Efficiency of time series homogenization: method comparison
with 12 monthly temperature test datasets

Peter Domonkos¹, José A. Guijarro², Victor Venema³, Manola Brunet⁴,⁵;
Javier Sigró⁴

¹Tortosa, Spain, e-mail: dpeterfree@gmail.com
²State Meteorological Agency (AEMET), Unit of Islas Baleares, Palma, Spain
³Meteorological Institute, University of Bonn, Bonn, Germany
⁴Centre for Climate Change, Univ. Rovira i Virgili, Vila-seca, Spain
⁵Climatic Research Unit, University of East Anglia, Norwich, UK.

Early Online Release: This preliminary version has been accepted for publication in Journal of Climate, may be fully cited, and has been assigned DOI 10.1175/JCLI-D-20-0611.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.
Abstract

The aim of time series homogenization is to remove non-climatic effects, such as changes in station location, instrumentation, observation practices, etc., from observed data. Statistical homogenization usually reduces the non-climatic 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 9 versions, based on 5 homogenization methods (ACMANT, Climatol, MASH, PHA and RHtests) are presented and evaluated. The tests were executed with 12 synthetic/surrogate monthly temperature test datasets containing 100 to 500 networks with 5 to 40 time series in each. Residual centred 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 non-climatic 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), there 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. Generally, ACMANT gave the most accurate homogenization results. In the accuracy of individual time series ACMANT is closely followed by Climatol, while for the accurate calculation of mean climatic trends over large geographical regions both PHA and ACMANT are recommended.

Keywords: homogenization, benchmarking, temperature, ACMANT, Climatol, MASH, PHA, RHtests
1 INTRODUCTION

Technical changes of climate observations and environmental changes around meteorological instruments often cause non-climatic biases in the time series of climate records. These changes are usually referred to as inhomogeneities (Aguilar et al. 2003; Vincent et al. 2012; Sanchez-Lorenzo et al. 2015; etc.). 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), hence metadata sets 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 non-systematic 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 which 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 characterising directly the appropriateness of homogenized data for climate research (Venema et al. 2012; Domonkos 2013a). During the European project COST ES0601 (known as “HOME”, 2007-2011) 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 both according to test dataset properties and to homogenization methods, 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 Sec. 2.2). In this study, the experiments with twelve temperature test datasets are presented. These datasets
include 1900 station networks overall with 5 to 40 time series in each. Due to the large size of the test datasets, only automatic homogenization methods are tested (see Sec. 2.3). 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 Sec. 2.4).

2 DATA AND METHODS

2.1 Theoretical foundation

Test datasets including time series of monthly temperatures (X) are created (Eq. 1).

\[ X = x_1, x_2, \ldots, x_i, \ldots, x_{12n} \]  

In (1) \( n \) is the length of the time series in years. A time series is realistic if its elements \( x \) (omitting index \( i \)) includes all 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 non-systematic observational error \( (d) \), (Eq. 2).
\[ x = C_R + C_A + S + W + d \] (2)

For perfectly homogeneous series \( S = 0 \), while the most frequent and significant technical changes, such as station relocations or changes of instrumentation result in sudden shift 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 (Brunet et al. 2011; Dienst et al. 2017; etc.), moreover it can also contain non-seasonal, weather dependent variation, though 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 seasonal cycle, but except for linear trend IHs, their non-seasonal 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 it 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 modelled with a Gaussian white noise with expected value zero, (Eq. 3).

\[ x = C_R + C_A + S + \varepsilon \] (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 \( S \), as constant differences do not impact climatic trends and temporal variability, (Eq. 4).

\[
    x = C_R + S + \varepsilon
\]  (4)

Using the terminology of the HOME benchmark, datasets modelled by (3) are referred to as surrogate datasets, while those which are modelled by (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 vs. surrogate test datasets are presented in the Supplementary material.

2.2 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.
2.2.1. 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, while 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 Supplementary material.

2.2.2. 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 behaviour, which they called random deviations, for German temperature data, while IHs in temperature data from the USA 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 sinusoid in synthetic datasets, while in surrogate
datasets it is semi-sinusoid or irregular-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 – 1 break per 100 years and 3 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°C 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, i.e., their occurrences are evenly probable within a 30-year section of the time series, but all the time series are affected by them.

Systematic trend bias is defined as the absolute value of mean linear trend bias for all 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, as 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 / 100yr, while the systematic trend bias is less than 0.1 °C / 100yr in the other datasets.

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

2.2.3. Groups of test datasets

Based on 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, as well as the test datasets with semi-synchronous breaks form a further group.

• Group of high SNR test datasets [G1]: Y1, Y2, Y4, U2.

• Group of low SNR test datasets: [G2]: Y3, U1, U3, U4.
• Group of test datasets with semi-synchronous breaks [G3]: Y5, 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.

2.3 Homogenization methods

Nine versions of 5 homogenization methods are tested. The five methods are ACMANT (Adapted Caussinus-Mestre Algorithm for the homogenization of Networks of climatic Time series), Climatol, MASH (Multiple Analysis of Series for Homogenization), PHA (Pairwise Homogenization Algorithm, known also as USHCN method) and RHtests-PMT (Penalized Maximal t-Test of Relative Homogenization tests package). Two of the five methods, i.e., 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, i.e., in ACMANT, Climatol and MASH, and all the methods except RHtests have inbuilt 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).
2.3.1 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 Caussinus-Lyazrhi criterion is applied for break detection (Caussinus and Lyazrhi 1997) and ANOVA model 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. a) The fully automatic treatment is extended up to datasets of 5000 series (Domonkos and Coll 2019), b) Ensemble homogenization in the third (last) phase is incorporated from varied pre-homogenization scenarios, c) Inclusion of a weighted ANOVA model 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).

2.3.2 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, while 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 sub-series 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 centred values by the standard deviation, thereby yielding season-dependent adjustment terms.

2.3.3 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, semi-automatic (Szentimrey 2017). In MASH, the maximum number of reference series is 9, 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 from “MASH annual” [MSy] the monthly break detection is omitted.

2.3.4 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
to 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 correction-term in the reverse case (Menne and Williams 2009).

2.3.5 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 is without QM referred to as RHTests basic or
RHT, while the other version includes QM and 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 are maximised by 9 for RHtests.
2.4 Efficiency measures

We use efficiency measures which 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. 3-4) with $S \equiv 0$. 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 centred 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.

2.4.1 Centred 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, which only differ by a constant are considered to be the same, as the objective of the homogenization is to eliminate any non-climatic 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 it is shown by (5) for a time series of $n$ observed values.

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

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, 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 Sec. 2.4.3.

2.4.2 Trend bias

Trends are computed using least squares linear regression, and the trend slopes of $V$ and $Z$ are
denoted with $\alpha_v$ and $\alpha_z$, respectively. Trend bias of individual time series (Trb) are defined by
Eq. 6.

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

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 (Eq. 7).

\[ \text{SysTr} = \frac{1}{K} \left| \sum_{K} (\alpha_Z - \alpha_V) \right| \] (7)

In Eq. 7 upper stroke denotes the spatial average, while \( K \) is the number of networks in the dataset.

While \( \text{Trb} \) is calculated for the period with observed data that can be equal with 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.

2.4.3 Consistent estimation of network means

Sometimes the spatial averages must be calculated from a subnetwork of \( N^* \) series \( (N^* < N) \), due to 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 Eq. 8.

\[ \bar{x}_{N,i} = \frac{1}{N^*} \sum_{N^*} x_i + \frac{1}{n^*} \sum_{j=1}^{n^*} (\bar{x}_{N,j} - \bar{x}_{N^*,j}) \] (8)

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0611.1.
When $N^* < N$, this adjustment is always applied before the calculations of NetEy, NetTr and SysTr, both for series $V$ and $Z$.

### 2.4.4 Accuracy of interpolated values

The accuracy of interpolated monthly values of gap filling is evaluated by calculating the CRMSE for them, according to Sec. 2.4.1.

### 2.5. 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

3.1 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 5 datasets are characterised 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 (Fig. 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.
3.2. 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 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.

While for the raw data the mean RMSE_y is only 22% smaller than the mean RMSE_m, 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 7 of the 9 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.

The residual errors in individual time series (i.e. RMSE_m, RMSE_y 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 RMSE_m 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. Finally, the mean systematic trend bias averaged for the 12 test datasets is the smallest with PHA (Table 3).
3.3 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. 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).
3.4 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 were saved only for two datasets, i.e., for 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.

3.5 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 non-climatic 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 are 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
6 datasets, AC4 produces smaller residual errors than PHA in 2 datasets, while the difference
is less than 0.01°C / 100yrs in the remaining 4 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
Supplementary material), and they are not too much influenced by the variation of IH
structures either. The only characteristic which seems to markedly impact the rank order of
efficiencies is the SNR, which depends on several factors: the spatial correlations, the
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 mid-latitude temperatures.

A large number of methods are in use in climate data homogenization, but we could
only test 5 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 help 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 to use thoroughly tested methods
whenever it is possible, and to create 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 is 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, i.e., traditional hierarchic methods including both their detection and
IH adjustment parts on the one hand, and hierarchic break detection methods without other
steps of traditional homogenization algorithms on the other hand. Regarding the methods
examined here, 3 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 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 (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, therefore 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, 2017b, 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; Guijarro et al. 2019), however, in those tests much less statistically independent networks were used, 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 realisation 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
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 few years period, 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
metadata use within automatic homogenization procedures does not always result in
significant improvement in the accuracy (Gubler et al. 2017; Domokos 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 non-
climatic 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

Accepted for publication in Journal of Climate. DOI 10.1175/JCLI-D-20-0611.1.
al. 2020), the overall attention towards 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. Regarding
  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 to the 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 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.
ACKNOWLEDGEMENT

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

DATA AVAILABILITY

The test datasets and the homogenization results are accessible at:

https://zenodo.org/record/3934835#.XwTjF-dS_IU

REFERENCES


Accepted for publication in Journal of Climate. DOI:10.1175/JCLI-D-20-0611.1.


Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0611.1.


Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-20-0611.1.


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, \( \sigma \) – standard deviation of noise term (Eqs. 3-4), \( R \) – mean spatial correlation, \( r \) – ratio of missing data, * – characteristic is unknown for adapted datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>( K )</th>
<th>( N )</th>
<th>( n )</th>
<th>( \sigma )</th>
<th>( R )</th>
<th>( r )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y1</td>
<td>100</td>
<td>10</td>
<td>60</td>
<td>0.27</td>
<td>0.91</td>
<td>0</td>
</tr>
<tr>
<td>Y2</td>
<td>100</td>
<td>10</td>
<td>60</td>
<td>0.45</td>
<td>0.81</td>
<td>0</td>
</tr>
<tr>
<td>Y3</td>
<td>100</td>
<td>10</td>
<td>60</td>
<td>0.975</td>
<td>0.56</td>
<td>0</td>
</tr>
<tr>
<td>Y4</td>
<td>100</td>
<td>40</td>
<td>60</td>
<td>0.45</td>
<td>0.81</td>
<td>0</td>
</tr>
<tr>
<td>Y5</td>
<td>100</td>
<td>40</td>
<td>100</td>
<td>0.45</td>
<td>0.69</td>
<td>0.11</td>
</tr>
<tr>
<td>Y6</td>
<td>100</td>
<td>20</td>
<td>100</td>
<td>0.45</td>
<td>0.69</td>
<td>0.12</td>
</tr>
<tr>
<td>U1</td>
<td>100</td>
<td>10</td>
<td>60</td>
<td>*</td>
<td>0.87</td>
<td>0</td>
</tr>
<tr>
<td>U2</td>
<td>100</td>
<td>10</td>
<td>60</td>
<td>*</td>
<td>0.88</td>
<td>0</td>
</tr>
<tr>
<td>U3</td>
<td>100</td>
<td>10</td>
<td>60</td>
<td>*</td>
<td>0.68</td>
<td>0</td>
</tr>
<tr>
<td>U4</td>
<td>500</td>
<td>5</td>
<td>40</td>
<td>*</td>
<td>0.82</td>
<td>0</td>
</tr>
<tr>
<td>U5</td>
<td>300</td>
<td>10</td>
<td>100</td>
<td>*</td>
<td>0.91</td>
<td>0.11</td>
</tr>
<tr>
<td>U6</td>
<td>200</td>
<td>15</td>
<td>100</td>
<td>*</td>
<td>0.68</td>
<td>0.21</td>
</tr>
</tbody>
</table>
### TABLE 2. Inhomogeneity properties of the 12 test datasets. Br – break frequency, Tr – frequency of trend inhomogeneities (IHs), Pl – frequency of short platform IHs, Sybr – frequency of semi-synchronous breaks. All frequencies are in number per 100yr unit. $\sigma$M – 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 100yr), Ox – maximum of outlier values, Sh – shape of seasonality (Si – sinusoid, Ss – semi-sinusoid, Ir – irregular), Am – mean amplitude of seasonal cycles (°C). * indicates that IHs with magnitude of even distribution between 0 and 5°C have different properties than other IHs of dataset U6.

<table>
<thead>
<tr>
<th></th>
<th>Br</th>
<th>Tr</th>
<th>Pl</th>
<th>Sybr</th>
<th>$\sigma$M</th>
<th>M+</th>
<th>L–</th>
<th>L+</th>
<th>Ou</th>
<th>Ox</th>
<th>Sh</th>
<th>Am</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y1</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.0</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>Si</td>
<td>0.56</td>
</tr>
<tr>
<td>Y2</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.0</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>Si</td>
<td>0.56</td>
</tr>
<tr>
<td>Y3</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.0</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>Si</td>
<td>0.56</td>
</tr>
<tr>
<td>Y4</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.0</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>Si</td>
<td>0.56</td>
</tr>
<tr>
<td>Y5</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
<td>0.7</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>Si</td>
<td>0.08</td>
</tr>
<tr>
<td>Y6</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.0</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>Si</td>
<td>0.08</td>
</tr>
<tr>
<td>U1</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0.5</td>
<td>50</td>
<td>-1.2</td>
<td>1.2</td>
<td>2</td>
<td>5.0</td>
<td>Ss</td>
<td>0.28</td>
</tr>
<tr>
<td>U2</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>0.8</td>
<td>30</td>
<td>-1.0</td>
<td>3.0</td>
<td>5</td>
<td>5.0</td>
<td>Ss</td>
<td>0.64</td>
</tr>
<tr>
<td>U3</td>
<td>6</td>
<td>1</td>
<td>15</td>
<td>0</td>
<td>1.5</td>
<td>0</td>
<td>-4.0</td>
<td>6.0</td>
<td>10</td>
<td>6.0</td>
<td>Ir</td>
<td>0.56</td>
</tr>
<tr>
<td>U4</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0.5</td>
<td>50</td>
<td>-1.2</td>
<td>1.2</td>
<td>0</td>
<td>-</td>
<td>Ir</td>
<td>0.28</td>
</tr>
<tr>
<td>U5a</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0.8</td>
<td>30</td>
<td>-1.0</td>
<td>3.0</td>
<td>0</td>
<td>-</td>
<td>Ss</td>
<td>0.64</td>
</tr>
<tr>
<td>U5b</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0.5</td>
<td>30</td>
<td>-1.2</td>
<td>1.2</td>
<td>0</td>
<td>-</td>
<td>Ss</td>
<td>0.28</td>
</tr>
<tr>
<td>U6a</td>
<td>5</td>
<td>1</td>
<td>6</td>
<td>0</td>
<td>0.8</td>
<td>30*</td>
<td>-2.0</td>
<td>2.0*</td>
<td>0</td>
<td>-</td>
<td>Ir</td>
<td>0.20</td>
</tr>
<tr>
<td>U6b</td>
<td>5</td>
<td>1</td>
<td>6</td>
<td>0</td>
<td>0.5*</td>
<td>30*</td>
<td>-1.2</td>
<td>1.2*</td>
<td>0</td>
<td>-</td>
<td>Ir</td>
<td>0.12</td>
</tr>
</tbody>
</table>

*(U6)
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 °C, Trb, NetTr and SysTr are in °C / 100 yrs unit. “Raw” indicates errors without homogenization.

<table>
<thead>
<tr>
<th></th>
<th>RMSEm</th>
<th>RMSEy</th>
<th>Trb</th>
<th>NetTr</th>
<th>NetEy</th>
<th>SysTr</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AC4</td>
<td>0.341</td>
<td>AC4</td>
<td>0.199</td>
<td>AC4</td>
<td>0.443</td>
</tr>
<tr>
<td>2</td>
<td>AC3</td>
<td>0.367</td>
<td>AC3</td>
<td>0.219</td>
<td>AC3</td>
<td>0.492</td>
</tr>
<tr>
<td>3</td>
<td>Cl2</td>
<td>0.410</td>
<td>Cl2</td>
<td>0.238</td>
<td>Cl2</td>
<td>0.544</td>
</tr>
<tr>
<td>4</td>
<td>Cl1</td>
<td>0.463</td>
<td>MSy</td>
<td>0.252</td>
<td>Cl1</td>
<td>0.560</td>
</tr>
<tr>
<td>5</td>
<td>MSy</td>
<td>0.484</td>
<td>Cl1</td>
<td>0.258</td>
<td>MSy</td>
<td>0.611</td>
</tr>
<tr>
<td>6</td>
<td>PHA</td>
<td>0.499</td>
<td>PHA</td>
<td>0.285</td>
<td>RHT</td>
<td>0.632</td>
</tr>
<tr>
<td>7</td>
<td>RHT</td>
<td>0.501</td>
<td>RHT</td>
<td>0.286</td>
<td>RHQ</td>
<td>0.637</td>
</tr>
<tr>
<td>8</td>
<td>MSm</td>
<td>0.550</td>
<td>RHQ</td>
<td>0.291</td>
<td>MSm</td>
<td>0.646</td>
</tr>
<tr>
<td>9</td>
<td>Raw</td>
<td>0.707</td>
<td>MSm</td>
<td>0.299</td>
<td>PHA</td>
<td>0.709</td>
</tr>
<tr>
<td>10</td>
<td>RHQ</td>
<td>1.018</td>
<td>Raw</td>
<td>0.556</td>
<td>RHQ</td>
<td>1.439</td>
</tr>
</tbody>
</table>
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 bold. In 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.

<table>
<thead>
<tr>
<th></th>
<th>RMSEm</th>
<th>RMSEy</th>
<th>Trb</th>
<th>NetTr</th>
<th>NetEy</th>
<th>Intp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y1</td>
<td>AC4, AC3</td>
<td>AC4</td>
<td>AC4, AC3</td>
<td>AC4, AC3, PHA</td>
<td>AC4</td>
<td></td>
</tr>
<tr>
<td>Y2</td>
<td>AC4, AC3</td>
<td>AC4</td>
<td>AC4</td>
<td>AC4, PHA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y3</td>
<td>AC4, CI2, AC3, MSy</td>
<td>MSy, AC4, CI2, RHT, AC3, RHQ, CI1</td>
<td>RHT, RHQ, CI2, CI1, MSy, AC4</td>
<td>All methods</td>
<td>AC4, PHA, CI2, AC3, CI1, MSy</td>
<td>AC4, CI2</td>
</tr>
<tr>
<td>Y4</td>
<td>AC4, AC3</td>
<td>AC4</td>
<td>AC4</td>
<td>AC4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y5</td>
<td>AC4, CI2</td>
<td>AC4</td>
<td>AC4, AC3</td>
<td>PHA</td>
<td>AC4, PHA, AC3</td>
<td>MSy, MSm, AC4, CI2</td>
</tr>
<tr>
<td>Y6</td>
<td>AC4</td>
<td>AC4</td>
<td>AC4</td>
<td>AC4, MSy, PHA, AC3</td>
<td>AC4</td>
<td>MSy, MSm, AC4, CI2</td>
</tr>
<tr>
<td>U1</td>
<td>AC4, AC3</td>
<td>AC4</td>
<td>AC4</td>
<td>RHT, AC4, RHQ, CI2</td>
<td>RHT, CI2, MSm, AC4, RHQ</td>
<td>AC4, MSm, CI2</td>
</tr>
<tr>
<td>U2</td>
<td>AC4</td>
<td>AC4</td>
<td>AC4</td>
<td>AC4, PHA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>U3</td>
<td>AC4, AC3, CI1, CI2</td>
<td>AC4, RHT, RHQ, CI2, CI1</td>
<td>CI1, CI2, AC4</td>
<td>AC4, CI1, MSm, CI2</td>
<td>AC4, MSm, CI1, CI2</td>
<td></td>
</tr>
<tr>
<td>U4</td>
<td>AC4, AC3, CI1, MSy, CI2</td>
<td>AC4, AC3, CI2</td>
<td>AC4, AC3, RHT, CI2, RHQ, CI1</td>
<td>MSm, AC4, AC3, CI2, CI1, MSy, PHA, RHT, RHQ</td>
<td>AC4, AC3, CI2, CI1, MSy, MSm</td>
<td></td>
</tr>
<tr>
<td>U5</td>
<td>AC4</td>
<td>AC4</td>
<td>AC4, AC3</td>
<td>AC4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>U6</td>
<td>AC4, AC3</td>
<td>AC4</td>
<td>AC4, AC3</td>
<td>AC3, AC4</td>
<td>AC4, AC3</td>
<td></td>
</tr>
<tr>
<td>G1</td>
<td>AC4, AC3</td>
<td>AC4</td>
<td>AC4</td>
<td>AC4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G2</td>
<td>AC4, AC3, CI2</td>
<td>AC4, CI2, AC3, RHT</td>
<td>CI2, RHT, RHQ, CI1, AC4, AC3</td>
<td>MSm, AC4, CI1, CI2</td>
<td>AC4, CI2, CI1, AC3</td>
<td></td>
</tr>
<tr>
<td>G3</td>
<td>AC4</td>
<td>AC4</td>
<td>AC4, AC3</td>
<td>AC4, PHA, AC3</td>
<td>AC4, PHA, AC3</td>
<td>MSy, MSm, AC4, CI2</td>
</tr>
</tbody>
</table>
FIGURE 1. Mean residual errors after homogenization for groups of test datasets: the ranges of the results of the applied homogenization methods. The results are normalised with the raw data errors. Horizontal sticks on the bars show the mean errors of all homogenization methods. RMSEm – centred root mean squared error for monthly data, RMSEy – centred root mean squared error for annual data, Trb – bias of linear trend for the whole data period in a time series, NetEy – centred root mean squared error for network mean annual means, NetTr – network mean bias of linear trends for the whole study period of a dataset, - SysTr – systematic trend bias for entire datasets. a) Mean of the examined 12 test datasets, b) The same as (a), but with the exclusion of RHtests-QM, c) the same as (b), but for G1 (datasets of high SNR), d) the same as (b), but for G2 (datasets of low SNR).
FIGURE 2. Error bars between 2 and 98 percentiles (P02 and P98) for the homogenization results of all experiments. The section borders are at P10, P25, P50, P75 and P90. “x” indicates arithmetical mean.
FIGURE 3. The same as Fig. 2, but for the group of high SNR datasets (group G1).
FIGURE 4. The same as Fig. 2, but for the group of low SNR datasets (group G2).
FIGURE 5. The same as Fig. 2, but for the datasets with semi-synchronous breaks in time series (group G3).
FIGURE 6. CRMSE of interpolated monthly values in the gap filling of G3. The section borders of the error bars indicate the same error percentiles as in Fig. 2.