



A quality control procedure for long-term series of daily precipitation data in a semiarid environment

Javier Estévez¹ · Alba Llabrés-Brustenga² · M. Carmen Casas-Castillo³ · Amanda Penélope García-Marín¹ · Ricard Kirchner³ · Raül Rodríguez-Solà⁴

Received: 30 July 2021 / Accepted: 12 May 2022
© The Author(s) 2022

Abstract

The availability of quality precipitation records in the current climate situation is of great importance in the scientific-technical field but also for the public institutions that manage the meteorological networks. This work has implemented a comprehensive spatial quality control procedure in the semiarid region of Andalusia (Southern Spain), using precipitation time series from 1947 stations from three meteorological networks: Spanish Meteorological Agency (AEMET), Agroclimatic Information Network of Andalusia (RIA), and Phytosanitary Information Alert Network (RAIF). The method consists of three consecutive steps: basic, absolute, and relative quality control processes. The latter step compares data from neighboring stations taking into account their proximity, height difference, and correlation, leading to a complete evaluation of each daily value. Finally, the quality of each year at each station can be declared as acceptable, good, or excellent. The automatic weather station networks RIA and RAIF gave absolute quality index Q above 85% for almost 87% of their stations, while only 57% of AEMET network reached this percentage. However, one of the longest AEMET datasets, San Fernando-Cádiz, obtained, except for 1 year, Q values over 90% in all available years for more than a century of measurements, since 1870 until 2000. From a total of more than 15 million daily records, almost 82% was flagged as correct. Despite the limitations of Andalusia region (low density of stations and its structural water deficit), the complete quality control procedure has been satisfactorily applied. Finally, related to the number of outliers, no temporal trend was found across the region.

Keywords Quality control · Validation · Precipitation · Long-term series · Climate change

1 Introduction

Quality assurance of precipitation data is crucial for the reliability of applications and models that use them as input variables, as well as for an adequate representativeness of the statistics derived from the raw time series. Firstly, an

appropriate maintenance of the weather station is necessary, checking records at different timesteps in the field and periodically calibrating the sensors if necessary (Estévez et al. 2011). Then, the application of quality control test is required in order to validate meteorological data, identifying errors, doubtful, or suspect records, among other kind

✉ M. Carmen Casas-Castillo
m.carmen.casas@upc.edu

Javier Estévez
jestevez@uco.es

Alba Llabrés-Brustenga
alba.llabres@bsc.es

Amanda Penélope García-Marín
amanda.garcia@uco.es

Ricard Kirchner
ricard.kirchner@upc.edu

Raül Rodríguez-Solà
raul.rodriguez@upc.edu

¹ Engineering Projects Area, Universidad de Córdoba, Campus Rabanales, Nacional IV Km 396, 14071 Córdoba, Spain

² Barcelona Supercomputing Center - Centro Nacional de Supercomputación, Jordi Girona 29-31, 08034 Barcelona, Spain

³ Department of Physics, ESEIAAT, Universitat Politècnica de Catalunya-Barcelona Tech, Colom 1, 08222 Terrassa, Spain

⁴ Department of Physics, ETSAB, Universitat Politècnica de Catalunya-Barcelona Tech, Diagonal 647, 08028 Barcelona, Spain

of labels flagging each value. Although the World Meteorological Organization (WMO) recommends following the guidelines reported in WMO (2008), many weather station networks belonging to public organizations do not yet have reliable systems that guarantee the availability of quality precipitation long-term series (Morbidelli et al. 2020). Therefore, scientists are often the ones who implement validation procedures and apply them to precipitation data as a prerequisite for their use (Parchure and Gedam 2019; Medina-Cobo et al. 2017; García-Marín et al. 2015; Estévez et al. 2020; Yang et al. 2020; Lana et al. 2020). Most of these quality control methods are based on algorithms, rules, tests, or procedures reported in different works. For example, Feng et al. (2004) applied different checks (range, temporal, internal, and spatial consistency) in China; Estévez et al. (2015) proposed a new method to detect spurious precipitation signals using the transmissivity atmospheric coefficient in irrigated areas; You et al. (2007) developed a test based on the multiple intervals' gamma distribution assuming that normal distribution is not efficient for identifying precipitation outliers; Sciuto et al. (2009) used the comparison of rainfall values on target station to reference estimations using an artificial neural networks model; Durre et al. (2010) presented a comprehensive set of fully automated quality assurance procedures for observations of daily surface precipitation, among other meteorological magnitudes, to detect duplicate data, climatological outliers, and internal, temporal, and spatial inconsistencies; and Serrano-Notivoli et al. (2017) developed a validation method based on probabilistic predictions (occurrence of a wet or dry day and the amount of precipitation) from nearby stations. Moreover, the detection of erroneous precipitation data due to problems in tipping-bucket rain gauge measures has been studied (Michaelides 2008; Habib et al. 2001; Einfalt et al. 2006).

Basically, in order to detect potential precipitation erroneous data, there are two main categories of validation methods: those designed to use data from a single site and those based on the comparison of candidate station data to neighboring station data. The procedures based on statistical decisions have a great efficiency if long-term precipitation series are available (Hubbard et al. 2005; Llabrés-Brustenga et al. 2019; Vicente-Serrano et al. 2010) and an adequate density of locations if spatial methods are involved. In this work, and due to the great importance (forecasting, trends and non-linear analyses, etc.) of having long-term validated datasets, a complete spatial quality control procedure for daily precipitation data validation has been applied in the semiarid region of Andalusia (Southern Spain).

The main goal of this work is the validation of daily precipitation records from different networks in an extensive region (one of the largest regions in the European Union) by applying a spatial quality control procedure. The whole process includes three consecutive steps: basic, absolute, and

relative quality control methods. It is important to note that no similar works have been carried out in a heterogeneous region like Andalusia (mainly semiarid but also with arid and dry sub-humid areas) where different climates coexist. This region has important singularities such as water being a limiting factor or its low density of precipitation stations that make it critical to verify the applicability of this type of quality control mechanisms.

2 Data and methods

Daily precipitation data from 1947 stations of three meteorological networks operating in Andalusia (AEMET, RIA, and RAIF) were available. The AEMET network is the oldest one and it is managed by the Spanish Agency of Meteorology, while RAIF and RIA correspond to Agroclimatic Information Network and Phytosanitary Information Alert Network, respectively. Andalusia is a region of almost 90,000 km² with a high interannual variability in rainfall and located in the south of the Iberian Peninsula (see Fig. 1). The dataset had an average number of 8145 daily data per station (with a minimum of 6 days and a maximum of 45,798 days) and a total of 15,857,892 days/station data.

Figure 2 shows the temporal evolution of the number of available series of the Andalusian network from 1870 to the present. The first available data was taken in 1870 at the San Fernando station (Cádiz), the only operative station of the network for more than 30 years. At the beginning of the twentieth century some other few stations started to work: Jaén-Instituto, which began to operate in 1901; Granada (Cartuja) and Granada-Universidad, both started in 1902; and Huelva and Sanlúcar de Barrameda, in 1903. From 1910 the network began to grow, keeping the number of stations at a few tens until the interruption caused by the Spanish civil war (1936–1939). After war, the growth rate increased considerably and uninterruptedly until the 1970s. Around 2000, the RIA and RAIF networks of automatic weather stations started up. After 2010, a progressive dismantling of some of the manual stations left the network at the present in just over half of the operational stations it had in 2000.

The quality control procedure applied is based on the methodology developed by Llabrés-Brustenga et al. (2019) using rainfall datasets in Catalonia (NE Spain), and consists of three stages which can be followed in the scheme of Fig. 3. The main novelty in the present study is to check the behavior of this procedure to verify its extensibility in different geographical areas and with different meteorological characteristics. The objective of this procedure is to select series with the best quality while detecting possible errors, starting with the most serious ones and gradually increasing the level of detail. First stage is a basic quality control that consists in the detection and

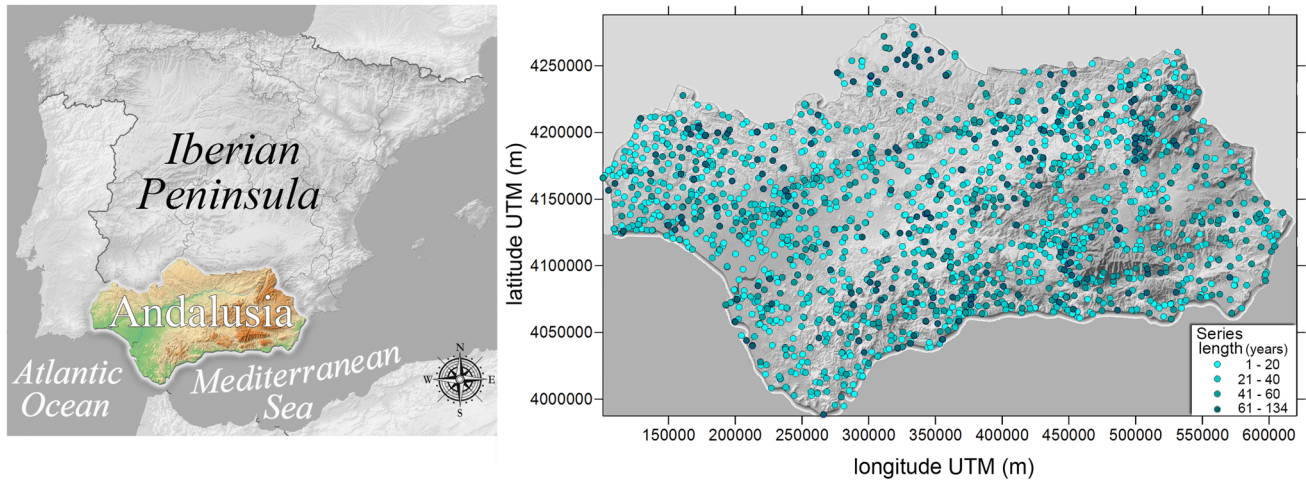


Fig. 1 Meteorological stations of Andalusia (Spain)

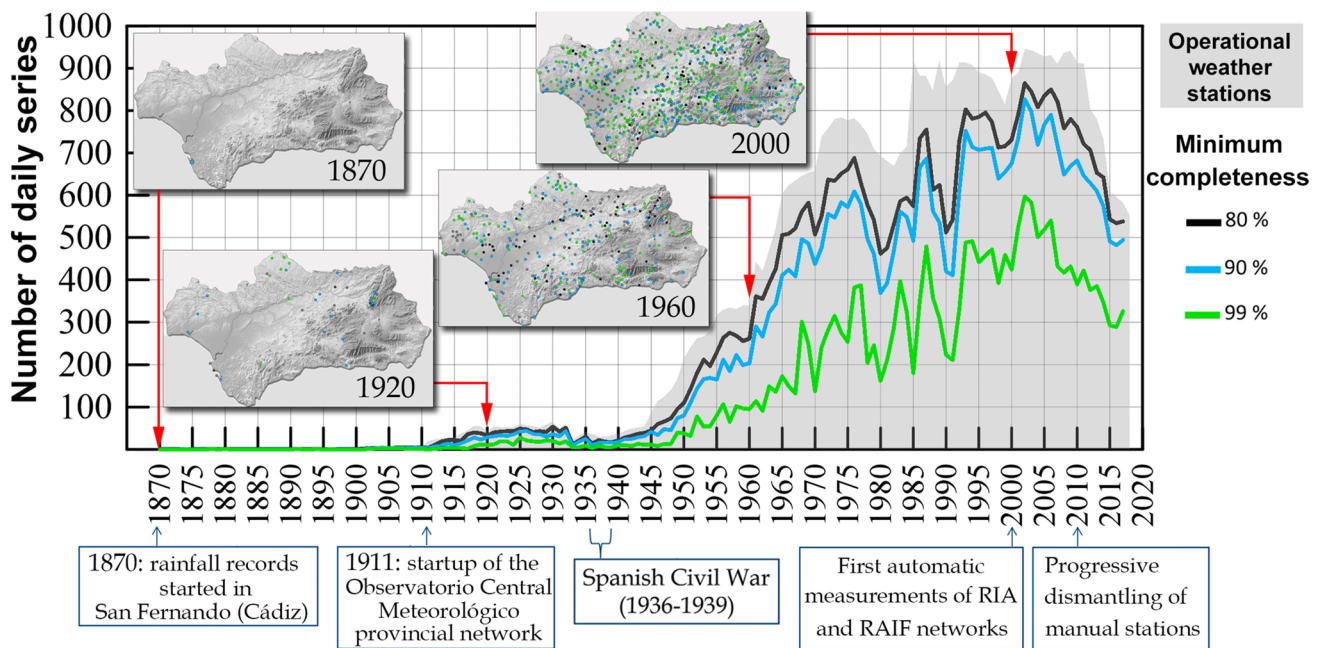
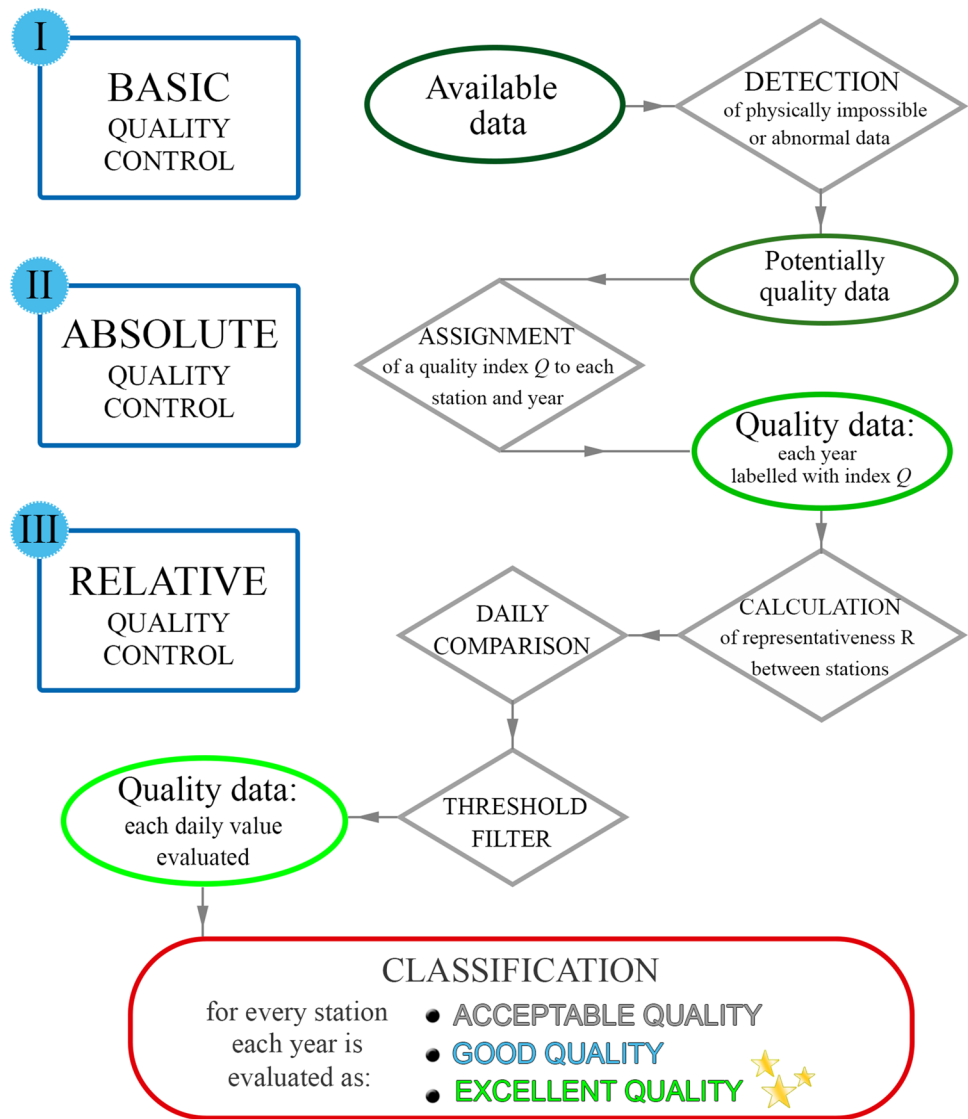


Fig. 2 Temporal evolution of the number of meteorological stations in Andalusia from 1870 to the present. Lines represent the number of series with a minimum annual percentage of available data

removal of certain erroneous values. That is the case of physically impossible rainfall amounts as negative ones or values too high for the climatology of the studied area. An upper threshold of 600 mm in 24 h was considered; this is the highest daily rainfall amount ever recorded in Andalusia, collected on October 19, 1973 in Zurgena (a locality of Almería, the easternmost province of Andalusia) during the extraordinary precipitation events that occurred in the southeast of the Iberian Peninsula (Capel Molina 1974), which gave rise to river and Mediterranean river bed overflows and floods, causing the disappearance

of crops, household goods, and the loss of human lives. In any case, all values above 300 mm were manually checked and verified. A filter to detect identical values repeated for consecutive days was also used, to further manual checking. Due to the common old procedure to write down equally distributed values for each day from a cumulative measure done for several days, incorrect daily values can be found. Series with this kind of errors could be used to calculate monthly totals, but these incorrect amounts should be removed from the daily dataset.

Fig. 3 Scheme of the quality control procedure



2.1 Absolute quality control

After the basic quality control, an absolute quality control has to be performed. At this stage every single series is evaluated individually and a quality index Q is calculated for every year at every station, taking into account the percentage of annual data, the distribution of gaps in the sample, the heterogeneity in the records of rainfall by day of the week, and the abundance of outliers.

Bearing in mind the most common problems in daily rainfall series, a quality index Q is used to categorize every series for their absolute and individual quality. This index is calculated using Eq. (1) and ranges from 0 to 100. Values below 50 indicate of a very poor quality while above 80 acceptable quality is considered.

$$Q = \frac{1}{4}(P + Q_{\text{gaps}} + Q_{\text{zero}}^w + Q_{\text{outliers}}) \tag{1}$$

The four terms in Eq. (1), contributing with the same weigh, evaluate the presence of the most common troubles in daily rainfall series. Parameter P measures completeness and is calculated dividing the number of daily data by 365 (or 366 in the case of leap years). Parameter Q_{gaps} evaluates the gap distribution, penalizing the presence of gaps interspersed with data more than single empty periods, and is calculated using Eq. (2). In Eq. (2) n_{gap} is the number of empty days from the total, n , and $L_{\text{gap}}^{\text{max}}$ is the length of the longest empty period, for every station-year.

$$Q_{\text{gaps}} = 100 - 100 \frac{2n_{\text{gap}} + L_{\text{gap}}^{\text{max}}}{n} \tag{2}$$

There is no significant evidence of weekly cycles that could favor or disfavor rainfall on a specific day of the week (Schultz et al. 2007; Stjern 2011), so the number of precipitation days in a given year should be approximately the same for every day of the week. The term Q_{zero}^w in Eq. (1) is able to detect if there is a much lower number of rainy days for a specific day of the week than the rest, with the aim to penalize series for which some days of the week would have been systematically filled by zeros due to not having any measure. For each year at every station, using an index i ranging from 1 to 7 (one value for each day of the week), the number n_i of rainy days for each day of the week is considered. Then, for every station-year, there will be a set of seven numbers, each of them indicating the number of rainy Mondays, Tuesdays, Wednesdays, etc., ideally similar between them, for which the coefficient of variation CV has to be calculated. Being CV the standard deviation divided by the mean value, it is supposed to be close to zero if the number of rainy days for each day of the week is similar. The term Q_{zero}^w is then calculated using Eq. (3), for every station-year.

$$Q_{\text{zero}}^w = 100 - 100 \text{ CV} \tag{3}$$

For the last term, Q_{outliers} , the proportion of number of days which have not exceeded the outliers' threshold with respect to the total number of days is calculated. For every month, the outliers' threshold is determined as the rainfall amount three times the interquartile range above the third quartile.

The quality control procedure applied by Llabrés-Brustenga et al. (2019) included also a term Q_{zero}^m intended to penalize series with many whole months with null precipitation, which could correspond in some areas to not operational time periods filled with false zeros. This is not the case of the Andalusia dataset, a region with several dry areas where it is common to have 2 or even 3 months each year without any precipitation. The automatic calculation of this term as it was defined cannot work properly in Andalusia; therefore, it has not been considered.

The value of this quality index Q can be a useful indicator for an appropriate selection of series, as it was the case of the works by Casas-Castillo et al. (2018) and Lana et al. (2020, 2021a, b), which used series extracted from the complete rainfall dataset of Catalonia (Llabrés-Brustenga et al. 2019) that fulfilled, among other quality requirements, to have an index Q higher than 80%.

An analysis of the values of the four terms of Eq. (1) found for all the years and all the stations of Andalusia has been performed, being shown in Section 3.

2.2 Relative quality control

The final stage of the control procedure is a relative quality control that takes into account and compares data from neighboring stations leading to a complete evaluation of each daily value. For every singular station some convenient auxiliary stations are selected to compare their registers. For this selection, a representativeness index R is then calculated for every couple studied station — possible auxiliary station. This index R takes into account the spatial coordinates, the altitude of the stations, and their daily records, and it is calculated using Eq. (4), where d is the distance (in km) between the two stations, h is the difference in altitude (in m), and C_{corr} is the Pearson's correlation coefficient between daily data from both stations. Llabrés-Brustenga et al. (2019) considered necessary the inclusion of the two first terms due to the complex orography of their area of study (Catalonia). The same consideration can be made to the Andalusian region, where big differences of altitude at short distances are possible, resulting in different local meteorological conditions and consequently different rainfall production.

$$R = \frac{100}{3} \left(\frac{D-d}{D} + 0.5^{h/500} + C_{\text{corr}} \right) \tag{4}$$

For the distance term of Eq. (4) a maximum radius of influence D has to be established, depending of the characteristics of the studied area and the density of stations. For instance, a value of $D = 50\text{km}$ was considered for Catalonia, with an averaged density of 5.3 stations per 100 km^2 . Then, values for this term range from 1 (for $d = 0$) to 0 (for $d = D$), decreasing linearly with distance. Values for the altitude term are also between 1 and 0 but decrease exponentially to penalize big differences of altitude more than small ones, with no upper limit considered.

The correlation term, which takes into consideration daily rainfall data, was designed to calibrate the similarity between the rainfall regime of the auxiliary station and that of the study station. Even though it is calculated from Pearson's correlation coefficient, no negative values were considered, assigning a null value to the correlation term C_{corr} in the case of negative coefficients.

To compare daily data from the studied station and the possible auxiliary station, a relative difference DIF is calculated using Eq. (5), which takes into account the daily value for the studied station, PPT, and that for the auxiliary station, PPT_{aux} , dividing by the average of both values and multiplying by the maximum value.

$$\text{DIF} = \frac{|\text{PPT} - \text{PPT}_{\text{aux}}|}{\text{Av}(\text{PPT}, \text{PPT}_{\text{aux}})} \text{Max}(\text{PPT}, \text{PPT}_{\text{aux}}) \tag{5}$$

To be considered valid, pairs of studied station-possible auxiliary station have to present DIF values lower than an established threshold T_m calculated using Eq. (6).

$$T_m = C_m \ln(101 - R) \tag{6}$$

The values of the coefficient C_m in Eq. (6) have to be determined from high-quality data series and are expected to be seasonal. The procedure to estimate these coefficients for every month is the following: (1) for each specific value of the index R , the relative differences between each monthly amount of each year collected in a station and the amounts of neighboring stations are calculated; (2) the 95% percentile as well as the interquartile range of these series of relative differences for each R are determined; and (3) the amounts corresponding to ten times the interquartile range above the 95% percentile (one point for each R) are fitted to Eq. (6) leading to the value of the coefficient C_m for the considered month. Llabrés-Brustenga et al. (2019) found a seasonal cycle for this coefficient, the highest values corresponding to summer months when local storms are likely to occur, being the lowest values for winter months. The lowest value of threshold T_m is 0, and corresponds to a value of representativeness R of 100; it would be the case, for instance, of two stations located at the same place with perfectly correlated data. If distance and difference of altitude between the two stations increase or data correlation decrease, this threshold increases logarithmically. One of the aims of the present work is to verify if the threshold coefficients C_m calculated from the best quality series of Andalusia perform a similar seasonal cycle and compare the obtained values. Their possible application in other areas, with similar climatology but unable to obtain their own coefficients due to the lack of an operational network of automatic and controlled stations, might be of great interest.

In case the DIF values for a specific day of the year were under the threshold, the couple of stations are labeled as “valid” for this day. If the threshold is exceeded, the couple is labeled as “invalid.” For each studied station, this relative comparison has to be made for every day of the year and every auxiliary station. Then, in order to be declared as a

valid station, the studied station must obtain more than a certain established value for a weighted arithmetic mean WM (see Eq. *(7)) which considers the representativeness of all available pairs, valid or not. In Eq. (7) L takes the value 1 for valid pairs and 0 for invalid pairs, and R_{\min} is the minimum value of representativeness considered.

$$WM = 100 \frac{\sum_{aux} (R - R_{\min})^2 L}{\sum_{aux} (R - R_{\min})^2} \tag{7}$$

The results of the relative quality control performed on the Andalusian database are shown in the next section.

3 Results and discussion

3.1 Absolute quality control

Regarding the annual completeness, Table 1 shows the number of stations with a certain percentage of completeness for a minimum number of years. Thus, for almost 90% of the stations at least 5 years of their records present a completeness of 50%, while only 3 stations have records with at least 50 years 100% complete. Figure 4 shows the number of available stations with at least 1 day with data for every year of the considered period (1870–2018).

The temporal evolution of the four terms contributing to the calculation of the quality index Q using Eq. (1) is shown in Fig. 5. Regarding the term P of completeness, some traits are remarkable: the only rainfall record working alone in the Andalusian set for more than 30 years since 1870 (San Fernando, Cádiz) had a very good completeness along this period, always over 93% for the available years. Regarding the other four stations which were put into operation in the early years of the twentieth century (Jaén-Instituto in 1901; Granada-Cartuja and Granada-Universidad, both in 1902; and Huelva and Sanlúcar de Barrameda, in 1903), due to the fact that their records did not start in January 1 but in the middle of the year, in any month, the term P suffers a visible drop in these years to values below 80% (Fig. 5). This

Table 1 Number of stations with a certain annual completeness of daily data for a given number of years

| | | Number of years with completeness | | | | | | | | | |
|---------------------|------|-----------------------------------|------|------|-----|-----|-----|-----|-----|-----|-----|
| | | 5 | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 |
| Annual completeness | 50% | 1698 | 1470 | 1249 | 940 | 728 | 565 | 448 | 384 | 314 | 231 |
| | 75% | 1635 | 1379 | 1156 | 835 | 629 | 479 | 390 | 312 | 239 | 161 |
| | 80% | 1629 | 1368 | 1144 | 820 | 617 | 470 | 381 | 299 | 232 | 154 |
| | 85% | 1585 | 1292 | 1062 | 713 | 528 | 400 | 317 | 255 | 174 | 121 |
| | 90% | 1565 | 1257 | 1009 | 677 | 491 | 384 | 305 | 238 | 154 | 105 |
| | 95% | 1405 | 1038 | 696 | 433 | 289 | 205 | 131 | 84 | 54 | 29 |
| | 100% | 1057 | 673 | 353 | 211 | 111 | 54 | 27 | 14 | 10 | 3 |

Fig. 4 Number of available stations with at least one daily data for every year

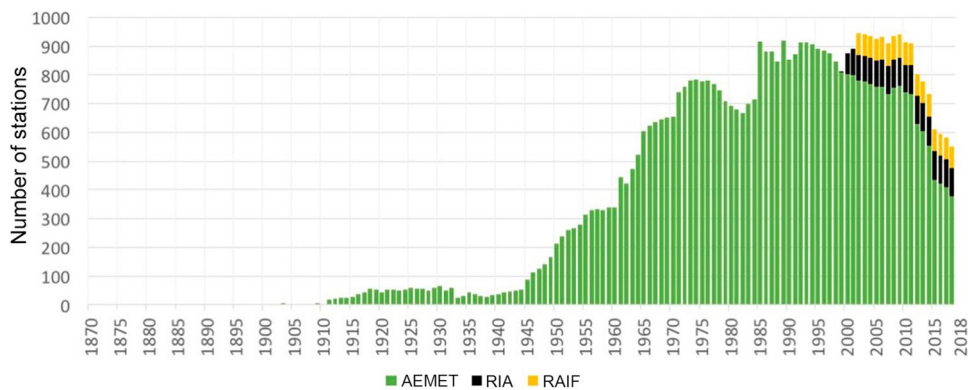
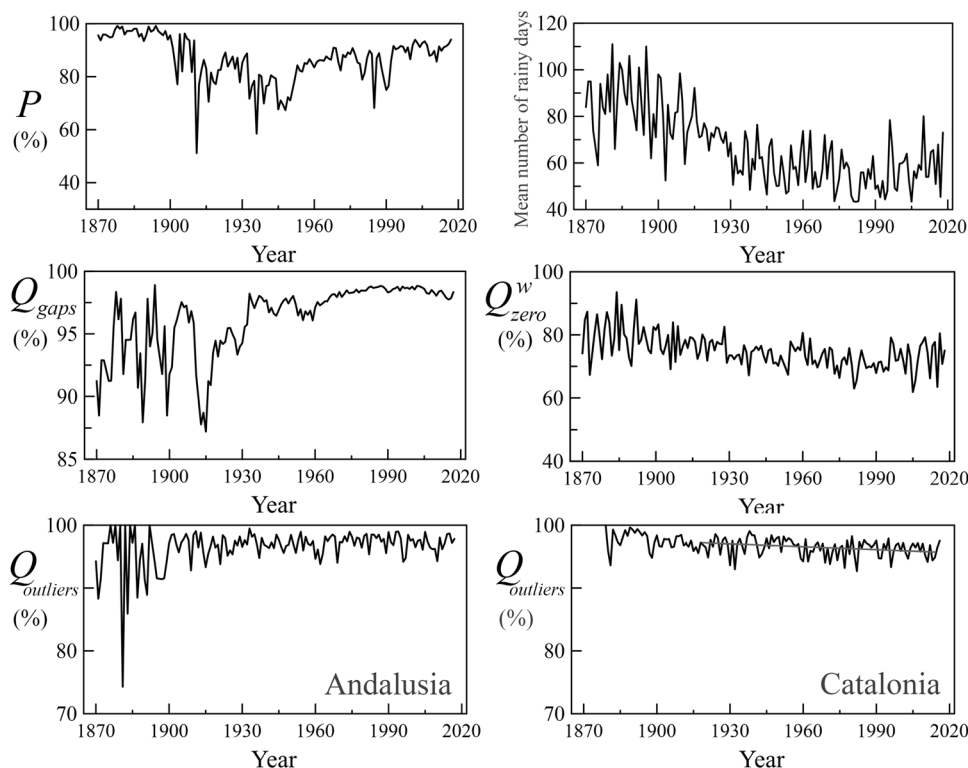


Fig. 5 Temporal evolution of the four quality terms contributing to the quality index Q . Every point has been drawn by assigning to each year the mean value obtained for all stations that had value this year. The temporal evolution of the mean number of rainy days in Andalusia, as well as that of the quality term $Q_{outliers}$ in Catalonia, have also been represented. The blue line indicates the significant decreasing trend ($p < 0.01$) of this term in Catalonia (slope -0.013% per year)



drop is even more pronounced a decade later, when in 1912 several new stations began to work and the total number of stations suddenly rose to more than 20. The fact that their first operational year was not complete makes the global term P to fall to values near to 50%. The Spanish Civil War (1936–1939) might be another cause contributing to the next sharp drop in the 30 s. This dropping effect in the graph of P due to new records with the first year incomplete is only noticeable when the number of stations is small. Then, later drops after 1960 might not be attributed to this effect.

Sharp fluctuations of the term Q_{gaps} can be seen at the beginning of the dataset, its values oscillating between 87 and 98% for the oldest stations. Even though, since 1915 this term begins to increase and after 1970 it remains almost steady around a value of 97–98%. This term and the term

$Q_{outliers}$ are both very sensitive to the number of operational stations, presenting anomalous behaviors at the beginning, when only few stations were working. The term $Q_{outliers}$ remains almost steady around a value of 95% after 1910, showing no temporal trend. On the contrary, the values of this term for the Catalonian dataset analyzed by Llabrés-Brustenga et al. (2019) do show a clear decreasing temporal trend (Fig. 5). A Mann–Kendall test performed on both series of $Q_{outliers}$ for the years after 1920, when there were almost 40 stations in operation, shows a significant decreasing trend for Catalonia ($p < 0.01$, slope -0.013% per year), which means an increase of the presence of outliers in the sample, while no trend is detected for Andalusia. When a more refined analysis is done by the calculation of this quality term by provinces, decreasing trends are also found for

the eastern Andalusian areas, more influenced by the Mediterranean Sea, as Catalonia is. These increases of the presence of outliers are in general accordance with the temporal increase of rainfall irregularity reported in the Mediterranean area during the most last part of the twentieth century (Lana et al. 2022; Casas-Castillo et al. 2018, 2022).

As expected, no evidence of weekly cycles favoring or disfavoring rainfall on a specific day of the week was found. The apparent general decrease over time of term Q_{zero}^w might be related to a decrease of the number of rainy days in certain zones. In this sense, a decreasing trend in seasonal precipitation has been observed in some areas of the Eastern Andalusia (Ramos-Calzado et al. 2008; Ruiz Sinoga et al. 2010). Apart from that, there have been well-known periods of drought affecting the whole region: 1978–1982, 1992–1995, and 2004–2010.

Figure 6 shows the distribution histogram of the mean values of the quality index Q for every meteorological network. As expected, automatic weather station networks RAIF and RIA obtained the best absolute quality indices, with almost 87% of the stations with values of Q above 85%, while only 57% of network AEMET, with older and mostly manual stations, reached this value.

The mean value of this global quality index Q for all the station/years of the entire database has found to be 86%. No spatial pattern has been detected for Q , indicating that this index seems not to depend on the diverse climatic characteristics of the study area. This absolute quality control has made it possible to select the best quality series from the Andalusian dataset to perform a simple scaling

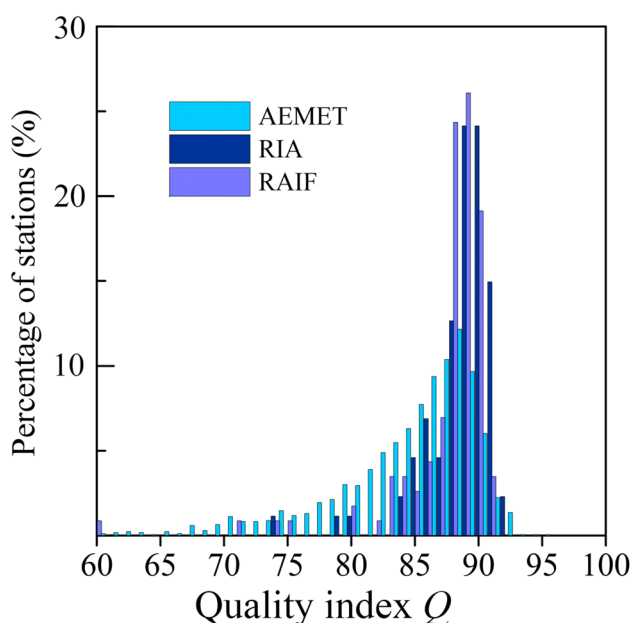


Fig. 6 Distribution histogram of the mean values of the quality index Q (%) for every meteorological network

analysis under the different precipitation regimes in the region (Casas-Castillo et al. 2022).

3.2 Relative quality control

The value of the representativeness index R has been calculated for every couple studied station — possible auxiliary station of the Andalusian database using Eq. (4), which considers location, altitude, and daily records from every station. Even though the density of stations in Andalusia is lower than for Catalonia, 2.3 per 100km², in Eq. (4) the maximum radius of influence between stations has been also set to $D = 50$ km. One of the possible consequences derived from the choice of an inappropriate D value in relation to the characteristics of the study area is that the number of possible auxiliary stations could be too low to carry out an appropriate relative quality control. In this case, 50 km has proven to be a suitable maximum radius of influence for the Andalusian network.

For a selection of couples of stations with a representativeness index R between them exceeding a value of 70, the relative differences (DIF) observed in daily data measured at each station have been calculated using Eq. (5). As commented in Section 2.2, couples with DIF values under an established threshold will be considered as valid. Threshold T_m has to be calculated using Eq. (6), for which the specific monthly coefficients C_m of the Andalusian region, for the maximum radius of influence chosen, have to be determined. It has been done using data from 97 automatic weather stations of the RIA network, all of them with values of the absolute quality index Q above 90%.

For all the RIA couples with representativeness above 70, the relative differences in daily measurements in the period 2008–2018 have been calculated. Then, following the procedure presented in Section 2, for every value of the index R , the relative differences corresponding to ten times the interquartile range above the 95% percentile have been determined for every month. Then, these monthly amounts have been fitted to the threshold functions (Eq. (6)) leading to the monthly coefficients C_m . Figure 7 shows four examples of observed relative differences and fitting threshold functions: those corresponding to September, the month with the highest relative differences, those to February, one of the months with lower differences, and the intermediate cases of April and November.

A seasonal cycle for the coefficients C_m has been obtained (Fig. 8; Table 2), with high values for late summer and beginning of the autumn (above 25 mm; 33.3 mm for September) and low values (around 10–15 mm) for winter (December to February). This outcome, very similar of that found by Llabrés-Brustenga et al. (2019) for Catalonia, is a consequence of the extremely irregular surface registers, even at close locations, for late summer storms and

Fig. 7 Relative difference observed in daily measurements for couples of RIA stations with a representativeness above 70. Blue points indicate the value of 10 times the interquartile range above the 95% percentile of the set of points corresponding to every value of representativeness, fitted to the threshold functions (red lines) for February (a), April (b), September (c), and November (d)

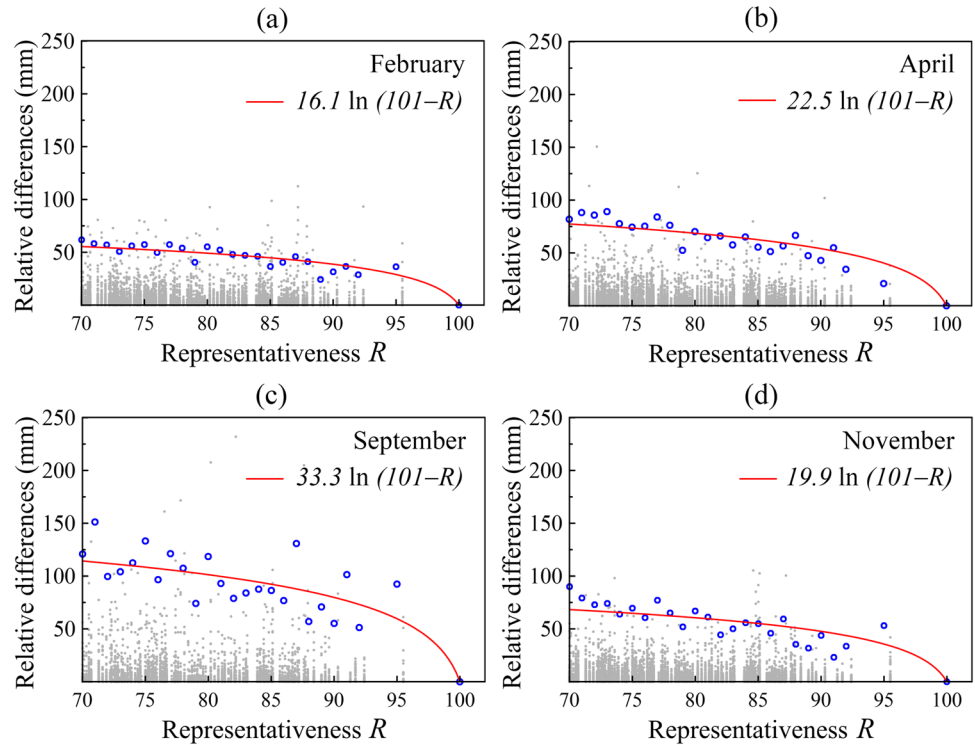


Fig. 8 Monthly C_m values of the threshold's function for Andalusia. The dashed line is a fitted third-order polynomial line drawn for a better visualization

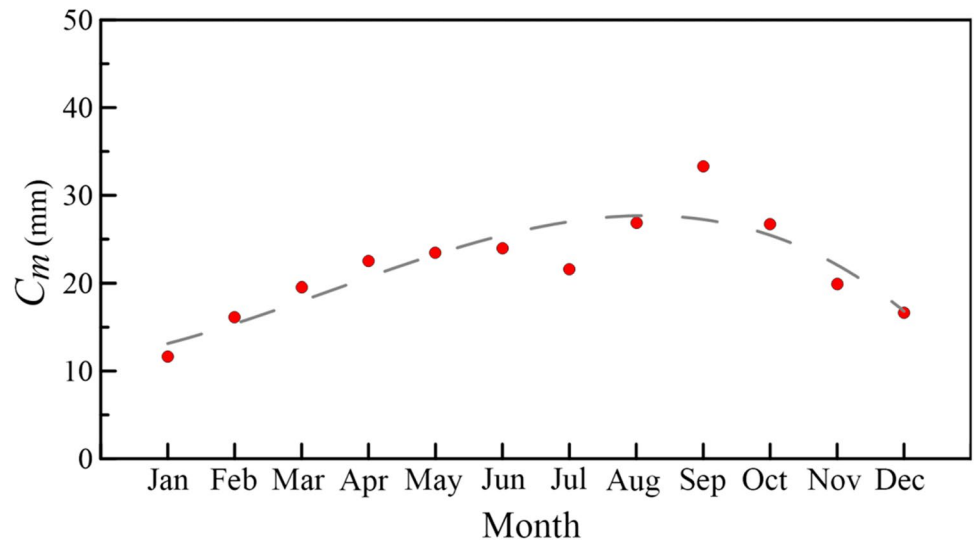


Table 2 Monthly values of the coefficient C_m of the threshold's function for Andalusia for a maximum radius of influence between stations of 50 km

| Month | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec |
|------------|------|------|------|------|------|------|------|------|------|------|------|------|
| C_m (mm) | 11.6 | 16.1 | 19.5 | 22.5 | 23.4 | 23.9 | 21.6 | 26.8 | 33.3 | 26.7 | 19.9 | 16.6 |

convective events in autumn, whereas for the continuous and more uniform winter rainfall episodes, often of synoptic origin, differences in measurements taken at distant stations are usually low. The values for early summer are not much

higher than in spring (even a slight decrease in July) due to the fact that, in the most part of Andalusia, rainfall is usually scarce in summer. Then, the observed rain differences cannot be very high, and there is a low dispersion of data

that causes the value of the interquartile to decrease. In fact, July is the month for which fewer coincident rainy days have been found, being also few the couples contributing to the analysis.

After a selection for which only the auxiliary stations with relative daily differences below the corresponding threshold remain, next step consists of tagging each station and each day of the whole dataset as valid/invalid/doubtful/insufficient information after relative comparison with its auxiliary stations. To finally declare a station as valid for a certain day, the studied station must achieve the value 50% for the arithmetic mean WM (Eq. (7)) for this day. If the value of WM is between 20 and 50% the station is labeled as doubtful for this day, and it is considered invalid for values below 20%. For a reliable comparison it was considered that a minimum of three auxiliary stations were necessary for every studied station and day. Days with less than three values to compare were considered to have insufficient information. A total of 15,857,892 days/station have been analyzed following this procedure (Table 3). The average number of auxiliary stations for every station was found to be 18.

In Table 3 are listed the numbers of days/station assigned to every of the relative quality control categories considered, together with their percentages of the total. The percentage of data declared invalid, i.e., detected errors, has been 0.14%. This a percentage similar to that found by Llabrés-Brustenga et al. (2019) for Catalonia (0.2%), to the rejected values by Vicente-Serrano et al. (2010) in the NE of Spain (0.1%), and those (1.58%) by Serrano-Notivoli et al. (2017) in the Spanish island of Mallorca. Similar is the percentage of doubtful cases (0.17%). From the rest, a substantial percentage (18%) have been determined to have insufficient information, a percentage slightly lower to that found by Llabrés-Brustenga et al. (2019), which was 25%. This is a consequence of the low density of stations in some places; not enough auxiliary stations under the considered requirements could be assigned to some stations for some days. Moreover, analyzing to which of the 1947 available series these cases with insufficient information belong, a percentage of almost 20% correspond to stations for which at least during their first 15 operational years it has not been possible to compare measurements with neighboring stations. In fact, in the period 1912–1932 some stations begin to have valid data (a maximum of 19 stations with valid data in 1924), but from 1933, in previous years and during the Spanish

Civil War (1936–1939), the number of operational stations dropped drastically. It is not until 1950 that there are again about 20 stations with valid data.

Bearing in mind both results from the absolute and the relative quality controls, the final quality of every year at each station have been assessed considering these three categories: (1) excellent quality, for series with a minimum absolute quality index Q of 90% and a minimum percentage of valid data after the relative control of 99%; (2) good quality, for series with a minimum absolute quality index Q of 85% and a minimum percentage of valid data of 95%; and (3) acceptable quality, for series with a minimum Q of 80%, and a minimum percentage of valid data of 90%. From a total of 49,789 station/years, 50% (24,778) have been declared of acceptable quality or better, 35% (17,350) of good or excellent quality, and only 17% can be considered of excellent quality. Notice that not all the available years/station correspond to complete years. If only the series from 1970 onward are considered, it would be a total of 39,338, with a 57% (22,443) of acceptable cases or better, 41% (16,111) of good or excellent ones, and 20% (7913) of excellent quality series. Figure 9 shows the temporal evolution of the number of series in each of these three quality categories, as well as the total available series.

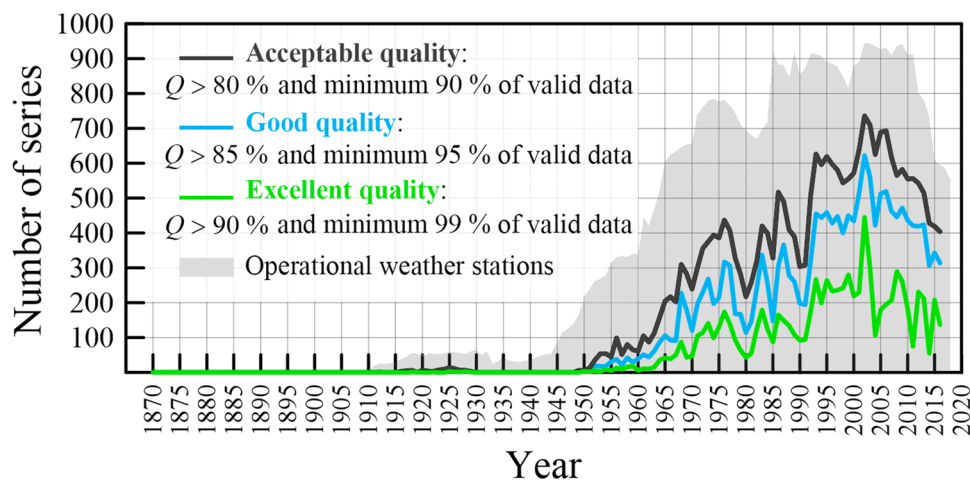
4 Conclusions

A quality control procedure has been applied to a large database of 15,857,892 daily rainfall measurements from 1947 stations of three meteorological networks (AEMET, RAIF, and RIA) operating in Andalusia, in the south of the Iberian Peninsula, from 1870 to the present. This procedure consists of three successive steps which obtain increasingly higher degrees of quality from the database. At the second stage a quality index Q is calculated for every year at every station individually, considering the percentage of annual data, the distribution of gaps in the sample, the homogeneity of rainfall records by day of the week, and the abundance of outliers. The mean value of this quality index Q for the whole database has found to be 86%, and no spatial pattern has been detected for it, indicating no dependence on the diverse climatic characteristics of the study area. The automatic weather stations networks RIA and RAIF obtained absolute quality index Q above 85% for almost 87% of their stations, while only 57% of AEMET network with older and mostly manual stations reached this value. However, it is remarkable that the only rainfall record working alone in the Andalusian set for more than 30 years since 1870 (San Fernando, Cádiz) had a very good completeness along this period, always over 93% for the available years. Regarding the presence of outliers over time, no temporal trend has been found after 1910, unlike what it can be found for

Table 3 Numbers of days/station and percentage of the total for every quality category considered

| Total of days/ station data | Valid | Doubtful | Invalid | Insufficient information |
|--------------------------------|----------------------|-----------------|-----------------|-----------------------------|
| 15,857,892 | 12,920,318 81.48% | 26,712 0.17% | 22,046 0.14% | 2,888,816 18.22% |

Fig. 9 Temporal evolution of the number of available and selected quality series



Catalonia: an increasing temporal trend of the presence of outliers possibly related to a reported increase of rainfall irregularity over time caused by climate change.

After the last stage, each daily value measured at each station is labeled among three possible categories (valid, doubtful, or invalid) according to the relative comparison between stations with sufficient representativeness between them. For this comparison, maximum possible differences under a certain threshold have been determined after the analysis of some reference series obtained from automatic weather stations of proven high quality (RIA, between 2008 and 2018). These maximum differences have turned out to be seasonal: higher values are observed for late summer and beginning of the autumn as a consequence of the extremely irregular surface records for late summer storms and convective rainfall events in autumn, even at nearby locations. For the more regular winter rainfall episodes, often of synoptic origin, lower differences are usually observed, even at distant stations. Monthly threshold functions have been found, with fitting coefficients very similar to those obtained in Catalonia (NE of the Iberian Peninsula): maximum values for September and minimum from December to January. The main exception is the coefficient relative to July, a month for which rainfall is usually scarce or null in the most part of Andalusia, whose value has resulted noticeable lower than in Catalonia. It can be concluded that for the use of this quality control procedure, it could be recommendable the determination of the threshold functions for the specific studied region, depending on its climate characteristics. The density of stations has to be considered too, eventually leading to different threshold functions that make possible the application of the methodology with enough stations to compare with each other.

A total of 15,857,892 days/station data have been processed, from which almost 82% have been declared to be correct. Only 0.14% of the cases have been found to be incorrect values, while the doubtful cases are 0.17%. For a

significant percentage of cases (18%) there was not enough information to perform the relative quality procedure under the established requirements. This is a reasonable percentage due the characteristics of the analyzed database, covering a large period which starts at the very beginning of the rainfall measurement instrumental history of Andalusia in 1870.

Three quality categories have been considered to evaluate every year at each station taking into account both the absolute and the relative controls: from a total of 49,789 available years/station since 1870, 50% have been declared of acceptable quality or better (a minimum absolute quality index Q of 80% and a minimum percentage of valid data after the relative control of 90%), 35% of good or excellent quality (minimum Q of 85% and a minimum percentage of valid data of 95%), and only 17% can be considered of excellent quality (minimum Q of 90% and minimum 99% of valid data).

The application of this type of complete quality control procedure has been very successful despite the limitations of the semiarid region of Andalusia, especially due to its low density of weather stations and its structural water deficit which acquires great relevance in the summer months.

Acknowledgements The authors thank the anonymous referees and the editor for their constructive comments and suggestions.

Author contribution The authors confirm contribution to the paper as follows: conceptualization, J. Estévez, A. P. García-Marín, R. Rodríguez-Solà, M. C. Casas-Castillo, and A. Llabrés-Brustenga; methodology, R. Rodríguez-Solà, M. C. Casas-Castillo, and A. Llabrés-Brustenga; software, R. Rodríguez-Solà, A. Llabrés-Brustenga, and R. Kirchner; validation, J. Estévez, A. P. García-Marín, R. Rodríguez-Solà, M. C. Casas-Castillo, and A. Llabrés-Brustenga; formal analysis, R. Rodríguez-Solà, M. C. Casas-Castillo, and A. Llabrés-Brustenga; investigation, J. Estévez, A. P. García-Marín, R. Rodríguez-Solà, M. C. Casas-Castillo, and A. Llabrés-Brustenga; resources, J. Estévez and A. P. García-Marín; data curation, J. Estévez and A. Llabrés-Brustenga; writing — original draft preparation, M. C. Casas-Castillo and J. Estévez; writing — review and editing, M. C. Casas-Castillo, J. Estévez, A. P. García-Marín, R. Rodríguez-Solà, and

A. Lladrés-Brustenga; figures and visualization, M. C. Casas-Castillo and R. Rodríguez-Solà; supervision, J. Estévez, A. P. García-Marín, R. Rodríguez-Solà, and M. C. Casas-Castillo; funding acquisition, J. Estévez and A. P. García-Marín. All authors have read and agreed to the published version of the manuscript.

Funding This research was funded by the Spanish Ministry of Science, Innovation and Universities (grant number AGL2017-87658-R).

Data availability Rainfall data used to support the findings of this study were supplied by the Spanish Agency of Meteorology (AEMET), available by request.

Code availability Not applicable.

Declarations

Ethics approval Not applicable.

Consent to participate All authors consent to participate into the study.

Consent for publication All authors consent to publish the study in a journal article.

Conflict of interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Capel Molina JM (1974) Génesis de las inundaciones de octubre de 1973 en el sureste de la península ibérica. Cuadernos geográficos de la Universidad de Granada, N.º 4, 1974, pages 149–166, ISSN 0210–546. Available (in Spanish) at <https://dialnet.unirioja.es/descarga/articulo/2195195.pdf>
- Casas-Castillo MC, Lladrés-Brustenga A, Rius A, Rodríguez-Solà R, Navarro X (2018) A single scale parameter as a first approximation to describe the rainfall pattern of a place: application on Catalonia. *Acta Geophys* 66(3):415–425. <https://doi.org/10.1007/s11600-018-0122-5>
- Casas-Castillo MC, Rodríguez-Solà R, Lladrés-Brustenga A, García-Marín AP, Estévez J, Navarro X (2022) A simple scaling analysis of rainfall in Andalusia (Spain) under different precipitation regimes. *Water* 14:1303. <https://doi.org/10.3390/w14081303>
- Durre I, Menne MJ, Gleason BE, Houston TG, Vose RS (2010) Comprehensive automated quality assurance of daily surface observations. *J Appl Meteorol Climatol* 49:1615–1633. <https://doi.org/10.1175/2010JAMC2375.1>
- Einfalt T, Jessen M, Quirnbach M (2006) Can we check raingauge data automatically? In: Proceedings 7th International Workshop on Precipitation in Urban Areas, St. Moritz, Switzerland, 7–10 December, ISBN 3–909386–65–2
- Estévez J, Gavilán P, Giráldez JV (2011) Guidelines on validation procedures for meteorological data from automatic weather stations. *J Hydrol* 402(3–4):144–154. <https://doi.org/10.1016/j.jhydrol.2011.02.031>
- Estévez J, Gavilán P, García-Marín AP, Zardi D (2015) Detection of spurious precipitation signals from automatic weather stations in irrigated areas. *Int J Climatol* 35:1556–1568. <https://doi.org/10.1002/joc.4076>
- Estévez J, Bellido-Jiménez JA, Liu X, García-Marín AP (2020) Monthly precipitation forecasts using wavelet neural networks models in a semiarid environment. *Water* 12(7):1909. <https://doi.org/10.3390/w12071909>
- Feng S, Hu Q, Qian W (2004) Quality control of daily meteorological data in China 1951–2000: a new dataset. *Int J Climatol* 24(7):853–870. <https://doi.org/10.1002/joc.1047>
- García-Marín AP, Estévez J, Medina-Cobo MT, Ayuso J (2015) Delimiting homogeneous regions using the multifractal properties of validated rainfall data series. *J Hydrol* 529(1):106–119. <https://doi.org/10.1016/j.jhydrol.2015.07.021>
- Habib E, Krajewski WF, Kruger A (2001) Sampling errors of tipping-bucket rain gauge measurements. *J Hydrol Eng* 6(2):159–166. [https://doi.org/10.1061/\(ASCE\)1084-0699\(2001\)6:2\(159\)](https://doi.org/10.1061/(ASCE)1084-0699(2001)6:2(159))
- Hubbard KG, Goddard S, Sorensen WD, Wells N, Osugi TT (2005) Performance of quality assurance procedures for an applied climate information system. *J Atmos Oceanic Technol* 22:105–112. <https://doi.org/10.1175/JTECH-1657.1>
- Lana X, Rodríguez-Solà R, Martínez MD, Casas-Castillo MC, Serra C, Kirchner R (2020) Multifractal structure of the monthly rainfall regime in Catalonia (NE Spain): evaluation of the non-linear structural complexity. *Chaos* 30:073117. <https://doi.org/10.1063/5.0010342>
- Lana X, Casas-Castillo MC, Rodríguez-Solà R, Serra C, Martínez MD, Kirchner R (2021) Rainfall regime trends at annual and monthly scales in Catalonia (NE Spain) and indications of CO₂ emissions effects. *Theoret Appl Climatol* 146:981–996. <https://doi.org/10.1007/s00704-021-03773-z>
- Lana X, Rodríguez-Solà R, Martínez MD, Casas-Castillo MC, Serra C, Kirchner R (2021) Autoregressive process of monthly rainfall amounts in Catalonia (NE Spain) and improvements on predictability of length and intensity of drought episodes. *Int J Climatol* 41(S1):E3178–E3194. <https://doi.org/10.1002/joc.6915>
- Lana X, Casas-Castillo MC, Rodríguez-Solà R, Prohom M, Serra C, Martínez MD, Kirchner R (2022) Time trends, irregularity, multifractal structure and effects of the CO₂ emissions on the monthly rainfall regime at Barcelona City, NE Spain, years 1786–2019. *Int J Climatol* (in press).
- Lladrés-Brustenga A, Rius A, Rodríguez-Solà R, Casas-Castillo MC, Redaño À (2019) Quality control process of the daily rainfall series available in Catalonia from 1855 to the present. *Theor Appl Climatol* 137:2715–2729. <https://doi.org/10.1007/s00704-019-02772-5>
- Medina-Cobo M, García-Marín AP, Estévez J, Jiménez-Hornero F, Ayuso J (2017) Obtaining homogeneous regions by determining the generalized fractal dimensions of validated daily rainfall data sets. *Water Res Manag* 31:2333–2348. <https://doi.org/10.1007/s11269-017-1653-2>
- Michaelides SC (2008) Precipitation: advances in measurement, estimation and prediction. Springer: Vienna, 540 p. ISBN 978–3–540–77655–0
- Morbideilli R, García-Marín AP, Mamun AA, Atiqur RM, Ayuso-Muñoz JL, Taouti MB, Baranowski P, Bellocchi G, Sangüesa-Pool

- C, Bennett B, Oyunmunkh B, Bonaccorso B, Brocca L, Caloiero T, Caporali E, Caracciolo D, Casas-Castillo MC, Catalini CG, Chetih M, Kamal Chowdhury AFM, Chowdhury R, Corradini C, Custò J, Dari J, Diodato N, Doesken N, Dumitrescu A, Estévez J, Flammini A, Fowler HJ, Freni G, Fusto F, García-Barrón L, Manea A, Goenster-Jordan S, Hinson S, Kanecka-Geszke E, Kar KK, Kasperska-Wołowicz W, Krabbi M, Krzyszczak J, Llabrés-Brustenga A, Ledesma JLJ, Liu T, Lompi M, Marsico L, Mascaro G, Moramarco T, Newman N, Orzan A, Pampaloni M, Pizarro-Tapia R, Puentes Torres A, Rashid MM, Rodríguez-Solà R, SepúlvedaManzor M, Siwek K, Sousa A, Timbadiya PV, Filippos T, Vilcea MG, Viterbo F, Yoo C, Zeri M, Zittis G, Saltalippi C (2020) The history of rainfall data time-resolution in a wide variety of geographical areas. *J Hydrol* 590:125258. <https://doi.org/10.1016/j.jhydrol.2020.125258>
- Parchure AS, Gedam SK (2019) Self-organising maps for rain event classification in Mumbai City, India. *ISH J Hydraul Eng*. <https://doi.org/10.1080/09715010.2019.1581099>
- Ramos-Calzado P, Gómez-Camacho J, Pérez-Bernal F, Pita-López MF (2008) A novel approach to precipitation series completion in climatological datasets: application to Andalusia. *Int J Climatol* 28(11):1525–1534. <https://doi.org/10.1002/joc.1657>
- Ruiz Sinoga JD, GarcíaMarín R, Martínez Murillo JF, GabarrónGaleote MA (2010) Precipitation dynamics in southern Spain: trends and cycles. *Int J Climatol* 31(15):2281–2289. <https://doi.org/10.1002/joc.2235>
- Schultz DM, Mikkonen S, Laaksonen A, Richman MB (2007) Weekly precipitation cycles? Lack of evidence from United States surface stations. *Geophys Res Lett* 34:L22815. <https://doi.org/10.1029/2007GL031889>
- Sciuto G, Bonaccorso B, Cancelliere A, Rossi G (2009) Quality control of daily rainfall data with neural networks. *J Hydrol* 364(1):13–22. <https://doi.org/10.1016/j.jhydrol.2008.10.008>
- Serrano-Notivoli R, de Luis M, Saz MA, Beguería S (2017) Spatially based reconstruction of daily precipitation instrumental data series. *Clim Res* 73(3):167–186. <https://doi.org/10.3354/cr01476>
- Stjern CW (2011) Weekly cycles in precipitation and other meteorological variables in a polluted region of Europe. *Atmos Chem Phys* 11:4095–4104. <https://doi.org/10.5194/acp-11-4095-2011>
- 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(8):1146–1163. <https://doi.org/10.1002/joc.1850>
- WMO (2008) Guide to meteorological instruments and methods of observation. World Meteorological Organization. Weather Climate Water (WMO-No. 8, the CIMO Guide)
- Yang SC, Wu MC, Kao HM, Yang TH (2020) An automated anomaly detection procedure for hourly observed precipitation in near-real time application. In: Gourbesville P., Caignaert G. (eds) *Advances in Hydroinformatics*. Springer Water, Springer, Singapore.
- You J, Hubbard KG, Nadarajah S, Kunkel KE (2007) Performance of quality assurance procedures on daily precipitation data. *J Atmos Oceanic Technol* 24:821–834. <https://doi.org/10.1175/JTECH2002.1>

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