Efficiency of Time Series Homogenization: Method Comparison with 12 Monthly Temperature Test Datasets

PETER DOMONKOS,[a] JOSÉ A. GUIJARRO,[b] VICTOR VENEMA,[c] MANOLA BRUNET,[d,e] AND JAVIER SIGRÓ[d]

[a] *Tortosa, Spain*
[b] *State Meteorological Agency (AEMET), Unit of Islas Baleares, Palma, Spain*
[c] *Meteorological Institute, University of Bonn, Bonn, Germany*
[d] *Centre for Climate Change, Universitat Rovira i Virgili, Vila-seca, Spain*
[e] *Climatic Research Unit, University of East Anglia, Norwich, United Kingdom*

(Manuscript received 3 August 2020, in final form 20 December 2020)

ABSTRACT: The aim of time series homogenization is to remove nonclimatic effects, such as changes in station location, instrumentation, observation practices, and so on, from observed data. Statistical homogenization usually reduces the nonclimatic effects but does not remove them completely. In the Spanish ''MULTITEST'' project, the efficiencies of automatic homogenization methods were tested on large benchmark datasets of a wide range of statistical properties. In this study, test results for nine versions, based on five homogenization methods—the adapted Caussinus-Mestre algorithm for the homogenization of networks of climatic time series (ACMANT), ''Climatol,'' multiple analysis of series for homogenization (MASH), the pairwise homogenization algorithm (PHA), and ''RHtests''—are presented and evaluated. The tests were executed with 12 synthetic/surrogate monthly temperature test datasets containing 100–500 networks with 5–40 time series in each. Residual centered root-mean-square errors and residual trend biases were calculated both for individual station series and for network mean series. The results show that a larger fraction of the nonclimatic biases can be removed from station series than from network-mean series. The largest error reduction is found for the long-term linear trends of individual time series in datasets with a high signal-to-noise ratio (SNR), where the mean residual error is only 14%–36% of the raw data error. When the SNR is low, most of the results still indicate error reductions, although with smaller ratios than for large SNR. In general, ACMANT gave the most accurate homogenization results. In the accuracy of individual time series ACMANT is closely followed by Climatol, and for the accurate calculation of mean climatic trends over large geographical regions both PHA and ACMANT are recommended.

KEYWORDS: Temperature; Algorithms; Climate records; Data quality control; Time series

1. Introduction

Technical changes of climate observations and environmental changes around meteorological instruments often cause nonclimatic biases in the time series of climate records. These changes are usually referred to as inhomogeneities (e.g., Aguilar et al. 2003; Vincent et al. 2012; Sanchez-Lorenzo et al. 2015). The removal of inhomogeneities (IHs) from climate data is important for the correct evaluation of past climate changes and climate variability. A large number of statistical homogenization methods are in use (Beaulieu et al. 2008; Domonkos et al. 2012; Ribeiro et al. 2016), and documented information on station histories (metadata) also helps homogenization. Even when climate records are accompanied by detailed station histories, the application of statistical homogenization methods is still recommended (e.g., Aguilar et al. 2003; Auer et al. 2005; Acquaotta and Fratianni 2014). The importance of statistical homogenization is enhanced by the fact that unintentional technical changes also occur (Thorne et al. 2016), and hence datasets of metadata are usually not complete. Note also that metadata usually do not quantify the size of IHs, and this shortcoming tends to limit their usefulness. However, no statistical homogenization is perfect, as the natural variability of local weather and nonsystematic observation errors limit the accuracy of detecting and correcting IHs.

For achieving high-quality homogenization, climatologists should use the most appropriate homogenization methods. The correct method selection needs objective knowledge about method efficiencies, therefore the efficiency of homogenization methods must be tested. Efficiency tests need to be based on simulated test datasets, as the true positions and magnitudes of IHs are known only in such datasets. However, it is not a straightforward task to provide efficiencies that are valid for real data homogenization, because the variation of real data properties (including the properties of IHs in them) is not known sufficiently. Differences between real data properties and test data properties have been reported by Venema et al. (2006), Domonkos (2011a, 2013a), Willett et al. (2014), and Gubler et al. (2017). To improve the reliability of the efficiency tests of homogenization methods, often varied scenarios of IH properties are applied (Menne and Williams 2005; Williams et al. 2012; Domonkos 2013b; Killick 2016). It is also important to focus on efficiency measures characterizing directly the appropriateness of homogenized data for climate research (Venema et al. 2012; Domonkos 2013a). During the

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/JCLI-D-20-0611.s1.

Corresponding author: Peter Domonkos, dpeterfree@gmail.com

DOI: 10.1175/JCLI-D-20-0611.1

European project COST ES0601 (known as "HOME"; 2007–11) a sophisticated simulated temperature and precipitation test dataset was created, and most of the widely used homogenization methods were tested (Venema et al. 2012). However, because manual methods were also tested, the size of the HOME benchmark was relatively small (i.e., only 15 temperature and 15 precipitation networks were used). Automatic methods can easily be tested on much bigger test datasets, which allows for much more accurate comparisons. HOME tests showed that the differences between method efficiencies are large according to both test dataset properties and homogenization methods, and therefore more tests are needed to obtain more profound and more detailed knowledge about the efficiencies of homogenization methods. New tests are also needed because of the fast development of new homogenization methods.

This study presents a part of the efficiency tests made within the Spanish "MULTITEST" project (http://www.climatol.eu/MULTITEST/). The aim of the project was to test statistical homogenization methods on large monthly air surface temperature and precipitation test datasets with a range of climatic and IH properties (see section 2b). In this study, the experiments with 12 temperature test datasets are presented. These datasets include 1900 station networks overall with 5–40 time series in each. Because of the large size of the test datasets, only automatic homogenization methods are tested (see section 2c). The success of homogenization is evaluated with the residual errors of the homogenized data. Such errors occur in various spatial and temporal scales. Perhaps the most important and frequently discussed error type is the bias in regional- and global-scale temperature data (Menne et al. 2009; Hua et al. 2017; Lindau and Venema 2018, etc.). However, considering that homogenized data are used for various research objectives from the reconstruction and understanding of past climate variability to climate impact studies, six different error types will be monitored (see section 2d).

2. Data and methods

a. Theoretical foundation

Test datasets including time series of monthly temperatures $\mathbf{X}$ are created:

$$\mathbf{X} = x_1, x_2, \ldots, x_i, \ldots, x_{12n}. \tag{1}$$

In Eq. (1) n is the length of the time series in years. A time series is realistic if its elements x (omitting index i) include all of the components of real temperature records. These are the regionally representative climate C_R, local climate anomaly relative to the regional mean C_A, station effect including IHs S, random anomaly due to the weather W, and nonsystematic observational error d:

$$x = C_R + C_A + S + W + d. \tag{2}$$

For perfectly homogeneous series $S = 0$, whereas the most frequent and significant technical changes, such as station relocations or changes of instrumentation, result in sudden shifts in S. Although certain factors result in gradual changes of S (e.g., growing vegetation near the instrument), the most frequently applied statistical model of S is a step function, or sometimes the combination of steps and linear trend sections (Wang 2003; Reeves et al. 2007). The station effect often has a marked annual cycle (e.g., Brunet et al. 2011; Dienst et al. 2017); moreover, it can also contain nonseasonal, weather-dependent variation, although the latter is more attenuated for monthly data than for daily temperatures.

In our test datasets three kinds of IHs are included: (i) sudden shift of S (break), (ii) linear change of S (trend IH), and (iii) short-term platform-shaped change (pair of breaks of the same size and opposite sign) of S. The inclusion of platform-shaped IHs is reasoned by theoretical considerations and experimental results (Domonkos 2011a; Rienzner and Gandolfi 2011). Biases related to an IH of our test dataset often have a seasonal cycle, but, except for linear trend IHs, their nonseasonal variation is zero.

The difference between C_A and W in Eq. (2) is that while the former depends on low-frequency processes and thus has memory, the latter is a temporally independent contributor on a monthly scale. Optimally, climatic records should be free from components S and d. Outlier values of temperature records for large d can be identified and removed from time series by quality control procedures, and then the role of d is minor. Both W and d, and thus also their common contribution, can be modeled with a Gaussian white noise with expected value zero:

$$x = C_R + C_A + S + \varepsilon. \tag{3}$$

In a given climatic zone, the temporal variation of C_A is often small, and then the constant part of C_A can be merged with S, because constant differences do not impact climatic trends and temporal variability:

$$x = C_R + S + \varepsilon. \tag{4}$$

Using the terminology of the HOME benchmark, datasets modeled by Eq. (3) are referred to as surrogate datasets, whereas those that are modeled by Eq. (4) are referred to as synthetic datasets. Both for synthetic and surrogate datasets, the climate signal may have any temporal evolution, and usually all of the climatic components and station effect include seasonal cycles. Spatial correlations of weather and climate anomalies are intended to be realistic.

Note that the generation of homogeneous test datasets and the possible impacts of using synthetic versus surrogate test datasets are presented in the online supplemental material.

b. Test datasets

We examine the efficiency of homogenization methods with 12 of our own developed test datasets, half of them are synthetic (Y1, Y2, . . . , Y6), and the other half are surrogate data (U1, U2, . . . , U6). The size of the datasets is large; each of them includes at least 100 networks of 5–40 time series to reduce sampling error. Spatial correlations and IH properties are widely varied between the test datasets in order to examine the functioning and efficiency of homogenization methods in varied homogenization tasks.

TABLE 1. Properties of homogeneous test datasets: K = number of networks, N = number of time series per network, n = length of time series in year, σ = standard deviation of noise term [Eqs. (3) and (4)], R = mean spatial correlation, r = ratio of missing data, and an asterisk indicates that a characteristic is unknown for the adapted datasets.

Dataset	K	N	n	σ	R	r
Y1	100	10	60	0.27	0.91	0
Y2	100	10	60	0.45	0.81	0
Y3	100	10	60	0.975	0.56	0
Y4	100	40	60	0.45	0.81	0
Y5	100	40	100	0.45	0.69	0.11
Y6	100	20	100	0.45	0.69	0.12
U1	100	10	60	*	0.87	0
U2	100	10	60	*	0.88	0
U3	100	10	60	*	0.68	0
U4	500	5	40	*	0.82	0
U5	300	10	100	*	0.91	0.11
U6	200	15	100	*	0.68	0.21

1) HOMOGENEOUS TEST DATASETS

Table 1 shows several properties of the homogeneous test datasets. Each dataset consists of 100–500 networks of 5–40 time series. The length of time series varies between 40 and 100 years, and the mean spatial correlation R is between 0.56 and 0.91 (Table 1).

The generation of surrogate data is presented in Willett et al. (2014) and Domonkos and Coll (2017a). More details about the homogeneous test datasets can be found in the online supplemental material.

2) INHOMOGENEOUS SECTION OF TEST DATASETS

The inhomogeneous time series include monthly outlier values, shifts of the section means (breaks), gradually increasing deviations of the mean (trend IHs), and short, platform-shaped IHs [as defined by Domonkos (2013b)], yet not all of these IHs occur in each dataset. The mean frequency of any kind of IH is specific for a given dataset, but it varies between time series, and IH positions are fully random. Inhomogeneity magnitudes are characterized by a normal distribution.

In synthetic datasets the inserted breaks are a change relative to the mean station effect (i.e., relative to the homogeneous case), while in surrogate datasets they are added to the bias produced by previous IHs, although the accumulated bias is limited (limited random walk; Domonkos and Coll 2017a). Lindau and Venema (2019) found the former behavior, which they called random deviations, for German temperature data, while IHs in temperature data from the United States were a mixture of both random deviations and random walks. For cases typical in climatology and similar sizes and frequency of the breaks, Brownian motion leads to larger trend errors than random deviations (Lindau and Venema 2020). In some datasets the absolute values of positive and negative bias limits differ (Table 2), which raises the probability of significant network mean trend bias.

The generated length of trend IHs has a uniform distribution between 5 and 99 years. However, long trend IHs are rarer than short ones because parts of their periods often fall outside the limits of the time series. The length of platform-shaped IHs varies between 1 and 120 months, and the frequency quadratically declines with growing length. In some datasets these IHs have elevated magnitudes relative to single breaks (Table 2). The magnitude of outliers has a uniform distribution between 0 and a defined maximum. Note that although platform-shaped IHs might have only 1 month duration, in Table 2 they are considered to be platforms (and not outliers).

The seasonal cycle of station effect is sinusoidal in synthetic datasets, while in surrogate datasets it is semi-sinusoidal or irregularly shaped with the characteristics presented by Domonkos and Coll (2017a). In datasets U5 and U6 the properties partly differ between the first 50 years (U5a and U6a in Table 2) and the last 50 years (U5b and U6b) of the time series.

Some IHs of dataset U6 have differing properties from the general ones. These special break IHs have a magnitude between 0° and 5°C, their frequency decreases linearly with their magnitude, and their random walk is unlimited. Such IHs—one break per 100 years and three short-term platforms per 100 years—were added for U6 to obtain time series with more frequent large IHs and large accumulated biases.

Semi-synchronous breaks with the same sign shifts are included in datasets Y5 and Y6. Break sizes are randomly drawn from a uniform distribution function between 0° and 1°C. In Y5 the timings of these breaks are concentrated in a period of two years, and half of the time series of a given network are affected by them. In Y6, the semi-synchronous breaks are spread over a much longer period; that is, their occurrences are evenly probable within a 30-yr section of the time series, but all of the time series are affected by them.

Systematic trend bias is defined as the absolute value of mean linear trend bias for all of the time series of a given dataset. Occurrences of significant systematic trend biases are the consequences of asymmetry in the frequency and magnitude of positive and negative shifts in the means, because such breaks occur also in observed data (Parker 1994; Vose et al. 2003; Menne et al. 2009; Böhm et al. 2010; Brunet et al. 2011; Hausfather et al. 2013; Acquaotta et al. 2016). Five of our test datasets, namely Y5, Y6, U2, U3, and U6, have significant systematic trend bias in the range of 0.38°–0.84°C century^{-1}, and the systematic trend bias is less than 0.1°C century^{-1} in the other datasets.

Inhomogeneous test data without homogenization are referred also as raw data in the study.

3) GROUPS OF TEST DATASETS

From the spatial correlations and the frequency, magnitude, and type of IHs, we define the groups of high-SNR test datasets and low-SNR test datasets, and the test datasets with semi-synchronous breaks form an additional group:

1) group of high SNR test datasets (G1): Y1, Y2, Y4, and U2;
2) group of low SNR test datasets (G2): Y3, U1, U3, and U4; and

TABLE 2. Inhomogeneity properties of the 12 test datasets: Br = break frequency, Tr = frequency of trend inhomogeneities (IH), Pl = frequency of short platform IHs, and Sybr = frequency of semi-synchronous breaks (all frequencies are in number per 100 yr); σM = standard deviation of IH magnitudes with sign (°C), $M+$ = increment of IH magnitudes for short-term platforms (%), $L-$ = minimum limit of accumulated bias (°C), $L+$ = maximum limit of accumulated bias (°C), Ou = frequency of monthly outlier values (per 100 yr), Ox = maximum of outlier values, Sh = shape of seasonality (Si = sinusoid, Ss = semi-sinusoid, and Ir = irregular), and Am = mean amplitude of seasonal cycles (°C). An asterisk indicates that IHs with magnitude of even distribution between 0° and 5°C have different properties than other IHs of dataset U6 [see section 2b(2) for a fuller explanation].

	Br	Tr	Pl	Sybr	σM	$M+$	$L-$	$L+$	Ou	Ox	Sh	Am
Y1	5	0	0	0	1.0	—	—	—	0	—	Si	0.56
Y2	5	0	0	0	1.0	—	—	—	0	—	Si	0.56
Y3	5	0	0	0	1.0	—	—	—	0	—	Si	0.56
Y4	5	0	0	0	1.0	—	—	—	0	—	Si	0.56
Y5	5	0	0	0.5	0.7	—	—	—	0	—	Si	0.08
Y6	5	0	0	1	0.7	—	—	—	0	—	Si	0.08
U1	4	1	2	0	0.5	50	−1.2	1.2	2	5.0	Ss	0.28
U2	4	1	5	0	0.8	30	−1.0	3.0	5	5.0	Ss	0.64
U3	6	1	15	0	1.5	0	−4.0	6.0	10	6.0	Ir	0.56
U4	4	1	3	0	0.5	50	−1.2	1.2	0	—	Ir	0.28
U5a	4	1	3	0	0.8	30	−1.0	3.0	0	—	Ss	0.64
U5b	4	1	3	0	0.5	30	−1.2	1.2	0	—	Ss	0.28
U6a	5	1	6	0	0.8*	30*	−2.0*	2.0*	0	—	Ir	0.20
U6b	5	1	6	0	0.5*	30*	−1.2*	1.2*	0	—	Ir	0.12
*					2.0	0	—	—				

3) group of test datasets with semi-synchronous breaks (G3): Y5 and Y6.

When a group of test datasets or all test datasets are examined together, each participating test dataset is equally represented in the group, trimming the quantity of contributing networks and time series to the size of the smallest participating test dataset.

c. *Homogenization methods*

Nine versions of five homogenization methods are tested. The five methods are the adapted Caussinus-Mestre algorithm for the homogenization of networks of climatic time series (ACMANT), "Climatol," the multiple analysis of series for homogenization (MASH), the pairwise homogenization algorithm [PHA, also known also as the U.S. Historical Climatology Network (USHCN) method], and the penalized maximal t test of the relative homogenization tests (RHtests) package (RHtests-PMT)]. Two of the five methods, ACMANT (Domonkos and Coll 2017b) and MASH (Szentimrey 1999), are multiple-break methods in the sense that they detect and correct multiple break structures with joint operations. By contrast, Climatol (Guijarro 2018), PHA (Menne and Williams 2009), and RHtests (Wang et al. 2007) apply a hierarchic algorithm of break detection based on the standard normal homogeneity test (SNHT; Alexandersson 1986) and cutting algorithm (Easterling and Peterson 1995). Data gap filling is part of the homogenization procedure in three methods (ACMANT, Climatol, and MASH), and all of the methods except RHtests have built-in routines for filtering outlier values. All the nine method versions are fully automatic, except that the selection of reference series is not provided in the automatic procedure in MASH and RHtests. The tested methods are freely available (see http://www.climatol.eu/tt-hom/index.html and its links).

1) ACMANT

ACMANT was developed from "PRODIGE" (Caussinus and Mestre 2004), keeping its principal detection and correction routines but adding new features. In ACMANT, the candidate series are compared with composite reference series (Peterson and Easterling 1994; Domonkos 2011b); step function fitting with the Caussinus–Lyazrhi criterion is applied for break detection (Caussinus and Lyazrhi 1997), and the ANOVA model is used for the correction of IHs (Caussinus and Mestre 2004; Mamara et al. 2014; Lindau and Venema 2018). See further details of ACMANTv3 (AC3) in Domonkos and Coll (2017b). The most recent version, ACMANTv4 (AC4; Domonkos 2020), has several novelties. The most important ones are as follows: (i) The fully automatic treatment is extended up to datasets of 5000 series (Domonkos and Coll 2019), (ii) ensemble homogenization in the third (last) phase is incorporated from varied prehomogenization scenarios, and (iii) a weighted ANOVA model is included for the assessment of correction terms (Szentimrey 2010; Domonkos 2017), which considers spatial differences in the regional climatic changes [component C_A of Eq. (3)].

2) CLIMATOL

The Climatol homogenization package (Guijarro 2018) performs its process in three main stages, with many iterations within them. The first two stages are devoted to removing unwanted outliers and splitting the series into two fragments at the position where the SNHT statistic is the highest. Successive iterations refine the process until neither outliers nor breaks are detected over preset thresholds. In the first stage, SNHT is

applied for overlapping windows along the series to reduce possible masking problems when several shifts are present, and in the second stage SNHT is applied over the whole series, getting the full power of the test. These outlier rejection and shift detection steps are performed over the series of anomaly differences between the observed data and a composite reference series built from a number of nearby sites, in both cases in normalized form. The third stage is dedicated to assigning synthetic values to all missing data in all series and subseries originated in the splitting process, with spatial interpolation, using the data of nearby stations. Two Climatol parameterizations are tested, referred to as Climatol-1 (Cl1) and Climatol-2 (Cl2), respectively. Both normalize the time series with the removal of the long-term mean, but only Cl2 divides the centered values by the standard deviation, thereby yielding season-dependent adjustment terms.

3) MASH

MASH applies multiple reference series for time series comparison and hypothesis tests to find the most likely break structure. The significance thresholds are calculated with Monte Carlo simulation (Szentimrey 1999). The correction of IHs is iterative: confidence intervals of break sizes are calculated, and the minimums of these confidence intervals are applied as adjustment terms in an iteration step. In the monthly homogenization, breaks are searched independently for each calendar month. Although the manual of the software (Szentimrey 2014) includes the description of a fully automatic algorithm, the recommended use of MASH is interactive and semiautomatic (T. Szentimrey 2017, personal communication). In MASH, the maximum number of reference series is nine, and the set of reference series must be prepared before running the automatic program. Synchronous data gaps for all series are not allowed. In our tests, all the partner series (i.e., all the series of a given network but the candidate series; Domonkos and Coll 2017b) are used as reference series in networks of no more than 10 time series, while the sets of reference series are generated in the same way as for ACMANT in the reverse case, considering also the mentioned limitations of MASH.

We examine two automatic algorithms of MASHv3.3. "MASH monthly" (MSm) is identical with the manual's recommended algorithm, while in "MASH annual" (MSy) the monthly break detection is omitted.

4) PHA

As an initial step, networks of sufficiently correlated series are formed with up to 40 time series. Then difference series are calculated with pairwise comparisons using all possible pairs of time series, and the significant breaks of a difference series are ticked in both series. When the break detection in all difference series has been completed, the numbers of coincident detection results for time series and dates are summed, and the breaks with the largest numbers of coincident detection results are retained, while the potential breaks with the same date are cancelled from the other time series. For the assessment of a break size of the candidate series, differences relative to the other series are considered again one by one, using only homogeneous subsections of the other series in the network. Each of these comparisons results in one independent estimation for the break size, and then a confidence interval is calculated from the individual estimations. If the confidence interval includes zero, then the break is cancelled from the break list, while the median of the individual size estimations is applied as a correction term in the reverse case (Menne and Williams 2009).

5) RHTESTS

An automatic version of penalized maximal t test (PMT) of the RHtestsV4 software package (Wang and Feng 2013) was selected. This test differs from SNHT single-break detection (Alexandersson 1986) only in the significance thresholds (Wang et al. 2007). The present version modifies further the significance thresholds according to the estimated autocorrelation of difference series (Wang 2008), and trend IHs may be part of the detection results (Wang 2003). The hierarchic organization of break detection and the calculation of correction terms are the same as in the homogenization with SNHT (Moberg and Alexandersson 1997). One further option in RHtests is the application of quantile matching (QM; Wang et al. 2010). In our tests, two RHtests versions are applied: one without QM is referred to as RHtests basic or RHT, while the other version includes QM and is referred to as RHtests-QM or RHQ.

Running RHtests needs the previous preparation of reference series. In our tests, the reference series of Climatol are used also in RHtests, with the exception that the number of reference composites (partner series) is maximized at 9 for RHtests.

d. *Efficiency measures*

We use efficiency measures that directly show the appropriateness of homogenized time series for climate variability and climate impact studies. Let $\mathbf{V}$ and $\mathbf{Z}$ denote the vectors of homogeneous series and homogenized series, respectively; $\mathbf{V}$ is a special case of the general model [Eqs. (3) and (4)] with $\mathbf{S} \equiv 0$. The homogenization result $\mathbf{Z}$ is successful when S (referred as residual error) is generally small, and the other components are the same as in $\mathbf{V}$. Residual errors of homogenized time series are characterized by the centered root-mean-square error (CRMSE) and the bias of linear trend. These characteristics are calculated for individual time series and also for regional network-mean averages. For each of these efficiency measures the average error and the threshold for the largest 5% errors (P95) are calculated. The mean systematic trend bias for datasets is also monitored.

Note that although ACMANT, Climatol, and MASH fill the data gaps with spatial interpolation, these filled values are never used in the evaluation of the homogenization results. However, the accuracy of interpolated values is evaluated in a separate examination.

1) CENTERED ROOT-MEAN-SQUARE ERROR

The use of the CRMSE instead of the common root-mean-square error in measuring homogenization efficiency was introduced by Venema et al. (2012). The idea behind it is that two homogenization results that only differ by a constant are considered to be the same, as the objective of the homogenization is to eliminate any nonclimatic component of the temporal variability and is not (and usually cannot be) to assess climatic

means and spatial differences. Therefore, the mean difference is extracted before the calculation of quadratic errors, as is shown by the following equation for a time series of n observed values:

$$\mathrm{CRMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[z_i - v_i - \frac{1}{n}\sum_{j=1}^{n}(z_j - v_j)\right]^2}. \quad (5)$$

We use CRMSE for monthly values (RMSEm), CRMSE for annual values (RMSEy), and network mean annual CRMSE (NetEy). For obtaining NetEy, first the network mean average errors are calculated for every year, and then Eq. (5) is applied to them. In case of missing data occurrences, the spatial average of the observed data may differ from the representative network mean value; the treatment of this issue is described in section 2d(3).

2) TREND BIAS

Trends are computed using least squares linear regression, and the trend slopes of **V** and **Z** are denoted with α_v and α_z, respectively. The trend bias of individual time series (Trb) is defined by

$$\mathrm{Trb} = |\alpha_z - \alpha_v|. \quad (6)$$

For the calculation of network mean trend bias (NetTr), first the spatial averages of annual values are calculated, then Eq. (6) is applied on the annual means. The systematic trend bias for a dataset (SysTr) is the average of the network mean trend biases, where, differing from Eq. (6), the trend biases are summed with their signs:

$$\mathrm{SysTr} = \frac{1}{K}\left|\sum_{K}(\alpha_{\overline{Z}} - \alpha_{\overline{V}})\right|. \quad (7)$$

In Eq. (7), the overbar denotes the spatial average and K is the number of networks in the dataset.

While Trb is calculated for the period with observed data that can be equal to or shorter than n, NetTr and SysTr are always calculated for the entire period of n. If the length of individual time series in dataset is varied, or data gaps occur, spatial averages of the observed data may differ from the representative network mean value. The treatment of this issue is described in the following section.

3) CONSISTENT ESTIMATION OF NETWORK MEANS

Sometimes the spatial averages must be calculated from a subnetwork of N^* series ($N^* < N$) because of data gaps in some of the time series. Subnetwork means may differ from the means of the whole network due to the spatial differences of local climate. A consistent estimation for the network mean values can be provided with the help of the periods without missing data. Let n^* denote the number of years without missing data in all stations of a network. In every network of our test datasets $n^* \geq 40$. For year i of series **X** the consistent estimation is presented by

$$\overline{x}_{N,i} = \frac{1}{N^*}\sum_{N^*} x_i + \frac{1}{n^*}\sum_{j=1}^{n^*}(\overline{x}_{N,j} - \overline{x}_{N^*,j}). \quad (8)$$

When $N^* < N$, this adjustment is always applied before the calculations of NetEy, NetTr, and SysTr, for both series **V** and **Z**.

4) ACCURACY OF INTERPOLATED VALUES

The accuracy of interpolated monthly values of gap filling is evaluated by calculating the CRMSE for them, as in section 2d(1).

e. *Statistical significance of the lowest mean residual errors*

The stability of the rank order between the homogenization method with the lowest mean residual error and any other homogenization method is examined, where the rank order is based on the mean residual errors. A bootstrapping is performed in which subsamples of 100 time series (networks) for RMSEm, RMSEy, and Trb (NetTr and NetEy) are selected randomly 2000 times when the sample size m is larger than 200. The mean residual errors in a subsample are calculated for both of the compared homogenization methods, and the frequency of their rank order is inferred from the 2000 experiments. For $m \leq 200$, the samples are sorted to two equally large subsamples 1000 times, and both subsamples are used in the bootstrapping.

3. Results

a. *Variation of efficiency according to efficiency measures*

In Fig. 1 the mean residual errors are shown as ratios of raw data errors. The error bars show the range of the residual errors with the different homogenization methods, while their mean distance from the horizontal axis shows the normalized error magnitude. The perceived success of homogenization notably depends on the efficiency measure. Generally, Trb can be reduced most, while the reduction of NetTr is the least successful. The spread of the efficiencies according to homogenization methods is relatively narrow for RMSEy and Trb, while it is wider for the network mean errors. Note that the results for SysTr likely have large sampling errors, as no more than five datasets are characterized by significant systematic trend bias of the raw data. The long error bar of RMSEm in Fig. 1a is due to the exceptionally large mean residual error with RHtests-QM, which distorts the general picture; therefore, RHtests-QM is omitted from the other three panels of Fig. 1. Most of the mean residual errors are below 1, which means that the residual errors are generally smaller than the raw data errors. Exceptions for the means of the 12 datasets (Figs. 1a,b) occur only with RHtests, while for the group of low SNR datasets such large mean errors occur with some other methods.

For group G1 (high SNR) the error bars for network mean errors are relatively large, indicating that the efficiency of homogenization strongly depends on the homogenization method applied. By contrast, for group G2 (low SNR) the error bars are generally small, and the residual network mean errors do not differ markedly from the raw data error.

b. *Results for all test datasets*

All test datasets are examined together. Table 3 and Fig. 2 show the arithmetical means and Fig. 2 also shows some

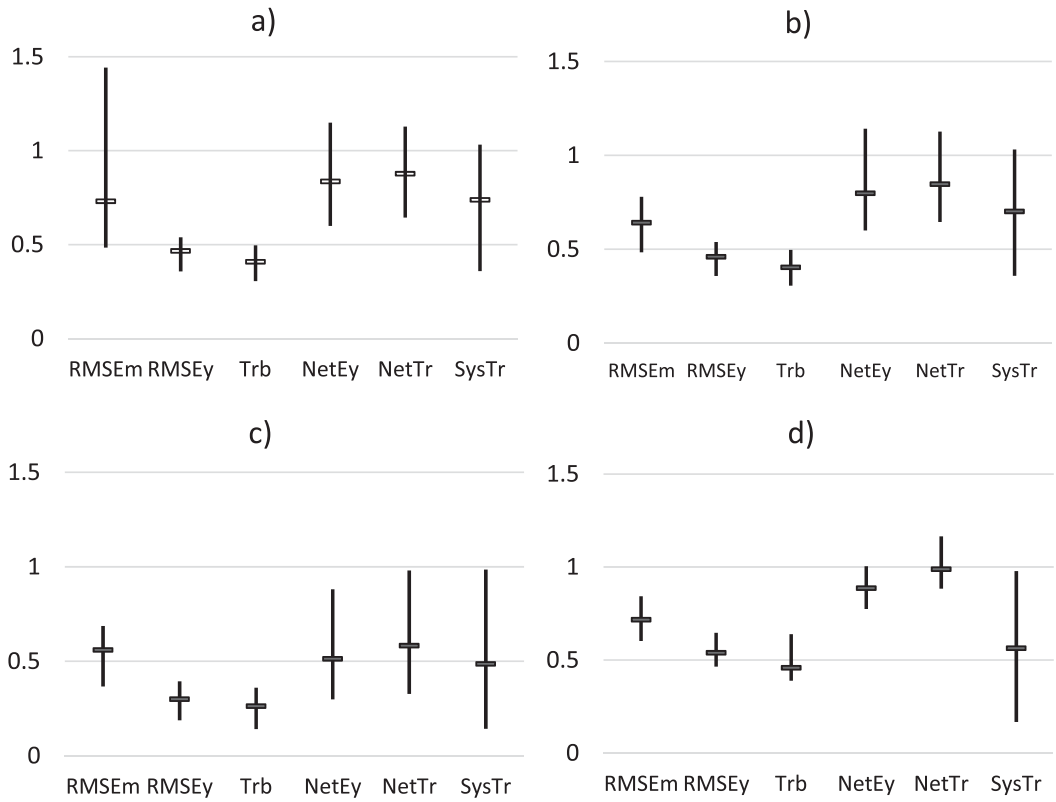

FIG. 1. Mean residual errors after homogenization for groups of test datasets, showing the ranges of the results of the applied homogenization methods. The results are normalized with the raw data errors. Horizontal sticks on the bars show the mean errors of all homogenization methods. RMSEm = centered root-mean-square error for monthly data, RMSEy = centered root-mean-square error for annual data, Trb = bias of linear trend for the whole data period in a time series, NetEy = centered root-mean-square error for network mean annual means, NetTr = network mean bias of linear trends for the whole study period of a dataset, and SysTr = systematic trend bias for entire datasets. (a) Mean of the examined 12 test datasets. (b) As in (a), but with the exclusion of RHtests-QM. (c) As in (b), but for G1 (datasets of high SNR). (d) As in (b), but for G2 (datasets of low SNR).

selected percentiles of the residual errors for each efficiency measure and each homogenization method. For SysTr only the arithmetical means are shown for the low sample size.

Whereas for the raw data the mean RMSEy is only 22% smaller than the mean RMSEm, this difference is 40%–48% for the homogenized data. By contrast, the error decrease from Trb to NetTr is larger for the raw data (62%) than for the homogenized data. In seven of the nine methods the mean NetTr is only 2%–20% smaller than the mean Trb. The exceptions are MASH monthly and PHA with 33% and 42% lower values of NetTr than Trb, respectively. This indicates that PHA and MASH monthly give better results for network mean series than for individual time series.

TABLE 3. Mean residual errors for all test datasets; the errors are ordered from the lowest to the highest. RMSEm, RMSEy, and NetEy are in degrees Celsius; Trb, NetTr, and SysTr are in degrees Celsius per 100 yr. "Raw" (in italics) indicates errors without homogenization.

	RMSEm		RMSEy		Trb		NetTr		NetEy		SysTr	
1	AC4	0.341	AC4	0.199	AC4	0.443	AC4	0.360	AC4	0.125	PHA	0.103
2	AC3	0.367	AC3	0.219	AC3	0.492	AC3	0.406	AC3	0.138	AC4	0.148
3	Cl2	0.410	Cl2	0.238	Cl2	0.544	PHA	0.412	PHA	0.147	AC3	0.156
4	Cl1	0.463	MSy	0.252	Cl1	0.560	MSm	0.436	MSm	0.162	MSy	0.172
5	MSy	0.484	Cl1	0.258	MSy	0.611	Cl2	0.492	Cl2	0.167	MSm	0.179
6	PHA	0.499	PHA	0.285	RHT	0.632	Cl1	0.495	Cl1	0.171	Cl1	0.229
7	RHT	0.501	RHT	0.286	RHQ	0.637	MSy	0.505	MSy	0.171	Cl2	0.232
8	MSm	0.550	RHQ	0.291	MSm	0.646	*Raw*	*0.550*	*Raw*	*0.207*	*Raw*	*0.286*
9	*Raw*	*0.707*	MSm	0.299	PHA	0.709	RHT	0.614	RHT	0.234	RHT	0.294
10	RHQ	1.018	*Raw*	*0.556*	*Raw*	*1.439*	RHQ	0.614	RHQ	0.235	RHQ	0.294

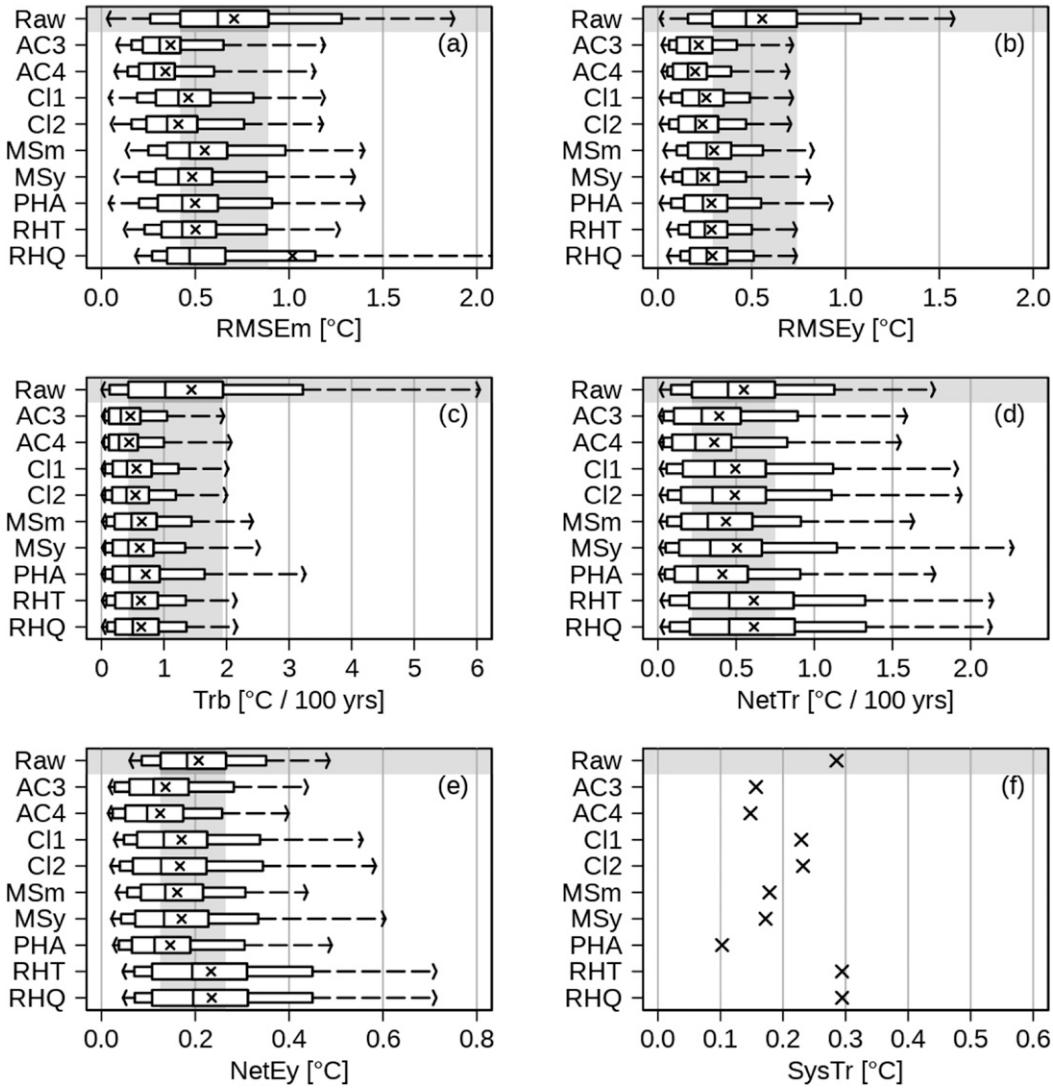

FIG. 2. The whiskers indicate the data between the 2nd and 98th percentiles (P02 and P98) for the homogenization results of all experiments. The section borders are at P10, P25, P50, P75, and P90; x indicates the arithmetic mean.

The residual errors in individual time series (i.e., RMSEm, RMSEy, and Trb) are the smallest with AC4, AC3, and Climatol-2 in this order, but note that the differences according to homogenization methods are generally not very large for these error types, except for the higher percentiles of RMSEm with RHQ method. Regarding the network mean errors, they are the smallest with AC4, PHA, and AC3, but for the higher percentiles the values are smaller with MASH monthly than with AC3 and PHA. The mean systematic trend bias averaged for the 12 test datasets is the smallest with PHA (Table 3).

c. *Results for groups of test datasets*

Figure 3 shows the residual errors for the high SNR group (G1). The errors for individual time series are markedly lower for AC4 and AC3 than for any other homogenization method. For network mean errors, again the AC4 errors are the smallest, although the error distributions of AC3 and PHA are somewhat similar to that of AC4.

Figure 4 shows the mean residual errors for the low SNR group (G2). Here, the reduction of raw data error is generally smaller than average (Fig. 2), and the differences according to homogenization methods are rather small, as was also shown in Fig. 1d. The errors are generally larger for PHA and MASH (both versions) than for the other methods, except for the network mean errors of MSm. The higher percentile errors of RMSEy and Trb are smaller for RHtests (both versions) and Climatol (both versions) than for the other methods. For network means, MASH monthly gives the smallest trend errors and AC4 gives the smallest CRMSE, but their advantage in comparison with the other methods is minor.

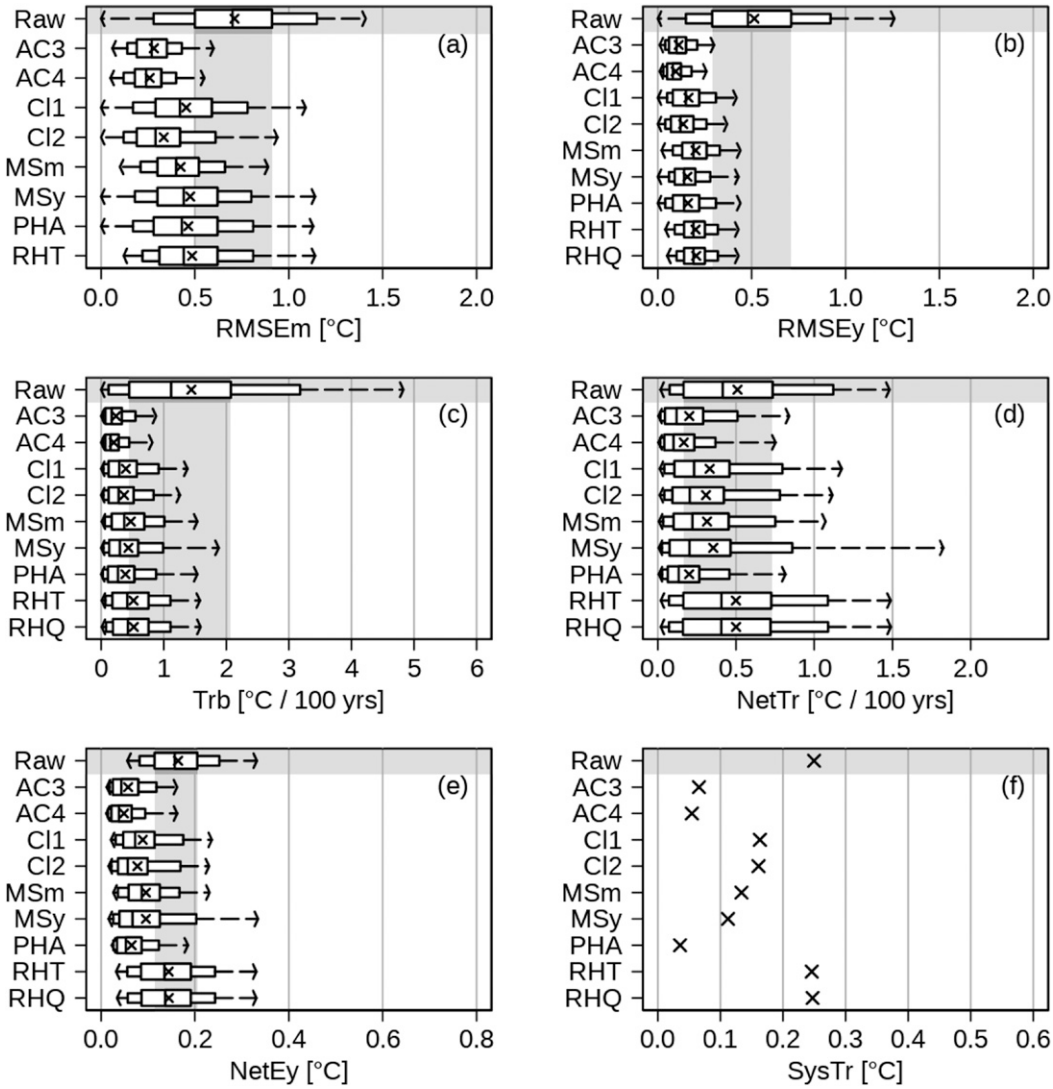

FIG. 3. As in Fig. 2, but for the group of high-SNR datasets (group G1).

Figure 5 shows the mean residual errors for the group of datasets with semi-synchronous breaks (G3). The errors for individual time series are the smallest with AC4, AC3, and MASH annual, while MASH monthly leaves markedly larger RMSEm than any other method. The reduction of Trb is generally lower here than for other test datasets because the removal of semi-synchronous inhomogeneities is partly unsuccessful. The residual network mean errors are the lowest with PHA in the lower half of the error distribution and with AC4 in the upper half of the error distribution. The residual errors are markedly larger than the raw data errors for the upper quartile of NetTr and NetEy with RHtests (both versions) and Climatol (both versions).

d. *Accuracy of data in gap filling*

Four of the 12 test datasets contain time series of varied lengths and missing data, but the homogeneous set without missing data was saved only for two datasets, Y5 and Y6. Therefore, these two datasets are used for calculating the CRMSE of the interpolated monthly data, and their results are shown in Fig. 6. In this task MASH annual, MASH monthly, AC4, and Climatol-2 provided the best results in this order, with small differences in their mean errors. By contrast, the mean errors of AC3 and Climatol-1 are notably larger, while PHA and RHtests do not provide completed time series.

e. *Stability of efficiency rank order*

The stability of rank order for the homogenization method with the lowest residual error is calculated for each test dataset and efficiency measure. The results are presented in Table 4. When the rank order is not significantly stable at the 0.05 significance level for two or more methods, all of the involved methods are shown as best method in Table 4. The results show that the advantage of AC4 in reducing

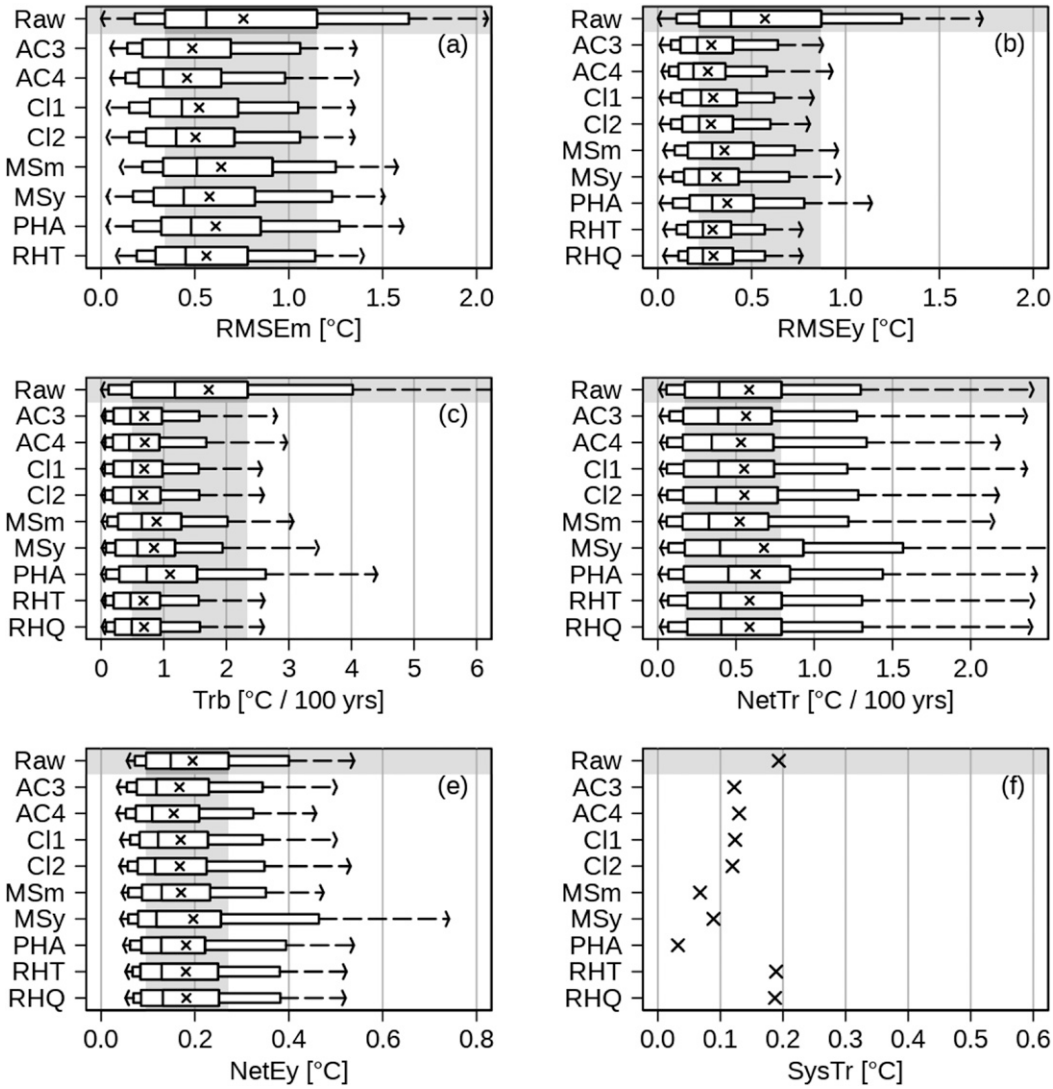

FIG. 4. As in Fig. 2, but for the group of low-SNR datasets (group G2).

nonclimatic biases is mostly statistically significant. Often relatively small differences in the error reductions are still significant, but in low SNR homogenization tasks (datasets of G2) the differences are often insignificant. The NetTr reduction in dataset Y5 is an exceptional result, there the PHA produces significantly smaller errors than any other method. Table 4 also shows that the accuracy of interpolated monthly values is significantly higher with MSy than with any other method.

The significance of SysTr results is not examined with bootstrapping for the low sample size (12). We have compared the results of the two best performing methods for SysTr (i.e., PHA and AC4) for each test dataset (not shown). PHA yields smaller SysTr than AC4 in six datasets, AC4 produces smaller residual errors than PHA in two datasets, while the difference is less than 0.01°C century^{-1} in the remaining four datasets. These together indicate that the advantage of PHA in comparison with AC4 is not significant statistically in the SysTr reduction.

4. Discussion

We have used 12 large size test datasets with varied climatic and highly varied IH properties. Although a large number of other combinations of climate and IH properties occur in nature, we think that the representativeness of the test datasets used here is sufficient, at least for mid- and high-latitude geographical areas. We base this hypothesis on the fact that the differences between method efficiencies are similar for synthetic and surrogate datasets (see the online supplemental material), and they are not too much influenced by the variation of IH structures either. The only characteristic that seems to markedly impact the rank order of efficiencies is the SNR, which depends on several factors: the spatial correlations, the

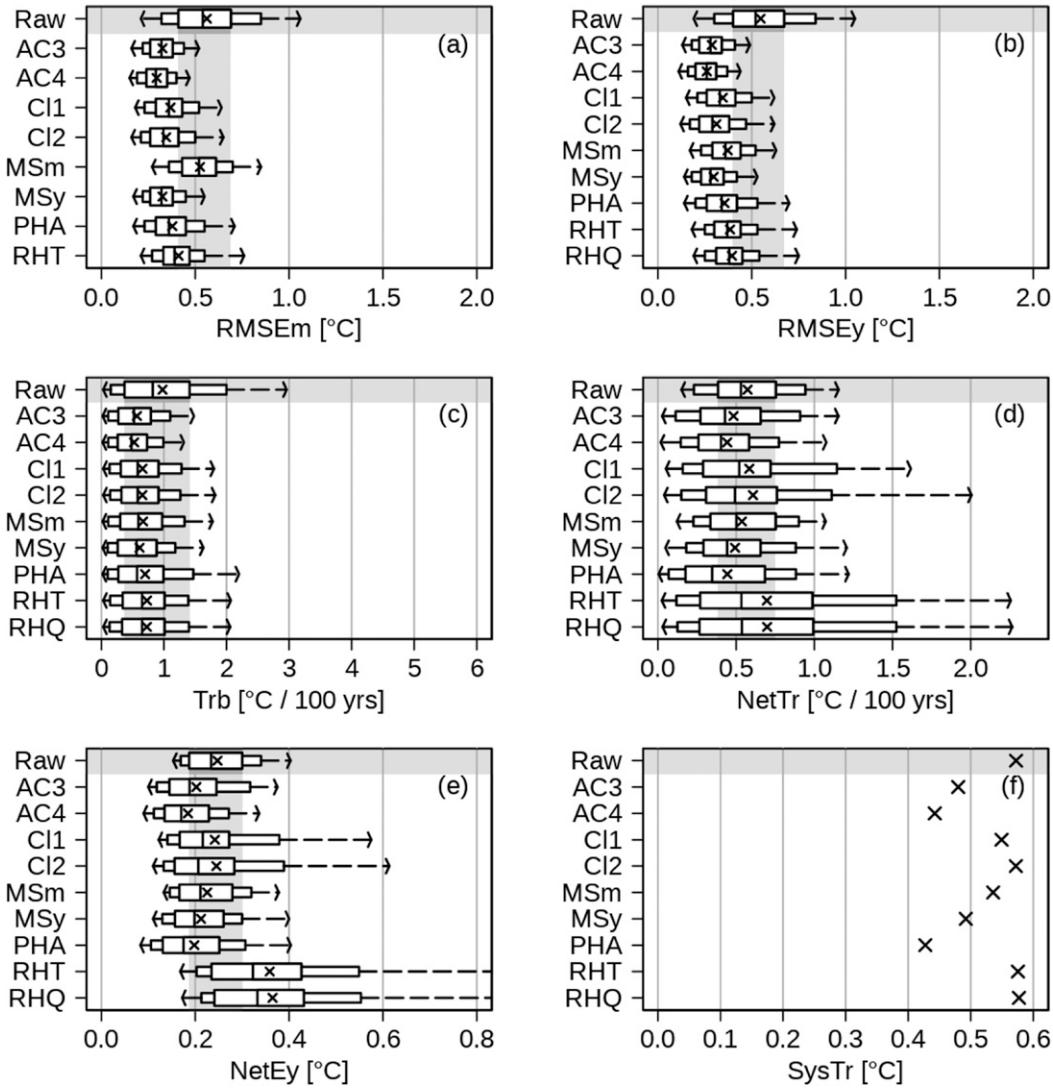

FIG. 5. As in Fig. 2, but for the datasets with semi-synchronous breaks in time series (group G3).

magnitude distribution of breaks, the number of time series, the completeness of time series, and the occurrence of outliers and short-term IHs. Therefore, despite the relatively good representativeness of the results, the involvement of further test datasets would be desirable. It would be important to develop and use validation datasets of tropical climates, as the spatio-temporal variation and spatial correlation of tropical temperatures differ from the characteristics of midlatitude temperatures.

A large number of methods are in use in climate data homogenization, but we could only test five methods. Only automatic methods can be tested on the large datasets we used, and the accessibility of the homogenization methods is also a constraint. We naturally cannot assess directly whether tested homogenization methods are more accurate than methods not subject to testing, but we have the general experience that tests helps to improve homogenization accuracy, due to the complexity of homogenization tasks. Therefore, we expect that homogenization methods objectively tested with high-quality test datasets tend to be more accurate than not-tested methods. We recommend use of thoroughly tested methods whenever it is possible, and we recommend creation of automatic versions of interactive homogenization methods to promote the performance of objective tests also for such methods. A further task is to test homogenization methods together with metadata use, but as metadata are generally not quantitative information, the development of such tests is a complicated task.

One conclusion of HOME was that multiple-break methods such as ACMANT and MASH generally perform better than hierarchic methods (Venema et al. 2012), and the theoretical advantages of multiple-break techniques are widely discussed (Lindau and Venema 2013; Szentimrey et al. 2014; Domonkos 2017). However, this study does not confirm the superiority of multiple-break techniques, at least not in all aspects. One should distinguish between two concepts: traditional hierarchic methods including both their detection and IH adjustment parts on the one hand, and hierarchic break detection methods

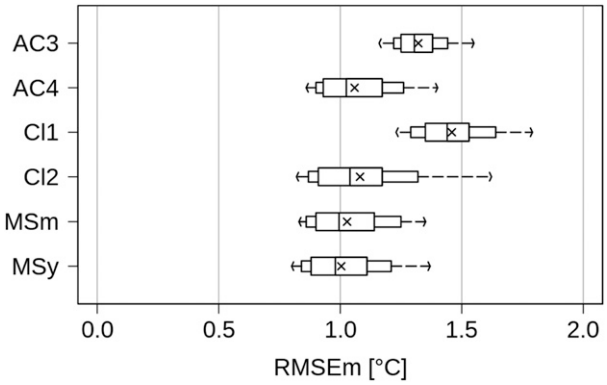

FIG. 6. CRMSE of interpolated monthly values in the gap filling of G3. The section borders of boxes and whiskers indicate the same error percentiles as in Fig. 2.

without other steps of traditional homogenization algorithms on the other hand. Regarding the methods examined here, three of them use the hierarchic break detection method of SNHT (i.e., Climatol, PHA, and RHtests), but only RHtests applies the traditional IH adjustment method of SNHT. Earlier tests showed that although the break detection with SNHT is less powerful than with optimal step function fitting, the difference of the efficiencies is small (Menne and Williams 2005; Domonkos 2011a, 2013b). By contrast, the traditional way of IH correction with SNHT-based methods often results in significantly larger residual errors than the joint correction of IHs with the ANOVA model (Domonkos et al. 2011). The tests presented here show that the combination of SNHT detection with some novel approaches others than the ANOVA correction may also produce very good results.

During MULTITEST, some automatic versions of the interactive homogenization method "HOMER" (see Mestre et al. 2013) were also tested, but we found the following problem: homogenization results of HOMER depend on the variation of the background climate C_R (Guijarro et al. 2017), which is not allowed in relative homogenization methods [this problem is reported also by Gubler et al. (2017)]. The dependence on C_R makes the HOMER results incompatible with the results of other methods, so we excluded HOMER from this study. Naturally, the deviation from the conditions of relative homogenization with HOMER has consequences on the practical use of HOMER (Domonkos 2017).

In most of the tests and efficiency measures presented here, AC4 shows the best results among the tested methods, and when AC4 is not the first in the efficiency rank order, its difference from the best performing method is small and not statistically significant with very few exceptions. The development of ACMANT is based on the incorporation of theoretically sophisticated methods and the continuous test of the method performance (Domonkos and Coll 2017a,b, 2019). AC4 is likely the most appropriate automatic homogenization method available at present when the accuracy of data is important both at station level and for larger geographical areas. Earlier tests showed insignificant differences between the accuracies of ACMANT and Climatol (Killick 2016;

TABLE 4. Homogenization method with the lowest residual error for each test dataset, group of test datasets, and efficiency measure. When the rank order is unstable for the method with the lowest residual error, more methods are shown as best method, and all of them are set as boldface. In the case of significant rank order but smaller than 10% difference in the residual error, the additional methods are shown in standard style. "Intp" means CRMSE of interpolated monthly values.

	RMSEm	RMSEy	Trb	NetTr	NetEy	Intp
Y1	**AC4**, AC3	**AC4**	**AC4**, **AC3**	**AC4**, **AC3**, **PHA**	**AC4**	
Y2	**AC4**, AC3	**AC4**	**AC4**	**AC4**, **PHA**	**AC4**	
Y3	**AC4**, **Cl2**, AC3, MSy	**MSy**, **AC4**, **Cl2**, **RHT**, **AC3**, **RHQ**, **Cl1**	**RHT**, **RHQ**, **Cl2**, **Cl1**, **MSy**, **AC4**	**All methods**	**AC4**, **PHA**, **Cl2**, AC3, Cl1, MSy	
Y4	**AC4**, AC3	**AC4**	**AC4**, AC3	**AC4**	**AC4**	
Y5	**AC4**, Cl2	**AC4**	**AC4**, **AC3**	**PHA**	**AC4**, **PHA**, AC3	**MSy**, MSm, AC4, Cl2
Y6	**AC4**	**AC4**	**AC4**, AC3	**AC4**, **MSy**, PHA, AC3	**AC4**	**MSy**, MSm, AC4, Cl2
U1	**AC4**, AC3	**AC4**	**RHT**, **AC4**, **RHQ**, **Cl2**	**RHT**, **Cl2**, **MSm**, **AC4**, RHQ	**AC4**, MSm, Cl2	
U2	**AC4**	**AC4**	**AC4**	**AC4, PHA**	**AC4**	
U3	**AC4**, AC3, Cl1, Cl2	**AC4**, **RHT**, **RHQ**, **Cl2**, Cl1	**Cl1**, **Cl2**, **AC4**	**AC4**, **Cl1**, **MSm**, **Cl2**	**AC4**, MSm, Cl1, Cl2	
U4	**AC4**, AC3, Cl1, MSy, Cl2	**AC4**, AC3, Cl2	**AC4**, **AC3**, **RHT**, **Cl2**, **RHQ**, **Cl1**	**MSm**, **AC4**, **AC3**, **Cl2**, **Cl1**, **MSy**, PHA, RHT, RHQ	**AC4**, **AC3**, **Cl2**, Cl1, MSy, MSm	
U5	**AC4**	**AC4**	**AC4**, **AC3**	**AC4**	**AC4**	
U6	**AC4**, AC3	**AC4**, AC3	**AC4**, **AC3**	**AC3**, **AC4**	**AC4**, AC3	
G1	**AC4**, AC3	**AC4**	**AC4**	**AC4**	**AC4**	
G2	**AC4**, AC3, Cl2	**AC4**, **Cl2**, AC3, RHT	**Cl2**, **RHT**, **RHQ**, **Cl1**, **AC4**, **AC3**	**MSm**, **AC4**, **Cl1**, **Cl2**	**AC4**, Cl2, Cl1, AC3	
G3	**AC4**	**AC4**	**AC4**, AC3	**AC4**, **PHA**, AC3	**AC4,** PHA, AC3	**MSy**, MSm, AC4, Cl2

Guijarro et al. 2019); however, in those tests much less statistically independent networks were used, and therefore network mean errors were not examined. On the other hand, the accuracy of station-level data of Climatol is close to that of ACMANT also in our results (e.g., 0.07°C mean difference for RMSEm; see Table 3). In low-SNR experiments, we have found relatively low station-level errors for Climatol and RHtests in the range of the higher percentiles of the error distribution. However, the realization of this advantage in real world homogenization tasks is uncertain for the following reason: In the creation of the composite reference series with Climatol (included also in the RHtests performed in our study), each partner series of a given network has the same weight, independently from its geographical distance or spatial correlation. This is likely a good approach for networks of highly correlated time series, but might result in larger errors when the spatial correlations are highly varied and adjacent stations might represent different climatic areas. In our test datasets every time series of a network includes the same climate signal, which is not necessarily true in real world homogenization tasks when the spatial correlations are low. This deviation from the true dataset properties might affect the results of the Y3, U3, and U6 datasets and group G2.

The performance of the pairwise homogenization algorithm in reducing network mean errors is generally close to that of AC4. When synchronous breaks occur within a short period (as in dataset Y5), the network mean trends are significantly more accurate with PHA than with ACMANT. In this specific case the pairwise comparison (of PHA) is a more powerful tool than the composite reference series (of AC4), at least when the SNR allows the algorithm to find the homogeneous sections of partner series during the pairwise comparisons. In the MULTITEST project we made tests also with larger synchronous breaks than in this study, applying break magnitudes of up to 2.0°C (not shown). When large magnitude breaks were concentrated within a period of a few years, PHA provided higher error reduction than AC4 in all efficiency measures. Another positive characteristic of PHA is the low residual systematic trend bias for test datasets. A likely reason of the higher accuracy for larger networks seen in the PHA results is that the PHA algorithm does not use iteration. Iterations tend to facilitate more accurate results for station-level data, but as the same pieces of information including error terms are repeatedly used in iterations, they may cause error accumulations in the area-averaged data.

Among the tested methods only ACMANT does not use metadata, which may appear as an important drawback of ACMANT. However, some recent studies indicate that the use of metadata within automatic homogenization procedures does not always result in significant improvement in the accuracy (Gubler et al. 2017; Domonkos et al. 2020).

5. Summary and conclusions

Nine versions of five automatic monthly temperature homogenization methods were tested with 12 large test datasets. The frequency and size of inhomogeneities and other properties of the test datasets are varied, and the vast majority of the results confirm that biases due to nonclimatic effects can be notably reduced with time series homogenization.

The instrumental temperature record is a product of the diligent work of several generations all over the world. Homogenization is a key step to turn this enormous effort into accurate climate change data products. Appropriate computer programs for the automatic homogenization of climatic time series are usually the result of a development work of several years. Although some benchmark datasets have been developed and tests have been conducted in the recent years as a part of national projects or international initiatives (Williams et al. 2012; Rennie et al. 2014; Willett et al. 2014; Chimani et al. 2018; Squintu et al. 2020), the overall attention toward the development and testing of homogenization methods is small in comparison to its value.

The main conclusions are as follows:

- Homogenization improves the data accuracy in the vast majority of the examined cases.
- Mostly ACMANTv4 provides the most accurate homogenization. With respect to the accuracy of individual time series, the advantage of ACMANTv4 in comparison with the second-best method Climatol-2 is generally small but statistically significant. For the accuracy of network mean characteristics, the advantage of ACMANTv4 in comparison with the second-best method PHA is generally small but often statistically significant.
- In low-SNR exercises, a small improvement in data accuracy still can be achieved, and the accuracy of ACMANTv4 ties in first place with Climatol-2 and sometimes also with some other homogenization methods.
- When semi-synchronous breaks occur within a period of a few years in a large portion of the time series, PHA provides the most accurate network mean trends.
- The systematic trend bias for entire homogenized test datasets is the smallest with PHA, although the advantage of PHA in comparison with ACMANTv4 is not statistically significant.

We recommend the use of the ACMANTv4 homogenization method when the number of time series and their spatial correlations allow the use of automatic homogenization. For the accurate calculation of climatic trends over large geographical areas we recommend both the PHA and ACMANTv4 methods.

Acknowledgments. This research was funded by the Spanish MULTITEST project (Ministry of Economics and Competitiveness, CGL2014-52901-P).

Data availability statement. The test datasets and the homogenization results are accessible online (https://zenodo.org/record/3934835#.XwTjF-dS_IU).

REFERENCES

Acquaotta, F., and S. Fratianni, 2014: The importance of the quality and reliability of the historical time series for the study of climate change. *Rev. Bras. Climatol.*, **14**, 20–38, https://doi.org/10.5380/abclima.v14i1.38168.

——, ——, and V. Venema, 2016: Assessment of parallel precipitation measurements networks in Piedmont, Italy. *Int. J. Climatol.*, **36**, 3963–3974, https://doi.org/10.1002/joc.4606.

Aguilar, E., I. Auer, M. Brunet, T. C. Peterson, and J. Wieringa, 2003: Guidelines on climate metadata and homogenization. WMO Tech. Doc. WCDMP-53, WMO/TD-1186, 51 pp.

Alexandersson, H., 1986: A homogeneity test applied to precipitation data. *J. Climatol.*, **6**, 661–675, https://doi.org/10.1002/joc.3370060607.

Auer, I., and Coauthors, 2005: A new instrumental precipitation dataset for the greater Alpine region for the period 1800–2002. *Int. J. Climatol.*, **25**, 139–166, https://doi.org/10.1002/joc.1135.

Beaulieu, C., O. Seidou, T. B. M. J. Ouarda, X. Zhang, G. Boulet, and A. Yagouti, 2008: Intercomparison of homogenization techniques for precipitation data. *Water Resour. Res.*, **44**, W02425, https://doi.org/10.1029/2006WR005615.

Böhm, R., P. D. Jones, J. Hiebl, D. Frank, M. Brunetti, and M. Maugeri, 2010: The early instrumental warm-bias: A solution for long central European temperature series 1760–2007. *Climatic Change*, **101**, 41–67, https://doi.org/10.1007/s10584-009-9649-4.

Brunet, M., and Coauthors, 2011: The minimization of the screen bias from ancient western Mediterranean air temperature records: An exploratory statistical analysis. *Int. J. Climatol.*, **31**, 1879–1895, https://doi.org/10.1002/joc.2192.

Caussinus, H., and F. Lyazrhi, 1997: Choosing a linear model with a random number of change-points and outliers. *Ann. Inst. Stat. Math.*, **49**, 761–775, https://doi.org/10.1023/A:1003230713770.

——, and O. Mestre, 2004: Detection and correction of artificial shifts in climate series. *J. Roy. Stat. Soc.*, **53**, 405–425, https://doi.org/10.1111/j.1467-9876.2004.05155.x.

Chimani, B., V. Venema, A. Lexer, K. Andre, I. Auer, and J. Nemec, 2018: Inter-comparison of methods to homogenize daily relative humidity. *Int. J. Climatol.*, **38**, 3106–3122, https://doi.org/10.1002/joc.5488.

Dienst, M., J. Lindén, E. Engström, and J. Esper, 2017: Removing the relocation bias from the 155-year Haparanda temperature record in northern Europe. *Int. J. Climatol.*, **37**, 4015–4026, https://doi.org/10.1002/joc.4981.

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

——, 2011b: Adapted Caussinus-Mestre Algorithm for Networks of Temperature series (ACMANT). *Int. J. Geosci.*, **2**, 293–309, https://doi.org/10.4236/ijg.2011.23032.

——, 2013a: Measuring performances of homogenization methods. *Időjárás*, **117**, 91–112.

——, 2013b: Efficiencies of inhomogeneity-detection algorithms: Comparison of different detection methods and efficiency measures. *J. Climatol.*, **2013**, 390945, https://doi.org/10.1155/2013/390945.

——, 2017: Time series homogenisation with optimal segmentation and ANOVA correction: Past, present and future. *Proc. Ninth Seminar for Homogenization and Quality Control in Climatological Databases and Fourth Conf. on Spatial Interpolation Techniques in Climatology and Meteorology*, WMO WCDMP-85, Geneva, Switzerland, OMSZ, 29–45.

——, 2020: ACMANTv4: Scientific content and operation of the software. Tech. Doc., 71 pp., https://github.com/dpeterfree/ACMANT.

——, and J. Coll, 2017a: Time series homogenisation of large observational datasets: The impact of the number of partner series on the efficiency. *Climate Res.*, **74**, 31–42, https://doi.org/10.3354/cr01488.

——, and ——, 2017b: Homogenisation of temperature and precipitation time series with ACMANT3: Method description and efficiency tests. *Int. J. Climatol.*, **37**, 1910–1921, https://doi.org/10.1002/joc.4822.

——, and ——, 2019: Impact of missing data on the efficiency of homogenization: Experiments with ACMANTv3. *Theor. Appl. Climatol.*, **136**, 287–299, https://doi.org/10.1007/s00704-018-2488-3.

——, V. Venema, and O. Mestre, 2011: Efficiencies of homogenisation methods: Our present knowledge and its limitation. *Proc. Seventh Seminar for Homogenisation and Quality Control in Climatological Databases*, WMO-WCDMP-78, Geneva, Switzerland, OMSZ, 19–32.

——, ——, I. Auer, O. Mestre, and M. Brunetti, 2012: The historical pathway towards more accurate homogenisation. *Adv. Sci. Res.*, **8**, 45–52, https://doi.org/10.5194/asr-8-45-2012.

——, J. Coll, J. Guijarro, M. Curley, E. Rustemeier, E. Aguilar, S. Walsh, J. Sweeney, 2020: Precipitation trends in the island of Ireland using a dense, homogenized, observational dataset. *Int. J. Climatol.*, **40**, 6458–6472, https://doi.org/10.1002/joc.6592.

Easterling, D. R., and T. C. Peterson, 1995: A new method for detecting undocumented discontinuities in climatological time series. *Int. J. Climatol.*, **15**, 369–377, https://doi.org/10.1002/joc.3370150403.

Gubler, S., and Coauthors, 2017: The influence of station density on climate data homogenization. *Int. J. Climatol.*, **37**, 4670–4683, https://doi.org/10.1002/joc.5114.

Guijarro, J. A., 2018: Homogenization of climatic series with Climatol. Tech. Doc., 22 pp., http://www.climatol.eu/homog_climatol-en.pdf.

——, J. A. López, E. Aguilar, P. Domonkos, V. Venema, J. Sigró, and M. Brunet, 2017: Comparison of homogenization packages applied to monthly series of temperature and precipitation: The MULTITEST project. *Proc. Ninth Seminar for Homogenization and Quality Control in Climatological Databases and Fourth Conf. on Spatial Interpolation Techniques in Climatology and Meteorology*, WMO WCDMP-85, 46–62.

——, E. Aguilar, P. Domonkos, J. Sigró, P. Štěpánek, V. Venema, and P. Zahradníček, 2019: Benchmarking results of the homogenization of daily Essential Climatic Variables within the INDECIS project. *Proc. 21st EGU General Assembly,* Vienna, Austria, EGU, 10896, https://meetingorganizer.copernicus.org/EGU2019/EGU2019-10896-1.pdf.

Hausfather, Z., M. J. Menne, C. N. Williams, T. Masters, R. Broberg, and D. Jones, 2013: Quantifying the effect of urbanization on U.S. Historical Climatology Network temperature records. *J. Geophys. Res. Atmos.*, **118**, 481–494, https://doi.org/10.1029/2012JD018509.

Hua, W., S. S. P. Shen, A. Weithmann, and H. Wang, 2017: Estimation of sampling error uncertainties in observed surface air temperature change in China. *Theor. Appl. Climatol.*, **129**, 1133–1144, https://doi.org/10.1007/s00704-016-1836-4.

Killick, R. E., 2016: Benchmarking the performance of homogenisation algorithms on daily temperature data. Ph.D. thesis, University of Exeter, 249 pp.

Lindau, R., and V. Venema, 2013: On the multiple breakpoint problem and the number of significant breaks in homogenization of climate records. *Időjárás*, **117**, 1–34.

——, and ——, 2018: On the reduction of trend errors by the ANOVA joint correction scheme used in homogenization of climate station records. *Int. J. Climatol.*, **38**, 5255–5271, https://doi.org/10.1002/joc.5728.

——, and ——, 2019: A new method to study inhomogeneities in climate records: Brownian motion or random deviations? *Int. J. Climatol.*, **39**, 4769–4783, https://doi.org/10.1002/joc.6105.

——, and ——, 2020: Random trend errors in climate station data due to inhomogeneities. *Int. J. Climatol.*, **40**, 2393–2402, https://doi.org/10.1002/joc.6340.

Mamara, A., A. A. Argiriou, and M. Anadranistakis, 2014: Detection and correction of inhomogeneities in Greek climate temperature series. *Int. J. Climatol.*, **34**, 3024–3043, https://doi.org/10.1002/joc.3888.

Menne, M. J., and C. N. Williams Jr., 2005: Detection of undocumented changepoints using multiple test statistics and composite reference series. *J. Climate*, **18**, 4271–4286, https://doi.org/10.1175/JCLI3524.1.

——, and ——, 2009: Homogenization of temperature series via pairwise comparisons. *J. Climate*, **22**, 1700–1717, https://doi.org/10.1175/2008JCLI2263.1.

——, ——, and R. S. Vose, 2009: The U.S. Historical Climatology Network monthly temperature data, version 2. *Bull. Amer. Meteor. Soc.*, **90**, 993–1008, https://doi.org/10.1175/2008BAMS2613.1.

Mestre, O., and Coauthors, 2013: HOMER: Homogenization software in R—Methods and applications. *Időjárás*, **117**, 47–67.

Moberg, A., and H. Alexandersson, 1997: Homogenization of Swedish temperature data. II: Homogenized gridded air temperature compared with a subset of global gridded air temperature since 1861. *Int. J. Climatol.*, **17**, 35–54, https://doi.org/10.1002/(SICI)1097-0088(199701)17:1<35::AID-JOC104>3.0.CO;2-F.

Parker, D. E., 1994: Effects of changing exposure of thermometers at land stations. *Int. J. Climatol.*, **14** (1), 1–31, https://doi.org/10.1002/joc.3370140102.

Peterson, T. C., and D. R. Easterling, 1994: Creation of homogeneous composite climatological reference series. *Int. J. Climatol.*, **14**, 671–679, https://doi.org/10.1002/joc.3370140606.

Reeves, J., J. Chen, X. L. Wang, R. Lund, and Q. Lu, 2007: A review and comparison of changepoint detection techniques for climate data. *J. Appl. Meteor. Climatol.*, **46**, 900–915, https://doi.org/10.1175/JAM2493.1.

Rennie, J. J., and Coauthors, 2014: The International Surface Temperature Initiative Global Land Surface Databank: Monthly temperature data release description and methods. *Geosci. Data J.*, **1**, 75–102, https://doi.org/10.1002/gdj3.8.

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

Rienzner, M., and C. Gandolfi, 2011: A composite statistical method for the detection of multiple undocumented abrupt changes in the mean value within a time series. *Int. J. Climatol.*, **31**, 742–755, https://doi.org/10.1002/joc.2113.

Sanchez-Lorenzo, A., M. Wild, M. Brunetti, J. A. Guijarro, M. Z. Hakuba, J. S. Calbó, S. Mystakidis, and B. Bartok, 2015: Reassessment and update of long-term trends in downward surface shortwave radiation over Europe (1939–2012). *J. Geophys. Res. Atmos.*, **120**, 9555–9569, https://doi.org/10.1002/2015JD023321.

Squintu, A. A., G. van der Schrier, P. Štěpánek, P. Zahradníček, and A. Klein Tank, 2020: Comparison of homogenization methods for daily temperature series against an observation-based benchmark dataset. *Theor. Appl. Climatol.*, **140**, 285–301, https://doi.org/10.1007/s00704-019-03018-0.

Szentimrey, T., 1999: Multiple Analysis of Series for Homogenization (MASH). *Proc. Second Seminar for Homogenization of Surface Climatological Data*, WMO WCDMP-41, Geneva, Switzerland, OMSZ, 27–46.

——, 2010: Methodological questions of series comparison. *Proc. Sixth Seminar for Homogenization and Quality Control in Climatological Databases*, WMO-WCDMP-76, Geneva, Switzerland, OMSZ, 1–7.

——, 2014: Manual of homogenization software MASHv3.03. Hungarian Meteorological Service Doc., 69 pp.

——, M. Lakatos, and Z. Bihari, 2014: Mathematical questions of homogenization and quality control. *Proc. Eighth Seminar for Homogenization and Quality Control in Climatological Databases and Third Conf. on Spatial Interpolation Techniques in Climatology and Meteorology*, WMO WCDMP-84, Geneva, Switzerland, OMSZ, 5–22.

Thorne, P. W., and Coauthors, 2016: Reassessing changes in diurnal temperature range: A new data set and characterization of data biases. *J. Geophys. Res. Atmos.*, **121**, 5115–5137, https://doi.org/10.1002/2015JD024583.

Venema, V., S. Bachner, H. W. Rust, and C. Simmer, 2006: Statistical characteristics of surrogate data based on geophysical measurements. *Nonlinear Processes Geophys.*, **13**, 449–466, https://doi.org/10.5194/npg-13-449-2006.

——, and Coauthors, 2012: Benchmarking monthly homogenization algorithms. *Climate Past*, **8**, 89–115, https://doi.org/10.5194/cp-8-89-2012.

Vincent, L. A., X. L. Wang, E. J. Milewska, H. Wan, F. Yang, and V. Swail, 2012: A second generation of homogenized Canadian monthly surface air temperature for climate trend analysis. *J. Geophys. Res.*, **117**, D18110, https://doi.org/10.1029/2012JD017859.

Vose, R. S., C. N. Williams Jr., T. C. Peterson, T. R. Karl, and D. R. Easterling, 2003: An evaluation of the time of observation bias adjustment in the U.S. Historical Climatology Network. *Geophys. Res. Lett.*, **30**, 2046, https://doi.org/10.1029/2003GL018111.

Wang, X. L., 2003: Comments on "Detection of undocumented changepoints: A revision of the two-phase regression model." *J. Climate*, **16**, 3383–3385, https://doi.org/10.1175/1520-0442(2003)016<3383:CODOUC>2.0.CO;2.

——, 2008: Accounting for autocorrelation in detecting mean-shifts in climate data series using the penalized maximal *t* or F test. *J. Appl. Meteor. Climatol.*, **47**, 2423–2444, https://doi.org/10.1175/2008JAMC1741.1.

——, and Y. Feng, 2013: RHtestsV4 user manual. Tech. Doc., 29 pp., https://github.com/ECCC-CDAS/RHtests.

——, Q. H. Wen, and Y. Wu, 2007: Penalized maximal *t* test for detecting undocumented mean change in climate data series. *J. Appl. Meteor. Climatol.*, **46**, 916–931, https://doi.org/10.1175/JAM2504.1.

——, H. Chen, Y. Wu, Y. Feng, and Q. Pu, 2010: New techniques for detection and adjustment of shifts in daily precipitation data series. *J. Appl. Meteor. Climatol.*, **49**, 2416–2436, https://doi.org/10.1175/2010JAMC2376.1.

Willett, K. M., and Coauthors, 2014: A framework for benchmarking of homogenisation algorithm performance on the global scale. *Geosci. Instrum. Methods Data Syst.*, **3**, 187–200, https://doi.org/10.5194/gi-3-187-2014.

Williams, C. N., M. J. Menne, and P. Thorne, 2012: Benchmarking the performance of pairwise homogenization of surface temperatures in the United States. *J. Geophys. Res.*, **117**, D05116, https://doi.org/10.1029/2011JD016761.