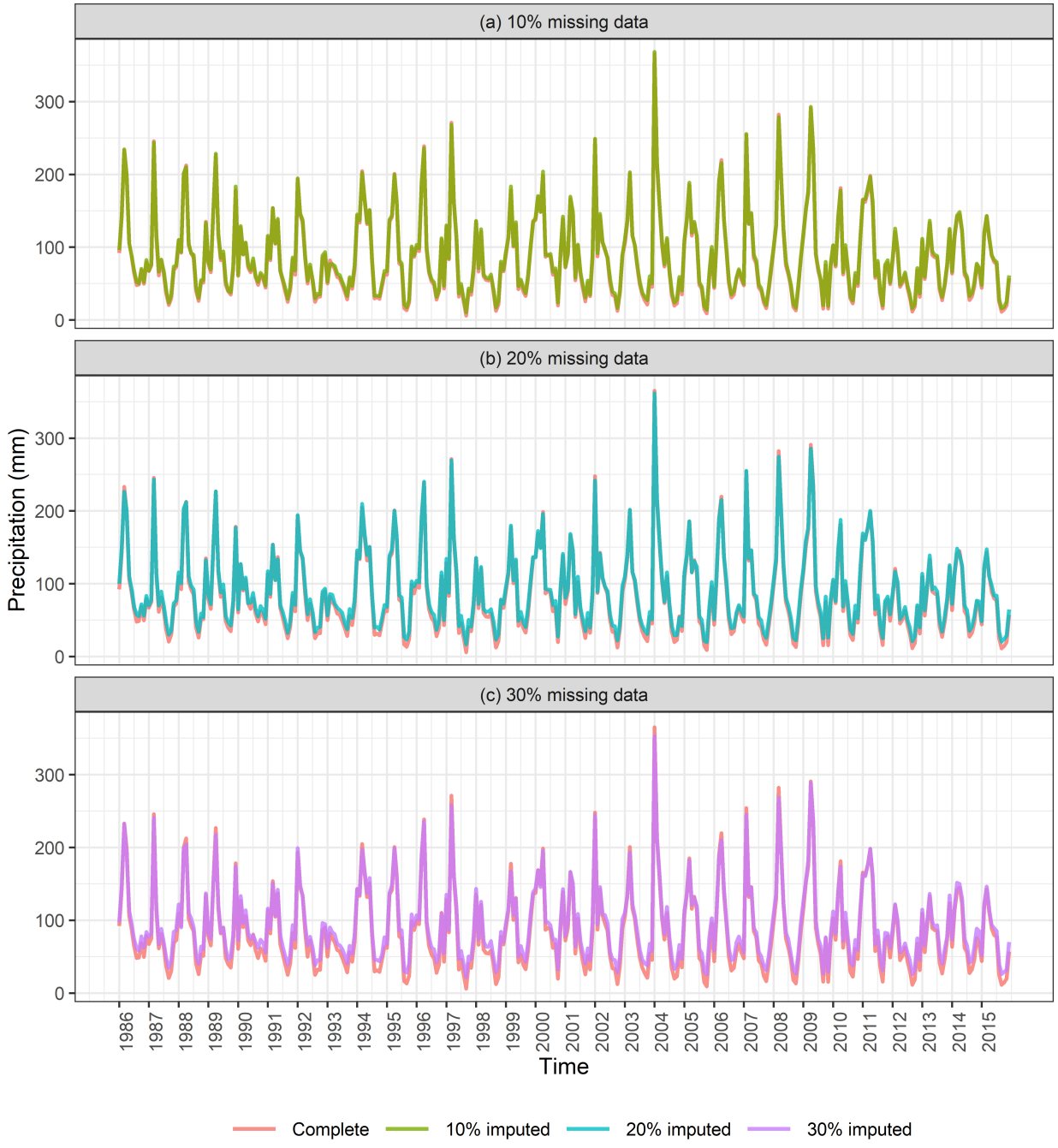

Figure 8. Time series of monthly precipitation in NEB for scenario: (a) 10% missing data; (b), 20% missing data; and (c) 30% missing data, from January 1986 to December 2015.

reconstruct daily precipitation and temperature data over the western United States found that the quality of the reconstructed series varied according to the time of year. Regarding NEB, the biggest difference between the complete and imputed datasets is observed in the third scenario, in the SON period, as shown in Figure 9c.

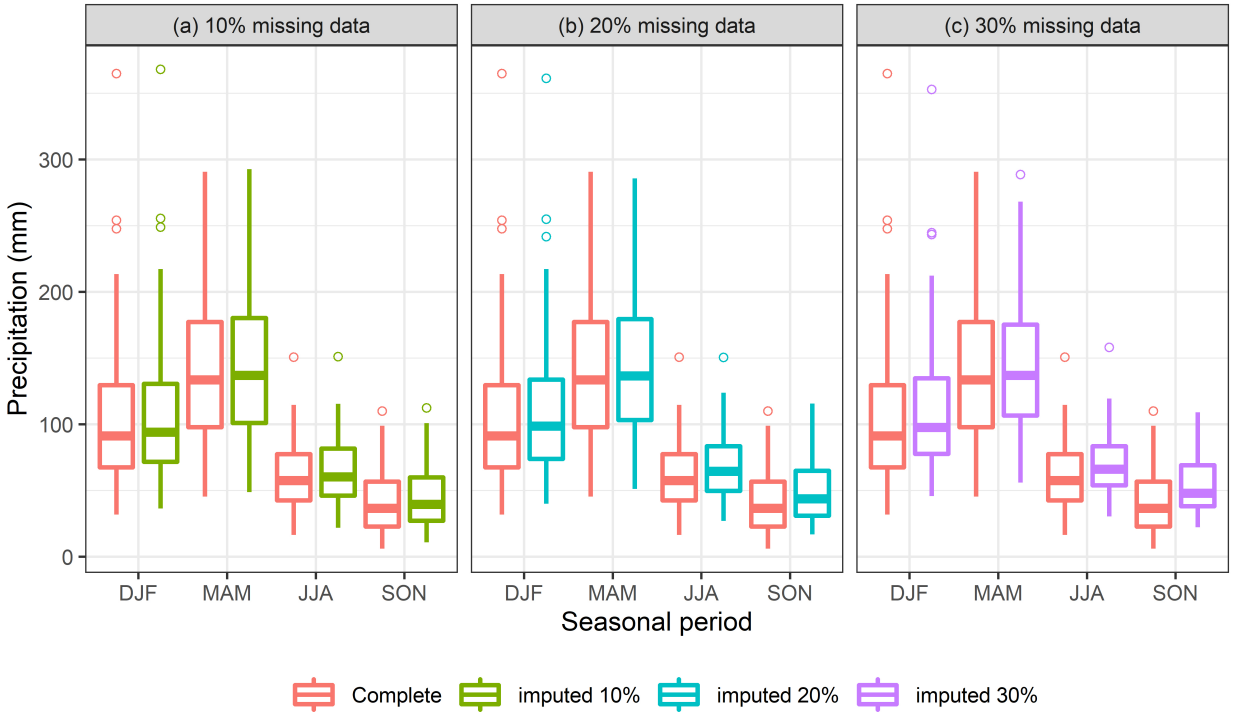

Figure 9. Boxplot of precipitation in NEB by seasonal period for scenario: (a) 10% missing data; (b), 20% missing data; and (c) 30% missing data.

Figure 10 shows the distribution of average daily precipitation over NEB by seasonal period from the complete and imputed database for scenarios of 10, 20 and 30% missing data. In general, the average daily rainfall between the complete dataset and imputed dataset (Figure 10) are similar during the study period. The spatial variability of rain is captured in the three scenarios, with more intense daily rainfall (typically above 4.0 mm/day) in the northwest and east coast of NEB and with less intensity (below 2.0 mm/day) in the semiarid region of NEB. The lowest values are observed in the JJA and SON periods. According to Rao et al. (1993), August, September and October are the months with the lowest accumulated rainfall in the NEB region. These include the JJA and SON periods. The exception occurs in eastern NEB, which has the highest cumulative rainfall between April and July (Oliveira et al. 2013, Rodrigues et al. 2019). The bias between imputed and full data varies between -1.3 and 1.3 mm (Figure 11). In the first scenario (10% missing data) the bias is close to zero in all seasonal periods (Figure 11a). In the second scenario (20% missing data) for most of the region the bias is still close to zero (Figure 11c), but in the worst scenario (30% missing data) the bias increases especially in the coast and west region of the NEB (Figure 11c).

With regard to the spatial distribution of extreme events, it can be seen in Figure 12 that the imputed data presented the 99% percentile (P99) satisfactorily compared to the observed data. The imputed data (Figure 12) express the spatial variability of the values represented by the 99% percentile (considered extreme precipitation values) when compared to those observed according to the complete dataset (Figure 12). The coast is the region with the highest values of the 99% percentile, Figure 12. This result corroborates those of Palharini & Vila (2017), Palharini et al. (2020)

Figure 10. Spatial distribution of the average daily precipitation in NEB by seasonal period for (a) complete data and imputed data for scenario: (b) 10% missing data; (c), 20% missing data; and (d) 30% missing data.

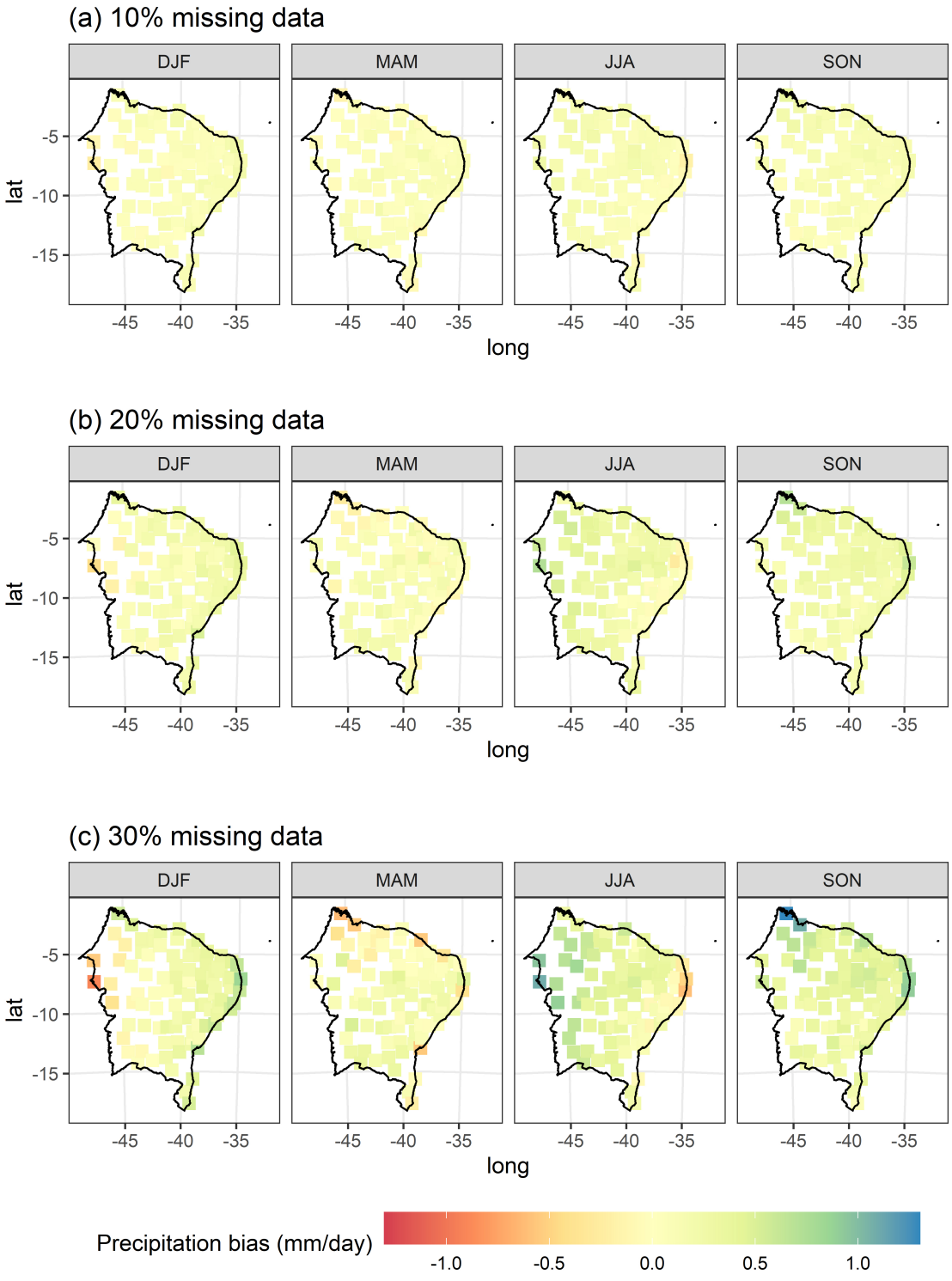

Figure 11. Spatial distribution of the daily precipitation bias in NEB by seasonal period for imputed data for scenario: (a) 10% missing data; (b), 20% missing data; and (c) 30% missing data.

and Rodrigues et al. (2020), who observed that the coast is the region with the most intense extreme daily precipitation in NEB. The NEB coast is characterized by high daily precipitation intensities, while the semiarid region is characterized by low rainfall values Oliveira et al. (2017) and Rodrigues et al. (2019). In general, the data imputed by the BootEM method capture these characteristics, Figure 12. The spatial distribution of the extreme values of the imputed data (Figure 12) are close to the extremes observed in the complete data (Figure 12), especially in the best scenario (Figure 12a). The highest percentile, 74.4 mm/day, is in the city of São Luís, located on the north coast of NEB. The percentiles of data imputed by the BootEM method for the city of São Luís are 71.84 mm, 68.57 mm and 64.66 mm, respectively for the scenarios of 10%, 20% and 30% missing data.

The MAM and JJA periods has the largest range of extreme precipitation on the coast of NEB (above 70 mm/day). The highest precipitation values are observed on the NEB east coast during its rainy season, between April and July. During this period, the rainfall of this region is mainly influenced by the EWD (Torres & Ferreira 2011, Gomes et al. 2019), which frequently acts in the tropical range of the globe, in the area of influence of the moving trade winds. The bies range between -10.2 and 10.2 mm (Figure 13). It is observed that the higher the percentage of missing data, greater the bias (Figure 13). In the most part of NEB the bias is negative, that is, the imputation method underestimates the extreme precipitation. Overestimation occurs in JJA and SON in some locations on the north coast and west region of the NEB (Figure 13).

CONCLUSION

In many regions, meteorological time series exhibit a considerable number of missing data. In the present study we evaluate four statistical methods of fault filling using multiple imputation, based on the database about daily precipitation (mm) from 94 meteorological stations distributed in Northeast Brazil. The results presented in this study show that the adequate use of the multiple imputation statistical methods proved to be efficient to fill in missing data in daily precipitation series in NEB. The average bias between the complete and the imputed database is close to zero in all the methods of imputation studied. It was possible to verify that the lower the percentage of missing data in the data series is, the greater the ability of multiple imputation methods to fill the gaps in correctly.

In general, the bootstrap expectation maximization algorithm, BootEM method, showed low values of bias and lower values of standard deviation and mean square error in relation to the results of the other methods. The correlation coefficients between the complete series and the imputed series were higher when the BootEM method was used, even in the scenario where the series had a large percentage of missing data, 30%. The spatial correlations showed spatial distribution without preferential areas, which leads us to conclude that the method works independently of the precipitation regime in NEB. At the same time, a higher percentage of missing data reduced the value of the Pearson correlation. On average 0.96, 0.91 and 0.86 respectively for 10%, 20% and 30% missing data.

It was possible to evaluate the databases imputed by the BootEM method taking into account the spatial and temporal seasonality of the NEB. This result is extremely important, especially for researchers who wish to work with a specific period or subregion, since the results characterize the spatial and temporal variability present in NEB according to precipitation. In general, the average bias

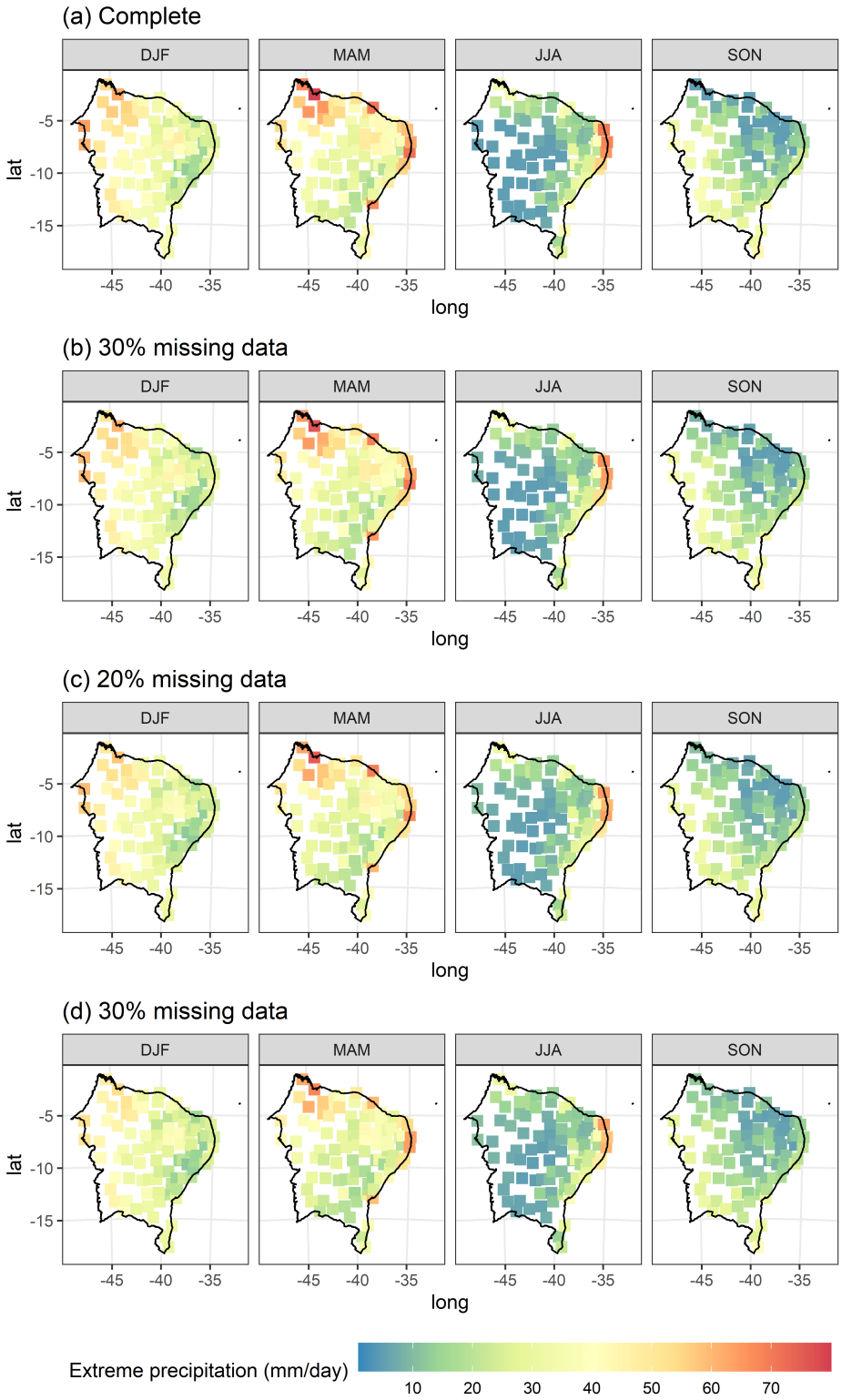

Figure 12. Spatial distribution of the extreme (P99) daily precipitation in NEB by seasonal period for (a) complete data and imputed data for scenario: (b) 10% missing data; (c), 20% missing data; and (d) 30% missing data.

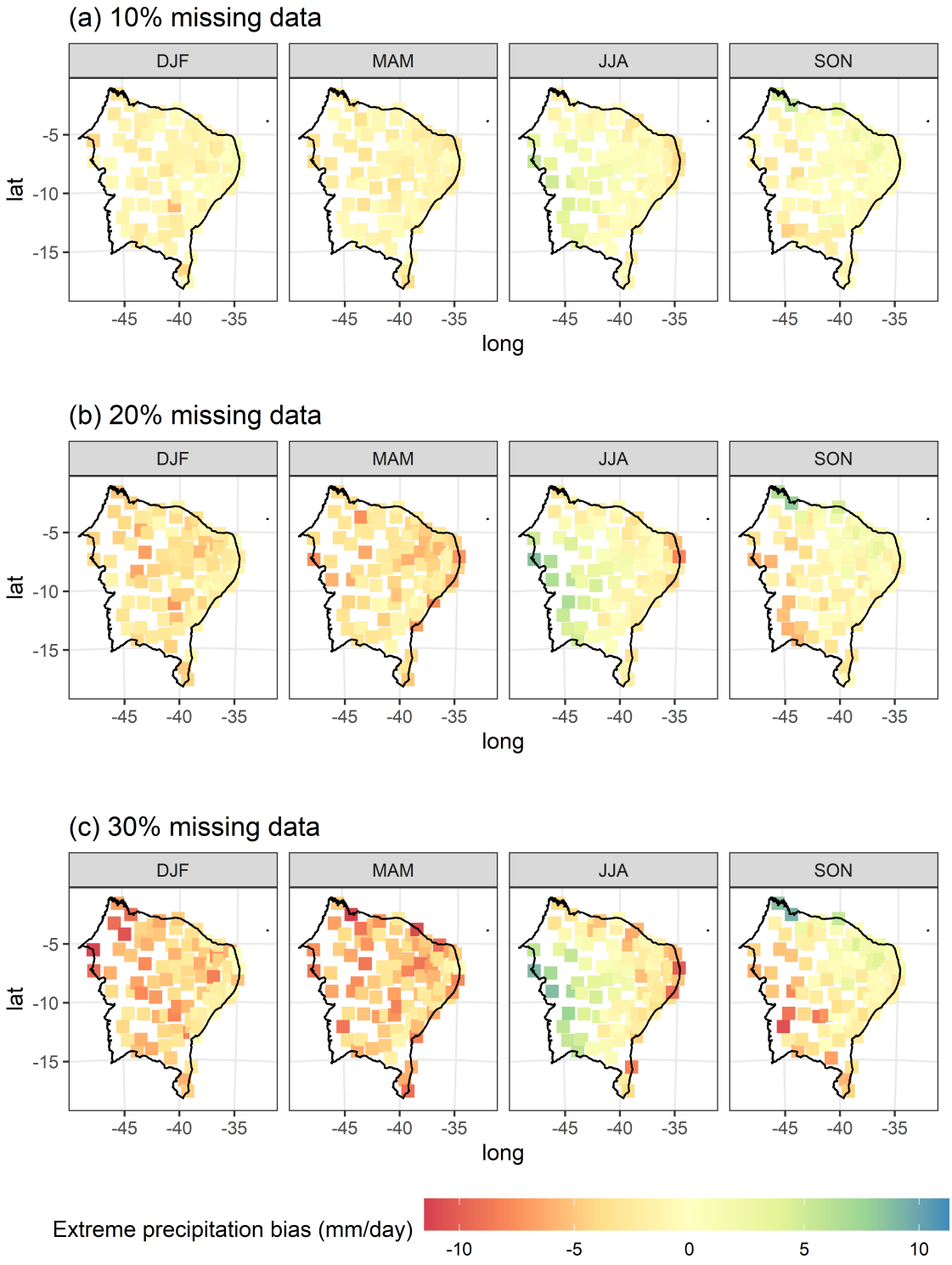

Figure 13. Spatial distribution of the extreme (P99) daily precipitation bias in NEB by seasonal period for imputed data for scenario: (a) 10% missing data; (b) 20% missing data; and (c) 30% missing data.

between the complete series and the imputed series are low in all seasonal periods and locations of the NEB, values ranging between -0.91 and 1.30 mm/day.

The imputed database was also evaluated according to extreme precipitation values, 99th percentile. The extreme values quality coming from the imputed database depends on the location and seasonal period of the NEB. The average bias ranged between -5.60 and 5.65 mm/day for series with 10% missing data and between -11.44 and 9.33 mm/day for series with 30% missing data. Further studies involving the quality of the multiple imputation method when estimating extreme precipitation values are recommended.

In view of the observed results, we can conclude that the use of synthetic series imputed by the BootEM method can be a tool to support the reconstruction of historical series of climatic data, to assist in the monitoring and planning of water resources. The results obtained will give users of the multiple imputation method prior knowledge regarding the quality of the imputaded database in relation to the precipitation, spatially and seasonally, in NEB. The researchers will have a previous understanding of imputaded database quality, which also depends do percentage of missing data.

REFERENCES

ACOCK AC. 2005. Working with missing values. J Marriage Fam 67: 1012-1028.

ALVARES CA, STAPE JL, SENTELHAS PC, DE MORAES G, LEONARDO J & SPAROVEK G. 2013. Köppen's climate classification map for Brazil. Meteorologische Zeitschrift 22(6): 711-728.

ARMINA R, MOHD ZAIN A, ALI NA & SALLEHUDDIN R. 2017. A Review On Missing Value Estimation Using Imputation Algorithm. J Phys: Conf Ser 892: 012004. doi:10.1088/1742-6596/892/1/012004.

AYBAR C, FERNÁNDEZ C, HUERTA A, LAVADO W, VEGA F & FELIPE-OBANDO O. 2019. Construction of a high-resolution gridded rainfall dataset for Peru from 1981 to the present day. Hydrol Sci J 65(5): 770-785. doi:10.1080/02626667.2019.1649411.

BOMBARDI RJ, TRENARY L, PEGION K, BENJAMIN C, DELSOLE T & KINTER III JL. 2018. Seasonal Predictability of Summer Rainfall over South America. J Clim 31(20): 8181-8195. URL https://doi.org/10.1175/JCLI-D-18-0191.1.

BUHI ER, GOODSON P & NEILANDS TB. 2008. Out of sight, not out of mind: strategies for handling missing data. Am J Health Behav 32: 83-92. doi:10.5993/AJHB.32.1.8.

BUUREN SV & OUDSHOORN CGM. 2011. mice: Multivariate imputation by chained equations in R. J Stat Softw 45(1): 1-67.

CHEN L, XU J, WANG G & SHEN Z. 2019. Comparison of the multiple imputation approaches for imputing rainfall data series and their applications to watershed models. J Hydrol 572: 449-460. doi:10.1016/j.jhydrol.2019.03.025.

COSTA RL, DE MELLO BAPTISTA GM, GOMES HB, DOS SANTOS SILVA FD, DA ROCHA JÚNIOR RL, DE ARAÚJO SALVADOR M & HERDIES DL. 2020. Analysis of climate extremes indices over northeast Brazil from 1961 to 2014. Weather Clim Extremes 28: 100254. doi:10.1016/j.wace.2020.100254.

DE MEDEIROS FJ, DE OLIVEIRA CP & SANTOS E SILVA CMEA. 2020. Numerical simulation of the circulation and tropical teleconnection mechanisms of a severe drought event (2012-2016) in Northeastern Brazil. Clim Dyn 54: 4043-4057. URL https://doi.org/10.1007/s00382-020-05213-6.

DI ZIO M & GUARNERA U. 2009. Semiparametric Predictive Mean Matching. AStA Adv Stat Anal 93: 175-186. URL https://doi.org/10.1007/s10182-008-0081-2.

DIKBAS F. 2017. Frequency based imputation of precipitation. Stoch Environ Res Risk Assess 31(9): 2415-2434. URL https://doi.org/10.1007/s00477-016-1356-x.

EISCHEID JK, PASTERIS PA, DIAZ HF, PLANTICO MS & LOTT NJ. 2000. Creating a Serially Complete, National Daily Time Series of Temperature and Precipitation for the Western United States. J Appl Meteor 39(9): 1580-1591. doi:10.1175/1520-0450(2000)039<1580: cascnd>2.0.co;.

GILLILAND JM & KEIM BD. 2018. Surface wind speed: trend and climatology of Brazil from 1980-2014. Int J Climatol 38: 1060-1073. URL https://doi.org/10.1002/joc.5237.

GOMES HB, AMBRIZZI T, DA SILVA BFP, HODGES K, DIAS PLS, HERDIES DL, SILVA MCL & GOMES HB. 2019. Climatology of easterly wave disturbances over the tropical South Atlantic. Clim Dyn 51(3-4): 1393-1411. doi:10.1007/s00382-019-04667-7.

HAYLOCK MR ET AL. 2006. Trends in total and extreme South American rainfall in 1960-2000 and links with sea surface temperature. J Clim 19: 1490-1512.

HONAKER J, KING G & BLACKWELL M. 2011. Amelia II: A Program for Missing Data. J Stat Softw 45: 1-47. doi:10.18637/jss.v045.i07.

HORTON NJ & LIPSITZ SR. 2001. Multiple imputation in practice: comparison of software packages for regression models with missing variables. Am Stat 55: 244-54. URL https://doi.org/10.1198/000313001317098266.

IBGE - INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. 2010. Sinopse do Censo Demográfico 2010. URL http://www.censo2010.ibge.gov.br.

IZZO M, AUCELLI PPC & MARATEA A. 2020. Historical trends of rain and air temperature in the Dominican Republic. Int J of Climatol doi:10.1002/joc.6710.

JADHAV A, PRAMOD D & RAMANATHAN K. 2019. Comparison of Performance of Data Imputation Methods for Numeric Dataset. Appl Artif Intell 33(10): 913-933. doi:10.1080/08839514.2019.1637138.

JAHAN F, SINHA NC, RAHMAN MM, RAHMAN MM, MONDAL MSH & ISLAM MA. 2018. Comparison of missing value estimation techniques in rainfall data of Bangladesh. Theor Appl Clim 136: 1115-1131. doi:10.1007/s00704-018-2537-y.

JUNNINEN H, NISKA H, TUPPURAINEN K, RUUSKANEN J & KOLEHMAINEN M. 2004. Methods for imputation of missing values in air quality data sets. Atmos Environ 38(18): 2895-2907. doi:10.1016/j.atmosenv.2004.02.026.

KAPLAN D & YAVUZ S. 2019. An Approach to Addressing Multiple Imputation Model Uncertainty Using Bayesian Model Averaging. Multivariate Behav Res 55(4): 553-567. doi:10.1080/00273171.2019.1657790.

KOUSKY VE. 1979. Frontal influences on northeast Brazil. Month Weather Rev 107(9): 1140-1153. doi:10.1175/1520-0493(1979)107<1140:FIONB>2.0.CO;2.

LI KH, RAGHUNATHAN TE & RUBIN DB. 1991. Large-sample significance levels from multiply imputed data using moment-based statistics and an F reference distribution. J Am Stat Assoc 86(416): 1065-1073.

LITTLE RJ & RUBIN DB. 2014. Statistical Analysis with Missing Data. Vol. 333. New Jersey: J Wiley & Sons.

LITTLE RJA & RUBIN DB. 2002. Statistical analysis with missing data. 2nd ed. New Jersey: J Wiley & Sons.

LITTLE RJA & RUBIN DB. 2019. Statistical analysis with missing data. 3rd ed. New Jersey: J Wiley & Sons.

LO PRESTI R, BARCA E & PASSARELLA GA. 2010. A methodology for treating missing data applied to daily rainfall data in the Candelaro River Basin (Italy). Environ Monitor Assess 160: 1-22. doi:10.1007/s10661-008-0653-3.

MARENGO JA, TORRES RR & ALVES LM. 2017. Drought in Northeast Brazil—past, present, and future. Theor Appl Clim 129(3-4): 1189-1200. URL https://doi.org/10.1007/s00704-016-1840-8.

MARINHO KFS, ANDRADE LDMB, SPYRIDES MHC, SANTOS E SILVA CM, DE OLIVEIRA CP, BEZERRA BB & MUTTI PR. 2020. Clim Prof Braz Microreg Atmos 11(11): 1217. URL ttps://doi.org/10.3390/atmos11111217.

MCKNIGHT PE, MCKNIGHT KM, SIDANI S & FIGUEREDO AJ. 2007. Missing data: a gentle introduction. New York: The Guilford Press.

MICHOT V, ARVOR D, RONCHAIL J, CORPETTI T, JEGOU N, LUCIO PS & DUBREUIL V. 2019. Validation and reconstruction of rain gauge-based daily time series for the entire Amazon basin. Theoretical and Applied Climatology URL https://doi.org/10.1007/s00704-019-02832-w.

MOURA AD & SHUKLA J. 1981. On the dynamics of droughts in Northeast Brazil: observations, theory and numerical experiments with a general circulation model. J Atmos Sci 38(12): 2653-2675. URL https://doi.org/10.1175/1520-0469(1981)0382.0.CO;2.

NUNES LN, KLUCK MM & FACHEL JMG. 2009. Uso da imputação múltipla de dados faltantes: uma simulação utilizando dados epidemiológicos. Cad Saúde Pública 25: 268-278.

OLIVEIRA PT, SANTOS E SILVA CM & LIMA KC. 2017. Climatology and trend analysis of extreme precipitation in subregions of Northeast Brazil. Theor Appl Clim 130(1-2): 77-90. URL https://doi.org/10.1007/s00704-016-1865-z.

OLIVEIRA PTD, E SILVA SC & LIMA KC. 2013. Synoptic environment associated with heavy rainfall events on the coastland of Northeast Brazil. Adv Geosci 35: 73-78. URL https://doi.org/10.5194/adgeo-35-73-2013.

PALHARINI RSA & VILA DA. 2017. Climatological Behavior of Precipitating Clouds in the Northeast Region of Brazil. Adv Meteorol 2017: 5916150. doi:10.1155/2017/5916150.

PALHARINI RSA, VILA DA, RODRIGUES DT, QUISPE DP, PALHARINI RC, SIQUEIRA RA & DE AFONSO JMS. 2020. Assessment of the Extreme Precipitation by Satellite Estimates over South America. Rem Sens 12(13): 2085. doi:10.3390/rs12132085.

RAO VB, DE LIMA MC & FRANCHITO SH. 1993. Seasonal and interannual variations of rainfall over eastern Northeast Brazil. J Clim 6(9): 1754-1763. URL https://doi.org/10.1175/1520-0442(1993)0062.0.CO;2.

RC TEAM. 2020. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, Austria. URL https://www.R-project.org/.

RODRIGUES DT, GONÇALVES WA, SPYRIDES MHC, ANDRADE LMB, DE SOUZA DO, DE ARAUJO PAA, DA SILVA ACN & SANTOS E SILVA CM. 2021. Probability of occurrence of extreme precipitation

events and natural disasters in the city of Natal, Brazil. Urb Clim 35: 100753. URL https://doi.org/10.1016/j.uclim.2020.100753.

RODRIGUES DT, GONÇALVES WA, SPYRIDES MHC & SANTOS E SILVA CM. 2019. Spatial and temporal assessment of the extreme and daily precipitation of the tropical rainfall measuring Mission satellite in Northeast Brazil. Int J Remote Sens 41(2): 549-572. URL https://doi.org/10.1080/01431161.2019.1643940.

RODRIGUES DT, GONÇALVES WA, SPYRIDES MHC, SANTOS E SILVA CM & DE SOUZA DO. 2020. Spatial distribution of the level of return of extreme precipitation events in Northeast Brazil. Int J Climatol 40(12): 5098-5113. URL https://doi.org/10.1002/joc.6507.

RUBIN DB. 1987. Multiple Imputation for Nonresponse in Surveys. New York: J Wiley & Sons.

RUBIN DB. 1994. Comment on "Missing Data, Imputation, and the Bootstrap" by Bradley Efron. J Am Statist Assoc 89: 475-478.

RUBIN DB. 1996. Multiple imputation after 18+ years. J Am Stat Assoc 91: 473-89.

SCHAFER JL. 1997. Analysis of Incomplete Multivariate Data. Florida: Chapman & Hall.

SCHAFER JL & GRAHAM JW. 2002. Missing data: our view of the state of the art. Psych Methds, Am Psychol Assoc 7: 147-177.

SCOTT DW. 2015. Multivariate density estimation: theory, practice, and visualization. 2nd ed. New Jersey: J Wiley & Sons.

SELUCHI M. 2004. Previsão de chuvas com distribuição irregular no período de março a maio de 2004 para o Nordeste do Brasil. INFOCLIMA, Boletim de Informações Climáticas 2(10): 2004.

SEMIROMI MT & KOCH M. 2019. Reconstruction of groundwater levels to impute missing values using singular and multichannel spectrum analysis: application to the Ardabil Plain, Iran. Hydrol Sci J 64(14): 1711-1726. doi:10.1080/02626667.2019.1669793.

SHEATHER SJ & JONES MC. 1991. A reliable data-based bandwidth selection method for kernel density estimation. J R Stat Soc Ser B 53: 683-690.

SILVERMAN BW. 1998. Density Estimation for statistics and data analysis. New York: Chapman and Hall/CRC.

SUDENE - SUPERINTENDÊNCIA DO DESENVOLVIMENTO DO NORDESTE. 2007. Delimitação do Semiárido. URL http://www.sudene.gov.br/delimitacao-do-semiarido. Accessed on 15 Oct 2019.

TAKAHASHI M. 2017. Multiple ratio imputation by the EMB algorithm: Theory and simulation. Journal of Modern Appl Stat Methods 16(1): 34.

TEEGAVARAPU RSV & CHANDRAMOULI V. 2005. Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. J Hydrol 312: 191-206. URL https://doi.org/10.1016/j.jhydrol.2005.02.015.

TEEGAVARAPU RSV & CHANDRAMOULI V. 2014. Missing precipitation data estimation using optimal proximity metric-based imputation, nearest-neighbour classification and cluster-based interpolation methods. Hydrol Sci J 59(11): 2009-2026. doi:10.1080/02626667.2013.862334.

TORRES RR & FERREIRA NJ. 2011. Case studies of easterly wave disturbances over Northeast Brazil using the Eta Model. Weather For 26(2): 225-235.

TORRES RR, LAPOLA DM, MARENGO JA & LOMBARDO MA. 2012. Socio-climatic hotspots in Brazil. Clim Change 115(3-4): 597-609. doi:10.1007/s10584-012-0461-1.

TORRES RR & MARENGO JA. 2014. Climate change hotspots over South America: from CMIP3 to CMIP5 multi-model datasets. Theor Appl Climatol 117(3-4): 579-587. URL https://doi.org/10.1007/s00704-013-1030-x.

VENABLES WN & RIPLEY BD. 2002. Modern Applied Statistics with S. New York: Springer.

VINCENT L & YUMIKO M. 2006. Trends in total and extreme South American rainfall 1960-2000 and links with sea surface temperature. J Clim 19: 1490-1512.

XAVIER AC, KING CW & SCANLON BR. 2016. Daily gridded meteorological variables in Brazil (1980-2013). Int J Climatol 36: 2644-2659. URL https://doi.org/10.1002/joc.4518.

YANTO BL & RAJAGOPALAN B. 2017. Development of a Gridded Meteorological Dataset over Java Island, Indonesia 1985-2014. Sci Data 4(1): 170072. doi:10.1038/sdata.2017.72.

YENDRA R, JEMAIN AA, ZAHARI M & WAN ZIN WZ. 2013. Methods on handling missing rainfall data with Neyman-Scott rectangular pulse modeling. AIP Conference Proceedings 1522: 1213-1220. URL https://doi.org/10.1063/1.4801269.

YOZGATLIGIL C, ASLAN S, IYIGUN C & I B. 2013. Comparison of missing value imputation methods in time series: the case of Turkish meteorological data. Theor Appl Climatol 112: 143-167. URL https://doi.org/10.1007/s00704-012-0723-x.

ZHOU XH, ECKERT GJ & TIERNEY WM. 2001. Multiple imputation in public health research. Stat Med 20: 1541-1549. URL https://doi.org/10.1002/sim.689.

How to cite

RODRIGUES DT, GONÇALVES WA, SILVA CMS, SPYRIDES MHC & LÚCIO PS. 2023. Imputation of precipitation data in northeast Brazil. An Acad Bras Cienc 95: e20210737. DOI 10.1590/0001-3765202320210737.

Manuscript received on May 18, 2021;
accepted for publication on November 13, 2021

DANIELE T. RODRIGUES[1]
https://orcid.org/0000-0003-4307-2832

WEBER A. GONÇALVES[2]
https://orcid.org/0000-0002-5073-8527

CLÁUDIO MOISÉS S. E SILVA[2]
https://orcid.org/0000-0002-2251-7348

MARIA HELENA C. SPYRIDES[2]
https://orcid.org/0000-0001-8087-1962

PAULO SÉRGIO LÚCIO[2]
https://orcid.org/0000-0002-8170-934X

[1]Graduação em Estatística, Universidade Federal do Piauí, Departamento de Estatística, Av. Campus Universitário Ministro Petrônio Portella, s/n, Ininga, Teresina, PI, Brazil
[2]Programa de Pós-Graduação em Ciências Climáticas, Universidade Federal do Rio Grande do Norte, Departamento de Ciências Climáticas e Atmosféricas, Av. Senador Salgado Filho, 3000, Lagoa Nova, 59078-970 Natal, RN, Brazil

Correspondence to: **Daniele Tôrres Rodrigues**
E-mail: mspdany@yahoo.com.br

Author contributions

DANIELE TÔRRES RODRIGUES, Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft. WEBER ANDRADE GONÇALVES, CLÁUDIO MOISÉS SANTOS E SILVA, Formal analysis, Supervision, Writing – review & editing. MARIA HELENA CONSTANTINO SPYRIDES, PAULO SÉRGIO LÚCIO, Supervision, Writing – review & editing.