Theoretical and Applied Climatology (2019) 138:759–775
https://doi.org/10.1007/s00704-019-02832-w

ORIGINAL PAPER

# Validation and reconstruction of rain gauge–based daily time series for the entire Amazon basin

Véronique Michot[1,2] · Damien Arvor[1] · Josyane Ronchail[2,3] · Thomas Corpetti[1] · Nicolas Jegou[4] · Paulo Sérgio Lucio[5] · Vincent Dubreuil[1]

Received: 21 March 2018 / Accepted: 1 March 2019 / Published online: 13 April 2019
© Springer-Verlag GmbH Austria, part of Springer Nature 2019

## Abstract

Monitoring the spatio-temporal variability of rainfall regimes in the Amazon basin is difficult because (1) time series of remote sensing–based rainfall estimates are still too short for long-time variability analysis and (2) rain gauge time series are not fully reliable and operational in their current state due to frequent gaps and zero values. The objective of this paper is to introduce a quality control and reconstruction procedure designed to produce a robust database of rain gauge–based daily rainfall in the Amazon basin. Despite the low density and heterogeneous spatial distribution of the rain gauges network, we eliminated unexpected values and produced accurate estimates using spatial and mathematical relationships with neighboring rain gauges. Three reconstruction methods were tested: the nearest neighbor approach (NN), the arithmetic mean with neighboring stations (AM), and the multiple imputation by chained equations used with the predictive mean matching procedure (MICE). The quality of the reconstruction has been assessed through the mean annual rainfall and the mean annual number of rainy days. We concluded that the AM approach performed better at the scale of the whole Amazon basin. This method has then been preferred to reconstruct the whole database of rainfall time series.

## 1 Introduction

Analyzing the seasonal and interannual spatio-temporal variability of rainfall in the Amazon basin is a complex issue that requires long and complete daily time series. Whereas remote sensing–based estimates are widely used to monitor rainfall regimes at a regional scale, daily rainfall observations from rain gauges still present major advantages by providing (1) more accurate precipitation measures (Liebmann and Allured 2005) on (2) longer time series, thus matching with the recommendations of the World Meteorological Organization (WMO) to use at least 30-year time series to analyze climatic trends (WMO 2011). This is why, even if estimated products exist and are provided by the National Atmospheric Agencies, it is of great interest to try to find a reliable reconstruction method based on observed data.

Unfortunately, the Amazon basin, covered by more than six million km$^2$ across six countries (Brazil, Venezuela, Ecuador, Colombia, Peru, and Bolivia), presents a very heterogeneous rain gauge network characterized by a poor density, a heterogeneous spatial distribution, and a large number of erroneous measurements. Potential sources of errors in precipitation time series refer to human mistakes or transmission errors, missing values, missing values recorded as zero, hidden accumulated values, unexpected high values, and inhomogeneity (Williams et al. 2005; WMO 2007). The homogenization of such data is a tricky task because it implies a risk of removing a potential real climatic signal (Aguilar et al. 2005; WMO 2007; Mestre et al. 2011). This is especially true for precipitations characterized by a large spatial and temporal variability in opposition to temperatures whose linear spatial distribution allows applying homogenization processes (Caussinus and Mestre 2004).

All these issues can affect climatological studies (Glasson-Cicognani and Berchtold 2010) so that data quality needs to be checked carefully and gaps must be reconstructed before

✉ Véronique Michot
veronique.michot@univ-rennes2.fr

[1] Univ-Rennes, LETG, French National Center for Scientific Research (CNRS), UMR 6554, F-35000 Rennes, France

[2] UMR 7159 Laboratoire d'Océanographie et du Climat (LOCEAN, CNRS-IRD-MNHN-SU), Institut Pierre Simon Laplace (IPSL), Université Pierre et Marie Curie, 4 place Jussieu, 75005 Paris, France

[3] Université Paris Diderot, Sorbonne Paris Cité, Case Courrier 7001, 75205 Paris Cedex 13, France

[4] Univ Rennes, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France

[5] Programa de Pós-Graduação em Ciências Climáticas, Universidade Federal do Rio Grande do Norte, Natal, Brazil

any climatic analysis. However, implementing such a procedure for data quality control and reconstruction is a challenging task because the lack of metadata often hinders the implementation of automatic approaches such as the ETCCDI (Expert Team on Climate Change Detection and Indices) provided by the Canadian Centre for Climate Modeling and Analysis for instance. In addition, some errors are very difficult to identify. For example, high values due to extreme rainy events are difficult to discriminate from those due to errors. Similarly, dry spells or days with zero precipitations can be confused with sequences of missing values recorded as zero. These issues often remain unanswered and are rarely taken into account in reconstruction software. The high spatial and temporal variability is problematic for time series reconstruction because this variability can be compared with something like discontinuities and heterogeneity in the series and then it turns hard to report on the complex distribution of the precipitations. Variability generates uncertainties and then requires transversal methods capable of recreating values similar to those observed in the same spatial and temporal context. The effectiveness of the quality control and reconstruction procedure is called into question by the geographical characteristics of the study area, particularly in the case of the large Amazon basin (more than 6 million km$^2$ between 5° N and 20° S). In addition, it is characterized by lowlands and highlands: the Guiana massif in the north and the Brazilian plateau in the south, while in the west, the Andes mountain range reaches a height of more than 6000 m. Orography plays an important role in water vapor flows as the highlands and Andes divert monsoon flows from the northeast to south, thus contributing to the redistribution of rainfall in the south of the AB and in South America.

In this context, the objective of this paper is to introduce a procedure to control and reconstruct daily precipitation time series measured at rain gauges in the Amazon Basin. We specially search for a method able to recreate high spatial and temporal (daily to seasonal) variability of precipitations. Section 2 presents the original database of daily rainfall time series. Section 3 describes the methods and results of the quality control procedure. Section 4 introduces three reconstruction methods and associated results. Section 5 provides an assessment of the quality of the reconstructed database. Finally, results are discussed in Section 6.

## 2 The original database and the data selection procedure

We acquired daily rainfall data from national meteorological agencies responsible for rain gauge networks in the Amazon basin: the National Water Agency (ANA) and the National Meteorological Institute (INMET) in Brazil, the National Meteorological and Hydrological Institute (INAMHI) in Ecuador, the Hydrological Meteorological and Environmental Studies Institute (IDEAM) in Colombia, and the National Hydrological and Meteorological Service (SENAMHI) in Peru and Bolivia. Unfortunately, no data have been collected in Venezuela. The international rain gauge network is heterogeneous with stations mainly located in historical human settlements along rivers and roads (for example, the North-South transamazonian BR163 road), at the estuary of the Amazon River and in the Peruvian Andes (Ronchail et al. 2002). Data quality is also unevenly distributed across the stations with more complete time series at stations from the Peruvian Andes and the mouth of the Amazon River (Fig. 2b). Finally, it is worth noting that some periods of political and economic crisis (e.g., the early 1990s in Brazil or 2003 in Ecuador) are strongly affected by data gaps due to the reduced capacity to collect meteorological data.

Because this original database contains many erroneous data, we applied a three-step selection procedure to focus on the analysis on relevant stations (Table 1). First, we selected stations with at least three-decade records (WMO 1989, 2007) beginning between 1981 and 1983 and ending between 2009 and 2013, i.e., the period with the greatest number of available data. After this process, the dataset is made of $N = 533$ rain gauges (400 in Brazil and 133 in the Andean countries as shown in Table 1). Second, we discarded all stations with more than 20% missing values or not available values (NA). This threshold, although lower than the one set by Campozano et al. (2015) in Ecuador, is high but necessary to maintain a sufficient network. The remaining stations were $N = 346$ (225 in Brazil and 121 in the Andean countries; Table 1). Third, we deleted 141 suspicious time series based on a visual inspection with, for instance, repeated values during long time periods were discarded (Fig. 1). At the end, the final rain gauge network is composed of 205 stations (Table 2 and Fig. 2a–c). It is worth noting that it includes a few stations located outside the Amazon basin (mostly in Brazil) in order to take into account stations close to the Atlantic Ocean and to fill spatial gaps.

## 3 Quality control

The quality control procedure aims at detecting unexpected high values and unexpected dry sequences.

### 3.1 Control of high values

High values in rainfall time series can result from real extreme climatic events or from mistakes during the data acquisition or transcription. Discriminating real from erroneous high values is a challenging task since it requires to set an appropriate threshold, which usually depends on the study area. In East Africa for instance, Boyard-Micheau (2013) set this threshold at 400 mm of daily precipitation, corresponding to the

**Table 1** Workflow of the selection, control, and reconstruction procedures to produce complete daily precipitation time series in the Amazon basin

| Source of data | Brasil | Colombia | Ecuador | Bolivia | Peru | Total |
|---|---|---|---|---|---|---|
| (1) Data selection: | (a) Selection of rain gauges with at least 30 years record | | | | | |
| number of stations by country | 400 | 10 | 23 | 31 | 69 | 533 |
| | (b) Selection of rain gauge with less than 20% of not available values (NA) | | | | | |
| | 225 | 10 | 22 | 20 | 69 | 346 |
| | (c) Deletion of inhomogeneous time series, after visual inspection | | | | | |
| | 145 | 2 | 12 | 17 | 29 | 205 |
| (2) Quality control | (d) Checking of improbably large values | | | | | |
| | (e) Control of wrong zero values: | | | | | |
| | -Detection of potential wrong values | | | | | |
| | -Validation of potential wrong values as wrong values: | | | | | |
| | * Clustering of rain gauges | | | | | |
| | * Validation by the nearest neighbor approach | | | | | |
| (3) Reconstruction | (f) Test of 3 different reconstruction methods | | | | | |
| | -MICE | | | | | |
| | -Arithmetic mean | | | | | |
| | -Nearest neighbor approach | | | | | |
| (4) Validation of reconstruction methods | (g) Computation of the relative RMSE | | | | | |
| | (h) Comparison of daily mean after and before the reconstruction | | | | | |
| | (i) Comparison of the number of rainy days per year after and before the reconstruction | | | | | |

maximum observed in this region. In the Brazilian Amazon, Santos et al. (2015) presented an analysis of the return periods of maximum daily precipitation according to extreme value theory, in the Brazilian Legal Amazon during the period 1983–2012. Their work particularly highlights a difference between the northwest and the rest of the territory. Using the

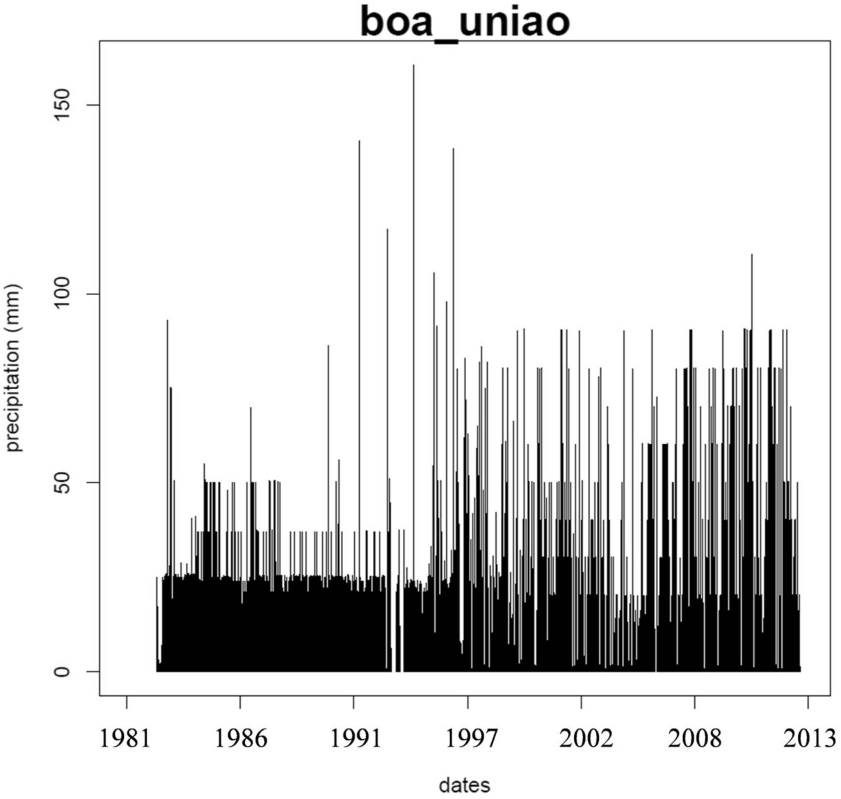


**Fig. 1** Example of a daily rainfall time series with a suspicious structure due to the repetition of the same value

**Table 2** Names, coordinates, and altitude of the 205 stations of the study

| Rain gauge name | Latitude | Longitude | Elevation (m) | Rain gauge name | Latitude | Longitude | Elevation (m) |
|---|---|---|---|---|---|---|---|
| Acampamento idbf | − 1.79 | − 51.43 | 2 | Cruzeiro do sul inmet | − 7.6 | − 72.66 | 200 |
| acanaui | − 1.82 | − 66.6 | 63 | Cucui | 0.78527778 | − 66.8522222 | 90 |
| Agrapecuaria cajabi | − 10.7461111 | − 54.5461111 | 342 | Cumaru | − 0.6 | − 63.4 | 41 |
| Aguaytia | − 9.03 | − 75.51 | 285 | Cunumbuque | − 7.35 | − 75.01 | 134 |
| Alao | − 1.53 | − 78.29 | 2674 | Cupari | − 4.18 | − 55.43 | 66 |
| Alao perou | − 6.53 | − 76.73 | 429 | Divinea | − 12.9397222 | − 51.8263889 | 390 |
| Almeirim | −1.53 | −52.58 | 15 | El alto | − 16.48 | − 68.17 | 4144 |
| Alo brasil | − 12.1641667 | − 51.6969444 | 339 | El pangui | − 3.646 | − 78.567 | 840 |
| Altamira inmet | − 3.21 | − 51.21 | 89 | El trompillo | − 17.8 | − 63.17 | 424 |
| Anthony br 364 | − 9.26055556 | − 63.1619444 | 98 | Envira | − 7.43 | − 70.02 | 134 |
| Apalai | 1.22 | − 54.66 | 311 | Escola caramuru | − 10.51 | − 63.65 | 181 |
| Aporema | 1.23 | − 50.9 | 16 | Estirao da santa cruz | − 4.29 | − 65.2 | 49 |
| Arapari | − 1.77 | − 54.4 | 21 | Faz parana | 1.13 | − 60.4 | 74 |
| Areias | − 1.21 | − 51.26 | 0 | Faz rio branco | − 9.89 | − 62.99 | 118 |
| Artunduaga | 1.35 | − 75.33 | 268 | Faz sao jao | 3.66 | − 61.38 | 96 |
| Assis brasil | − 10.93 | − 69.57 | 278 | Fazenda agrochapada | − 13.4466667 | − 54.2805556 | 428 |
| Badajos amazonas | − 3.42 | − 62.68 | 17 | Fazenda itauba | − 11.4713889 | − 56.4333333 | 348 |
| Badajos para | − 2.51 | − 47.77 | 12 | Fazenda paranacre | − 7.95 | − 71.48 | 198 |
| Balsa do rio urubu | − 2.91 | − 59.04 | 19 | Fazenda sheffer | − 8.33 | − 65.72 | 113 |
| Bambamarca | − 6.68 | − 78.53 | 2657 | Fe e esperanca | 2.87 | − 61.44 | 78 |
| Barcelos inmet | − 0.96 | − 62.91 | 17 | Fez esetrela do norte | − 3.87 | − 50.46 | 110 |
| Barra do sao manuel | − 7.34 | − 58.16 | 128 | Fontanilhas | − 11.3416667 | − 58.3383333 | 242 |
| Barreira do campo | − 9.18 | − 50.21 | 184 | Fonte boa | − 2.53 | − 66.16 | 81 |
| Barreirinha | − 2.79 | − 57.06 | 13 | Foz do breu | − 9.4 | − 72.7 | 261 |
| Belterra | − 2.63 | − 54.95 | 97 | Francisco orellana | − 3.42 | − 72.77 | 87 |
| Benjamin constant | − 4.38 | − 70.03 | 78 | Giron | − 3.9 | − 78.9 | 1157 |
| Boa vista inmet | 2.82 | − 60.66 | 61 | Guajara mirim | − 10.79 | − 65.35 | 118 |
| Boa vista roraima | 2.82 | − 60.65 | 61 | Huancabamba | − 5.25 | − 79.55 | 3177 |
| Boca do inferno | − 1.5 | − 54.87 | 39 | Huangacocha | − 7.94 | − 78.07 | 3759 |
| Brasil novo | − 3.62 | − 52.54 | 117 | Humboldt | − 10.1752778 | − 59.4516667 | 233 |
| Brasileia | − 11.02 | − 68.74 | 194 | Iauarete | 0.61 | − 69.18 | 122 |
| Caarapo | − 14.3841667 | − 58.2344444 | 585 | Ipixuna | − 7.05 | − 71.68 | 157 |
| Cachachi | − 7.46 | − 78.27 | 3320 | Itacoatiara | − 3.13 | − 58.43 | 30 |
| Cachoeira | − 7.7 | − 66.05 | 72 | Itaituba | − 4.82 | − 56 | 261 |
| Cachoeira da porteira | − 1.09 | − 57.05 | 18 | Jacareacanga | − 6.24 | − 57.78 | 108 |
| Cachoeira morena | − 2.11 | − 59.34 | 23 | Jacas chico | − 9.88 | − 76.5 | 3668 |
| Cachoeira uaupes | − 0.11 | − 67 | 99 | Jaen | − 5.68 | − 78.78 | 633 |
| Cafelandia do leste | − 11.6511111 | − 55.7025 | 303 | Jamalca | − 5.9 | − 78.24 | 1294 |
| Cajueiro | − 5.65 | − 54.52 | 201 | Jarilandia | − 1.12 | − 52 | 19 |
| Camiri | − 20.04 | − 63.52 | 812 | Jaru | − 10.45 | − 62.47 | 140 |
| Canutama | − 6.54 | − 64.39 | 49 | juruti | − 2.15 | − 56.09 | 26 |
| Capinota | − 17.71 | − 65.27 | 2599 | Jusante foz peixoto de azevedo | − 9.64333333 | − 56.0186111 | 240 |
| Caracarai | 1.82 | − 61.12 | 45 | Km 1027da br 163 | − 7.51 | − 55.26 | 232 |
| Carvoeiro | − 1.39 | − 61.98 | 14 | Km 1326br 163 | − 5.18 | − 56.06 | 100 |
| Chiquian | − 10.15 | − 77.15 | 3344 | Km 1385 br 163 | − 4.75 | − 56.08 | 128 |
| Cipoal | − 2.79 | − 50.45 | 25 | Km 947br 163 | − 8.19 | − 55.12 | 241 |
| Coari | − 4.08 | − 63.13 | 34 | Km zero pa 70 | − 4.29111111 | − 47.5652778 | 261 |
| Cobija | − 11.03 | − 68.76 | 219 | Livramento | − 0.29 | − 66.15 | 47 |

**Table 2** (continued)

| Rain gauge name | Latitude | Longitude | Elevation (m) | Rain gauge name | Latitude | Longitude | Elevation (m) |
|---|---|---|---|---|---|---|---|
| Cochabamba | − 17.41 | − 66.17 | 2564 | Macapa | − 0.05 | − 51.07 | 0 |
| Colonia do Taiano | 3.29 | − 61.09 | 141 | Magdalena | − 13.25 | − 64.05 | 148 |
| Concepcion | − 16.13 | − 62.03 | 496 | Maloca do contao | 4.17 | − 60.53 | 108 |
| Costa rica | − 10.7986111 | − 55.4486111 | 313 | Maloca sao tome | 0.18 | − 67.95 | 71 |
| Manaus | − 3.12 | − 59.95 | 76 | Salinopolis | − 0.65 | − 47.55 | 21 |
| Maracacuera Florestal | − 2.25 | − 51.18 | 37 | San borja | − 14.86 | − 66.74 | 195 |
| Marco rondon | − 12.02 | − 60.86 | 264 | San ignacio mo | − 14.97 | − 65.63 | 163 |
| Milpo | − 9.88 | − 77.23 | 4468 | San ignacio ve | − 16.37 | − 60.95 | 387 |
| Missao icana | 1.07 | − 67.59 | 87 | San joaquin | − 13.05 | − 64.67 | 140 |
| Monte alegre do xingu | − 4.67 | − 52.72 | 144 | San marcos | − 7.32 | − 78.17 | 2294 |
| Moura | − 1.45 | − 61.63 | 12 | San pablo | − 6.81 | − 76.58 | 275 |
| Moyobamba | − 6 | − 76.97 | 832 | Sangay | − 1.42 | − 77.57 | 598 |
| Namora | − 7.2 | − 78.34 | 2779 | Santarem | − 2.44 | − 54.71 | 47 |
| Naranjillo | − 5.83 | − 77.39 | 913 | Santarem sucunduri | − 6.8 | − 59.04 | 33 |
| Nauta | − 4.52 | − 73.6 | 102 | Sao francisco | − 0.57 | − 52.58 | 37 |
| Navio | − 0.4 | − 51.42 | 3 | Sao paulo de olivenca | − 3.46 | − 68.91 | 83 |
| Nhamunda | − 2.19 | − 56.71 | 1 | Sayausi | − 2.866 | − 79.067 | 2794 |
| Nova california | − 9.76 | − 66.61 | 156 | Selviria | − 11.73 | − 51.9888889 | 374 |
| Nova maringa | − 13.0661111 | − 57.1133333 | 307 | Seringal 70 | − 10.24 | − 62.63 | 221 |
| Nova olinda do norte | − 3.88 | − 59.09 | 8 | Seringal boa fe | − 7.24 | − 72.33 | 176 |
| Novo airao | − 2.62 | − 60.95 | 29 | Seringal fortaleza | − 7.72 | − 66.98 | 104 |
| Novo aripuana | − 7.2 | − 60.38 | 154 | Seringal jenipapo | − 6 | − 60.19 | 19 |
| Nucleo colonial rio ferro | − 12.5177778 | − 54.9125 | 349 | Seringal moreira | − 5.11 | − 63.98 | 54 |
| Nuevo rocafuerte | − 0.917 | − 75.417 | 183 | Serra do moa | − 7.44 | − 73.65 | 237 |
| Oeiras do para | − 2 | − 49.86 | 18 | Serra do navio | 0.88 | − 52.01 | 92 |
| Oriximina | − 1.76 | − 55.86 | 66 | Shanusi | − 6.07 | − 76.25 | 153 |
| Osorio fonseca | − 3.82 | − 58.29 | 21 | Sigsig | − 3.048 | − 78.784 | 2746 |
| Palmeiras do javari | − 0.17 | − 72.81 | 219 | Sitio sao pedro | − 3.89 | − 54.32 | 74 |
| Paranatinga | − 14.4177778 | − 54.0494444 | 485 | Sondorillo | − 5.34 | − 79.41 | 1776 |
| Parecis | − 14.1563889 | − 56.9330556 | 501 | Sta maria do boiacu | − 0.51 | − 61.79 | 25 |
| Pari cachoeira | 0.25 | − 69.78 | 118 | Sto antonio do ica | − 2.1 | − 67.94 | 109 |
| Passagem da br 309 | − 14.6119444 | −53.9986111 | 546 | Sto dumont | − 6.44 | − 68.24 | 92 |
| Pedras negras | − 12.85 | − 62.9 | 163 | Sucre | − 19.01 | − 65.3 | 2912 |
| Picota | − 6.95 | − 76.34 | 213 | Tabajara | − 8.93 | − 62.05 | 80 |
| Pilahuin | − 1.18 | − 78.44 | 4153 | Tamshiyacu | − 4 | − 73.16 | 98 |
| Pilluana | − 6.78 | − 76.27 | 222 | Tapuruquara | − 0.42 | − 65.02 | 33 |
| Pimenta bueno | − 11.68 | − 61.19 | 181 | Taraqua | 0.13 | − 68.54 | 77 |
| Piscicola chirimichay | − 2.46 | − 78.1 | 1094 | Taumaturgo | − 8.94 | − 72.79 | 227 |
| Pongo de caynarachi | − 6.33 | − 76.3 | 227 | Tefe | − 3.83 | − 64.7 | 71 |
| Pontes e lacerda | − 15.2155556 | − 59.3536111 | 227 | Tingo maria | − 9.29 | − 76 | 641 |
| Porculla | − 5.85 | − 79.51 | 1944 | Torixoreu | − 9.94166667 | − 57.1330556 | 272 |
| Prainha | − 1.8 | − 53.48 | 35 | Trinidad | − 14.82 | − 64.92 | 155 |
| Pto leguizamo | − 1 | − 74.46 | 174 | Tuluce | − 5.49 | − 79.37 | 2056 |
| Pucallpa huimbayoc | − 6.46 | − 75.85 | 173 | Tunui | 1.39 | − 68.15 | 90 |
| Quebo | − 14.6525 | − 56.1238889 | 212 | Umanapana | − 1.89 | − 62.44 | 46 |
| Rest poteira do amazonas | − 9.5 | − 67.28 | 168 | Uruara | − 3.68 | − 53.55 | 78 |
| Riberalta | − 11.01 | − 66.08 | 138 | Urucara | − 7.20083333 | − 59.8922222 | 152 |
| Ricaurte cuenca | − 2851 | − 78.935 | 2520 | Vila do apui | − 7.2 | − 59.89 | 153 |
| Rio branco ana | − 9.98 | − 67.8 | 132 | Vila sao benedito | − 1.99 | − 50.37 | 23 |

**Table 2** (continued)

| Rain gauge name | Latitude | Longitude | Elevation (m) | Rain gauge name | Latitude | Longitude | Elevation (m) |
|---|---|---|---|---|---|---|---|
| Rio branco inmet | − 9.96 | − 67.8 | 138 | Vila sao jose do xingu | − 10.8072222 | − 52.7461111 | 339 |
| Rio mazar rivera | − 2.574 | − 78.65 | 2319 | Vista alegre rondonia | − 11.44 | − 61.48 | 167 |
| Robore | − 18.34 | − 59.72 | 376 | Xapuri | − 10.63 | − 68.51 | 198 |
| Rosario oeste | − 11.9291667 | − 54.9980556 | 301 | Xavantina | − 14.6722222 | − 52.3547222 | 266 |
| Rumipampa salcedo | − 1.1 | − 78.35 | 3605 | | | | |
| Ruropolis presidente medici | − 4.09 | − 54.9 | 103 | | | | |
| Rurrenabaque | − 14.43 | − 67.5 | 216 | | | | |

"generalized extreme value" (GEV) method, the author estimates a 10-year return of the daily maximum values of 234.2 mm in the south of the basin and 169.1 mm in the northwest. In the northwestern part of the basin, wetter than the rest of the BA (Espinoza Villar et al. 2009; Figueroa and Nobre 1990), the lower extreme daily values are due to the regularity of the rainfall throughout the year. Conversely, in the rest of the basin, the rainfall that mainly contributes to the year-to-date total is concentrated in the rainy season, during which much more intense events occur due to deep atmospheric convection. These results are consistent with those of Brito et al. (2014), which show that extreme events are (a) relatively rare and are not the main factor in the annual accumulation of rainfall in the northwestern part of the AB, (b) less frequent in the South than in the North, and (c) more intense in the South and Northeast of Brazil.

Taking into account the two Barbosa's thresholds, daily values higher than 169.1 mm in the northwest of the BA between − 2° S and 5° N − 80° W and − 87° W were deleted and replaced by a missing value. In the rest of the BA, the threshold was set at 234.2 mm per day. A total of 34 values above these thresholds were detected for the entire BA, including 15 in the northwest and 19 in the rest of the basin. Each station concerned recorded only one extreme value, so the addition of a missing value had little effect on the quality of the series concerned.

## 3.2 Control of dry sequences

Similar than high values, dry sequences (i.e., 0 values) can also correspond to real dry days or to errors, for instance when missing observations are recorded as 0. To control dry sequences, our strategy consisted in comparing them with rainfall measures from neighboring stations belonging to the same climatic region (Cressie and Chan 1989). Indeed, the probability to measure dry sequences at a given station depends on the average rainfall regime observed in the corresponding region. For instance, real and long dry sequences are more likely to occur in regions characterized by a long dry season. For this reason, our complete procedure to control dry sequences includes two steps: (1) the regionalization of major climatic regions and (2) the implementation of rules to identify unexpected dry sequences.

### 3.2.1 Rainfall regionalization

Two stations are considered as neighbors when they belong to the same rainfall regime (characterizing the global consistency of the stations over the region) and when the dissimilarity between them is small. Precipitation profiles have then to be regionalized into consistent clusters. To carry out these clusterization series without missing data is desirable. In order to temporarily overcome this difficulty, the monthly average for each series was calculated on the incomplete dataset. Moreover, since the relationship between the variables is more constant and strong when the time step increases (WMO 2011), this aggregation also makes it possible to strengthen the quality of the clustering. However, such calculation also may be in contradiction with WMO recommendations that indicate that cumulative or average precipitation totals or averages should not be produced when a large number of missing values exist in the month. This is relatively problematic since each of the problems requires the simultaneous resolution of the other.

Because of the complexity of related precipitations associated with the local variability of the topography, there is no reason why the separation between clusters should be linear. We therefore exploit a non-linear clustering approach, namely the spectral clustering, in order to allow a better climatological regionalization. The main principle of spectral clustering is to represent all rain gauges in separate nodes of a connected graph whose vertexes express the similarity between two nodes. The spectral analysis of this graph enables to isolate its main consistent groups. To compute the connection between two nodes into this graph, the basic solution consists in computing a simple Euclidian distance. However, in order to estimate clusters separated in a non-linear way, we exploit the kernel trick. The idea consists in projecting data in another space than the usual one (represented by multi-variate vectors where each component is a value of precipitation) where the separation between clusters is linear. Under some specific properties (see Camps-Valls and Camps-Valls and Bruzzone

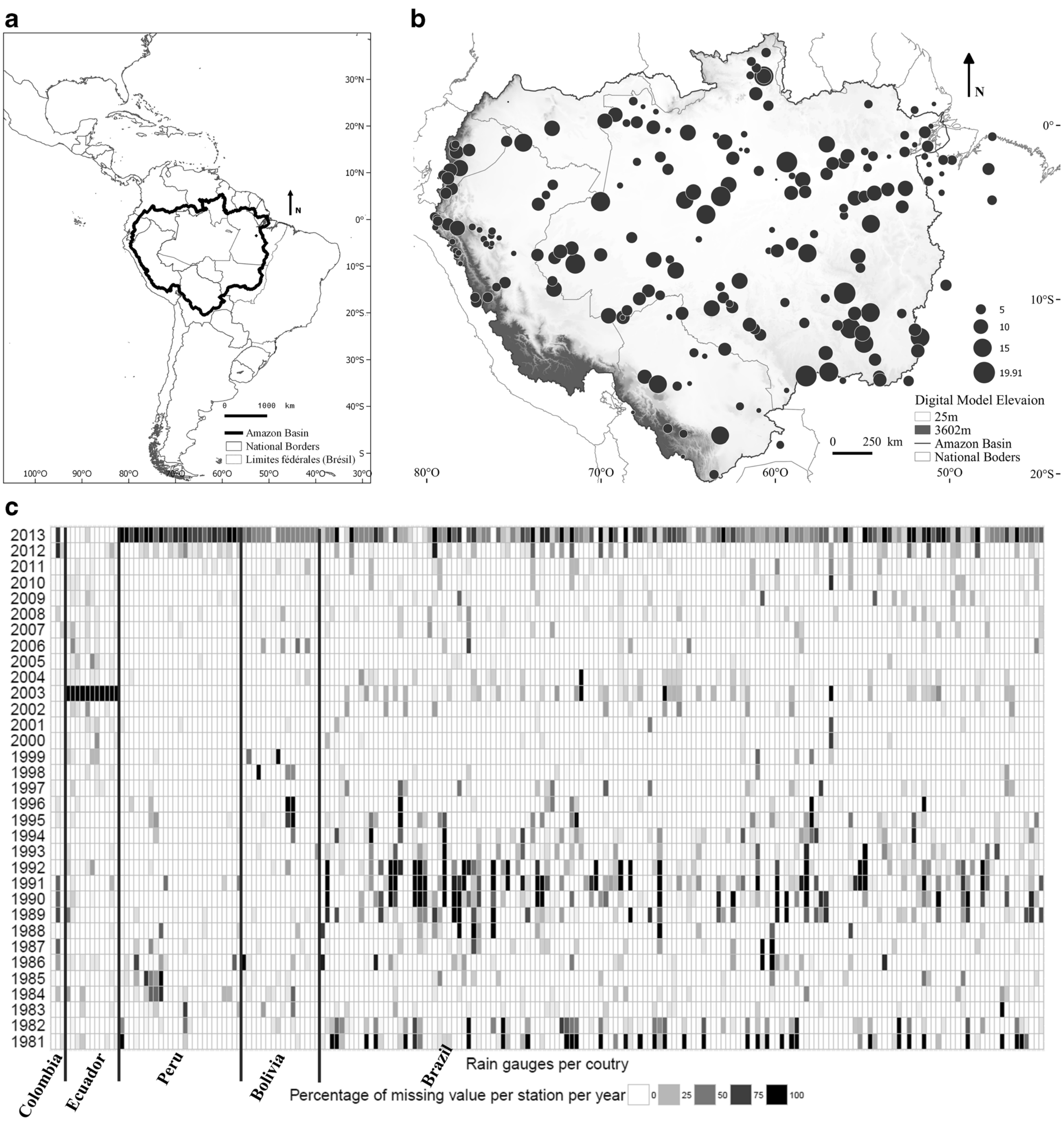


**Fig. 2** **a** Location of the Amazon basin. **b** Spatial distribution of missing values per rain gauge from 1981 to 2013, in percentage. **c** Temporal distribution of missing values, per station. Source of data: see Section 2. Digital elevation model, in meters, source: DEM GTOPO30 (USGS)

2009 for a complete theory of kernels), this projection can simply be done by changing the way one computes the connection between nodes. In practice, this is done using a Gaussian kernel where the connection between two rain gauges $x_1$ and $x_2$ is:

$$K(x_1, x_2) = e^{-\frac{\|x_1 - x_2\|}{\sigma^2}}$$

where $\sigma$ is a parameter to fix. It has been proven that this kernel enables to efficiently separate highly non-linear clusters.

The determination of the optimal number of clusters is an open problem for which no sound solution exists at the moment. In this study, we rely on the intra/inter inertia. More precisely, a reliable clustering should reveal both homogeneities inside clusters (all stations of the same group are similar) and heterogeneity between averaged clusters (all clusters represent different groups). Therefore, the ratio between the inertia among (averaged) clusters and the internal inertia (sum on inertia inside all groups) should be maximal. This ratio is depicted in Fig. 3a, and as one can see, the maximal value is reached for a number of 12 clusters that are used in practice. The resulting regionalization in Fig. 3b is consistent with former studies, showing for example the separation between tropical and equatorial regions and between highlands and lowlands (Barbosa Santos et al. 2015; Delahaye et al. 2015; Espinoza Villar et al. 2009; Figueroa and Nobre 1990).

### 3.2.2 Identification of unexpected dry sequences

Once clusters are determined, the detection of unexpected dry sequences in rainfall time series is done in four steps. First, the average duration of dry (i.e., 0 mm/day) sequences is calculated for each month of the year. Second, all dry sequences are compared with the average duration of dry sequences of the corresponding month. Third, the dry sequences longer than the average duration are flagged as doubtful. Four, these sequences are compared with the values measured at the two nearest neighboring stations located in the same cluster during the same time period (Vicente-Serrano et al. 2010). If more

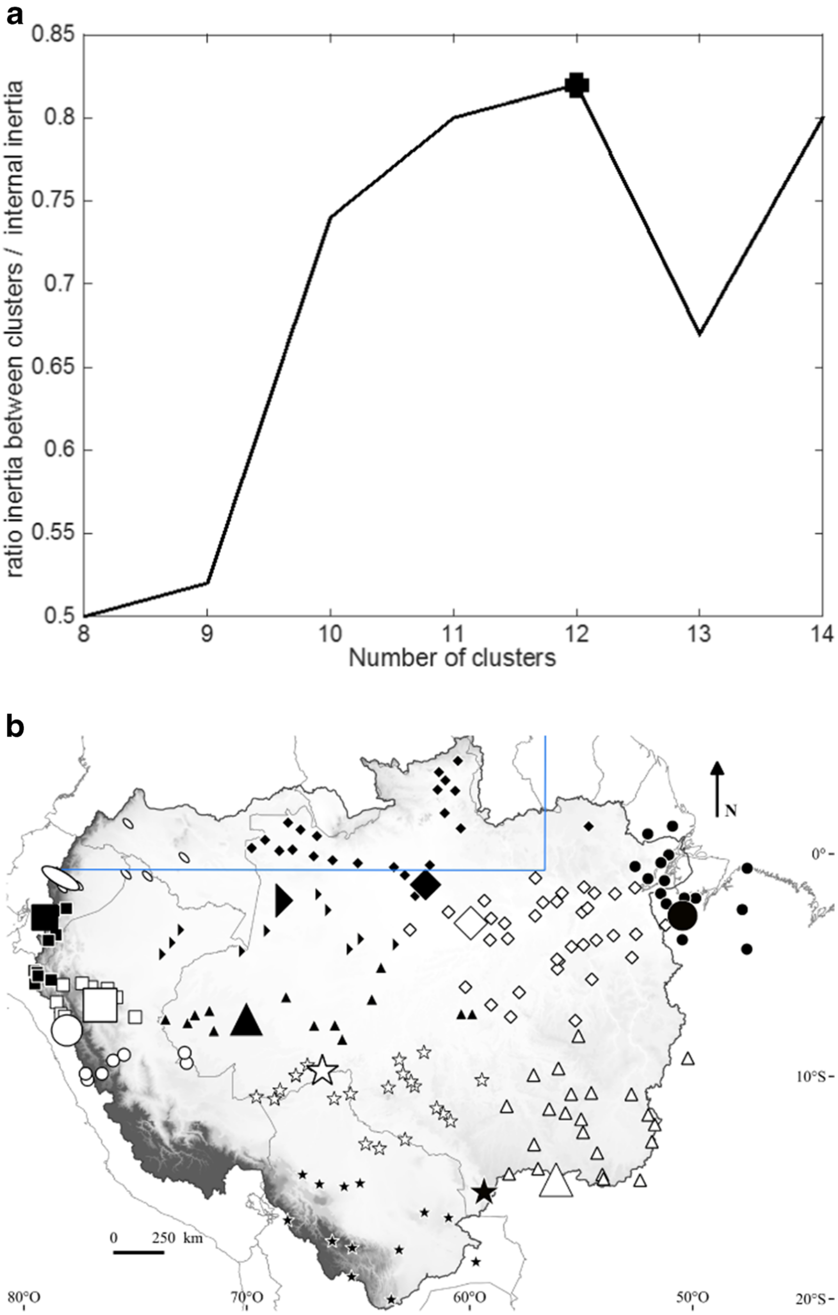


**Fig. 3** **a** Ratio between inertia among clusters and internal inertia for the estimation of the optimal number of clusters. The line represents the ratio and the cross represents the maximum value (12 clusters). **b** Identification of neighbor stations using the spectral clustering. Each symbol represents a cluster. The larger symbols represent the stations chosen as illustrations (in Section 4.3.1). The blue lines represent the northwestern region for the maximum daily precipitation (in Section 3.1)

than 20% of the days of the doubtful sequence are also recorded as 0 mm in the neighboring stations, then the doubtful dry sequence is definitely considered as real. If not, the entire period is considered as unexpected and 0 values are replaced by NA values. In practice, with a threshold of 20%, the identification of NA in the entire time series grows slowly (between 0.1 and 3.5%). With a value lower than 20%, no significant change in the results is visible. On the contrary, when its value increases (until 35%), we observed that too many dry periods were removed and that entire dry seasons could be wrongly removed. Therefore, the choice of a threshold of 20% appears rational. However, we underline that this point really depends on the area context, that is why we recommend to determine this threshold under the caution of a good knowledge of the climate of the study region by the authors, as recommended by the WMO (2007, 2011) when a method or protocol does not already exist.

## 4 Reconstruction of times series

After these two correction steps (unexpected high values and dry sequences replaced by NA), data imputation methods have been tested to re-estimate rainfall values and thus reconstruct complete daily time series. Data imputation is challenging due to the spatial and temporal variability of rainfall, especially in such a large catchment area as the Amazon basin. Several procedures have been developed in order to homogenize and fill the gaps in meteorological variables like temperatures and precipitations. Different methods can be used to reconstruct time series depending on the final objectives (Boyard-Micheau 2013). For example, many methods use the probability of intensity and rainy day distribution in the time series (Brunetti et al. 2006, 2006; Moron et al. 2007), reanalyses data (Hansen et al. 2006), or probabilistic models based on the maximum likelihood method (ML) (Dempster et al. 1977; Makhuvha et al. 1997a, 1997b). Multiple linear regressions at monthly (Camberlin et al. 2012) or daily timescale (Boyard-Micheau 2013; Eischeid et al. 2000; Vicente-Serrano et al. 2010) have also been frequently used. In general, statistical analyses show that ML, multiple imputation by predictive mean matching (PMM) best perform to fill the gaps (Glasson-Cicognani and Berchtold 2010). Although many studies on the topic do exist, Vicente et al. (2010) consider that general guidance it does not exist to choose the best method in order to fill the gaps. Choices are related to the context of the study and the appreciation of the author.

In the present study, three reconstruction methods were tested in order to determine the most efficient one to reconstruct missing rainfall values for the whole Amazon basin: the nearest neighbor approach (NN), the arithmetic mean using neighboring stations (AM), and the multiple imputation by chained equations (MICE).

### 4.1 Nearest neighbor approach

The NN approach consists in using the nearest station from the same climatic region as a predictor of missing values (Eischeid et al. 2000). The gaps in time series recorded at the station of interest were replaced by the values observed at the nearest station in the same cluster without any limit in the distance between two stations (Campozano et al. 2015; Vicente-Serrano et al. 2010). When records at the nearest station were also missing, it was necessary to consider further neighboring stations (up to the sixth nearest station).

### 4.2 Arithmetic mean using neighboring stations

The arithmetic mean (AM) consists in replacing the gaps in rainfall time series at a station of interest by the average of precipitation values measured at the neighboring stations (Fig. 4). Although this approach has been criticized because it may lead to high over- or underestimations (Glasson-Cicognani and Berchtold 2010), it is also considered to perform better for data “Missing completely at Random” (MCAR), which means that the lack of data is totally at random, which is the case of missing values in precipitation time series (Glasson-Cicognani and Berchtold 2010; Little and Rubin 2002).

Defining the optimal neighboring stations is not trivial because of the high spatial rainfall variability in the Amazon (Campozano et al. 2015; Espinoza et al. 2015) and because there are no guidelines to establish objective criteria. It thus depends on the expertise and knowledge of the author (WMO 2007, 2011). Here, we defined the neighboring stations as the four nearest stations located less than 500 km away from the station of interest and classified them in the same climatic region (i.e., cluster). We have chosen this distance based on the guidelines of the WMO (2011), which consider that the maximum spacing between rain gauge should be 500 km. Sometimes, measures recorded at the neighboring stations were also missing during the time period to be reconstructed. In that case, if less than three values were available at the neighboring stations, we replaced the missing values with other rainfall estimates provided by the Unified Gauge-Based Analysis of Global Daily Precipitation of the NOAA Climate Prediction Center (CPC). The CPC data is a gridded interpolated data of daily rainfall since 1979 to present, combining rain gauges and remote sensing data (Chen et al. 2008). As all gridded products, its major advantage is to provide spatially homogeneous information even in areas without rain gauges and its main limitation is its accuracy since rainfall is spatially smoothed and sometimes underestimated, especially in the Andes (Silva et al. 2007). However, Carvalho et al. (2012) and Getirana et al. (2011) compared several precipitation datasets in the Amazon basin and validated the good

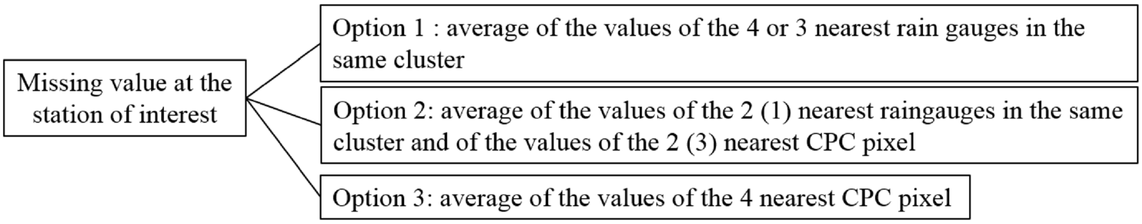

Missing value at the station of interest
Option 1 : average of the values of the 4 or 3 nearest rain gauges in the same cluster
Option 2: average of the values of the 2 (1) nearest raingauges in the same cluster and of the values of the 2 (3) nearest CPC pixel
Option 3: average of the values of the 4 nearest CPC pixel

**Fig. 4** Options for infilling gaps with the arithmetic mean

performance of CPC data. Juárez et al. (2007) also considered CPC data as the best daily rainfall dataset in the region.

### 4.3 Multiple imputation by chained equations

The multiple imputation by chained equations (MICE, van Buuren and Groothuis-Oudshoorn 2011) computes several plausible stochastic values for each missing data. These values result from regressions between the time series of interest and (1) its own values and (2) the values of the four neighboring stations. As for the AM approach, when necessary, the missing values in the neighboring stations were replaced by CPC data. In this study, we tested two procedures, i.e., bootstrap and predictive mean matching (MICE-PMM). The first one has been rejected because of the generation of negative data. The second one gives more coherent results because it does not directly impute the modeled value but a real observed one closest to the modeled one thus avoiding outliers (van Buuren and Groothuis-Oudshoorn 2011).

## 5 Validation of reconstructed time series

In order to assess the performance of the reconstruction methods, we selected a sample of 12 stations (Figs. 3b and 5) with good quality time series. The stations are spread across the 12 climatic regions (i.e., one station per cluster) in order to represent the diversity of rainfall regimes in the Amazon basin: tropical regimes with rainy and dry season in the southern, eastern, and northern regions (Fig. 5d, j, k); equatorial regime with two rainy seasons in the Andes; and regimes without a dry season in the northwest (Fig. 5a–f) (Espinoza Villar et al. 2009). These stations have no missing data during a common period that runs from the 1st August 1986 until the 31st July 1990. For each station, we then artificially created sequences of missing values. Since the gap's duration can influence the quality of the reconstruction (Cardenas and Krainski 2011), sequences of 5, 30, 60, 180, 240, and 365 successive missing values were generated at the same periods for the 12 time series. This procedure was iterated 100 times in order to get more robust results. We finally imputed new values for these gaps and compared them with the original values by computing the relative root mean square error (relative RMSE):

$$\text{relative RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(Po,i{-}Pe,i)2}{n}}/\overline{Po}$$

where *Pe* is the prevision (the imputed value), *Po* is the observation (measured value), and $\overline{Po}$ is the mean observed precipitation in the whole period. The RMSE is the square root of the ratio between the sum of the square differences in observations and the estimated number of days. It is a frequently used criterion to evaluate the performance of a predictor. The closer the relative RMSE is to zero, the better is the reconstructed value.

## 6 Results of validation of the reconstruction

### 6.1 Comparison of the different methods

Figure 6a shows the distribution of 600 values (6 different durations of missing sequences × 100 iterations) of relative RMSE computed for each sampled station for each reconstruction method. The AM method showed the best results for ten stations out of 12. For these stations, the median relative RMSE is notably lower (then better) with the AM than with the two other methods. The method works especially well for tropical stations in general (in particular the Envira station) where we observed the greatest difference between the AM and the two other methods (Fig. 6g–l), and for two equatorial stations with high precipitations levels and without a dry season (Fig. 6e, f).

In stations located in the western part of the basin (i.e., the Andes), the results are more balanced between the three approaches. The AM method led to the lowest results at Rumipampa Salcedo station and MICE performed best at San Pablo station. Sayausi and Huangacocha stations showed better reconstruction results with the AM approach but results from the three methods are almost identical. Except for Huangacocha station, these Andean stations present weak annual rainfall without a dry season (Fig. 6a–c). In conclusion, the performance of the methods depends on the rainfall regime of the station. The AM method provided lower reconstruction results in climatic regions characterized by low annual precipitations without a dry season (Fig. 5b) and better results in

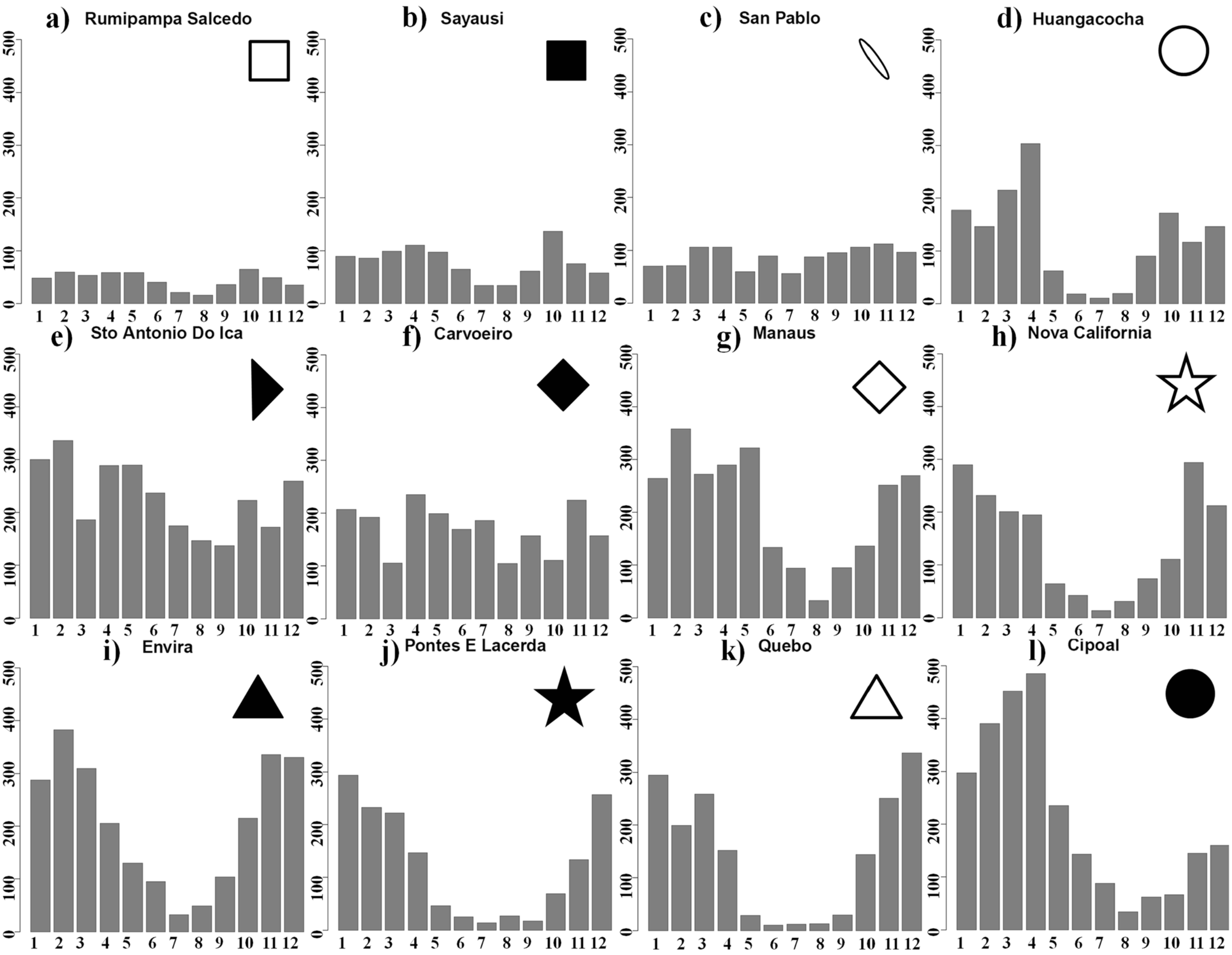


**Fig. 5** **a**–**l** Precipitation regime for each rain gauge of the sample. The mean of each month (axis *x*) is computed from August 1986 to July 1990. The symbols associated with each graphic represent the cluster to which belong

regions marked by a unimodal rainfall regime associated with a dry season.

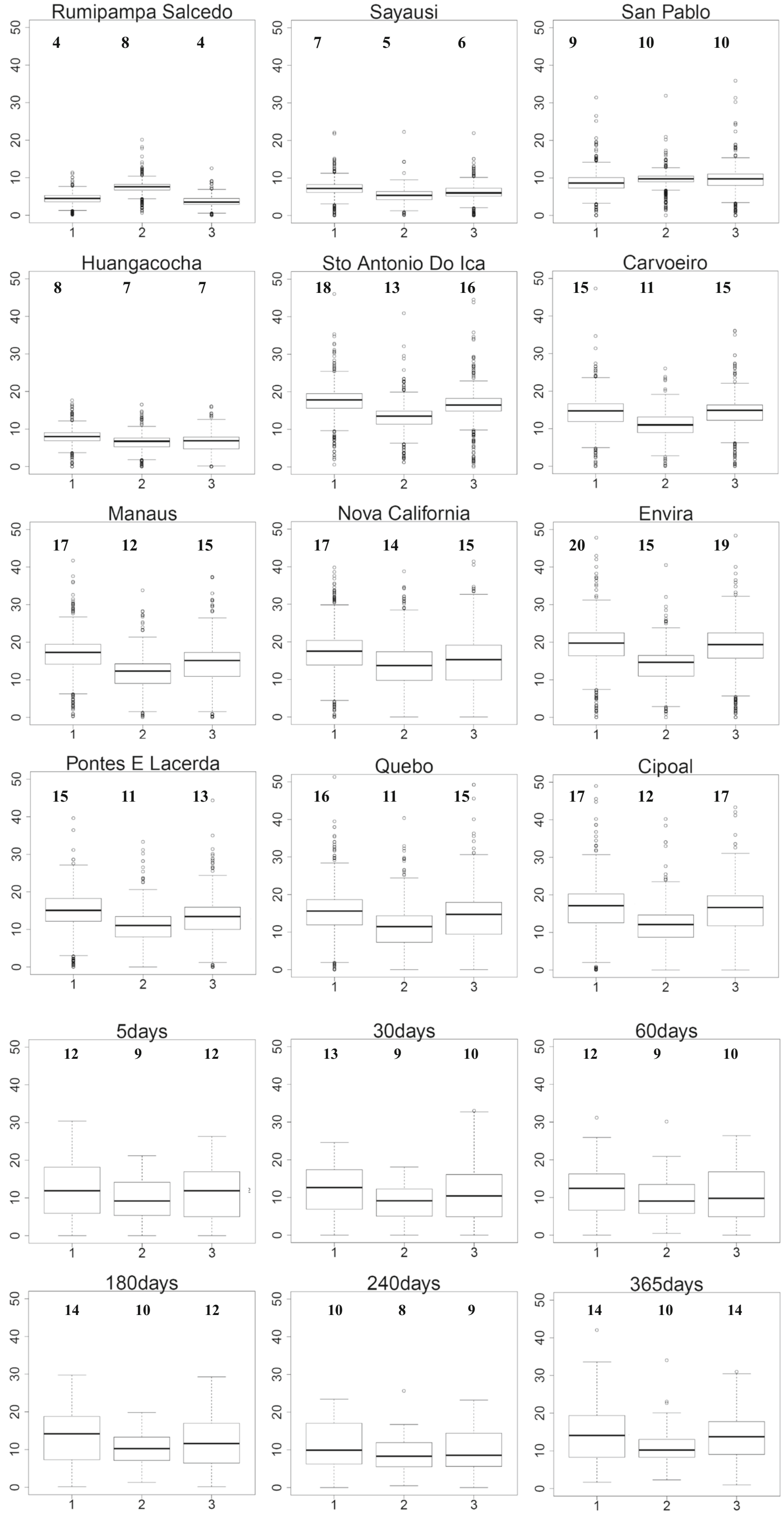
Figure 6b represents for each sequence and the three reconstruction methods, the distribution of the 1200 values of relative RMSE (12 stations × 100 iterations). The median shows that the AM method also provides the lower relative RMSE while MICE-PMM still provides the poorest results (Fig. 6b). Moreover, the sequence lengths do not impact the estimated values, whatever the method.

A time evaluation of the three methods was also computed, with the aim to assess if the results are seasonally dependent. However, the seasonality is not uniform across the AB; the equatorial regions (Figs. 5e, f and 3b) have rainfall throughout the year, while in the tropical regions (Figs. 5d, g–l and 3b), a rainy season alternates with a dry season, which can be opposite between the north and the south. Then, the assessment was done on classes of the quantity of monthly rainfall. This approach makes it possible to analyze the monthly amount of precipitation that each method allows to better reconstruct.

For the 12 stations of the sample, monthly rainfall was computed from the 1st August 1986 until the 31st July 1990. Based on the distribution of the 576 months of these time series (Fig. 7), each month was attributed to one of the six following classes: 0–50 mm, 50–100 mm, 100–200 mm, 200–300 mm, 300–450 mm, and 450–+ mm.

Subsequently, the reconstruction methods of AM, NN, and MICE were applied to 5-day sequences of missing values, because the monthly time scale avoids using the other sequence lengths tested before. Figure 8 shows that for each class of monthly precipitations, the medians are close, with a good score (0 mm) in class (0, 50). The reconstruction by the three methods is pretty better for the driest months than for the rainiest (200, + 450) which maximum and range of values are higher. However, there are more outliers for the driest than for the rainiest months. Even if there is a slight advantage of the AM to reconstruct the rainiest months, these results show that the performance of the reconstruction depends on monthly rainfall amount and not on the method used.

**Fig. 6** Root mean square error for each reconstruction method per station or per sequence length of missing value. In each graphic, the *x*-axis represents the three methods of reconstruction 1: MICE, 2: arithmetic mean, 3: nearest neighbor. **a** Represents the relative RMSE for each station of the sample. **b** Represents the relative RMSE for each length of gap. The number above each boxplot indicate the median



To conclude, even if the station regime can influence the performance of the method, the arithmetic mean is frequently the best method to reconstruct missing values, regardless of the length of the gap. The arithmetic mean performs notably better in the plain of the AB while the differences between the three methods are

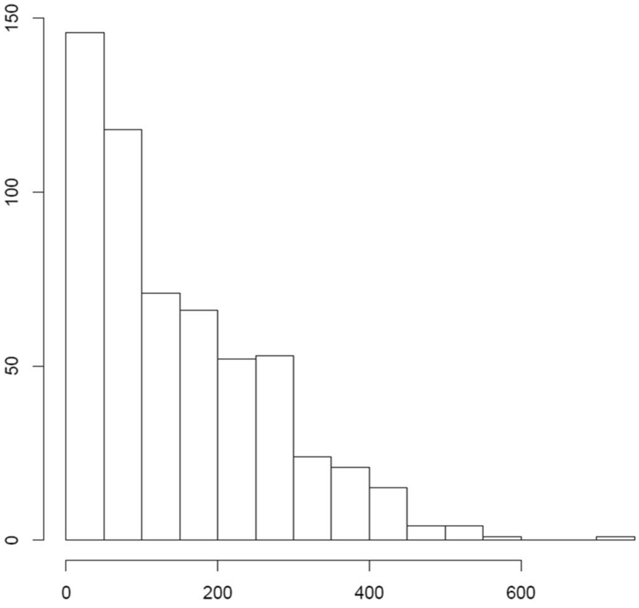


**Fig. 7** Distribution of the monthly rainfall of the 576 months of the sample time series. The *x*-axis represents the precipitations in mm, and the *y*-axis, the number of month

less significant in the Andean regions. Consequently, as a method is researched for the whole AB, the arithmetic means seems to be the most appropriate among the three tested methods.

## 6.2 Assessment of the quality of the reconstruction of time series by the arithmetic mean

Figure 9a–d compares the mean annual precipitation and the mean annual numbers of rainy days after and before the data reconstruction.

In Fig. 9a that shows the number of rainy days per year after the reconstruction, the spatial pattern is consistent with expected (Espinoza Villar et al. 2009; Simões Reibota et al. 2010). There are numerous rainy days near the equator, in the northwest and in the Andes, where precipitations occur all year long (see for example stations in Fig. 6a–f, l) and southward, the number of rainy days decreases (see for example stations in Fig. 6h–k). The difference in the number of rainy days per year after and before reconstruction (Fig. 7b) is consistent with the spatial distribution of missing values per year (Fig. 10) that are more important in the southeast, in the center of the AB, and in the extreme north of the Andean regions. This means that the AM method allows reconstructing time series with unimodal regimes. Indeed, all missing values are not necessarily substituted by a value higher than 0. Then, this method can provide regional conditions in a given moment.

Figure 9c shows the mean annual precipitation in each station after the reconstruction of the time series. As for

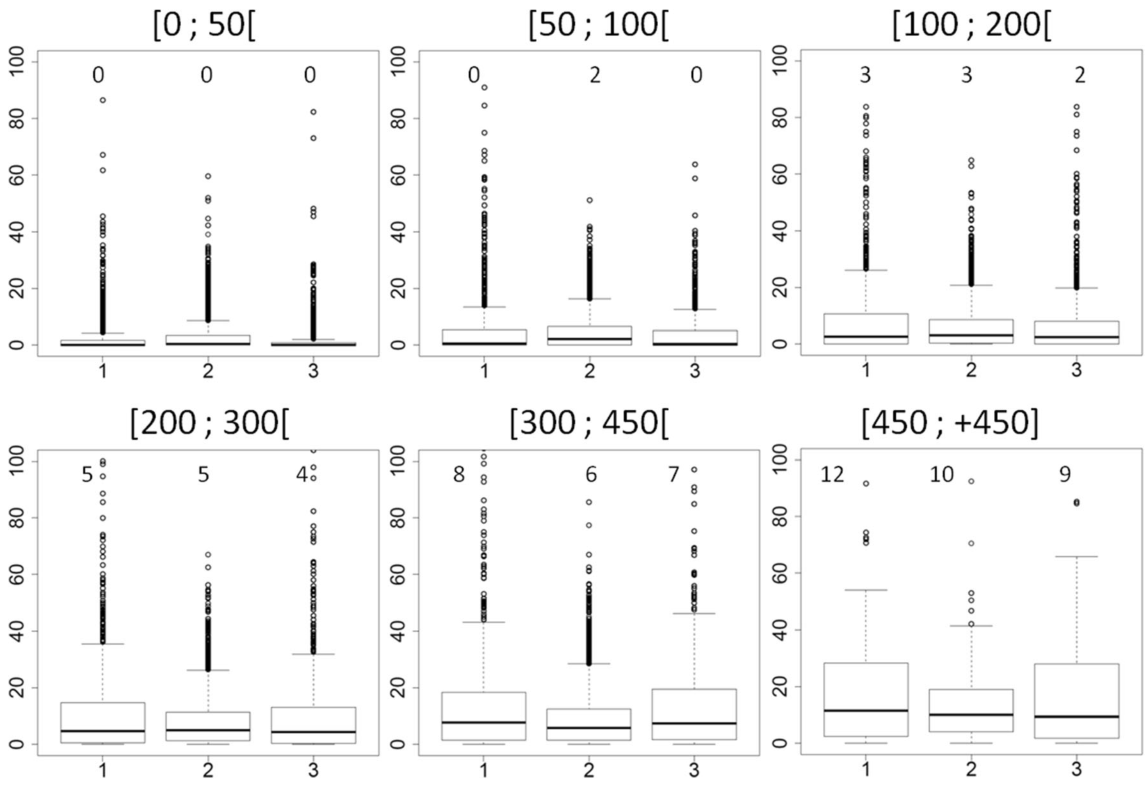


**Fig. 8** Root mean square error for each reconstruction method per class of monthly rainfall. In each graphic, the *x*-axis represents the three methods of reconstruction 1: MICE, 2: arithmetic mean, 3: nearest neighbor. The number above each boxplot indicate the median

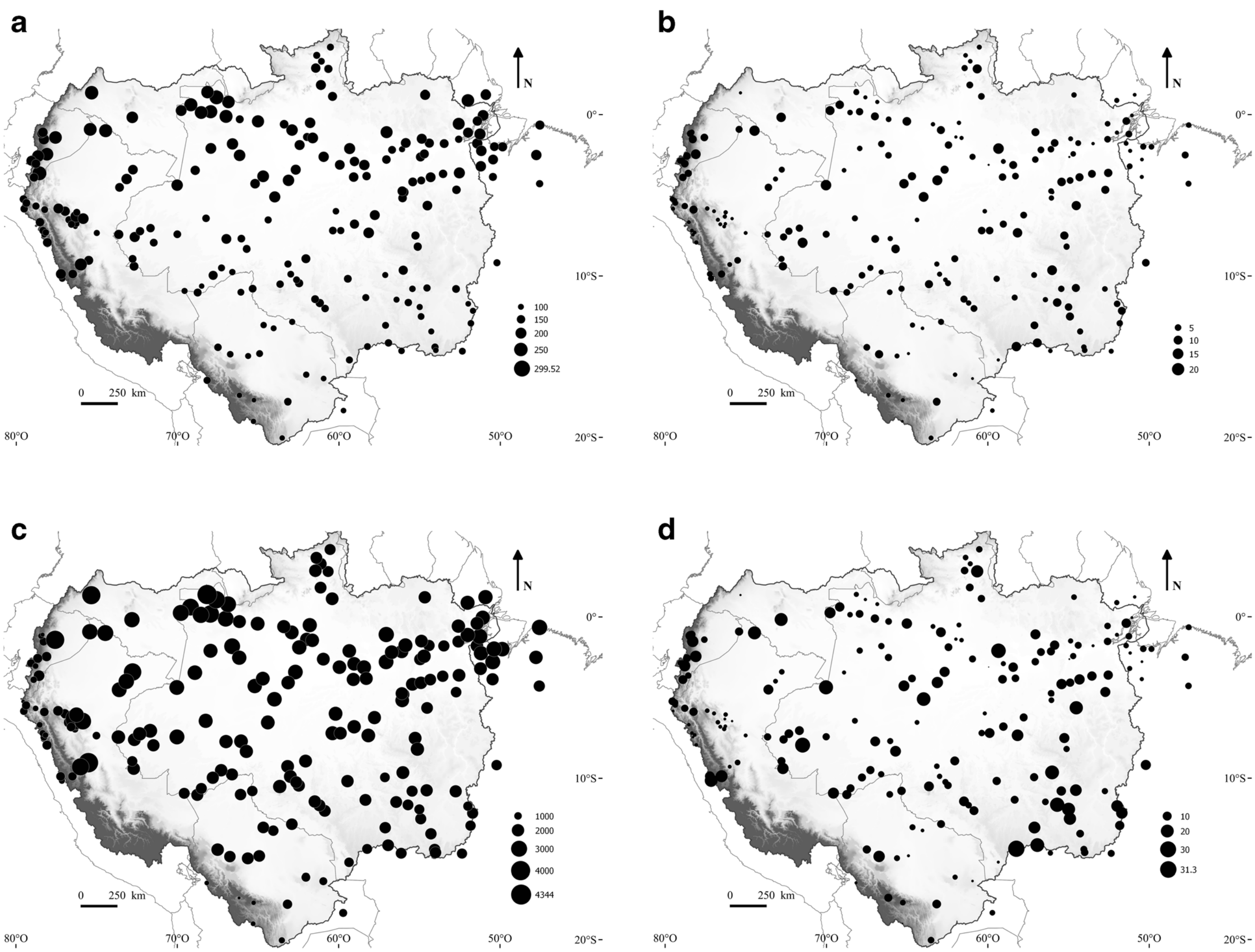


**Fig. 9** **a** Number of rainy days after reconstruction by the arithmetic mean for each station of the dataset (100 to 300). **b** Difference between the mean of rainy days per year after and before the reconstruction, in percentage (5 to 20). **c** Mean annual precipitation after reconstruction by the arithmetic mean, in millimeter (1000 to 4344 mm). **d** Difference between the mean annual rainfall after and before the reconstruction, in percentage (10 to 31.1%)

the number of rainy day per year, the spatial pattern of the quantity of precipitation is consistent with the literature (Espinoza Villar et al. 2009; Figueroa and Nobre 1990; Liebmann and Marengo 2001; Simões Reibota et al. 2010) with higher rainfall near the equator line and in the northwest of the AB (for example stations in Fig. 6d–f, l), lower rainfall southward (for example stations in Fig. 6h–k) and northward toward the tropics, and the lowest in the Andean stations (for example stations in Fig. 6a–d). The increase of the annual average after the reconstruction (Fig. 9d, in percentage) is spatially heterogeneous, which is also the case of the percentage of missing value per year (Fig. 10) before the reconstruction. However, unlike the number of rainy days per year, in the southeast, rainfall increases intensely in a group of stations. This can mean that the increase in rainfall is concentrated in a short period.

## 7 Discussion and conclusion

Rain gauges are an important source of observed data for the Amazon basin since they provide long time series and give the possibility to better understand the variability of climate. However, the rain gauge network in the Amazon basin is very heterogeneous and mainly characterized by a poor density and numerous erroneous measurements, partly due to the accessibility in this territory. On the other side, the series of rainfall satellite estimations data are still too short and sometimes uncertain to completely replace observations. It is therefore of prime importance to construct a sound database from rain gauges, using advanced quality control and reconstruction methods valid for the entire Amazon basin.

Initially, 533 rain gauges have been gathered from 1981 to 2013. Among these stations, we have retained those with less than 20% of missing values and without an inhomogeneous

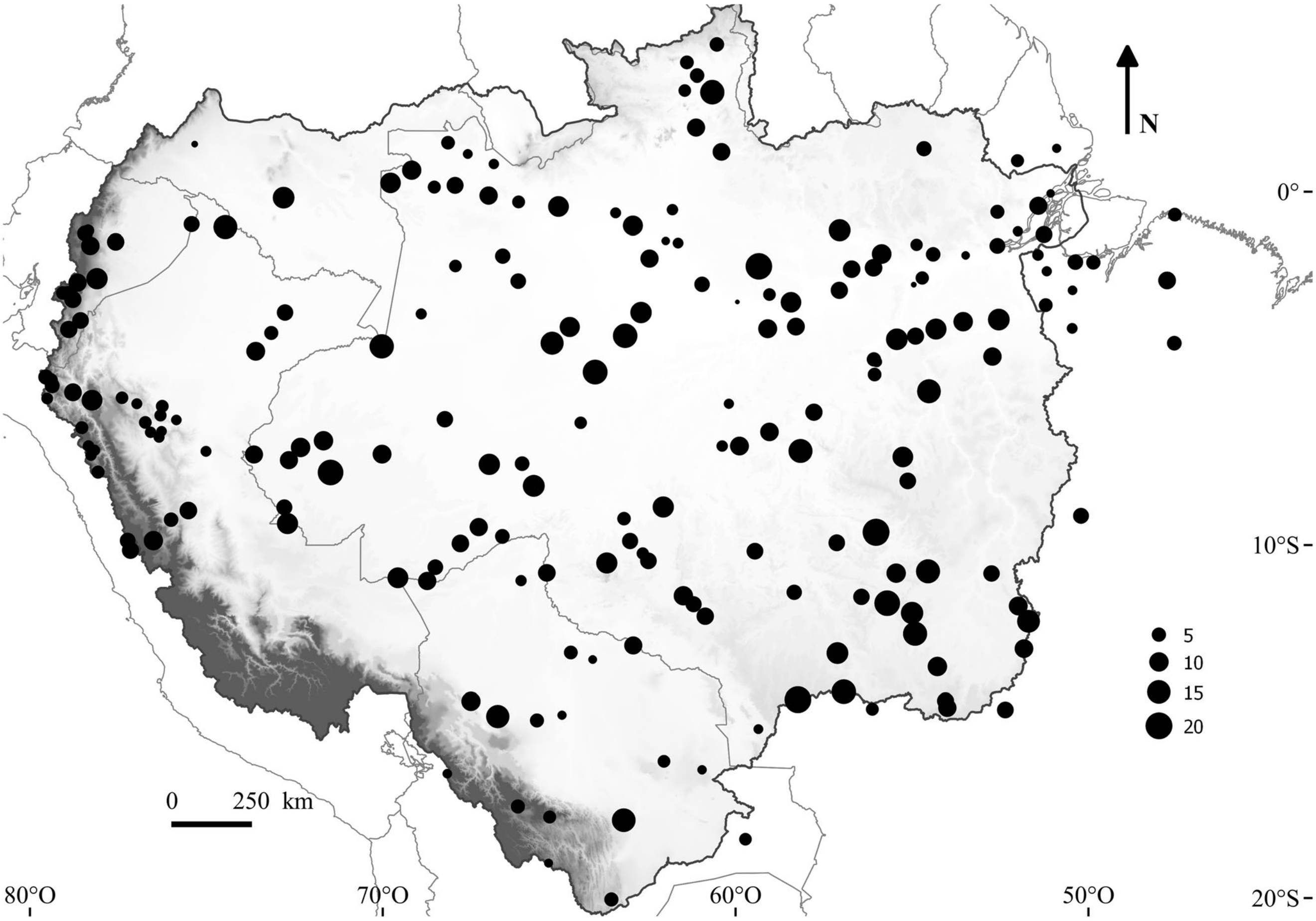


**Fig. 10** Mean annual percentage (5 to 20%) of missing values per rain gauge

structure. Finally, only 205 rain gauges were selected. Afterward a control of quality allowed to remove the unexpected large values (34) and the wrong zero ones.

Among the methods for filling the gaps that would be valid for the whole basin, multiple imputation by chained equations with the predictive mean matching procedure (MICE-PMM), the nearest neighbor (NN) approach, and the arithmetic mean (AM) have been tested. The latter performs better, although the three methods have experimented rather similar results in the Andean regions. The AM was used to reconstruct the dataset and associated precipitation parameters (the daily mean rainfall and the number of rainy days) were used to assess the quality of the dataset reconstruction. We can conclude that for this database, the AM allows obtaining acceptable values to reconstruct long time series and to produce a useful dataset of daily precipitation for the whole AB. Of course, even if the AM gives better results than the other two methods, the relative validation of the RMSE shows that errors remain. However, this study shows that this method can help to improve results better than a simple method such as NN and works better at the AB level than a more sophisticated method such as MICE. In addition, as there is not yet a consensus on how to reconstruct the observed daily precipitation data, this work aims to contribute to the development of a methodology.

The efficiency of the methods tested in our work appears to really be related to the region of study, the associated rainfall regimes, and the density of rain gauges. For example, Eischeid et al. (2000) show that the reconstruction of precipitation in the USA depends on the location and the precipitation regime and that the quality of infilling can vary with the seasonality. In the Amazon basin, the NN method performs less well than the AM, while Vicente-Serrano et al. (2010) were able to have a better quality precipitation reconstruction with the NN than with the linear regression in northeast Spain; this was due to the fact that they have at their disposal 286 rain gauges with a high density and radius neighboring less than 15 km. Conversely, Campozano et al. (2015) filled the gaps for 14 precipitation time series in Ecuador and showed that complex methods based on linear regression perform better than the AM and the nearest neighbor approaches. It can be noted in this latest work that AM is more appropriate during months with few precipitations that seems consistent with the observations of the present work which shows that the more efficient method for rain gauges with dry season is the arithmetic mean. Cardenas and Krainski (2011) also tested several methods of reconstruction data for 41 precipitation time series in Brazil, in the State of Parana. The imputations results, estimated among others from MICE-PMM and for several lengths of missing values, were among the worst with this method, but tended to be slightly better when the gap is long.

This work pointed out the great challenge to obtain a useful and robust rainy database in the AB. Soon, the rainfall satellite data will be long enough and then will be able to provide an alternative to ground-based rainfall data. This remote sensing technique can offer a wide geographical coverage and a good resolution. But, as satellite data are rainfall estimations, it is necessary to assess their quality, which is done by ground validation. Thus, a good network rain gauge will remain crucial.

**Acknowledgments** The authors would like to express their special thanks to Naurinete J. C. Barreto and George Ulguim Pedra for their help and discussions from the Meteorology Department of the Brazilian National Institute for the Space Research (INPE) and Cláudio Moisés Santos E. Silva from the Centro de Ciências Exatas e da Terra (CCET) in Natal University (Brazil).

## References

Aguilar E, Peterson TC, Obando PR, Frutos R, Retana JA, Solera M, Soley J, García IG, Araujo RM, Santos AR et al (2005) Changes in precipitation and temperature extremes in Central America and northern South America, 1961–2003. J Geophys Res 110(D23):107

Barbosa Santos E, Sérgio Lucio P, Silva CM (2015) Precipitation regionalization of the Brazilian Amazon. Atmos Sci Lett 16:185–192

Boyard-Micheau J (2013) Prévisibilité potentielle des variables climatiques à impact agricole en Afrique de l'est et application au sorgho dans la région du mt kenya. Thèse de doctorat. Université de Bourgogne, France

Brito, A.L., Paix, J.A., Yoshida, M.C.,et al (2014). Extreme rainfall events over the Amazon basin produce significant quantities of rain relative to the rainfall climatology. Atmos Climate Sci, 4: 179–191

Brunetti, M., Maugeri, M., and Nanni, T. (2006). Trends of the daily intensity of precipitation in Italy and teleconnections. Il Nuovo Cimento C 105

Camberlin P, Boyard-Micheau J, Philippon N, Baron C, Leclerc C, Mwongera C (2012) Climatic gradients along the windward slopes of Mount Kenya and their implication for crop risks. Part 1: climate variability. Int J Climatol 34:2136–2152

Campozano L, Sánchez E, Aviles A, Samaniego E (2015) Evaluation of infilling methods for time series of daily precipitation and temperature: the case of the Ecuadorian Andes. Maskana 5:99–115

Camps-Valls G, Bruzzone L (2009) Kernel methods for remote sensing data analysis. John Wiley & Sons, United Kingdom

Cardenas, R., and Krainski, E.T. (2011). Preenchimentos de falhas em bancos de dados meteorologicos diarios: comparação de abordagens. XVII Congresso Brasileiro de Agrometeorologia, Guarapari-Brasil

Carvalho LM, Jones C, Posadas AN, Quiroz R, Bookhagen B, Liebmann B (2012) Precipitation characteristics of the South American monsoon system derived from multiple datasets. J Clim 25:4600–4620

Caussinus H, Mestre O (2004) Detection and correction of artificial shifts in climate series. J R Stat Soc: Ser C: Appl Stat 53:405–425

Chen J, Del Genio AD, Carlson BE, Bosilovich MG (2008) The spatiotemporal structure of twentieth-century climate variations in observations and reanalyses. Part II: Pacific pan-decadal variability. J Clim 21:2634–2650

Cressie N, Chan NH (1989) Spatial modeling of regional variables. J Am Stat Assoc 84:393–401

Delahaye F, Kirstetter P-E, Dubreuil V, Machado LA, Vila DA, Clark R (2015) A consistent gauge database for daily rainfall analysis over the Legal Brazilian Amazon. J Hydrol 527:292–304

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc 39:1–38

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 Meteorol 39: 1580–1591

Espinoza Villar JC, Ronchail J, Guyot JL, Cochonneau G, Naziano F, Lavado W, De Oliveira E, Pombosa R, Vauchel P (2009) Spatio-temporal rainfall variability in the Amazon basin countries (Brazil, Peru, Bolivia, Colombia, and Ecuador). Int J Climatol 29:1574–1594

Espinoza JC, Chavez S, Ronchail J, Junquas C, Takahashi K, Lavado W (2015) Rainfall hotspots over the southern tropical Andes: spatial distribution, rainfall intensity, and relations with large-scale atmospheric circulation. Water Resour Res 51:3459–3475

Figueroa SN, Nobre CA (1990) Precipitation distribution over central and western tropical South America. Climanalise 5:36–45

Getirana AC, Espinoza JCV, Ronchail J, Rotunno Filho OC (2011) Assessment of different precipitation datasets and their impacts on the water balance of the Negro River basin. J Hydrol 404:304–322

Glasson-Cicognani, M., and Berchtold, A. (2010). Imputation des données manquantes: Comparaison de différentes approches. In 42èmes Journées de Statistique, Marseille-France

Hansen JW, Challinor A, Ines A, Wheeler T, Moron V (2006) Translating climate forecasts into agricultural terms: advances and challenges. Clim Res 33:27–41

Juárez RIN, Hodnett MG, Fu R, Goulden ML, von Randow C (2007) Control of dry season evapotranspiration over the Amazonian Forest as inferred from observations at a southern Amazon Forest site. J Clim 20:2827–2839

Liebmann B, Allured D (2005) Daily precipitation grids for South America. Bull Am Meteorol Soc 86:1567–1570

Liebmann B, Marengo J (2001) Interannual variability of the rainy season and rainfall in the Brazilian Amazon basin. J Clim 14:4308–4318

Little RJA, Rubin DB (2002) Statistical analysis with missing data. John Wiley & Sons, Inc, USA

Makhuvha T, Pegram G, Sparks R, Zucchini W (1997a) Patching rainfall data using regression methods: 1. Best subset selection, EM and pseudo-EM methods: theory. J Hydrol 198:289–307

Makhuvha T, Pegram G, Sparks R, Zucchini W (1997b) Patching rainfall data using regression methods. 2. Comparisons of accuracy, bias and efficiency. J Hydrol 198:308–318

Mestre O, Gruber C, Prieur C, Caussinus H, Jourdain S (2011) SPLIDHOM: a method for homogenization of daily temperature observations. J Appl Meteorol Climatol 50:2343–2358

Moron V, Robertson AW, Ward MN, Camberlin P (2007) Spatial coherence of tropical rainfall at the regional scale. J Clim 20:5244–5263

Ronchail J, Cochonneau G, Molinier M, Guyot J-L, De Miranda Chaves AG, Guimarães V, de Oliveira E (2002) Interannual rainfall variability in the Amazon basin and sea-surface temperatures in the equatorial Pacific and the tropical Atlantic Oceans. Int J Climatol 22:1663–1686

Silva V, Kousky V, Shi W, Higgins RW (2007) An improved gridded historical daily precipitation analysis for Brazil. J Hydrometeorol 8: 847–861

Simões Reibota M, Gan MA, Porfirio da Rocha R, Ambrizzi T (2010) Regimes de precipitacao na America do sul. Rev Bras Meteorol 25: 185–204

van Buuren S, Groothuis-Oudshoorn K (2011) MICE: multivariate imputation by chained equations in R. J Stat Softw 45:1–68

Vicente-Serrano SM, Beguería S, López-Moreno JI, García-Vera MA, Stepanek P (2010) A complete daily precipitation database for Northeast Spain: reconstruction, quality control, and homogeneity. Int J Climatol 30:1146–1163

Williams E, Dall' Antonia A, Dall' Antonia V, de Almeida JM, Suarez F, Liebmann B, Malhado ACM (2005) The drought of the century in the Amazon basin: an analysis of the regional variation of rainfall in South America in 1926. Acta Amazon 35:231–238

WMO (1989). Calculation of monthly and annual 30 year standard normal (World Meteorological Organization)

WMO (2007). Guide to the global observing system (World Meteorological Organization)

WMO (2011). Guide des pratiques climatologiques (World Meteorological Organization)

The artwork was created with R-cran, QGis, and Matlab software.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.