Comparison of homogenization methods for daily temperature series against an observation-based benchmark dataset

Antonello Angelo Squintu^{1,2} · Gerard van der Schrier¹ · Petr Štěpánek^{3,4} · Pavel Zahradníček^{3,4} · Albert Klein Tank^{1,2,5}

Received: 25 June 2019 / Accepted: 30 September 2019 / Published online: 7 January 2020
© The Author(s) 2020

Abstract

Homogenization of daily temperature series is a fundamental step for climatological analyses. In the last decades, several methods have been developed, presenting different statistical and procedural approaches. In this study, four homogenization methods (together with two variants) have been tested and compared. This has been performed constructing a benchmark dataset, where segments of homogeneous series are replaced with simultaneous measurements from neighboring homogeneous series. This generates inhomogeneous series (the test set) whose homogeneous version (the benchmark set) is known. Two benchmark datasets are created. The first one is based on series from the Czech Republic and has a high quality, high station density, and a large number of reference series. The second one uses stations from all Europe and presents more challenges, such as missing segments, low station density, and scarcity of reference series. The comparison has been performed with pre-defined metrics which check the statistical distance between the homogenized versions and the benchmark. Almost all homogenization methods perform well on the near-ideal benchmark (maximum relative root mean square error (rRMSE): 1.01), while on the European dataset, the homogenization methods diverge and the rRMSE increases up to 1.87. Analyses of the percentages of non-adjusted inhomogeneous data (up to 39%) and substantial differences in the trends among the homogenized versions helped identifying diverging procedural characteristics of the methods. These results add new elements to the debate about homogenization methods for daily values and motivate the use of realistic and challenging datasets in evaluating their robustness and flexibility.

1 Introduction

Homogenization of climatic data is a fundamental step in any climatological analyses. Human intervention on measuring stations induces sharp or gradual changes to temperature time series, affecting the reliability of the climatological analyses.

In the last decades, several homogenization methods have been developed. While initially the focus has been on the adjustment of annual, seasonal, or monthly series (thus, mainly studying the changes on the first moments of the temperature distribution) in the last two decades, several studies have focused attention on daily values and on the effects on extreme events, which have more effect on higher moments.

Homogenization procedures are divided into two phases: the break detection, that is, generally applied to series that are aggregated to the yearly, seasonal, or monthly time scale, and the adjustment calculations. The first step is not part of the focus of this paper. The latter is suitable to be adapted to a daily resolution and has been subject of a vivid debate. Several studies have stated that the use of metadata and the eventual subjective assessments and decisions are preferable for a thorough homogenization (Venema et al. 2013; Pérez-Zanón et al. 2015; Gubler et al. 2017; Delvaux et al. 2018). Nevertheless, the size of certain datasets and the

✉ Antonello Angelo Squintu
squintu@knmi.nl

¹ Koninklijk Nederlands Meteorologisch Instituut, Utrechtseweg 297, De Bilt, Netherlands

² Wageningen University, Wageningen, Netherlands

³ GCRI, Brno, Czech Republic

⁴ CHMI, Brno, Czech Republic

⁵ MetOffice, Hadley Centre, Exeter, UK

need for objectivity and reproducibility require the use of completely automated processes (Venema et al. 2013). The currently existing automated approaches for the adjustment of daily series range from methods that are based on the detection of a yearly cycle (Vincent et al. 2002), to the parametrization of the distribution (Caussinus and Mestre 2004; Della-Marta and Wanner 2006; Mestre et al. 2011), or to more heuristic approaches (Menne and Williams 2009; Štěpánek et al. 2009; Trewin 2013; Squintu et al. 2019). The different mathematical and statistical theories behind the methods and their various procedural aspects (e.g., separation between break detection and adjustment calculation, number of iteration of the process, and different number of references) imply that the homogenized version of an inhomogeneous series will be different depending on the method that is used to homogenize it.

The comparison of the efficiency and capability of these methods is then a powerful tool for the detection of their strong and weak spots, contributing to the development of more capable homogenization systems.

In several cases, the comparison among methods has been performed as a validation for newly introduced homogenization methods (Caussinus and Mestre 2004; Menne and Williams 2005; Della-Marta and Wanner 2006; Mestre et al. 2011; Trewin 2013; Vincent et al. 2018). On the other hand, some outstanding studies have inspected the efficiency of automated, semi-automated, and manual homogenization procedures (Venema et al. 2013; Domonkos 2013; Li et al. 2016), focusing on break detection skills and on the performance on monthly data (Venema et al. 2013).

Every comparison of homogenization methods must ensure transparency and objectivity. The establishment of a benchmark dataset and the definition of clear metrics and evaluation parameters, determined before the examination of the results, give a fundamental contribute to this purpose.

The benchmark series are series that are meant to carry only the climatic signal, which can be either real or simulated, and are used as truth to be reproduced by a homogenization method. These records, which must be homogeneous and lack any quality issue, are intentionally perturbed in order to create fictitious inhomogeneous test series that will be the input of the evaluated homogenization procedures. The creation of the test series has usually been implemented with the insertion in the benchmarks of missing values, outliers, trends, noise, and inhomogeneities (Menne and Williams 2005; Mestre et al. 2011; Williams et al. 2012; Venema et al. 2013; Domonkos 2013; Lindau and Venema 2016). The temporal frequency and amplitude used for the insertion of such events is chosen following ad-hoc studies, in order to reproduce as much as possible realistic situations. An evolution of this approach considers the creation of ensembles (Domonkos 2011; Vincent et al.

2018; Domonkos and Coll 2017) of perturbed test series. This is done by varying amplitude and frequency of the breaks, allowing to test the efficiency of homogenization methods under particular conditions. Though such an approach is not able to reproduce all the possible artificial signals that might be introduced to inhomogeneous series (Vincent et al. 2018), giving more relevance to those signals that are easier to be studied and recognized. Moreover, the events and phenomena that can affect a recording station are so various, complex, and unexpected, that recent works have considered the opportunity to rely only on validated data (instead of simulated ones) for the creation of the test series (Venema et al. 2013; Gubler et al. 2017; Vincent et al. 2018). This approach has been recently developed by who elaborated the idea of joining records from close-by stations. This method allows to test homogenization methods on real data (thus, carrying all the possible signals). Nevertheless, with this choice, there is no knowledge on the truth (i.e., knowing the series as it was affected by only the climatic signal), since the benchmark, even after a preliminary homogenization and quality check, will never be completely free from artificial signals.

Whatever method is chosen for the creation of a benchmark dataset, the considered homogenization methods are applied to the test series with the aim to reproduce the benchmarks. The *distance* between the homogenized versions and the benchmarks, evaluated with transparent and clear metrics that must be defined in advance, indicates the performance of the homogenization. The most common used metrics involve calculations of relative RMSE (rRMSE) (Mestre et al. 2011; Venema et al. 2013; Domonkos 2013; Trewin 2013; Domonkos and Coll 2017; Gubler et al. 2017; Vincent et al. 2018), changes in the trends of indices (Domonkos 2013; Venema et al. 2013; Pérez-Zanón et al. 2015; Domonkos and Coll 2017), and countings of number of days within fixed absolute thresholds around benchmark values (e.g., PD05, see Section 2.4) (Trewin 2013; Vincent et al. 2018). Finally, the calculation of network mean biases of the trends (Domonkos 2013; Domonkos and Coll 2017) are a good indication of systematic behaviors related to certain homogenization methods and provide a good general estimation of the performance. In the specific case of evaluation of homogenization methods of daily data, it is important to focus on the effect on trends of indices that are related to extreme values (Trewin 2013), such as 10p (10th percentile of a daily series calculated on annual basis) or 90p (same but for the 90th percentile).

Benchmarking and evaluating methods helps a scientific community to develop and get mature (Venema et al. 2013). On one hand, users which are performing climatological studies are assisted in the evaluation of how the methods fit their needs. On the other hand, the researchers

that develop homogenization software can improve the methods including new features and facing statistical and procedural challenges with more awareness. Furthermore, this stimulates the debate on the comparison methods themselves, pushing in the direction of the development of more reliable system for the creation of benchmark datasets and more sophisticated metrics.

The aim of the present work is to evaluate the performances of the homogenization methods that have been considered and to highlight criticisms in their statistical and procedural features. An important focus will be given to the construction of a reliable benchmark and the description of clear metrics which are mandatory steps to assess solidity and transparency of any study of this kind. This analysis is limited to methods that can be applied to large datasets, thus completely automated. Several methods have been developed in the last decades (Della-Marta and Wanner 2006; Menne and Williams 2009; Štěpánek et al. 2009; Wang et al. 2010; Mestre et al. 2011; Trewin 2013; Domonkos and Coll 2017; Guijarro 2018; Squintu et al. 2019). In the framework of Copernicus Project (C3S.311a-Lot4), only the four listed below have been tested. The involvement of more methods is planned and will be implemented within future works.

1.1 Compared methods

The Quantile Matching, developed within the European Dataset Climate and Assessment (ECA&D), calculates adjustments via the comparison of the distribution of temperatures before and after a detected break. This happens taking the difference of similar quantiles (5th, 10th) in the two periods, identifying the changes in the distribution in the two target segments. The same comparison is performed on contemporary segments in a set of reference series, identifying which part of the difference is due to the climatic background. The obtained adjustment estimates depend on the quantile, on the month, and on the corresponding reference series, their median value is then applied to the daily values. Homogeneous references are chosen among the highest-correlated surrounding homogeneous series, with threshold on distance and correlation. A series is corrected if at least 3 homogeneous reference series with at least 5 years of overlap are available. See Squintu et al. (2019) for more details.

The DAP (Distribution Adjusted by Percentile) method (Štěpánek et al. 2009; Štěpánek et al. 2013) works as well on empirical distributions. The process calculates the day-by-day difference between candidate and reference series (obtained as average of 5 highly correlated neighboring series), such values are binned according to the percentiles of the candidate series. This is performed separately for the portions of the series before and after the break and

iterated for each month individually, including adjacent months. Finally, the differences of the results before and after the break are calculated and smoothed by a low-pass filter. For each datum before the break, its quantile is determined via interpolation of the percentiles and the corresponding adjusting factor is then applied. The tails of the distribution receive a special treatment: if the correlation after correction does not improve by at least 0.05% (considering each month individually), the adjustment is not applied and the original data are preserved. Note that even if both QM and DAP are based on quantiles, the first one manages the reference series with a “pairwise comparison” approach (Menne and Williams 2009; Trewin 2013) while the latter in this work calculates the average of the selected references, coherently with all the previous applications of DAP performed by GCRI. Finally, there are several different procedural differences, as the low-pass filter and the correlation improvement threshold used by DAP.

The third tested method was developed by Della-Marta and Wanner (2006) and is widely known as high-order moments (HOM). This works comparing the candidate with an overlapping reference series. A model is developed comparing candidate and reference after the break. The same model, applied to the reference segment before the break, is used to create a predicted series. The values of difference between observed and predicted are binned in deciles according to the quantile of the observed series. The values related to each decile are fitted with a local estimated scatter plot smoothing (LOESS) function, that determines a smooth set of adjustments depending on the quantile of the candidate series.

Finally, SPLIDHOM (Mestre et al. 2011) considers as well only one highly correlated reference series, which can also be inhomogeneous. Regressions based on cubic splines are calculated in order to describe the relationship between the candidate and the reference in the two overlapping periods before and after the break. The difference of these two models allows to determine the adjustments basing on the values of the reference before the break. The values of this last segment can be obtained from the candidate values with the inversion of the regression introduced above. Thus, starting from the value of the candidate before the break, the adjustment can be calculated knowing the regression relationships between the candidate itself and a reference series.

This last method, differently from the previous ones, does not make use of quantile bins, but it completely relies on regression methods. On the other hand, while HOM couples regression and quantile binning, Quantile Matching and DAP are solely based on empirical features.

HOM and SPLIDHOM have been tested twice, the first time with their original configuration, and the second time using procedural approaches similar to DAP. For example,

the application of a low-pass filter with coefficients retrieved from a Gaussian distribution (more reliable than a moving average (Mitchel et al. 1966)) in order to smooth the adjustments and a threshold on the improvement of correlation have been applied, leaving unchanged those data that do not fulfill such constraint. Thus, these variants keep the statistical aspects of HOM and SPLIDHOM, coupling them with the more conservative approaches typical of DAP.

2 Data and method

2.1 Construction of a benchmark dataset

A benchmark dataset is a fundamental tool in the comparison of homogenization methods. It is based on so-called benchmark series (*benchmarks*), which must be homogeneous and lack serious quality issues. In this work, the data of the benchmarks are partially replaced by homogeneous segments of data coming from surrounding series (*perturbers*), following Trewin (2013). Such intervention generates test series (*tests*), which are inhomogeneous, but composed of homogeneous segments, and need to be adjusted. The homogenization of such test series, adjusting the perturbing segments, is performed by the selected methods with the help of a set of homogeneous reference series (*references*), whose density and data availability can change. The different homogenized versions are compared through established metrics.

The challenges of a benchmark dataset are in the inhomogeneities themselves, which can have various amplitudes, structure, and timing. At the same time, the data availability and the station density has a great influence on the performance of homogenization (Caussinus and Mestre 2004; Domonkos 2013; Gubler et al. 2017). In this study, two different benchmark dataset have been created for these reasons. The first one consists of a very dense dataset in the Czech Republic where 20 benchmark series are accompanied by more than 200 references. Here, the methods can easily work and their results can be compared in favorable conditions. The second one aims at testing the capability of the homogenization methods to work in difficult conditions where the number of homogeneous reference series is low. This implies that test series are required to be used as reference for each others homogenization and that in some cases the homogenization is impossible due to data scarcity. In the following sections, the two benchmark datasets are explained in detail.

2.2 The Czech benchmark dataset

The Czech dataset aims at testing the methods on a nearly ideal situation where station density and series continuity is

high and where the group of series to be adjusted represent a very small percentage of the whole dataset (less than 10%), while the remainder includes homogeneous series.

In such favorable conditions, the statistical features of the homogenization methods can be compared independently from any thresholds on data availability and data quality that are used in the software scripting and reflect the views of the developer of the minimal conditions for which homogenization is meaningful.

The benchmark series (green dots in Fig. 1) span the 1970–1999 period and have been perturbed in the first 15 years with data coming from homogeneous neighboring stations (black dots), so that the breaks are all set to the same date (1st of January 1985). More than 200 homogeneous reference series (purple dots) are provided to help the homogenization. The used series are all provided by CHMI and have been preliminarily checked in their quality and in their homogeneity with Proclim DB software (Štěpánek et al. 2009; Štěpánek et al. 2013).

2.3 The European benchmark datasets

The European benchmark dataset is composed of original series extracted from the European Climate and Assessment Dataset (ECA&D, Klein Tank et al. (2002)) and are based on actual observations that did not undergo homogeneity adjustments. Benchmarks are chosen among the set of already homogeneous series of the ECA&D. Their homogeneity is tested through the agreement of three tests (Prodige (Caussinus and Mestre 2004), RHtest (Wang et al. 2007) and GAHMDI (Toreti et al. 2012)) according to the method documented by Kuglitsch et al. (2012) and Squintu et al. (2019). All the series which have been assessed as homogeneous in the 1970–1999 period are considered to provide the benchmarks. For each of the

Fig. 1 Geographical distribution of benchmark and test series (green dots), perturbing series (black) and reference homogeneous series (purple dots) for the Czech dataset. Benchmark series have same location for minimum and maximum temperatures

potential benchmarks, the neighborhood (radius of 20 km and maximum elevation difference of 50 m) is inspected in order to find series whose homogeneous segments will be used to perturb the benchmarks. A nearby series with at least 6 years of homogeneous data between 1970 and 1993 is found become eligible as a perturbing series (*perturber*). All homogenization procedures adjust the earlier part of the series consistent to their most recent part (Aguilar et al. 2003). Hence, the last 6 years of the considered period (1994–1999) are preserved for the data coming from the benchmark and will be used as basis for the calculation of the adjustments. The potential perturbing segments are checked to be sure they are not total or partial duplicates of the benchmarks, analyzing correlation, and absolute mean of daily differences. If no neighbors following such constraints are found, the potential benchmark is discarded. This process is performed separately for minimum (TN) and maximum (TX) temperatures. In order to generate a dataset of moderate size, only 50 TN and TX benchmark series (among circa 70 series per each element) are chosen. This selection has given priority to those stations that provide series to both datasets (more or less 35). The remaining ones are chosen in order to approximately reproduce the same geographical distribution for both variables, with few interesting exceptions (Belarus for TN, Serbia for TX, see Fig. 2).

Once benchmark and its perturbers are identified, the test series are created, replacing the data of the benchmark with the data of the perturber. In case only one perturber is related to a certain benchmark, the whole (homogeneous) perturbing segment is used to replace the benchmark data. Thus the change point will be at the end of the perturbing segment. If more perturbers for the same benchmark are

Fig. 3 Schematic diagram describing how a test series is generated. Each of the perturbing segments is required to be at least 6 years long and after 1994 only benchmark data are used

found, these have been combined in order to donate segments of at least 6 years of data. In case the identified perturbers are homogeneous all through the whole period, the change point has been set on the 1st January of 1985. This has allowed to create a sub-set of test series with simultaneous breaks, which is a great challenge for the homogenizers in current times (Venema et al. 2013). Figure 3 illustrates the building of the test dataset. The generated change points (*breaks*), whose distribution can be observed in Fig. 4, have been transcribed, together with coordinates and elevations of the benchmarks and their corresponding perturbers.

The set of 50 test series for each variable has been integrated with 40 homogeneous series, chosen among the discarded potential benchmarks. The selection has been performed on a geographical uniform pattern. Since the percentage of homogeneous series observed in the ECA&D dataset is 35%, the goal has initially been to reproduce such proportion between homogeneous and non-homogeneous series. This choice would have caused a lack of series in some areas (such as Spain, UK, France) and would have

Fig. 2 Geographical distribution of benchmark and test series (green dots) and reference homogeneous series (purple dots) for minimum (top) and maximum (bottom) temperatures in the European dataset

Fig. 4 Distribution of inserted breaks in the European Benchmark dataset and follows from the data availability and automatic combination. The peak in 1985 is the result of the adding of simultaneous

breaks (see text). No breaks are allowed before January 1976 and after December 1993 (dashed vertical lines), since each segment in the test must be at least 6 years long

lead to the likely possibility not being able to homogenize at all. For this reason, the references have been selected so that each benchmark has at least 3 homogeneous references in its surrounding area (see Fig. 2).

Figure 5 shows the distribution of daily differences for each pair of test and corresponding benchmark for the Czech and European tests, and for both daily maximum and minimum temperature. The amplitude of the breaks shows a strong variation, as demonstrated by the interquartile range and the 5th and 95th quantiles. In the figure, although the centers of the distribution tend to be more in the negative half of the plot, for almost all the series, the distributions of the breaks cover negative and positive values. This indicates the need of particular care in the homogenization process and confirms the complexity of inhomogeneities.

2.4 Definition of metrics

A reliable comparison of homogenization methods requires a set of metrics, defined in advance. The most commonly used metric is the relative RMSE (rRMSE) (Mestre et al. 2011; Venema et al. 2013; Domonkos 2013; Trewin 2013; Vincent et al. 2018; Domonkos and Coll 2017) which allows to measure the distance between the homogenized and the benchmark version of a series. Then, an average of the results on all the benchmark series and an analysis of their distribution allow to compare all the tested methods. A powerful metric, developed specifically for daily series, is the PD05: number of days where the day-by-day difference between the homogenized version and the benchmark is

lower than 0.5°C (Trewin 2013; Vincent et al. 2018). This is useful in defining a day-by-day accuracy of the homogenization, although it may be positively biased in cases of breaks with low magnitude.

Even though these metrics give a general overview of the power of the tested methods, they lack the information

Fig. 5 Distributions, series per series, of the daily differences between test series and corresponding benchmark series. Thick line indicates interquartile range, while dotted lines stand for the tails of the distribution up (down) to 95th (5th) quantile

Table 1 Acronym, references, and colors associated to the homogenization methods

Method	Reference	Acronym	Color
Test series (inhomogeneous)	See Section 2.1	inh	black
Quantile Matching	Squintu et al. (2019)	QM	red
DAP	Štěpánek et al. (2009)	DAP	blue
HOM	Della-Marta and Wanner (2006)	HOMHOM	yellow
SPLIDHOM	(Mestre et al. 2011)	SPLIDHOM	green
HOM with “DAP” parameters	See Section 1.1	HOMDAP	purple
SPLIDHOM with “DAP” parameters	See Section 1.1	SPLIDHOMDAP	orange

Colors are used for figures from 6 onwards

about variability of the series and about the effect of the homogenization methods on extreme values and interannual variance, which are of great importance (Trewin 2013). The trends assessment, inspecting how the homogenized versions get close to the benchmark trend, are one of the metrics that allow a thorough comparison of the methods (Domonkos 2013; Venema et al. 2013; Pérez-Zanón et al. 2015; Domonkos and Coll 2017). In this paper, trends on 10th, annual mean, and 90th percentile are compared.

The distance of the obtained trends from the ones of the benchmark is measured with a homogenization indicator (*hom_ind*) defined by the following expression:

$$\text{hom_ind} = 1 - \frac{\text{hom} - \text{ben}}{\text{tst} - \text{ben}} \quad (1)$$

where *hom* is the homogenized value (which can also be daily value, annual mean, annual percentile, etc.), *ben* is the value related to the benchmark (the goal of the process), and *tst* is the value related to the test series (the starting point).

Such indicator takes the following values:

- 0 when the adjustments, if any, have no effect on the series

- 1 when the adjustment perfectly reproduce the benchmark value
- Above 1 when the series is over-adjusted
- Between 0 and 1 when the series is partially adjusted;
- Negative values when the series is adjusted in the wrong direction

This metric allows an estimation of the accuracy of the homogenization and is more powerful than a simple difference between homogenized and benchmark values that would be difficult to analyze in the general context, due to the specific features of each benchmark series. Nevertheless when the difference between *tst* and *ben* is close to zero, singularities appear, limiting the significance of the indicator in such eventualities.

In the case of the European benchmark dataset, the scarcity of data may cause that one or more of the tested methods does not apply adjustments to the test series (or part of it). For this reason, a preliminary evaluation of the number of non-effective adjustments is made. The fraction of non-adjusted data is reported as ratio between the number of daily data where the difference between test value and the

Fig. 6 Histograms and density plots of rRMSE for the comparison of methods on the Czech dataset

Table 2 Results of methods comparison on the Czech dataset

Dataset (CZ)	Mean of RMSE (°C)		Mean of PD05 (%)		Mean of hom_ind of trends of					
	TN	TX	TN	TX	ann mean		10p		90p	
					TN	TX	TN	TX	TN	TX
Test series	1.31	1.17	39.01	37.44						
Quantile matching	0.96	0.73	52.79	65.48	1.37	1.08	0.71	0.81	1.09	0.63
DAP	0.96	0.74	51.74	64.89	1.03	1.12	0.83	0.70	0.90	0.62
HOM	0.99	1.03	50.59	62.82	0.86	1.01	0.77	0.68	0.73	0.42
SPLIDHOM	0.97	0.74	51.50	64.76	0.89	1.12	0.86	0.70	0.80	0.63
HOM (DAP)	0.98	0.74	50.86	64.38	0.99	1.10	0.59	0.51	0.79	0.47
SPLIDHOM (DAP)	0.96	0.74	51.82	64.93	1.09	1.12	0.88	0.71	0.81	0.48

Italicized values are discussed in the text

homogenized value is zero and the number of daily data that has to be adjusted (i.e., parts of the test series that belong to the perturbors). In this work, the homogenization of a test series is considered fruitless when the percentage of non-adjusted data is above 80%. Such threshold is high enough to avoid considering fruitless series with a high amount of data whose calculated adjustment is actually zero or series with high percentage of missing data.

3 Results

3.1 Czech dataset

The introduced metrics have been used to evaluate the performance of the homogenization methods on the Czech dataset. As expected, due to the good quality of the dataset,

all the methods managed to homogenize all the series in all of their parts. In the following section, the color code displayed in Table 1 is used.

First inspection is performed with the rRMSE, taking the benchmark version as reference, here lowest values indicate the best performances.

In Fig. 6, all the methods show similar behaviors except HOM (yellow line). The deviation of this method in comparison to other methods is confirmed in Table 2, where HOM shows a high average of rRMSE for the homogenization of TX. The high value of mean rRMSE is due to the performance on two series, one of these can be observed in Fig. 7. Here it is clear how the results of HOM do not correctly reproduce the annual mean of the benchmark (gray thick line). The anomalous behavior of HOM is confirmed by plotting the homogenization indicator (Eq. 1) applied to the yearly values (bottom panel) where

Fig. 7 Comparison of homogenized versions of the benchmark series of Semčice, Czech Republic. Color code is set according to Table 1, black items represents the test series, gray items stand for the value of the benchmark. Top panel shows annual means (points) with running mean (lines), the more the homogenized values get close to the gray line, the better the homogenization has been. This distance is better displayed in the bottom panel, applying the homogenization indicator, explained in Eq. 1, to each yearly value

ind trends_90p hist, TN, dataset:CZ

Fig. 8 Histograms and density plots of hom_ind applied to the trends of the 90th percentile (90p) of minimum temperature series (TN). Vertical lines define the central area of the histogram, that needs smaller

bins in order to better describe the distribution. Values out of such limits are plotted in single columns on the external areas, helping identifying outliers

all the methods (except QM) show to be over-adjusting the series ($\text{hom_ind} > 1$), while HOM under-adjusts it and in some years the annual mean is adjusted in the wrong direction (negative hom_ind). In this example, the very similar values of the test and the benchmark in the first two years reveal a weak spot of such tool, easily affected by singularities. This aspect will be treated in the concluding remarks. Finally, when checking the trends (which are presented in the legend), HOM performs better than the other methods, being the only one not over-correcting the trends and being the second closest to the benchmark value.

Table 2 also indicates that the performance on PD05 is approximately the same among the six methods. At the same time, the last six columns, which display the hom_ind applied to the trends, show only a general overestimation of the trends on annual mean of TX. For the annual mean of TN an even distribution around the ideal value (1) is found with the exception of the anomalous value of QM (1.37). The homogenization of the extreme values shows underestimation of almost all the method and all indices. This is observed especially for the 90th percentile of TX with values ranging between 0.42 and 0.63. Furthermore,

Fig. 9 Scatter plots with contours of half height of density distributions of homogenization indicator calculated on trends of various indices, Czech dataset. Left: TN— hom_ind on trends of 10th percentile

vs. annual mean. Right: TX— hom_ind on trends of 90th percentile vs. annual mean. N.b. values which lay out of the graph domain are plotted on the boundaries. Color code follows Table 1

Table 3 Table reporting the amount of non-adjusted series and non-adjusted data

Dataset(Euro)	Number of series with more than 80% of non-adjusted data		Mean of percentage of non-adjusted data [%]	
	TN	TX	TN	TX
Quantile Matching	2	3	17.43	20.73
DAP	3	1	34.22	39.69
HOM	0	0	33.43	27.27
SPLIDHOM	0	0	13.27	17.86
HOM(DAP)	3	3	36.22	40.60
SPLIDHOM(DAP)	4	2	35.18	40.79

the performance on the trends of the 10th percentile by HOMDAP appears to be very low for both TN and TX. Figure 8 shows the distributions that generate these averages. The higher value for QM is due to a distribution that peaks at values larger than 1, indicating a good performance with a slight overestimation. On the other side, the lower averages observed for the other methods are related to a more spread distribution and a higher presence of outliers, which are plotted in single columns out of the vertical solid lines.

A more exhaustive description of the performance of the methods in reproducing the trends on average and extreme indices is shown in Fig. 9. Here, the evaluation has been made via scatter plots that allow to compare at the same time

the hom_ind results about trends of extreme indices and on the mean values.

Left (right) panel shows the comparison of hom_ind of trends of annual mean vs. 10th (90th) of minimum (maximum) temperatures: all methods show a scattered behavior. Nevertheless, the lines (contours of half height of the two-dimensional probability density function) help identifying that all the methods homogenize generally well, since they are centered around the black diamond (coordinates (1,1), corresponding to an ideal homogenization on both indices). Thus, the anomalous behaviors mentioned above are isolated and not systematic.

In both cases, the distribution is more spread along the x -axis, indicating a lower performance in the homogenization

Fig. 10 Series whose failing rate is above 80%. Color code follows Table 1. Sizes of the circles change for graphical purposes and do not represent different magnitudes

Fig. 11 Histograms and density plots representing the percentage of non-adjusted data for each series in the European dataset. Color code follows Table 1

of extreme values. At the same time, there is less ordered behavior in the contours of the distributions in the case of minimum temperature which is probably due to the more local characteristics of such variable.

3.2 European dataset

The scarce geographical density of test and reference series of the European benchmark dataset and the presence of multiple and non-simultaneous breakpoints make that the availability and length of homogeneous sub-series that could be used for the homogenization process are limited. These challenging conditions make that some methods abstain from homogenizing some tests and apply no adjustments on portions of the series. Table 3 reports the number of series with a fruitless homogenization (see Section 2.4 for the definition) indicating an average rate of 1 to 4 failures (out of the 50 test series) for all methods but HOM and SPLIDHOM. These latter two methods adjust

a sizeable part of all test series, despite the scarcity of references. Figure 10 shows the areas where the methods fail more often: regions with low density (Spain, Italy) and occasionally those on the edges of the denser areas.

Though, while the number of failed series are similar, significant differences are observed on the average of non-adjusted data. The mean of the percentage of non-adjusted data (last two columns of Table 3) show different behaviors. In the histograms of Fig. 11 QM, SPLIDHOM, and HOM present relative narrow distribution, whose low peak values indicate successful performances. Nevertheless, QM presents 5 series with 100% of non-adjusted and the high value of the peak of HOM indicates the possibility of issues when homogenizing some of the series. At the same time, the other methods show spread distributions indicating high percentage of sparsely non-adjusted data, whose features will be inspected further in the text. This preliminary analysis has allowed to understand how the methods approach the low density of stations, since

Table 4 Results of methods comparison on the European dataset

Dataset(Euro)	Mean of		Mean of		Mean of hom.ind of trends of					
	rRMSE [°C]		PD05[%]		annmean		10p		90p	
	TN	TX	TN	TX	TN	TX	TN	TX	TN	TX
Test series	1.60	1.28	39.51	46.37						
Quantile matching	1.25	1.16	43.72	50.41	0.65	0.82	1.01	0.33	0.78	1.23
DAP	1.34	1.21	42.10	48.43	0.55	0.56	0.86	0.91	0.36	1.56
HOM	2.00	1.74	38.20	44.80	1.46	0.31	0.80	1.92	2.00	3.14
SPLIDHOM	1.27	1.21	42.77	47.50	0.87	0.59	0.95	0.79	0.30	1.49
HOM (DAP)	1.42	1.24	40.68	47.85	0.58	0.37	0.63	0.65	0.29	1.63
SPLIDHOM (DAP)	1.35	1.22	41.93	48.51	0.57	0.49	0.69	0.80	0.32	1.51

Relevant negative results are highlighted in italics

RMSE hist, TN&TX, dataset:Euro

Fig. 12 Histograms and density plots of rRMSE (TN and TX together) for the comparison of methods on the European dataset. Lines represent corresponding probability density functions

some methods (as QM) leave the most problematic series unchanged, while others (ProclimBD, HOMDAP, and SPLIDHOMDAP) manage to homogenize portions of their daily data.

rRMSE, PD05, and accuracy in the reproduction of the trends have been then evaluated on the homogenized versions, including those series (and those portions of the series) that have not been adjusted. Table 4 highlights an evident lower performance of the tested methods compared to the case of the Czech dataset (Table 2), which can be expected. The reduction of rRMSE after the application of the homogenization methods has considerably a lower magnitude and, at the same time, the PD05 shows a lower increase of percentages. While almost all histograms and pdf show a good improvement (see Fig. 12) compared to the inhomogeneous series, the HOM histogram (yellow) presents a clearly anomalous behavior, which was anticipated by high averages of Table 4 and repeats what was already observed in the context of the Czech dataset.

The comparison of the obtained trends with the ones of the corresponding benchmark has revealed a high difficulty in the reproduction of the original trends within the European dataset. While the averages of indicators related to HOM versions are almost always outside of the 0.5–1.5 range (which indicates 50% distance from the benchmark trend), all the other methods present similar behaviors, with few exceptions. Worst performances are observed on the trends of the warm extremes (90th percentile). The underestimation of the trends for TN (of all calculated indices) are related to a large presence of very negative results (related to adjustments in the wrong direction) as shown in the top panel of Fig. 13, where a high number

of series lie on the left of the boundaries of the histogram. The opposite happens for trends of 90p for TX, where the overestimation is due to a high number of overcorrected series, which lie on the right of the boundaries of the histogram (bottom panel of Fig. 13). In both cases, the very high and very low values of the homogenization indicator are related to singularities when the amplitude of the inhomogeneity was very low (i.e., below 0.5 °C). Such a situation implies a low value in the denominator of Eq. 1, thus large absolute value of the indicator.

An example of such fact is displayed in Fig. 14, where the low magnitude of the breaks is identified but all methods but HOM that over-corrects the first portion of the series (large positive values of the homogenization indicator on yearly 90p of TX). This has a consequence on the calculation of hom_ind on the trend of HOM, which in this case has a value of 6, indicating a correction 6 times bigger than the one needed to reproduce the trend of the benchmark.

Special attention has to be given to the particular shape of the distribution of DAP, HOMDAP, and SPLIDHOMDAP in the bottom panel of Fig. 13. This is related to the conservative philosophy of DAP, which is to preserve original data rather than risking wrong adjustments. Consequence of this choice is the high percentage of non-adjusted data observed in Fig. 11. The peak close to 0, which is related to trends which are not altered during the homogenization, indicates a high number of non-changed data in the warmest records of these series. The same behavior is found in the scatter plots of the right panel of Fig. 15, where these three methods present a bimodal structure, with one of the two maxima that tends to get close to the (0,0) point, which represents

Fig. 13 Histograms and density plots of hom_ind calculated on trends of 90p of TN (top) and TX (bottom) for the European dataset. Same graphical features as Fig. 8

uncorrected data. This scatter plot, together with the one related to TN temperatures (left panel), displays a very scattered distribution as well as broader and less defined densities, reflecting the difficulty in homogenizing such dataset. Nevertheless, QM and especially SPLIDHOM, appear to perform well, moving the peak in the distribution near (1,1).

4 Discussion

The higher density and correlations of the Czech dataset (see Table 5) makes that almost all the methods perform well, homogenizing the whole series with similar results.

Nevertheless, even under favorable conditions, a method like HOM has been affected by statistical issues. The causes of this might be inspected in the software of this method; nevertheless, such research is not in the scope of the present work.

On the other side, the comparison of the methods on a more challenging dataset has helped to test the robustness of the procedures and has revealed different behaviors. The evaluation of the percentages of unchanged data has highlighted how approaches in the application of the adjustments vary among the considered methods. Those that use DAP settings (DAP, HOMDAP, and SPLIDHOM-DAP) require an improvement in the candidate-reference correlation (at least 0.05) in order to allow the adjustment

Fig. 14 Effects of the homogenization methods on the 90p of TX of Hamburg. Same structure and color code as Fig. 7. N.b. Good accordance among the results imply that some of the lines are overlapped

of a daily datum, preferring to leave the data unchanged when such constraint is not fulfilled. This is evident in Figs. 11, 13, and 15 and highlights how procedural choices lead to similar results even when statistical properties (see Section 1.1) of the methods are considerably different. The aforementioned choices are the effect of a more conservative strategy that can be adopted by homogenizers. This aims at giving (when in doubt) priority to original data over homogenized ones, which can be required for particular studies or projects.

The spread distribution of percentage of non-adjusted data in the series observed in Fig. 11 is the most evident effect of the conservative constraints. Though, this is considerably different from the distribution observed, for example, in the case of quantile matching. This last method implements its conservative strategy with a different approach. It evaluates the possibility of applying

adjustments considering the whole candidate and reference series: when correlation is too low or the overlapping period is too short, the whole series is left unchanged, explaining the peak of unchanged data at 100%.

Finally, while the non-adjustment criteria of QM affect the whole distribution of daily data, Figs. 13 and 15 have shown that the unchanged data related to the methods with DAP parameters are more concentrated around the warm extremes. This is probably due to a lower correlation gain when homogenizing summer extremes.

5 Summary and conclusions

A comparison between homogenization methods (Quantile Matching (Trewin 2013; Squintu et al. 2019), DAP (Štěpánek et al. 2009), HOM (Della-Marta and Wanner

Fig. 15 Same as Fig. 9 but for the European dataset

Table 5 Average distance and correlation in the two datasets

	Czech		European	
	TN	TX	TN	TX
Average distance [km]	40.4	38.3	198.0	220.1
Average correlation	0.896	0.928	0.629	0.615

Please note that the Czech dataset has an homogeneous distribution of the stations, while the density of stations in the European vary considerably among the regions

2006), and SPLIDHOM (Mestre et al. 2011)) has been performed on two benchmark datasets. The two dataset (Czech Republic, Europe) are built with analogous methods, replacing data of homogeneous series (*benchmarks*) with data coming from homogeneous segments of neighboring series, thus generating non-homogeneous series to be adjusted (*tests*). While the Czech dataset has high quality and station density, in the European dataset, stations are sparse and homogeneous reference series represent a minority (44% of the set, down to 20% in some areas). The performance of the homogenized versions (*candidates*) has been evaluated using metrics such as the rRMSE (RMSE of the difference series between candidate and corresponding benchmark) and PD05 (percentage of values in the difference series below 0.5 °C). These metrics have revealed a good performance of all the methods but HOM, which presents a high number of outliers in both datasets. Trends on indices such the 10th and 90th percentile (together with the annual mean) have then been tested. This has allowed to test performance of the homogenization methods in reproducing extreme values and related indices. Such comparison has been operated via the introduction of a homogenization indicator (*hom_ind*) that can be applied to any indices or measured values. This determines the goodness of the homogenization on a clear common scale: wrong adjustment (< 0), no adjustment (0), partial adjustment (between 0 and 1), ideal adjustment (1), and overcorrection (> 1). Plotting results of *hom_ind* helps identifying anomalous behaviors and systematic biases. In particular, it has been observed that, while in the Czech dataset all methods have performed well with minor exceptions, the more challenging European dataset has revealed several criticisms, among all in the homogenization of warm extremes for three methods (DAP, HOMDAP, SPLIDHOMDAP).

Though, the structure of the equation that defines *hom_ind* has shown to be vulnerable in case of low difference between test and benchmark values. These situations can lead to singularities (i.e., very large values of *hom_ind*) which in reality are related to reasonable values. At the same time, conditions of no difference between test

and benchmark are affected by a very low signal-to-noise ratio. This means that the homogenization process presents high values of uncertainty, making the result hard to be defined in terms of quality. Thus, a more sophisticated version of the Eq. (1) may be developed in order to avoid this issue, since inhomogeneities with low magnitude are important challenges of benchmark (and real) datasets and any homogenization method needs to be able to identify and account for them. Further changes will also be needed in the technique for the construction of benchmark datasets, for example, widening the size of the datasets, extending the covered time interval, inserting plateau breaks (as suggested by Domonkos and Coll (2017)), or giving more relevance to mountain areas where a minimal change of altitude may bring to large inhomogeneities. Nevertheless, the use of two different benchmark datasets, as done here, has shown to be a powerful tool in order to perform a complete analysis of the capabilities of the methods.

In this study, the methods are compared in their original version, without further elements but those described in their original papers. Though it is clear that the different approach in the selection of reference series affects the performance of the methods themselves. In particular, the use of a single or a set of references (real record or average of surrounding series) make the methods respond differently to the statistical features of the references. For this reason, a further study would be needed with the aim of comparing the homogenization methods in similar conditions of reference selection, analyzing how they would behave in all the possible situations (i.e., set of references, average of references or single reference). Nevertheless, this work has allowed a comparison under different aspects of the selected methods and has given clear indications about their performances.

It is fundamental to stress that the selection of the best homogenization method is not in the goals of this work: such choice depends on the needs of each individual user. Studies like this aim at stimulating the debate on the field of homogenization of temperatures, stressing criticisms of some methods, and indicating to future programmers which aspects (e.g., management of low density of stations, importance of extreme values) should be better developed to reach an always higher quality of homogenization on a daily basis. For example, particular care should be given in managing the adjustment of extreme values which have been proven to need a different treatment with respect to average values. On the other hand, homogenization methods must be able to account for the most various geographical patterns. This implies that the procedure is required to be flexible enough to adapt to conditions of areas with very high and very low data quality and correlation between series. At the same time, users should choose the most suitable homogenization method evaluating these aspects,

considering especially the station density and the statistical properties of the climate of the interested area (i.e., mountain, plain, coastal climates). For these reasons, the belief of the benefit (Venema et al. 2013) that publication of studies about the comparison of homogenization methods have on the field has become stronger.

Finally, a short consideration is needed about the specific case of the European Climate Assessment and Dataset. This includes areas with extreme high density (Germany, Sweden, etc.) and areas with low density (Northern Africa, Middle East, parts of Mediterranean Europe, etc.); thus, a method that is able to face both data-scarce as data-rich conditions is needed. Quantile Matching was applied earlier to the whole ECA&D dataset (Squintu et al. 2019) and the good performance of this method on both the Czech dataset and the European dataset, characterizing data-rich and data-scarce conditions respectively, and the similarity in the results to a widely known and recognized method as SPLIDHOM, gives confidence in the homogenization of ECA&D using Quantile Matching.

Acknowledgments We acknowledge the data providers in the ECA&D project (www.ecad.eu).

Funding information Funding has been received from the Copernicus Climate Change Service through contract C3S_311a_Lot4.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Aguilar E, Auer I, Brunet M, Peterson TC, Wieringa J (2003) Guidance on metadata and homogenization. Wmo Td 1186:53
- Caussinus H, Mestre O (2004) Detection and correction of artificial shifts in climate series. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 53(3):405–425
- Della-Marta P, Wanner H (2006) A method of homogenizing the extremes and mean of daily temperature measurements. *J Clim* 19(17):4179–4197
- Delvaux C, Ingels R, Vrábel V, Journée M, Bertrand C (2018) Quality control and homogenization of the belgian historical temperature data. *International Journal of Climatology*
- Domonkos P (2011) Efficiency evaluation for detecting inhomogeneities by objective homogenisation methods. *Theor Appl Climatol* 105(3–4):455–467
- Domonkos P (2013) Efficiencies of inhomogeneity-detection algorithms: comparison of different detection methods and efficiency measures. *J Climatol* 2013
- Domonkos P, Coll J (2017) Homogenisation of temperature and precipitation time series with ACMANTt3: method description and efficiency tests. *Int J Climatol* 37(4):1910–1921
- 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* 37(13):4670–4683
- Guijarro JA (2018) Homogenization of Climatic Series with Climatol. Reporte técnico State Meteorological Agency (AEMET), Balearic Islands Office, Spain
- Klein Tank AMG, Wijngaard JB, Können GP, Böhm R, Demarée G, Gocheva A, Milate M, Pashiardis S, Hejkrlik L, Kern-Hansen C, Heino R, Bessemoulin P, Müller-Westermeier G, Tzanakou M, Szalai S, Páldóttir T, Fitzgerald D, Rubin S, Capaldo M, Maugeri M, Leitass A, Bukantis A, Aberfeld R, van Engelen AFV, Forland E, Mielus M, Coelho F, Mares C, Razuvaev V, Nieplova E, Cegnar T, Antonio López J, Dahlström B, Moberg A, Kirchhofer W, Ceylan A, Pachaliuk O, Alexander L, Petrovic P (2002) Daily dataset of 20th-century surface air temperature and precipitation series for the european climate assessment. *Int J Climatol* 22(12):1441–1453
- Kuglitsch F-G, Auchmann R, Bleisch R, Brönnimann S, Martius O, Stewart M (2012) Break detection of annual swiss temperature series. *Journal of Geophysical Research: Atmospheres* 117 (D13):1–12
- Li Z, Cao L, Zhu Y, Yan Z (2016) Comparison of two homogenized datasets of daily maximum/mean/minimum temperature in China during 1960–2013. *Journal of Meteorological Research* 30(1):53–66
- Lindau R, Venema V (2016) The uncertainty of break positions detected by homogenization algorithms in climate records. *Int J Climatol* 36(2):576–589
- Menne MJ, Williams CNJr (2005) Detection of undocumented changepoints using multiple test statistics and composite reference series. *J Clim* 18(20):4271–4286
- Menne MJ, Williams CNJr (2009) Homogenization of temperature series via pairwise comparisons. *J Clim* 22(7):1700–1717
- Mestre O, Gruber C, Prieur C, Caussinus H, Jourdain S (2011) SPLIDHOM: a method for homogenization of daily temperature observations. *J Appl Meteorol Climatol* 50(11):2343–2358
- Mitchel JM, Dzerdzeevskii B, Flohn H, Hofmeyr WL, Lamb HH, Rao KN, Wallen CC (1966) Climatic Change. Technical Note, No. 79 World Meteorological Organization, Geneva
- Pérez-Zanón N, Sigró J, Domonkos P, Ashcroft L (2015) Comparison of HOMER and ACMANTt homogenization methods using a central Pyrenees temperature dataset. *Adv Sci Res* 12(1):111–119
- Squintu AA, van der Schrier G, Brugnara Y, Klein Tank A (2019) Homogenization of daily temperature series in the European climate assessment & dataset. *Int J Climatol* 39(3):1243–1261
- Štěpánek P, Zahradníček P, Skalák P (2009) Data quality control and homogenization of air temperature and precipitation series in the area of the Czech Republic in the period 1961–2007. *Adv Sci Res* 3(1):23–26
- Štěpánek P, Zahradníček P, Farda A (2013) Experiences with data quality control and homogenization of daily records of various meteorological elements in the Czech Republic in the period 1961–2010. *Időjárás* 117(1):123–141
- Toreti A, Kuglitsch FG, Xoplaki E, Luterbacher J (2012) A novel approach for the detection of inhomogeneities affecting climate time series. *J Appl Meteorol Climatol* 51(2):317–326
- Trewin B (2013) A daily homogenized temperature data set for Australia. *Int J Climatol* 33(6):1510–1529
- Venema VK, Mestre O, Aguilar E, Auer I, Guijarro JA, Domonkos P, Vertacnik G, Szentimrey T, Stepanek P, Zahradnicek P et al (2013) Benchmarking homogenization algorithms for monthly data. *1552(1):1060–1065*

- Vincent LA, Zhang X, Bonsal B, Hogg W (2002) Homogenization of daily temperatures over Canada. *J Clim* 15(11):1322–1334
- Vincent LA, Milewska EJ, Wang XL, Hartwell MM (2018) Uncertainty in homogenized daily temperatures and derived indices of extremes illustrated using parallel observations in Canada. *Int J Climatol* 38(2):692–707
- Wang XL, Wen QH, Wu Y (2007) Penalized maximal t test for detecting undocumented mean change in climate data series. *J Appl Meteorol Climatol* 46(6):916–931
- Wang XL, Chen H, Wu Y, Feng Y, Pu Q (2010) New techniques for the detection and adjustment of shifts in daily precipitation data series. *J Appl Meteorol Climatol* 49(12):2416–2436
- Williams CN, Menne MJ, Thorne PW (2012) Benchmarking the performance of pairwise homogenization of surface temperatures in the United States. *Journal of Geophysical Research: Atmospheres* 117(D5):1–16
- Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.