Post-print of: Int. J. Climatol., 19 pp, http://doi.org/10.1002/joc.8069

Homogenization of monthly series of temperature and precipitation: benchmarking results of the MULTITEST project

José A. Guijarro[1], José A. López[1], Enric Aguilar[2], Peter Domonkos[3], Victor K.C. Venema[4], Javier Sigró[2], Manola Brunet[2, 5]

1 *Retired from the State Meteorological Agency (AEMET), Spain*

2 *Universitat Rovira i Virgili, Tarragona, Spain*
3 *Tortosa, Spain, dpeterfree@gmail.com*
4 *Meteorological Institute, University of Bonn, Germany*
5 *Climatic Research Unit, University of East Anglia, Norwich, UK*

* *Corresponding author address*: Jose A. Guijarro, Pje. Particular Galicia, 3, 07181-Palmanova, Balearic Islands, Spain
E-mail: jaguijarro21@gmail.com

ABSTRACT:

The homogenization of climate observational series is a needed process before undertaking confidently any study of their internal variability, since changes in the observation methods or in the surroundings of the observatories, for instance, can introduce biases in the data of the same order of magnitude than the underlying climate variations and trends. Many methods have been proposed in the past to remove the unwanted perturbations from the climatic series, and some of them have been implemented in software packages freely available from the Internet. The Spanish project MULTITEST was intended to test their performance in an automatic way with synthetic monthly series of air temperature and atmospheric precipitation, in order to update inter-comparison results from former projects, especially those of the COST Action ES0601. Several networks representing different climates and station densities were used to test a variety of homogenization packages on hundreds of random samples. Results were evaluated mainly in form of Root Mean Squared Errors (RMSE) and errors in the trend of the series, showing that ACMANT, followed by Climatol, minimized these errors. However, other packages performed also relatively well, even outperforming them when there were simultaneous biases of the same sign in most or all the test series.

Key words (up to eight): homogenization, benchmarking, monthly series, temperature, precipitation

1. Introduction

Observational series are very important pieces of information to know local climates and their variability, but they are frequently affected by changes in the observing practices, instrumentation or in the environment of the observing site which often introduce biases in the measurements that mask or distort true climate variations. Therefore, climate series must be subjected to quality control and homogenization procedures to identify and remove any artifacts and guarantee that the variability in the resultant series is only driven by weather and climate changes (Conrad and Pollack, 1950; Aguilar et al., 2003; Hunziker et al., 2018).

Many methods have been developed over time to detect these inhomogeneities, either in form of abrupt changes (break-points) or as short-term gradual drifts (artificial trends). Comprehensive reviews of them have been published by Peterson et al. (1998), Aguilar et al. (2003), Ribeiro et al. (2016), WMO (2020) and Domonkos et al. (2022), but it is difficult to assess their relative performance with real series because the correct solution is unknown (Venema et al., 2012; Mamara et al., 2013; Coll et al., 2020). Conformation of results with metadata is often used for the assessment of the reliability of homogenization results, but our experience shows that metadata are often absent or incomplete, and not all events in the history of the observatories must necessarily produce significant biases in the series. Hence there is a need of implementing benchmarking approaches where known inhomogeneities are added to synthetic homogeneous series, and they can serve to measure the quality of the solutions returned by the tested methods. There have been several studies on the efficiency of some homogenization methods applied to simulated temperature series (Caussinus and Mestre, 2004; Menne and Williams, 2005; Domonkos, 2011). However, the most important effort undertaken before the MULTITEST project was done in the frame of the Action ES0601 "Advances in Homogenization Methods of Climate Series: An Integrated Approach" (alias "HOME"), funded by the COST office of the European Cooperation in Science and Technology, in which scientists from over 20 countries gathered periodically during 2006-2011 to compare and discuss their methodologies (Venema et al., 2012).

However, although many methods have been improving along time, repeating that inter-comparison exercise in the same way as in the HOME project is impractical because of the huge work involved. To partially infill this gap, the Spanish project MULTITEST (https://climatol.eu/MULTITEST/) was dedicated to compare performances of the latest versions of the publicly available homogenization software packages which can be run in automatic mode. This condition is necessary to facilitate the comparison of the benchmarking results when applied to many hundreds of simulated monthly datasets with many different characteristics following a Monte Carlo approach. This paper presents the main results reached from a unified methodology of numerical experiments in which the test series were affected by different kinds of inhomogeneities. Some late experiments of the MULTITEST project, in which 12 synthetic and surrogate temperature test datasets were used, were published in a separate paper (Domonkos et al., 2021).

Section 2 of this article explains the methodology used to apply and evaluate the homogenization software packages (often referred to as simply “packages” here) on the different benchmark datasets. Section 3 presents the results, which are subsequently discussed in section 4, followed by the conclusions in section 5.

2. Methodology

Several synthetic data-sets simulating homogeneous monthly series of atmospheric precipitation and air temperature with different characteristics were used as master networks from which random subsets could be drawn. Inhomogeneities were inserted in these sampled subsets, which were then homogenized by means of the homogenization software packages to be tested. These tests were repeated 100 times for each master network, package and type of homogenization problem, hence allowing the production of multiple homogenization results from each software, which were then compared with the original homogeneous test series to evaluate the method performances. Networks of 10 time series were used in most homogenization experiments, but the influence of the network size was also assessed with networks of 20, 40 and 80 series. R scripts (R Core Team, 2018) were programmed to automatize all the processing tasks, and the generation of pseudo-random numbers was reset before every test run to ensure that the problem series were the same for all tested packages.

The generation of synthetic master networks is based on earlier studies (Guijarro, 2011). Their characteristics, types of inhomogeneities introduced and tested homogenization packages details are presented in the following sub-sections.

2.1. Generation of the master networks

2.1.1. Precipitation

Three synthetic precipitation networks of 100 series of 720 monthly values (equivalent to 60 years of data) were generated simulating Atlantic temperate, Mediterranean and monsoonal climates. Real series from Ireland (198) and Majorca (107), and gridded series from Southwestern India (64) drawn from the Global Precipitation Climate Center at 0.5° resolution (Schneider et al., 2015) were used, respectively, as models of each type of climate. These series were homogenized and completed with Climatol 3.0 (Guijarro, 2016).

The statistical properties and correlation structure of these series were determined with the help of several functions of the R package “gstat” (Pebesma, 2019). After a data gaussianization and randomized treatment of zero values, the “variogram” function was used to obtain a spheric variogram for every dataset. These variograms served to obtain random gaussian values by means of the function “krige”, as well as estimations of the precipitation gamma parameters shape and scale through kriging of log-transformed data. The probability of zeros were assigned by inverse distance weight interpolation, and the effect of elevation on the scale parameter was accounted for by log-regression. Synthetic series were then produced using these parameters , which allowed the preservation of realistic spatial correlation structures in these three master precipitation networks (López et al., 2016). As the original series were only used to extract their statistical properties, the method applied to homogenize them is not expected to favor any benchmarking results.

It is clear that the three chosen areas may not be the best representatives of the intended climates, which have important variations depending on the geographical area. However, they suffice for our purpose of trying the homogenization packages with three different pluviometric regimes. Also the recourse to gridded data, due to not having found observational series from India at the time of data collection, may have altered the proportion of months with zero precipitation, but the 22 % occurrence in the gridded series seemed a good proportion for deriving the simulated tropical-monsoonal dataset.

Figure 1 shows the average monthly precipitation of the three climates and the correlation-distance scatter-plots of these master networks with correlations calculated on the first differences of the series.

2.1.2. Temperature

Three master networks of 100 series containing 720 monthly values of average temperature were also generated, this time with a simpler methodology. Random locations were assigned to 100 points in a 4x3° longitude-latitude geographic domain. The monthly mean temperature series of Valladolid station (Duero basin, Spain) for the period 1951-2010 was assigned to the first point, located near the center of the domain. This series was chosen because of i: its completeness in that period; ii: being representative of a mid-latitude site with a marked seasonal cycle; and iii: it was found homogeneous in a previous study (Guijarro, 2013), although this is irrelevant for relative homogenization when the rest of the series share the same long term variability. The same series plus white noise, multiplied by a constant factor C, was assigned to the closest point. This procedure was repeated for all the remaining points, always adding noise to the closest series available. Three master networks characterized by markedly differing cross correlation structures were generated by using different C factors (0.18, 0.30 and 0.65), which are referred to as Tm1, Tm2 and Tm3, respectively.

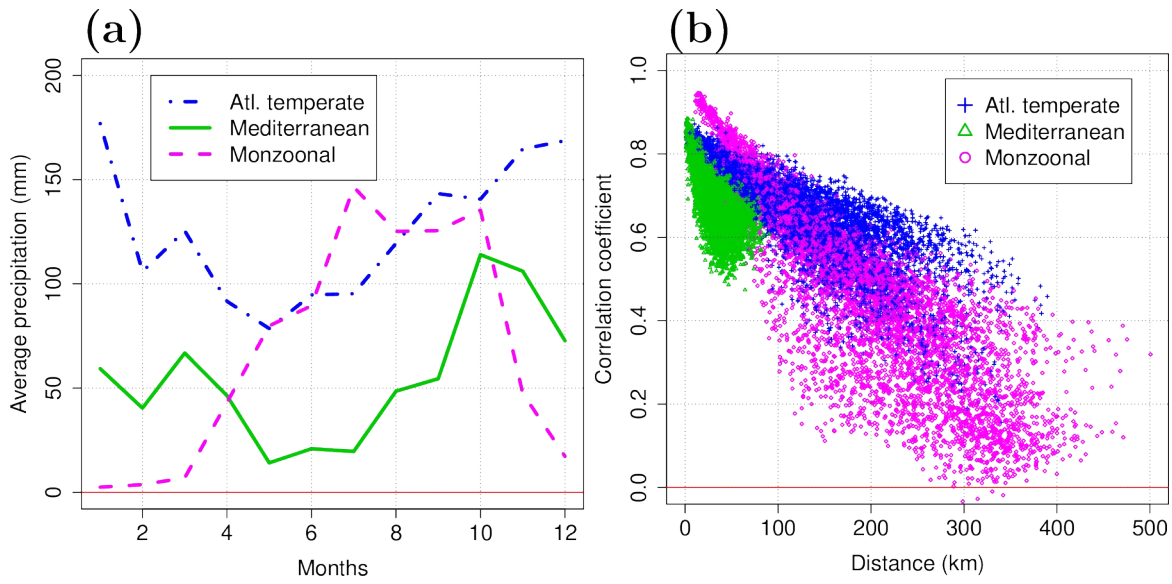


Figure 1. Average monthly precipitations of the three climates analyzed (a) and correlation-distance scatter-plots of these networks (b) calculated with the first differences of the series.

Finally, the amplitudes of their seasonal cycles were randomly varied up to ±20 % and, although irrelevant for relative homogenization, series were biased up to ±0,975 °C to simulate elevation differences of up to -/+150 m, and a 2 °C/century trend was added to all of them.

Additionally, a fourth master network, named Tr2, was built by the same procedure (with C=0.30) but using a complete synthetic series of the HOME benchmark as initial seed. This network, containing 1200 monthly data (100 years), was used to check the performance of the packages in conditions like those experimented in the HOME project. Figure 2 shows the correlation-distance scatter-plots of the aforementioned four master networks with correlations calculated on the first differences of the series.

2.2. Inserting inhomogeneities

Different kinds of inhomogeneities with increasing level of difficulty were added to each sample of the master networks to create the problem series on which the homogenization methods are tested. The last five years of the series (ten in the first three temperature experiments) were always kept untouched to allow finding reliable adjustments for inhomogeneities from the last homogeneous sub-period backwards.

2.2.1 Precipitation biases

Breaks (i.e. sudden shifts in the means) were inserted into the homogeneous series of the three master networks at randomly selected positions in any of the ten series of every sample, with a mean frequency of 1 per 20 years. The original values between two consecutive breaks were biased by multiplying them by a factor randomly drawn from a normal distribution with mean 1.0 and standard deviation 0.2 (equivalent to a variation of ±20% in precipitation). This factor was applied to all data in the series, without varying it seasonally.

No more variations of the testing procedure were done on the precipitation databases. However, we believe that many of the lessons derived from the much more extended variety of inhomogeneity problems tested on the synthetic temperature networks can be applied also to the homogenization of precipitation series.

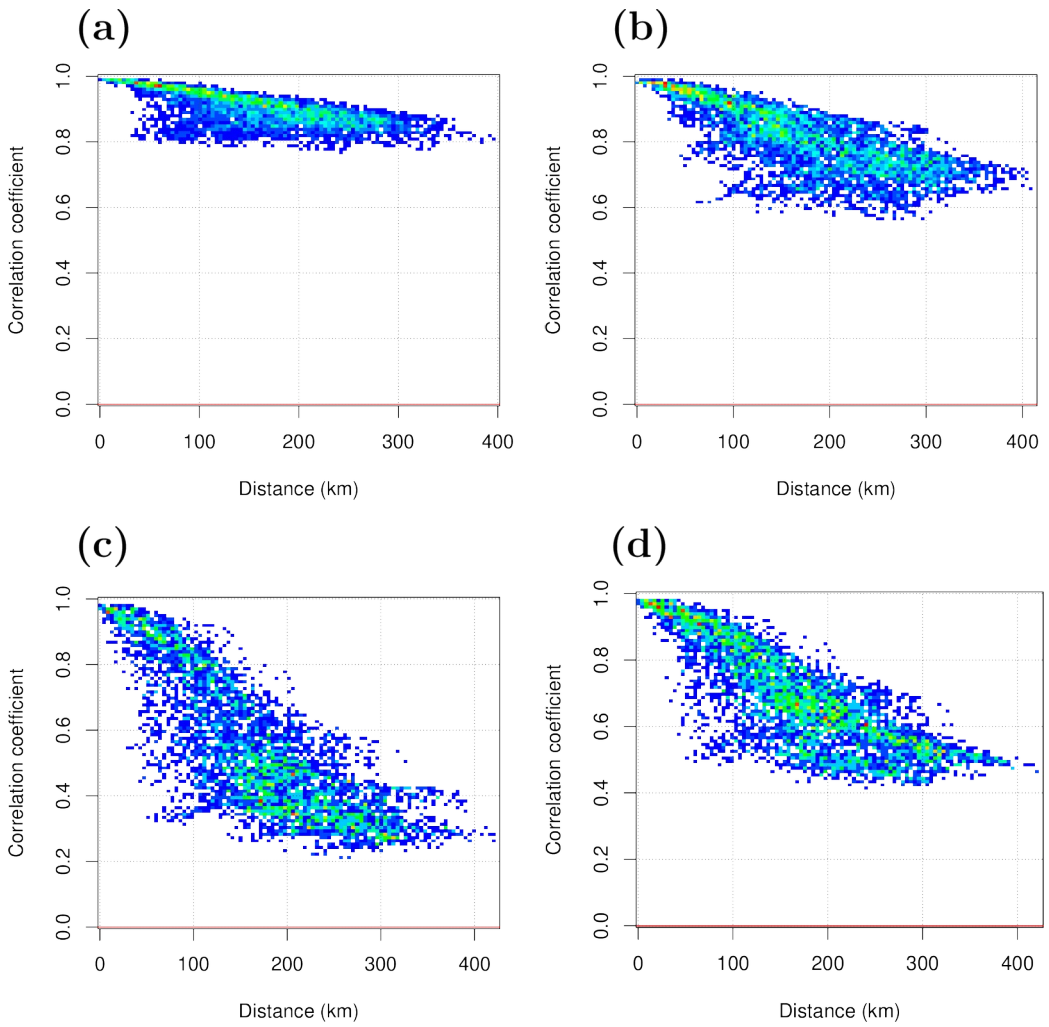


Figure 2. Correlation-distance scatterplots of the four master networks Tm1 (a), Tm2 (b), Tm3 (c) and Tr2 (d) of monthly temperature with correlations calculated on the first differences of the series.

2.2.2 Temperature biases

Tests with temperature datasets were performed in three rounds. In the first round, five experiments were performed with samples of 10 series by setting increasing degrees of difficulty. Big biases were inserted in the first three experiments in order to test the inhomogeneity correction ability of the packages when all the break-points can be easily detected, while biases of random size were introduced in the other two experiments. More precisely, the inhomogeneities applied to these first five experiments were (see examples in Figure 3):

1) Two shifts of 2 °C were imposed to the first three series at fixed positions, and one shift of the same magnitude was applied at random locations to series 4 and 5. No modification was done on the last five series, which remained as homogeneous references.

2) As in (1), but with shifts of 1.5 and 2 °C affected by a strong sinusoidal seasonality calculated from $\cos((x-7.16)*2*\pi/12)$, where x is the ordinal number of the values of the series. Seasonality in the inhomogeneities is expected to be caused by the different intensity of radiation fluxes along the year. The phase of the seasonal cycle chosen here corresponds to the phase of the seasonality of maximum average temperatures in the same area around the Valladolid station taken here as model for the generation of the synthetic temperature networks (Guijarro, 2011).

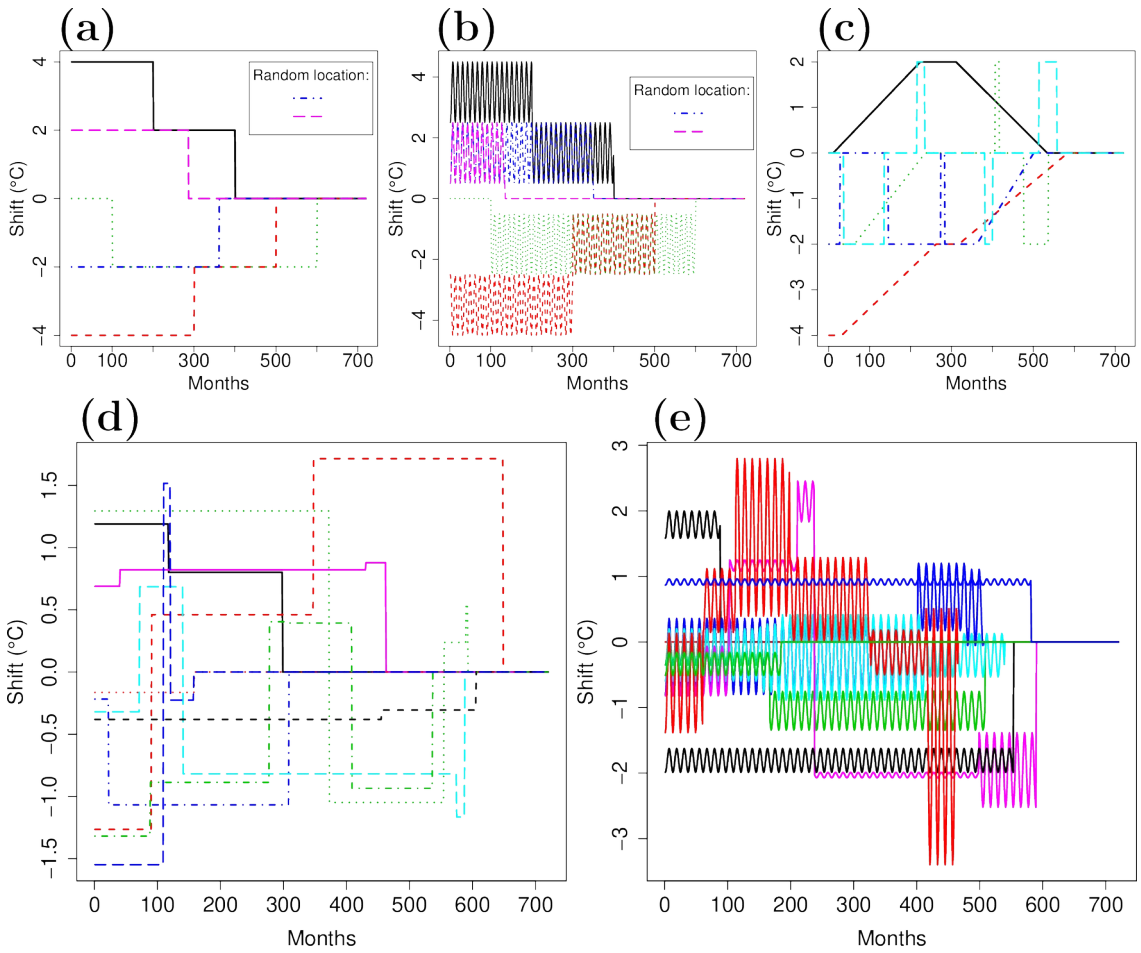


Figure 3. Examples of shifts applied to obtain the problem series in the first five temperature experiments: (a) One or two large shifts in the first five series; (b) the same as (a) but with an added strong sinusoidal seasonal variation; (c) large-size short-term biases and local trends in the first five series; (d) random number of biases of random magnitude and location in all 10 series; (e) the same as (d) but with sinusoidal seasonal variation of random amplitude.

3) Combinations of "short term platforms" (biases of small duration) and local trends (increments/decrements of 2 °C over short periods) were applied to the first five series of the samples.

4) A random number of shifts with random size and location was applied to any of the 10 series. The number of shifts was drawn from a binomial distribution with an average frequency of 5 per 100 years. Shift sizes were taken from a normal distribution with zero mean and one standard deviation.

5) As in (4) but with an added sinusoidal seasonality calculated as in (2). Shift values were then multiplied by 0.7 to compensate the increased deviations induced by the addition of the seasonal variation.

The second round of tests was done with the Tm2 master network to assess the performance of the methods with different seasonality shapes (two sinusoidal cycles per year and a squared seasonality consisting in an abrupt rise and ulterior sudden decrease to simulate potential effects in tropical savanna climates with a wet and a dry season), increasing network sizes (20, 40 and 80 series), nearly simultaneous shifts in 40, 70 and 100% of the series and networks of 40 series with many missing data.

In the third round, the methods were tested on networks of 40 series of 100 years length, from which data were deleted at the beginning of time series, around 1945 (simulating the lack of observations caused in Europe by World War II) and at random short spells elsewhere (see data availability in Figure 4). The number of shifts was 5 per century on average, and a sinusoidal variation of the biases with random amplitude was introduced with randomly located maximums between June and August.

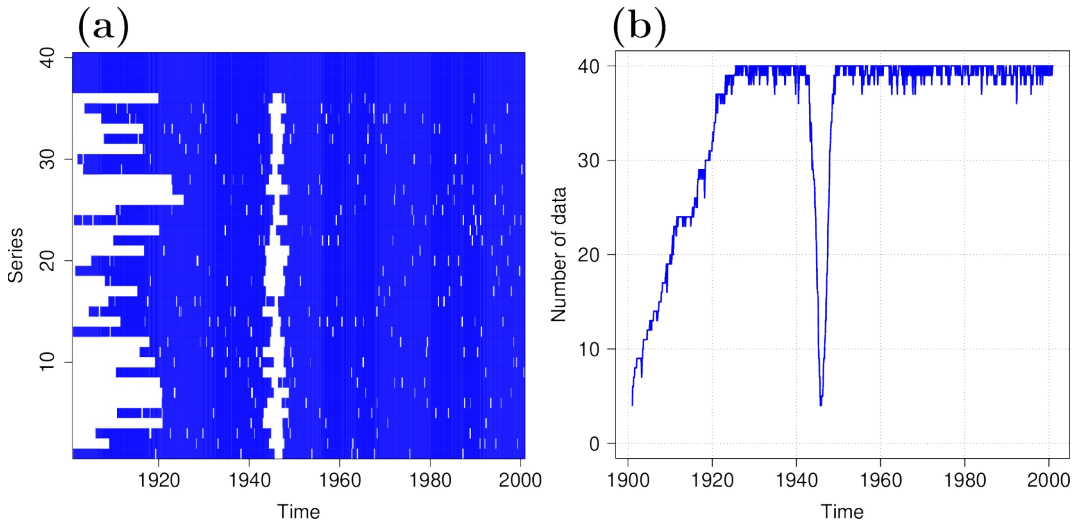


Figure 4. Data availability after a partial deletion in one of the samples of 40 series extracted from the master network Tr2. (a) Blue segments indicate the presence of data. (b) Total data availability along time. The last four series are free from missing data to ensure available references to infill the gaps present in the other series.

2.3. Tested homogenization packages

The tested homogenization packages were chosen among those more used in the literature, with the requirements that they should be freely available and able to be run automatically without any human intervention. Some of them could be tested in several ways, while others had only one mode of operation. In addition, the individual packages needed different implementations to run them in a uniform computer environment (a Linux PC by means of R and bash scripts). The list of tested packages and their running characteristics are as follows:

- ACMANT 4.3 (Domonkos, 2015 and 2020; Domonkos and Coll, 2017), in its versions for temperature (sinusoidal and irregular seasonality) and precipitation. These programs are MS-DOS executables, but can be run in Linux by means of the *wine* application (an API that allows Linux to run programs compiled for MS-Windows) in a rather straightforward way.
- Climatol 3.1 (Guijarro, 2016), with constant and variable corrections. Cubic root and logarithmic transformations were tested on the precipitation series. Written in R, no adaptation was needed.
- HOMER 2.6 (Mestre *et al.*, 2013) was the software outcome of HOME, but it could not be tested during that project. It is also written in R, but it is expected to be run interactively. Answers to the questions output by the program were redirected from a file, but this procedure only worked when these answers were supplied by means of the utility *expect*, simulating human intervention.
- MASH 3.03 (Szentimrey, 1999 and 2008) is a set of MS-DOS executables and batch scripts. The Manual of MASH specifies the required sequence of running programs for an unattended run, but the automatic connection between those programs is incomplete in the

package. Therefore, running MASH either needs manual interventions on the running process, or script editions before running. We did the necessary script editions to run MASH under Linux. The software functioned correctly, but with a too high computational time demand. Therefore, the running scripts had to be simplified by skipping the monthly adjustments and only results with yearly constant corrections are presented in this paper.

- RHtestsV4 (Wang, 2008; Wang and Feng, 2013) was applied in absolute and relative homogenization modes, both with and without quantile adjustment. This package is written in R, and therefore could be run without any adaptation. The problem in this case was that this software is designed to be applied to individual series, and it is the user who must provide a homogeneous reference series to perform a relative homogenization. Therefore, these tests were automated by using the closest homogeneous series as reference series in the first three temperature experiments, and the mean of all series in the other experiments.
- PHA v52d (Menne and Williams, 2005) implements a Pairwise Homogenization Algorithm which was developed for the homogenization of the temperature dataset of the United States Historical Climatology Network (USHCN). This package is available as Fortran sources (NCDC, 2012) which must be compiled, a process far from trivial because it depends on particular versions of Linux libraries. The version tested here is a bit outdated, but efforts to compile newer sources were unsuccessful or gave run-time errors. Sometimes this package did not infill missing data, returning the -9999 code instead. In these cases, those codes were substituted by their corresponding raw data.

In testing the packages, we had to tackle with some running errors: For instance, RHtests returned an error condition when no inhomogeneity was found in the tested series. In these cases, the problem series were taken as returned solutions. This strategy was generalized to all the methods whenever they failed to provide results. However, HOMER sometimes stopped with an irrecoverable error, yielding incomplete sets of solutions not readily comparable with those of the other methods.

2.4. Evaluation metrics

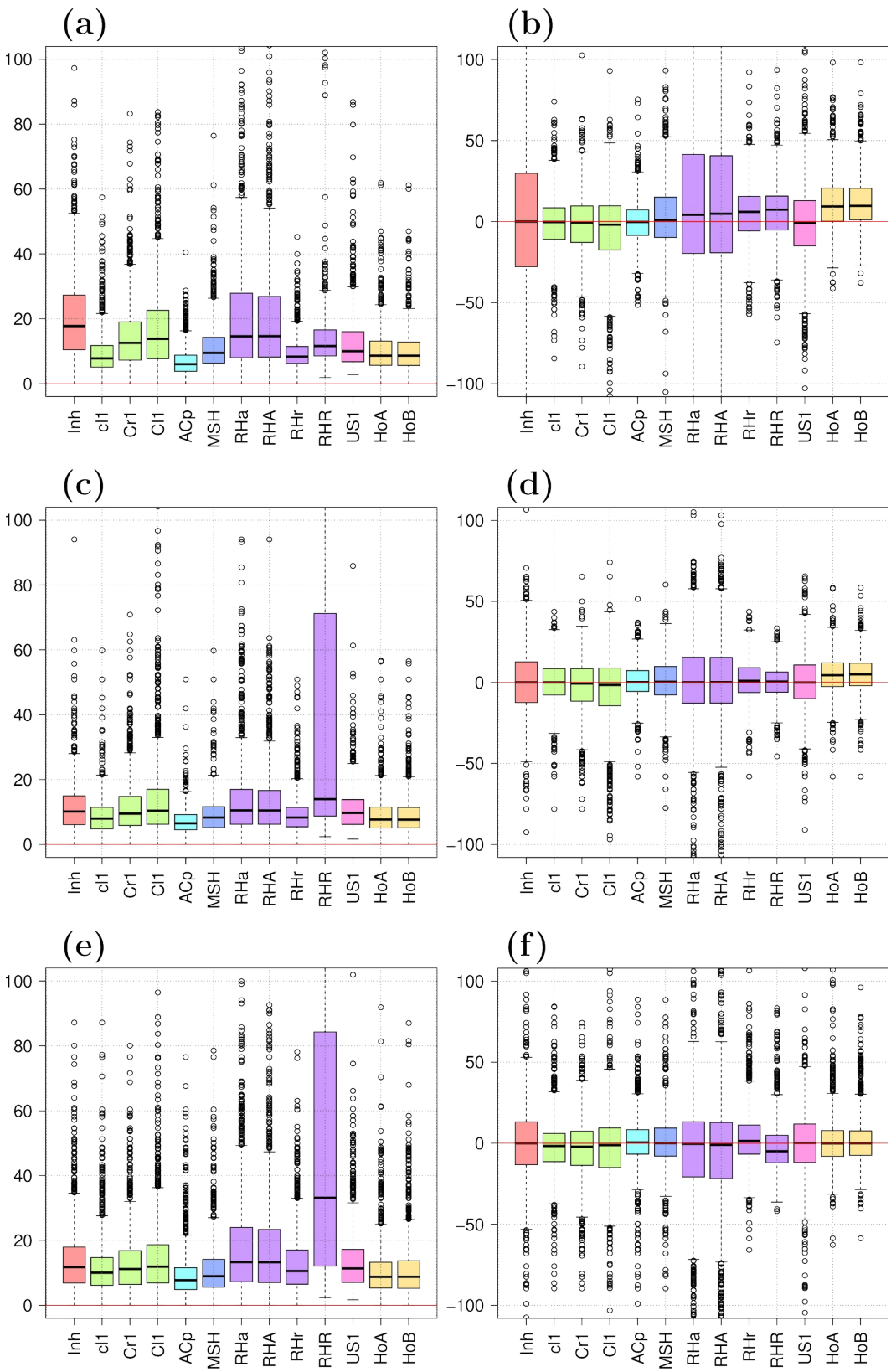
All tested packages were required to return homogenized solutions from each of the hundreds of inhomogeneous samples provided. These results were then compared with the homogeneous samples, calculating the Root Mean Squared Errors (RMSE) and the errors of the linear trends, means and standard deviations of the series. As the latter two metrics were not found to be relevant enough, only RMSE and trend errors will be discussed in this paper.

Having calculated these metrics hundreds of times, a convenient way of displaying them is with the use of box-whisker plots, which visualize both the more frequent values (the 50 % of data inside the box) and their whole range, outliers included. However, to avoid an excessive number of figures, other results will be shown in tables, ranking the tested packages according to their RMSE averages.

3. Results

3.1. Precipitation

Box-plots in Figure 5 summarize the results of the homogenization packages for the tested precipitation networks. The settings of packages applied in precipitation homogenization are presented in Table 1, and the mean RMSE of the homogenized series are shown in Table 2 by the rank order of this statistic. Most methods correct the inhomogeneities substantially in the Atlantic temperate precipitation regime, but not so well in the other two climates, where improvements are more modest, when existing.



Figures 5. RMSE (left column: a, c, e; in mm) and trend errors (right column: b, d, f; in mm per 100 years) of the homogenization of the three precipitation networks: Atlantic temperate (top row: a, b), Mediterranean (middle row: c, d) and Monsoonal (bottom row: e, f). (Fixed scale for a better comparison; outliers may lie outside the graphic limits.)

Table 1: Packages and settings tested on the three precipitation data-sets, with labels used in Figure 4 and Table 2.

Package	**Label**	**Settings**
None	Inh	(Inhomogeneous problem series)
Climatol	cl1	Normal ratio normalization of raw values (SNHT=15)
	Cr1	Full normalization of cubic root transformed values (SNHT=15)
	Cl1	Full normalization of log transformed values (SNHT=15)
ACMANT	ACp	Precipitation version
MASH	MSH	Multiplicative model and significance level of 0.01
RHTests	RHa	Absolute homogenization
	RHA	Absolute homogenization with quantile adjustment
	RHr	Relative homogenization
	RHR	Relative homogenization with quantile adjustment
PHA	US1	PHA method implemented for the USHCN data-set
HOMER	HoA	Pairwise detection, two rounds of joint detection/correction and month of change assessment
	HoB	HoA with an additional round of joint detection/correction

ACMANT scores first in all precipitation benchmarks, followed by the best settings of HOMER, Climatol, MASH and RHtests with little differences between the latter four methods. Although PHA was written to homogenize the USHCN temperature data-set, it could also improve the inhomogeneous precipitation series of the three simulated climates.

HOMER has a good performance with the two rounds of joint detection plus month assessment (HoA), but an additional round of joint detection (HoB) did not yield further improvement in their results.

Climatol only performs well with the normal ratio normalization of the raw values (cl1). The other two settings, in which data are transformed by a cubic root (Cr1) or a logarithmic function (Cl1) before being normalize with a full standardization (centering and scaling the data by their mean and standard deviation) produce much worse results. This is probably due to the amplification of errors produced when undoing the transformation through a cubic power or an exponential function respectively. Because of these results, the Climatol manual recommends not to use these options in the homogenization of variables with a highly skewed distribution of probabilities such as precipitation and wind (Guijarro, 2019). Due to the higher difficulty in the detection of inhomogeneities in precipitation series because of their high variability, the default threshold value of 25 for the Standard Normal Homogeneity Test (SNHT; Alexandersson, 1986) was lowered to 15 in these tests.

Table 2: Rank of the RMSE averages (mm) in the three climates Atlantic, Mediterranean and Monsoonal, and in the mean of all three. Errors of the inhomogeneous (Inh) problem series are enhanced in bold characters.

Rank	**Atlantic**		**Mediterranean**		**Monsoonal**		**Mean**	
1	ACp	7.24	ACp	7.46	ACp	10.43	ACp	8.38
2	RHr	9.64	cl1	9.05	HoA	11.39	HoA	10.39
3	cl1	9.68	HoA	9.27	MSH	11.61	HoB	10.41
4	HoB	10.07	HoB	9.33	HoB	11.83	cl1	10.45
5	HoA	10.51	MSH	9.39	cl1	12.62	MSH	10.82
6	MSH	11.46	RHr	9.48	RHr	13.74	RHr	10.95
7	US1	13.12	US1	11.27	Cr1	13.92	US1	12.89
8	Cr1	15.38	Cr1	11.78	US1	14.29	Cr1	13.69
9	Cl1	18.36	**Inh**	**11.98**	**Inh**	**14.81**	**Inh**	**15.96**
10	**Inh**	**21.08**	RHA	14.18	Cl1	16.29	Cl1	16.86
11	RHA	21.58	RHa	14.76	RHA	19.66	RHA	18.47
12	RHa	21.98	Cl1	15.93	RHa	20.69	RHa	19.14
13	RHR	21.98	RHR	38.15	RHR	54.59	RHR	38.24

MASH produced good results for all tested climates. Note that the lack of monthly correction implementation in our MASH scripts did not affect its results here, because no seasonality was introduced in the biases of the problem precipitation series. The sensitivity to detect inhomogeneities was also increased by using a significance level of 0.01.

RHtests gave good results when using reference series (relative homogenization mode, as in all the other tested methods) without quantile adjustment (RHr). RHtests absolute homogenization modes (RHa and RHA) were expected to produce bad results, but they were tested for confirmation. Quantile adjustments, when applied, worsened the results, especially in relative homogenization mode (RHR).

One of the expected benefits of homogenization is a greater spatial coherence of the long term trends of the series. Trend errors in Figure 5 (right column: b, d, f) show a clear reduction of the original error variability and unbiased mean trends in most of the homogenization results. The largest error reductions were achieved for the series of the Atlantic climate, with the exceptions of RHtests and HOMER. In addition, RHtests and HOMER sometimes produced notable systematic biases.

3.2. Temperature

As some settings of the packages differ according to climate variables, Table 3 lists them for temperature homogenization, updating the labels used in figures and tables when needed. These labels are common to all temperature benchmarking experiments, whose results are detailed in the following sub-sections.

Table 3: Packages and settings tested on the temperature benchmarking experiments, explaining the labels used in the related figures and tables.

Package	**Label**	**Settings**
None	Inh	(Inhomogeneous problem series)
Climatol	cl1	Centered raw values (no seasonality in the corrections)
	Cl1	Standardized raw values (seasonal variability in the corrections)
	Cl4	As Cl1, but using 30 additional short series
ACMANT	ACi	Irregular seasonal cycle of corrections
	ACs	Sinusoidal seasonal cycle of corrections
	Ai4	As ACi, but using 30 additional short series
	As4	As ACs, but using 30 additional short series
MASH	MSH	Without seasonality in the corrections. 0.05 significance level
RHTests	RHa	Absolute homogenization
	RHA	Absolute homogenization with quantile adjustment
	RHr	Relative homogenization
	RHR	Relative homogenization with quantile adjustment
PHA	US1	PHA method implemented for the USHCN temperature data-set
	US4	As US1, but using 30 additional short series
HOMER	Hoa	Pairwise detection, two rounds of joint detection/correction and month of change assessment
	Hob	Hoa with an additional round of joint detection/correction

3.2.1. First five experiments

Figure 6 shows the boxplots summarizing the RMSE of the tested methods in the first five experiments for the benchmark with intermediate difficulty (Tm2, with good and fair cross-correlations). The arrangement of panels (a) to (e) corresponds to that in Figure 3 for an easier interpretation.

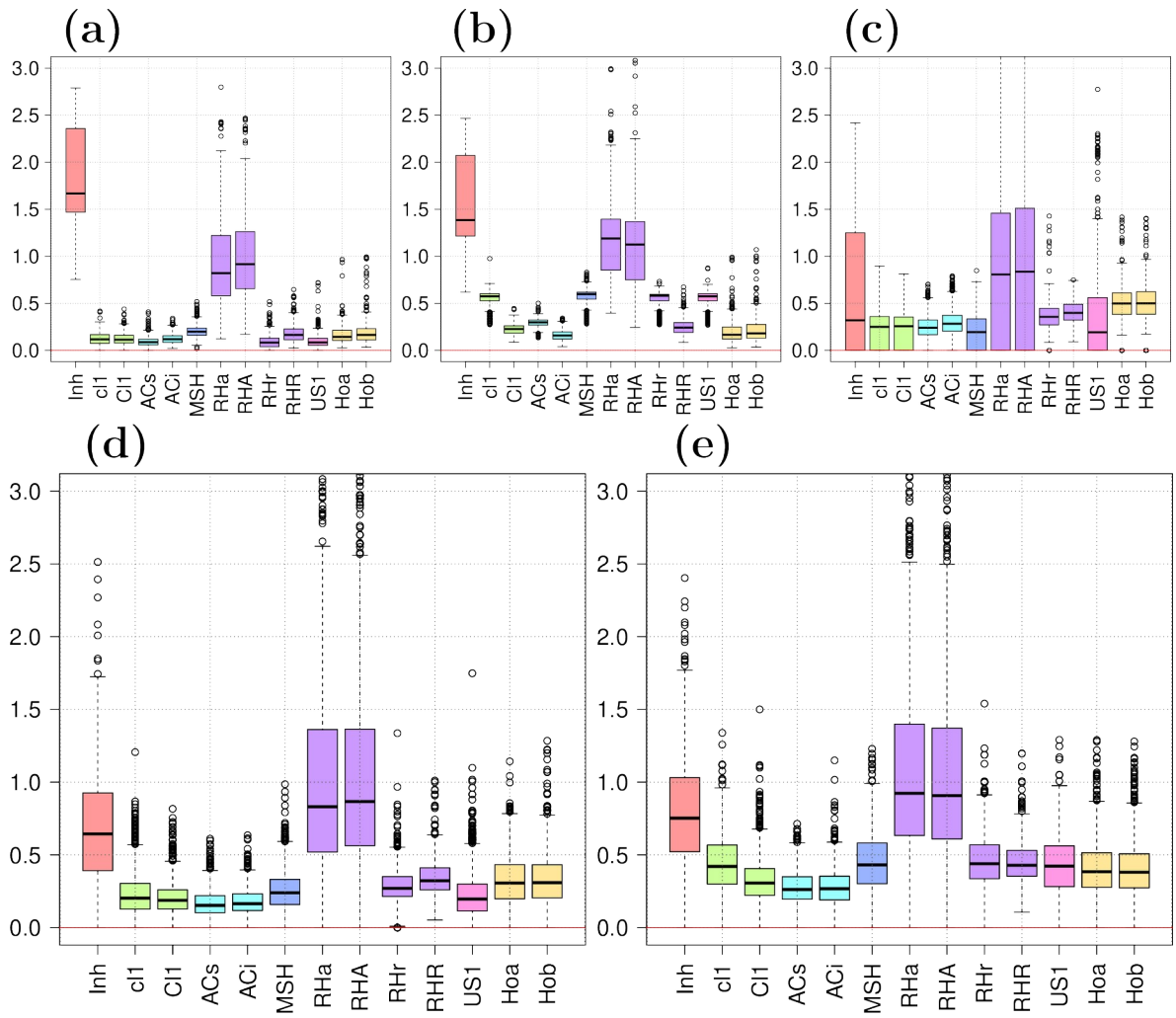


Figure 6. RMSE (°C) of the homogenization of temperature benchmark Tm2 (with good and fair cross-correlations) in the first five experiments, arranged as in Figure 3. (Fixed scale for a better comparison; outliers may lie outside the graphic limits.)

The few big shifts of the first two experiments (Figure 6a and 6b) are well detected and corrected by all relative homogenization methods. Even the absolute homogenization by RHtests corrected the series a bit, although this approach is discouraged (WMO, 2020). When the inhomogeneities have a strong seasonal component (Figure 6b), methods or settings allowing a seasonal variation of the corrections show substantially better performances. (As we could not test MASH with different monthly corrections, this method is penalized in Figures 6b and 6e.)

The third experiment, with large-size short term platforms and local trends, also shows good performances of all relative homogenization methods (Figure 6c), although with some notable differences between them. The lower RMSE averages were achieved by ACMANT, MASH, Climatol and RHtests.

Experiments four and five (Figures 6d and 6e) allow the presence of a random number of biases with random size in any of the ten series of each random sample network. They are therefore the most realistic experiments so far, especially experiment 5, in which sinusoidal seasonal variations of shifts with random amplitude were included. This seasonality penalized PHA and our MASH implementation, which perform very well when biases do not vary along the year. Anyway, the overall picture still shows the usefulness of all relative homogenization methods, with some differences between them that can be appreciated in the figures. Also worth noting is that, in all five

experiments, Climatol with full standardization of the data (Cl1) performed better than when data were only centered (cl1), even when biases have no seasonality.

The fifth experiment was also tried with samples of 40 series, keeping the first 10 series untouched and deleting between 30 to 50 years (50 to 83.3% of the data) in the other 30. The objective of this variation was to explore the potential benefit of using short lived series normally present in real databases but generally disregarded in homogenization practice, which tends to focus on the more complete and long records. Yet short series contain climate information that may improve the reliability of the homogenization of the longer series, especially when they are in their vicinity. In the experiments presented here, only ACMANT, Climatol and PHA could tolerate the large amount of missing data in the 30 short series. Their mean RMSE in all tests are presented in Table 4, with labels cl4, Cl4, Acs4, Aci4 and US4 in bold italic characters, to distinguish them from the other tests results. ACMANT gave the lowest mean RMSE in almost all tests, followed by Climatol. When the 30 additional short series were included, Climatol results improved and reached the level of ACMANT. ACMANT and PHA tolerated well the presence of the additional short series, but they almost never improved their RMSE reduction results for the 10 complete series by the added information of short series.

The trend errors of relative homogenizations in these five experiments are generally very low (not shown) and unbiased in the first two experiments. In the experiment with short term platform and trend inhomogeneities, the remaining trend errors are strongly positively biased (except for the HOMER results, which exhibit a strong but negative bias), although much less than the trend errors of the inhomogeneous series. For the last two more realistic experiments of this round, trend errors are again unbiased except for HOMER (negative again). The overall picture is that most methods (with the clear exception of absolute homogenization) reduce the spread of the trends in time series without introducing spurious biases.

3.2.2. Second round of experiments: variation of seasonal cycle, network size and coincidental breaks

Other seasonality shapes of biases than simple sinusoidal cycles have been tested introducing biases with double sinusoidal and squared cycles to the Tm2 master network. Table 5 compares the ranks and RMSE averages of these seasonality cycles with those of the simple sinusoidal results, suggesting that the shape of seasonality of biases has generally little impact on the rank order of the method performances. Again, settings that do not address differences in the corrections along the year appear penalized in the last places with higher errors, but only relative homogenization of RHtests or its quantile adjustment return greater errors than those in the problem series.

The remaining trend errors in the homogenized series are also similar for the cases of simple and double sinusoidal variation (not shown), but when biases have a squared seasonal variation, the biased trends in the problem series (Inh) are only partially corrected in the solutions returned by the tested methods. The lower remaining average trend bias is achieved by MASH, PHA, ACMANT, Climatol and RHtests in the case of double sinusoidal seasonality, with small differences between them, while in the case of a squared seasonality all returned solutions have biased trends, although those of PHA and ACMANT present smaller deviations than the other methods.

Another experiment was carried out to explore the influence of network size on the performance of the methods. Only samples of 10 series had been used so far (apart from the tests with 30 additional short series), but now their results are compared with those obtained with networks of 20, 40 and 80 series, always with the use of the Tm2 master network and the most realistic settings (i.e., random number, position and seasonality amplitude of the biases). RMSE averages for every network size and their overall means are displayed in Table 6, and the results suggest that network size has little effect on the rank order of method performances. An interesting feature is that the advantage of

ACMANT, Climatol and HOMER methods increases with growing network size. In the performed tests, RMSE of the three mentioned packages decreased by 30-35% when network size increased from 10 to 80 time series. By contrast, the RMSE reduction for growing network size was only 10–20% with the other homogenization methods and absent in some RHtests versions.

Table 4: Rank of the RMSE averages (°C) in the three temperature benchmarks Tm1, Tm2, Tm3, and in their means, for the experiment with random number, size and seasonality of biases. Errors of the inhomogeneous (Inh) problem series are enhanced in bold characters. Results using 30 additional short series are shown in bold italics (labels Ai4, As4, cl4, Cl4 and US4).

Rank	**Tm1**		**Tm2**		**Tm3**		**Mean**	
1	ACi	0.24	ACs	0.28	***Cl4***	***0.48***	ACs	0.34
2	ACs	0.24	ACi	0.29	***As4***	***0.49***	***Cl4***	***0.35***
3	***Cl4***	***0.26***	***Cl4***	***0.30***	ACs	0.50	ACi	0.35
4	***As4***	***0.28***	***As4***	***0.31***	***Ai4***	***0.52***	***As4***	***0.36***
5	Cl1	0.28	Cl1	0.34	ACi	0.53	Cl1	0.39
6	***Ai4***	***0.30***	***Ai4***	***0.34***	cl4	0.53	***Ai4***	***0.39***
7	Hob	0.36	Hob	0.41	Cl1	0.54	Hob	0.46
8	Hoa	0.37	Hoa	0.41	RHr	0.54	Hoa	0.46
9	***US4***	***0.40***	***cl4***	***0.44***	MSH	0.54	***cl4***	***0.47***
10	US1	0.40	US1	0.45	cl1	0.56	cl1	0.48
11	***cl4***	***0.42***	cl1	0.46	***US4***	***0.58***	MSH	0.48
12	cl1	0.43	MSH	0.47	Hob	0.60	US1	0.48
13	MSH	0.44	RHr	0.47	Hoa	0.60	RHr	0.49
14	RHr	0.45	***US4***	***0.53***	US1	0.60	***US4***	***0.50***
15	RHR	0.69	**Inh**	**0.82**	**Inh**	**0.82**	**Inh**	**0.82**
16	**Inh**	**0.82**	RHR	0.83	RHa	1.09	RHA	1.06
17	RHA	0.99	RHA	1.10	RHA	1.10	RHa	1.07
18	RHa	1.00	RHa	1.11	RHR	2.23	RHR	1.25

Table 5: Ranks and RMSE averages of the tested packages on Tm2 samples with seasonal cycles single sinusoidal, double sinusoidal and squared in the introduced biases. Errors of the inhomogeneous (Inh) problem series are enhanced in bold characters.

Rank	**Single**		**Double**		**Squared**	
1	ACs	0.28	ACs	0.17	ACi	0.31
2	ACi	0.29	ACi	0.19	ACs	0.35
3	Cl1	0.34	Cl1	0.21	Cl1	0.40
4	Hoa	0.41	cl1	0.24	Hoa	0.44
5	US1	0.45	US1	0.24	US1	0.45
6	cl1	0.46	MSH	0.27	MSH	0.49
7	MSH	0.47	RHr	0.29	cl1	0.49
8	RHr	0.47	Hoa	0.34	RHr	0.58
9	**Inh**	**0.82**	RHR	0.67	**Inh**	**0.92**
10	RHR	0.83	**Inh**	**0.70**	RHR	0.92
11	RHA	1.10	RHa	1.03	RHA	1.15
12	RHa	1.11	RHA	1.06	RHa	1.16

Trend errors (not shown) generally decrease slightly with growing network size, but HOMER is an exception: The high biases produced in the homogenization of 10-series networks vanished when larger networks were tested.

In the last experiment of this round, the Tm2 master network was used to check the performance of the homogenization packages when shifts in the mean of the series have the same sign and are concentrated in a short period of time (up to a decade). This situation may occur when observation practices change in a whole network to adapt to changes in technology (as, e.g., introducing the Stevenson shelter or changing manned stations to automatic observing systems). Figure 7 shows the RMSE box-plots when such simultaneous biases take place in 40, 70 or 100 % of the networks (4, 7 or all of the 10 series samples).

When near simultaneous changes affect only a 40 % of the network series, most methods still perform well, but when 70 or 100 % of the series are affected, the superiority of the pairwise time series comparison methods (PHA and HOMER) becomes evident, although ACMANT also show some resistance to produce biased results. The same considerations can be applied to the trend errors (right column in Figure 7). When all series of a network are simultaneously affected by the same bias, relative homogenization methods will not detect this change because it will be taken as real climate variation. In these extreme case absolute homogenization ability to detect biases will give better results.

Table 6: Ranks and RMSE averages of the tested packages on Tm2 samples of 10, 20, 40 and 80 series, plus the overall mean values. Errors of the inhomogeneous (Inh) problem series are enhanced in bold characters.

Rank	**10 series**		**20 series**		**40 series**		**80 series**		**Mean**	
1	ACs	0.28	ACs	0.23	ACs	0.22	ACi	0.19	ACs	0.23
2	ACi	0.29	ACi	0.24	ACi	0.23	ACs	0.20	ACi	0.24
3	Cl1	0.34	Cl1	0.27	Cl1	0.25	Cl1	0.22	Cl1	0.27
4	Hob	0.41	Hoa	0.29	Hob	0.28	Hob	0.27	Hob	0.31
5	Hoa	0.41	Hob	0.29	Hoa	0.28	Hoa	0.27	Hoa	0.31
6	US1	0.45	US1	0.39	MSH	0.39	MSH	0.38	cl1	0.41
7	cl1	0.46	MSH	0.40	cl1	0.39	cl1	0.38	MSH	0.41
8	MSH	0.47	cl1	0.40	US1	0.40	RHr	0.39	US1	0.41
9	RHr	0.47	RHr	0.42	RHr	0.40	US1	0.40	RHr	0.42
10	**Inh**	**0.82**	**Inh**	**0.79**	**Inh**	**0.78**	**Inh**	**0.78**	**Inh**	**0.79**
11	RHR	0.83	RHR	0.97	RHA	1.08	RHa	1.06	RHA	1.07
12	RHA	1.10	RHA	1.04	RHa	1.08	RHA	1.06	RHa	1.07
13	RHa	1.11	RHa	1.04	RHR	1.19	RHR	1.41	RHR	1.10

3.2.3. Third round of experiments: 100 years series

Homogenization packages were tested on networks of 20 and 40 series of 100 years length, with missing data mimicking the characteristics of the HOME benchmark. Figure 8 shows the RMSE (left panels: a, c) and trend errors (right panels: b, d) obtained from 100 runs. As MASH and HOMER gave errors due to the presence of missing data, their box-plots appear in white with the same dimensions as those of the problem series (Inh). Results from the 20 series samples (Figures 8a and 8b) show that ACMANT and PHA correct the problem series better, followed by Climatol. RHtests apply substantial corrections in absolute mode, but that is not the case when using reference series, probably due to some errors linked to the presence of missing data.

The bottom row of Figure 8 displays the results for the 40-series networks whose inhomogeneities are of a lower magnitude and complicated by concentrated positive shifts around 1975 in 36 of the series (90 %). In this case, the best results are achieved by ACMANT, Climatol and PHA.

Trend errors (Figure 8d) are reduced in all successfully tested methods, except for the absolute homogenization mode of RHtests. However, the network-wide trend induced by the concentrated positive shifts is considered to be a part of the climate signal, resulting in systematically biased trend estimations by all of the tested methods.

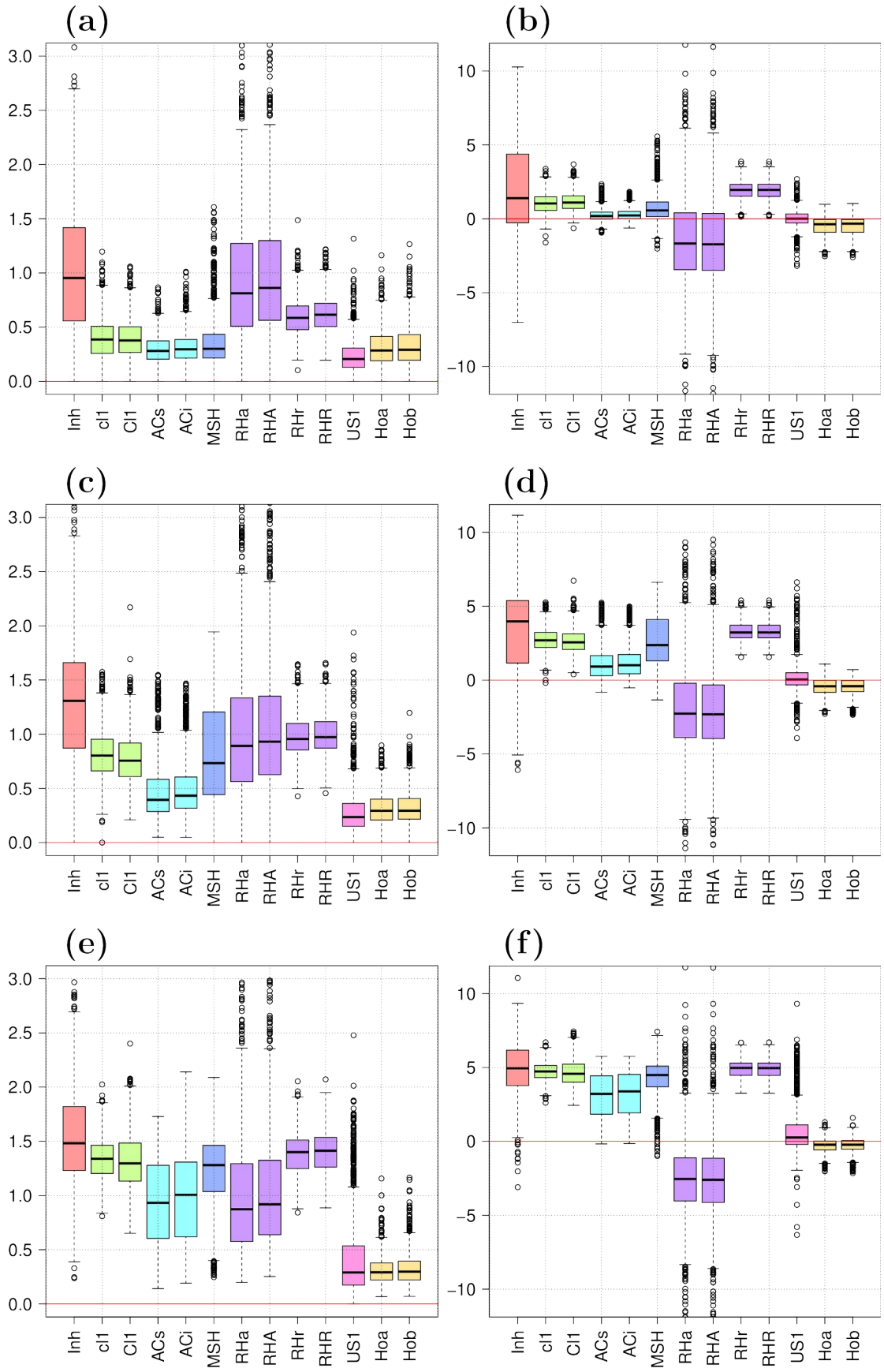


Figure 7. RMSE (left column: a, c, e; in °C) and trend errors (right column: b, d, f; in °C per 100 years) in case of concentrated biases in 40 (top row: a, b), 70 (middle row: c, d) and 100 % (bottom row: e, f) of the series.

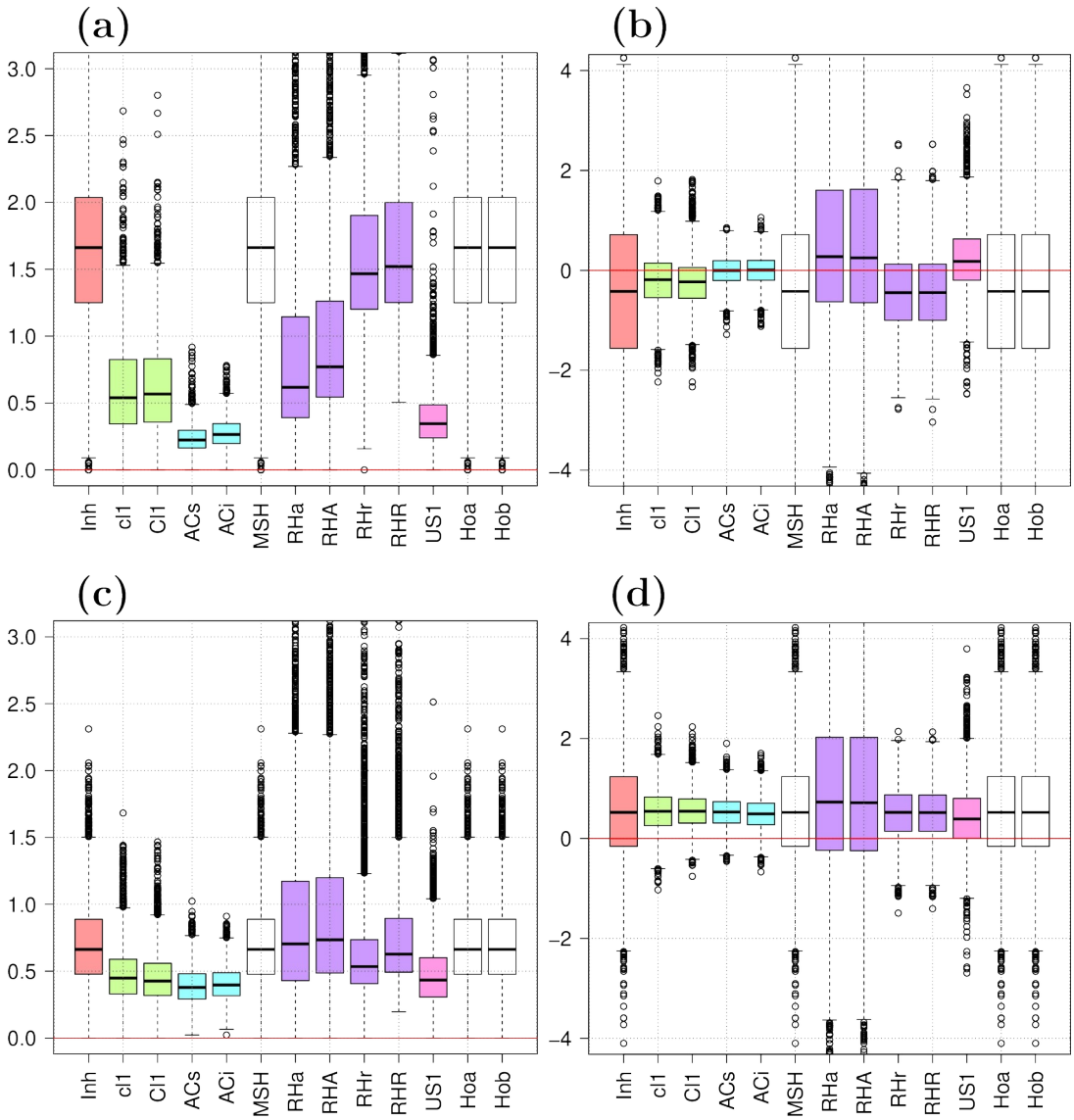


Figure 8. RMSE (left column: a, c; in °C) and trend errors (right column: b, d; in °C per 100 years) of the test results of the 100 year series with missing data. In the top row (a, b), networks contained 20 series, while in the bottom row (c, d) networks included 40 series and concentrated biases were applied to 90 % of them (36 series). (MASH and HOMER crashed because of the missing data, hence their boxplots replicate, in white color, those of the problem series, Inh.)

4. Discussion

In this work we updated some of the method intercomparison results of the HOME initiative, with two important differences:

(i) The characteristics of the homogeneous datasets in HOME were derived from Central European observing networks (Venema et al., 2012). As a result, cross-correlations were much better than those found in other geographical areas with different climates, more complex orography and less dense station distribution. Here, precipitation series mimic three different climates and the temperature datasets have different degrees of cross-correlations, including more adverse situations.

(ii) In order to repeat the tests many times, sometimes with different settings, only methods that could be run in a completely automatic way were tested. However, the tested methodologies include the best known and more used methods currently implemented as publicly available computer packages, including the software HOMER produced (but not tested) by HOME. The main missing

software was AnClim (Štěpánek, 2008), a too Windows oriented package as to be automated in a Linux environment. But more than a method it is a software suite that implements some of the methodologies evaluated here.

One of the praised characteristics of the HOME benchmark tests was its blind mode (Venema et al., 2012), since none of the participants homogenizing the problem datasets knew the homogeneous versions of the series. In our tests, the generation of the homogeneous master networks was the first step, and although the master networks were not hidden to the tester, automatic algorithms randomly sampled the homogeneous series and inserted inhomogeneities into them without using any "a priory" knowledge. Therefore, our experiments can also be included in the blind category.

One thing worth noting is that, apart from a few variations in some methods, they have been run using their default parameters. Some packages (e.g., Climatol) admit a high degree of tuning to adapt to different variables and time scales, while others run with no or few options of change. Moreover, to automate the application of some packages, especially Rhtests and HOMER, whose primary operation is manual, scripts had to be developed that may not be optimal. Therefore, in all cases an experienced user might obtain better results on applications of these software packages to specific problems than these automatic runs. However, these experiments illustrate several aspects of homogenization practices, as discussed in the following paragraphs.

RHtests results confirm the recommendation of avoiding absolute homogenization unless there are no reliable reference series available (Venema et al., 2012). This is often the case of remote islands and extremely data-sparse regions, but even in those cases it may be worth exploring the possibility of using reanalysis series as references.

In its relative homogenization mode, this package yields good results, although the quantile adjustment produced huge errors in the precipitation tests. For temperature, it improved the results very much when there was a marked seasonality in the inhomogeneities, otherwise the impact of the quantile adjustment was negligible. Note that a more correct evaluation of the performance of quantile adjustments would need tests focusing more on the homogenization accuracy of extreme values.

Most methods detect break-points with a significant threshold of $\alpha=0.05$. Although there are publications on critical values of SNHT for various significant levels (Alexandersson, 1986; Khaliq and Ouarda, 2007), Climatol application experiences showed that the values of this test are highly dependent of the climatic variable and geographic characteristics of the area under study, and therefore Climatol applies a conservative default value of SNHT=25 for monthly data, which the user can modify based on the anomaly graphs and histogram of residual SNHT of the homogenized series provided in its graphic output.

Very different tolerance to missing data has been observed during these experiments. HOMER and MASH refused to work even with a moderate proportion of gaps (Figure 8), and only ACMANT, Climatol and PHA could make use of additional short series (section 3.2.3 and Table 4). From these three methods, only Climatol will always provide estimations for all missing data. Although ACMANT and PHA avoid the estimation of missing data when reference series does not have a minimum correlation with the problem series (0.4 in the case of ACMANT) or the number of neighbor series is too low, Climatol does not impose such limits. As a result, the estimated data can be affected by substantial errors when correlations are poor (a condition met when the closest reference series are very far away), but the population of the filled in data are expected to have a probability distribution similar to that of the observed values.

The PHA algorithm and a similar pairwise procedure included in HOMER showed their increased ability to detect biases in the tested series when the same kind non-climatic shifts affects many time series of the network within a short period of time (Figure 7). This condition may arise when changes in observing practices are applied to the whole network (e.g., changes in the thermometric

shelter or from manual to automatic instruments). In these cases, they outperform any of the other methods, since concurrent variations are confounded with the true climate signal when a single combined series is used as reference. However, this does not preclude the use of the other methods in these cases provided that series unaffected by the same biases (e.g., from a neighbor country or from a reanalysis) are included as references. (Reanalysis has already been successfully used to homogenize climate series in works as those of Gonzalez et al., 2018 and Azorin-Molina et al., 2019.) Moreover, the use of composite reference series generally provides higher signal-to-noise ratio than pairwise comparisons, and this may be one reason why ACMANT and Climatol often outperformed PHA and HOMER. Also note that the joint detection routine of the HOMER method functions in absolute homogenization mode, which is a known problem (Mestre et al., 2013; Gubler et al., 2017) affecting the performance of HOMER.

Other metrics commonly used in homogenization studies have been the Centered Root Mean Squared Error (CRMSE; Gubler et al., 2017; Joelsson et al., 2022; Killick et al., 2021) and the Pearson correlation coefficient (Coscarelli et al., 2021) between the homogenized and the problem series. Both have been disregarded here because they can be misleading, since when a series is added to or multiplied by a big number it would still show a good result in terms of CRMSE or correlation coefficient, respectively.

Another consideration we would like to note is that the results and rankings here exposed may not be the only criteria to choose a homogenization program, since the user may also be influenced by other characteristics, such as their operation ways (underlying programming environment, automation, format of the input files, etc), openness of the code, tolerance to missing data, availability of guidance manuals and variety of output products. Moreover, the final decision must take into account the special characteristics of the network under study, climate variable, time resolution and other climatic and geographical features.

Finally, we wish to refer to the increasing interest on the homogenization of daily series (Szentimrey, 2013), which is needed for assessments on the variability and trends of extreme values. Therefore, benchmarking of homogenization methods applied to daily series is currently an active field of study (Killick, 2016 and 2021; Skrynyk et al., 2020; Guijarro, 2019b). While some of the experiences in the homogenization of monthly series can be applied also to the daily resolution, the metrics to evaluate the performance of the methods may not be all appropriate for the daily series. This is the case of the RMSE, which when correlations are low reaches higher scores when corrections converge to the mean value of the series, with the undesired effect of lowering its variability, hence hindering the usefulness of the series to reliably estimate the probability of extreme values.

5. Conclusions

The automatic application of the main publicly available homogenization software packages has allowed the comparison of their performances when applied to monthly precipitation and temperature synthetic series with different climate and inhomogeneity characteristics. The main conclusions of this study can be summarized as follows:

i ACMANT, followed by Climatol, gave the best results in almost all tested conditions and datasets. However, when concurrent substantial biases are concentrated in a short period of time, the pairwise algorithms of PHA and HOMER outperformed them.

ii Therefore, relative homogenization methods relying on synthetic reference series must take care of using series alien to the studied network when similar changes concentrated in time are suspected in most of the series. These additional series can be observational series of a neighboring region or series derived from a reanalysis.

iii HOMER and MASH appear as the less tolerant methods to a substantial presence of missing data in the series. Other packages admit a higher proportion of gaps, but not to the extent of Climatol, ACMANT and PHA, which can make use of the short series usually available in real observational meteorological networks.

iv Denser networks of time series generally facilitate more accurate homogenization results, as happened with ACMANT, Climatol and HOMER. The inclusion of short neighbor series in the homogenization of longer series resulted in notable improvement only in the accuracy of the Climatol results.

v Relative homogenization generally does not introduce biases into the series and by correcting inhomogeneities the spatial coherence of climate trends can be substantially improved.

vi The variety of results obtained in this project may serve as a guidance for choosing a homogenization method, but users should consider the special characteristics of a given homogenization task like network properties, climate variable, time resolution, metadata information and further climatic and geographical features.

Acknowledgements

This work was supported by the MULTITEST (Multiple verification of automatic software homogenizing monthly temperature and precipitation series; CGL2014-52901-P) project, funded by the Spanish Ministry of Economy and Competitivity. Thanks to Met Éireann for providing the Irish monthly precipitation series that served as model to synthesize the network of Atlantic Temperate precipitations. Mallorca monthly precipitations were taken from AEMET data bases, and monthly precipitations from SW India, gridded at 0.5° resolution, were obtained from the Global Precipitation Climate Center (GPCC). Many thanks for the advice received on the settings needed to run MASH (Tamás Szentimrey) and HOMER (Gregor Vertacnik, John Coll and Stefanie Gubler). The authors are grateful to the three anonymous reviewers for their constructive and helpful comments to the original manuscript.

References

Aguilar E, Auer I, Brunet M, Peterson TC, Wieringa J. 2003. Guidelines on Climate Metadata and Homogenization. World Meteorological Organisation, 52. http://www.wmo.int/datastat/documents/WCDMP-53_1.pdf (accessed 1 September 2018).

Alexandersson H. 1986. A homogeneity test to precipitation data. *Int. J. Climatol.* **6**(6): 661–675. doi: 10.1002/joc.3370060607.

Azorin-Molina C, Guijarro JA, McVicar TR, Trewin BC, Frost AJ, Chen D. 2019. An approach to homogenize daily peak wind gusts: An application to the Australian series. *Int. J. Climatol*., 39:2260-2277, DOI: 10.1002/joc.5949.

Caussinus H, Mestre O, 2004. Detection and correction of artificial shifts in climate series. J. Roy. Stat. Soc. C 53: 405-425. doi: 10.1111/j.1467-9876.2004.05155.x

Coll J, Domonkos P, Guijarro J, Curley M, Rustemeier E, Aguilar E, Walsh S, Sweeney J. 2020. Application of homogenization methods for Ireland's monthly precipitation records: comparison of break detection results. *Int. Jour. Climatol*., 40:6169-6188, DOI: 10.1002/joc.6575

Conrad V, Pollack LW. 1950. *Methods in Climatology*. Harvard Univ. Press, Cambridge-Massachusetts, xi+459 pp.

Coscarelli R, Caroletti Gn, Joelsson M, Engström E, Caloiero T. 2021. Validation metrics of homogenization techniques on artificially inhomogenized monthly temperature networks in Sweden and Slovenia (1950–2005). *Scientific Reports*, 11:18288, https://doi.org/10.1038/s41598-021-97685-7

Domonkos P. 2011. Efficiency evaluation for detecting inhomogeneities by objective homogenisation methods. *Theor. Appl. Climatol.*, 105, 455-467, doi: 10.1007/s00704-011-0399-7.

Domonkos P. 2015. Homogenization of precipitation time series with ACMANT. *Theor. Appl. Climatol.*, 122:303-314.

Domonkos P, Coll J. 2017. Homogenisation of temperature and precipitation time series with ACMANT3: method description and efficiency tests. Int. J. Climatol., 37:1910–1921, DOI: 10.1002/joc.4822

Domonkos P. 2020. ACMANTv4: Scientific content and operation of the software. 71 pp.

https://github.com/dpeterfree/ACMANT (Accessed on December 2022).

Domonkos P, Guijarro JA, Venema V, Brunet M, Sigró J. 2021. Efficiency of time series homogenization: method comparison with 12 monthly temperature test datasets. J. of Climate, 49 pp, DOI: https://doi.org/10.1175/JCLI-D-20-0611.1

Domonkos P, Tóth R, Nyitrai L. 2022. Climate observations: Data quality control and time series homogenization. Elsevier, 302p.

Gonzalez S, Vasallo F, Recio-Blitz C, Guijarro JA, Riesco J. 2018. Atmospheric Patterns over the Antarctic Peninsula. *Jour. of Climate*, 31:3597-3608, DOI: 10.1175/JCLI-D-17-0598.1.

Gubler S, Hunziker S, Begert M, Croci-Maspoli M, Konzelmann T, Brönnimann S, Schwierz C, Oria C, Rosas G. 2017. The influence of station density on climate data homogenization. Int. J. Climatol., 137:4670-4683. doi: 10.1002/joc.511

Guijarro JA. 2011. Influence of network density on homogenization performance. *Seventh Seminar for Homogenization and Quality Control in Climatological Databases jointly organized with the Meeting of COST ES0601 (HOME) Action MC Meeting*, Budapest, 24-27/October, WMO WCDMP-No. 78, pp. 11-18.

Guijarro JA. 2013. Temperature trends. In *Adverse weather in Spain* (García-Legaz C and Valero F, Eds.), AMV Ediciones, Madrid, ISBN 978-84-96709-43-0, pp. 297-306.

Guijarro JA. 2016. Package 'climatol', version 3.0. https://cran.r-project.org/src/contrib/Archive/climatol/climatol_3.0.tar.gz (Accessed in June 2019).

Guijarro JA. 2019. Homogenization of climate series with Climatol. http://www.climatol.eu/homog_climatol-en.pdf (Accessed in December 2019).

Guijarro JA. 2019b. Recommended Homogenization Techniques based on Benchmarking Results. INDECIS project, Work Package 3, Deliverable 3.2-b, 9 pp. http://www.indecis.eu/docs/Deliverables/Deliverable_3.2.b.pdf (Accessed in June 2022).

Hunziker S, Brönnimann S, Calle J, Moreno I, Andrade M, Ticona L, Huerta A, Lavado-Casimiro W. 2018. Effects of undetected data quality issues on climatological analyses. *Clim. Past,* 14:1-20. doi:10.5194/cp-14-1-2018

Joelsson LMT, Sturm C, Södling J, Engström E, Kjellström E. 2022. Automation and evaluation of the interactive homogenization tool HOMER. Int. J. Climatol. 42:2861-2880.

Khaliq MN, Ouarda TBMJ. 2007. On the critical values of the standard normal homogeneity test (SNHT). *Int. J. Climatol.*, **27**:681-687. doi: 10.1002/joc.1438

Killick RE. 2016. Benchmarking the Performance of Homogenisation Algorithms on Daily Temperature Data, PhD Thesis, University of Exeter, 249 pp. http://hdl.handle.net/10871/23095 (last accessed 1 September 2018).'

Killick R, Jolliffe It, Willett KM. 2021. Benchmarking the performance of homogenisation algorithms on daily temperature data. Int. J. Climatol., 41 pp., doi: 10.1002/joc.7462

López JA, Guijarro JA, Aguilar E, Domonkos P, Brunet M. 2016. Una propuesta metodológica para la generación de redes de precipitación simuladas a partir de redes de precipitación observadas en el marco del proyecto MULTITEST. In Olcina J, Rico AM, Moltó E (eds.): *Clima, sociedad, riesgos y ordenación del territorio*, University of Alicante (Spain), Asociación Española de Climatología, ISBN 978-84-16724-19-2, pp. 183-194. (In Spanish with an English abstract).

Mamara A, Argiriou AA, Anadranistakis M, 2013. Homogenization of mean monthly temperature time series of Greece. *Int. J. Climatol.*, 33:2649-2666, DOI: 10.1002/joc.3614.

Menne MJ, Williams CN Jr. 2005. Detection of undocumented changepoints using multiple test statistics and composite reference series. *J. Climate* 18:4271-4286.

Mestre O, Domonkos P, Picard F, Auer I, Robin S, Lebarbier E, Boehm R, Aguilar E, Guijarro J, Vertachnik G, Klancar M, Dubuisson B, Štěpánek P. 2013. HOMER: a homogenization software—methods and applications. *Időjárás* **117**(1):47–67.

NCDC. 2012. ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/v3/software/52i/phav52i.tar.gz (last accessed September 2019).

Pebesma EJ. 2019. The meuse data set: a brief tutorial for the gstat R package. https://cran.r-project.org/web/packages/gstat/vignettes/gstat.pdf

Peterson TC, Easterling DR, Karl TR, Groisman P, Nicholls N, Plummer N, Torok S, Auer I, Boehm R, Gullett D, Vincint L, Heino R, Tuomenvirta H, Mestre O, Szentimrey T, Salinger J, Førland EJ, Hanssen-Bauer I, Alexandersson H, Jones PD, Parker D. 1998. Homogeneity adjustments of in situ atmospheric climate data: a review. *Int. J. Climatol*. **18**(13):1493-1517. doi: 10.1002/(SICI)1097-0088(19981115)18:13<1493::AID-JOC329>3.0.CO;2-T

R Core Team. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/ (last accessed in June 2019).

Ribeiro S, Caineta J, Costa AC, 2016. Review and discussion of homogenisation methods for climate data. *Phys. Chem. Earth* Parts ABC 94, 167-179. https://doi.org/10.1016/j.pce.2015.08.007

Schneider U, Becker A, Finger P, Meyer-Christoffer A, Rudolf B, Ziese M. 2015. GPCC Full Data Reanalysis Version 7.0 at 0.5°: Monthly Land-Surface Precipitation from Rain-Gauges built on GTS-based and Historic Data. DOI: 10.5676/DWD_GPCC/FD_M_V7_050.

Štěpánek P. 2008. AnClim - software for time series analysis. Dept. of Geography, Faculty of Natural Sciences, Masaryk University, Brno, Czech Republic. Štěpánek

Szentimrey T. 1999. Multiple Analysis of Series for Homogenization (MASH). *Proc. Second Seminar for Homogenization of Surface Climatological Data*, Budapest, Hungary, WMO, WCDMP-No. 41, 27–46. http://www.dmcsee.org/uploads/file/331_2_mashmanual.pdf (last accessed 1 September 2018)

Szentimrey T. 2008. The manual of Multiple Analysis of Series for Homogenization (MASH). Hungarian Meteorological Service, Budapest, Hungary. http://www.met.hu/pages/seminars/seeera/ index.htm (last accessed 1 September 2018)

Szentimrey T. 2013. Theoretical questions of daily data homogenization. *Időjárás* **117**(1):113-122.

Skrynyk O, Aguilar E, Guijarro J, Randriamarolaza LYA, Bubin S. 2020. Uncertainty evaluation of Climatol's adjustment algorithm applied to daily air temperature time series. Int. J. Climatol., 25 pp, https://doi.org/10.1002/joc.6854

Venema V, Mestre O, Aguilar E, Auer I, Guijarro Ja, Domonkos P, Vertacnik G, Szentimrey T, Štěpánek P, Zahradnicek P, Viarre J, Müller-Westermeier G, Lakatos M, Williams Cn, Menne M, Lindau R, Rasol D, Rustemeier E, Kolokythas K, Marinova T, Andresen L, Acquaotta F, Fratianni S, Cheval S, Klancar M, Brunetti M, Gruber C, Prohom Duran M, Likso T, Esteban P and Brandsma T. 2012. Benchmarking homogenization algorithms for monthly data. *Clim. Past* **8**: 89-115. doi: 10.5194/cp-8-89-2012

Wang XL. 2008. Penalized maximal F test for detecting undocumented mean shift without trend change. *J. Atmos. Oceanic Technol.* **25**(3): 368–384. doi: 10.1175/2007JTECHA982.1

Wang XL, Feng Y. 2013. RHtestsV4 User Manual. http://etccdi.pacificclimate.org/software.shtml, 29 pp. (Accessed in June 2019).

WMO, 2020. Guidelines on Homogenization. World Meteorological Organization, WMO-No. 1245, 54 pp, Geneva.