Clim. Past, 8, 89–115, 2012
www.clim-past.net/8/89/2012/
doi:10.5194/cp-8-89-2012


Climate of the Past

# Benchmarking homogenization algorithms for monthly data

**V. K. C. Venema[1], O. Mestre[2], E. Aguilar[3], I. Auer[4], J. A. Guijarro[5], P. Domonkos[3], G. Vertacnik[6], T. Szentimrey[7], P. Stepanek[8,9], P. Zahradnicek[8,9], J. Viarre[3], G. Müller-Westermeier[10], M. Lakatos[7], C. N. Williams[11], M. J. Menne[11], R. Lindau[1], D. Rasol[12], E. Rustemeier[1], K. Kolokythas[13], T. Marinova[14], L. Andresen[15], F. Acquaotta[16], S. Fratianni[16], S. Cheval[17,18], M. Klancar[6], M. Brunetti[19], C. Gruber[4], M. Prohom Duran[20,21], T. Likso[12], P. Esteban[22,20], and T. Brandsma[23]**

[1]Meteorological institute of the University of Bonn, Germany
[2]Meteo France, Ecole Nationale de la Meteorologie, Toulouse, France
[3]Center on Climate Change (C3), Universitat Rovira i Virgili, Tarragona, Spain
[4]Zentralanstalt für Meteorologie und Geodynamik, Wien, Austria
[5]Agencia Estatal de Meteorologia, Palma de Mallorca, Spain
[6]Slovenian Environment Agency, Ljubljana, Slovenia
[7]Hungarian Meteorological Service, Budapest, Hungary
[8]Czech Hydrometeorological Institute, Brno, Czech Republic
[9]Czechglobe-Global Change Research Centre AS CR, v.v.i., Brno, Czech Republic
[10]Deutscher Wetterdienst, Offenbach, Germany
[11]NOAA/National Climatic Data Center, USA
[12]Meteorological and hydrological service, Zagreb, Croatia
[13]Laboratory of Atmospheric Physics, University of Patras, Greece
[14]National Institute of Meteorology and Hydrology – BAS, Sofia, Bulgaria
[15]Norwegian Meteorological Institute, Oslo, Norway
[16]Department of Earth Science, University of Turin, Italy
[17]National Meteorological Administration, Bucharest, Romania
[18]National Institute for R&D in Environmental Protection, Bucharest, Romania
[19]Institute of Atmospheric Sciences and Climate (ISAC-CNR), Bologna, Italy
[20]Grup de Climatologia, Universitat de Barcelona, Spain
[21]Meteorological Service of Catalonia, Area of Climatology, Barcelona, Catalonia, Spain
[22]Centre d'Estudis de la Neu i de la Muntanya d'Andorra (CENMA-IEA), Andorra
[23]Royal Netherlands Meteorological Institute, De Bilt, The Netherlands

*Correspondence to:* V. K. C. Venema (victor.venema@uni-bonn.de)

Received: 23 July 2011 – Published in Clim. Past Discuss.: 12 August 2011
Revised: 11 November 2011 – Accepted: 22 November 2011 – Published: 10 January 2012

**Abstract.** The COST (European Cooperation in Science and Technology) Action ES0601: advances in homogenization methods of climate series: an integrated approach (HOME) has executed a blind intercomparison and validation study for monthly homogenization algorithms. Time series of monthly temperature and precipitation were evaluated because of their importance for climate studies and because they represent two important types of statistics (additive and multiplicative). The algorithms were validated against a realistic benchmark dataset. The benchmark contains real inhomogeneous data as well as simulated data with inserted inhomogeneities. Random independent break-type inhomogeneities with normally distributed breakpoint sizes were added to the simulated datasets. To approximate real world conditions, breaks were introduced that occur simultaneously in multiple station series within a simulated network of station data. The simulated time series also contained outliers, missing data periods and local station trends. Further, a stochastic nonlinear global (network-wide) trend was added.

**Published by Copernicus Publications on behalf of the European Geosciences Union.**

Participants provided 25 separate homogenized contributions as part of the blind study. After the deadline at which details of the imposed inhomogeneities were revealed, 22 additional solutions were submitted. These homogenized datasets were assessed by a number of performance metrics including (i) the centered root mean square error relative to the true homogeneous value at various averaging scales, (ii) the error in linear trend estimates and (iii) traditional contingency skill scores. The metrics were computed both using the individual station series as well as the network average regional series. The performance of the contributions depends significantly on the error metric considered. Contingency scores by themselves are not very informative. Although relative homogenization algorithms typically improve the homogeneity of temperature data, only the best ones improve precipitation data. Training the users on homogenization software was found to be very important. Moreover, state-of-the-art relative homogenization algorithms developed to work with an inhomogeneous reference are shown to perform best. The study showed that automatic algorithms can perform as well as manual ones.

## 1 Introduction

Monitoring and analysis of our climate has received more and more attention following assessments that most of the temperature change observed over the last fifty years can be attributed to anthropogenic forcings (IPCC, 2007). To study climate change and variability, at the surface many long instrumental climate records are available. These datasets are essential since they are the basis for assessing century-scale trends, for the validation of climate models, as well as detection and attribution of climate change at a regional scale. The value of these datasets, however, strongly depends on the homogeneity of the underlying time series.

In essence, a homogeneous climate time series is defined as one where variations are caused only by variations in weather and climate. Long instrumental records are rarely if ever homogeneous. Results from the homogenization of instrumental western climate records indicate that detected inhomogeneities in mean temperature series occur at a frequency of roughly one per 15 to 20 yr. Moreover, the typical size of the breaks is often of the same order as the climatic change signal during the 20th century (Auer et al., 2007; Menne et al., 2009; Brunetti et al., 2006; Caussinus and Mestre; 2004, Della-Marta et al., 2004). Inhomogeneities are thus a significant source of uncertainty for the estimation of secular trends and decadal-scale variability.

homogenization is important at two spatial scales. homogenization should produce station series that more consistently reflect true variations in climate to allow for more reliable assessments of local climatic variability and change. If all inhomogeneities would be purely random perturbations of the climate records, collectively their effect on the mean climate signal for a large network and, especially, global average time series would be small. However, numerous studies indicate that inhomogeneities are not always independent, but can collectively lead to artificial biases in climate trends across large regions (Menne et al., 2010; Brunetti et al., 2006; Begert et al., 2005). For example, for the Greater Alpine Region a bias in the temperature trend between 1870s and 1980s of half a degree was found, which was due to decreasing urbanization of the network and systematic changes in the time of observation (Böhm et al., 2001). The precipitation records of the early instrumental period are biased by $-10\,\%$ due to the systematically higher installation of the gauges at the time (Auer et al., 2005). Other possible bias sources are new types of weather shelters (Brunet et al., 2011; Brunetti et al., 2006), the change from liquid and glass thermometers to electrical resistance thermometers (Menne et al., 2009), as well as the tendency to replace observers by automatic weather stations (Begert et al., 2005), the much discussed urban heat island effect (Hansen et al., 2001; Peterson 2003) and the transfer of many urban stations to airports (Trewin, 2010).

The most commonly used method to detect and remove the effects of artificial changes is the relative homogenization approach, which assumes that nearby stations are exposed to almost the same climate signal and that thus the differences between nearby stations can be utilized to detect inhomogeneities (Conrad and Pollak, 1950). In relative homogeneity testing, a candidate time series is compared to multiple surrounding stations either in a pairwise fashion or to a single composite reference time series computed for multiple nearby stations.

homogenization has a long tradition. In the early instrumental period, documented change-points have been removed with the help of parallel measurements. For example, biases due to changes in observing times were adjusted using multi-annual 24 h measurements (Kreil, 1854a, b). In the early 20th century Conrad (1925) made use of the Heidke criterion (Heidke, 1923) using ratios of two precipitation series. As a consequence, he recommended the use of additional criteria to test the homogeneity of series, dealing with the succession and alternation of algebraic signs, the Helmert criterion (Helmert, 1907) and the tedious Abbe criterion (Conrad, 1944). The use of Helmert's criterion for pairs of stations and Abbe's criterion still has been described as an appropriate tool in the 1940s (Conrad, 1944). Some years later the double-mass principle was popularized for break detection (Kohler, 1949).

Modern techniques were then developed using classical statistical tests (Alexandersson, 1986; Gullett et al., 1990), regression models (Easterling and Peterson, 1995; Vincent, 1998), or Bayesian approaches (Perreault et al., 2000). More recent procedures focus on methods specifically designed to detect and correct multiple change-points and work with inhomogeneous references (Szentimrey, 1999; Mestre, 1999; Caussinus and Mestre, 2004; Menne and Williams,

2009, among others). To stimulate the development of homogenization methods, the Hungarian Meteorological Service started a series of "Seminars for homogenization" in 1996 (HMS, 1996; WMO, 1999, 2004, 2006, 2011; OMSZ, 2001). A review on existing homogenization methods and national approaches for creating homogenized data sets was given by Peterson et al. (1998), a work complemented a few years later under the auspices of WMO by Aguilar et al. (2003). A recent review by Trewin (2010) focused on the causes of inhomogeneities.

An early intercomparison study by Buishand (1982) compared several classical homogenization methods for precipitation data. Reeves et al. (2007) compared various absolute (without using neighboring stations) homogenization methods with each other. A number of intercomparison studies for relative homogenization were inspired by the work of Easterling and Peterson (1995). This may have been the first peer reviewed validation of homogenization algorithms with candidate time series containing multiple break points. Their candidate and reference series were modeled as first-order autoregressive processes and represent one century of annual data. To the candidates breaks of 0.5 to 2.0 times the standard deviation of the candidate are added at fixed positions, which are at least 10 yr apart. This set-up, but with three homogeneous reference time series, was also combined with a multiple break-point candidate by Ducré-Robitaille et al. (2003) to examine eight different homogenization techniques. The comparison study by DeGaetano (2006) of seven homogenization methods, made this set-up more realistic by reproducing cross-correlations of real data, by varying the length of the data and decreasing the minimum break size to 0.11 °C. In their intercomparison study of homogenization techniques for precipitation, Beaulieu et al. (2008) used the same assumptions for the homogeneous data, but inserted one to three inhomogeneities with sizes determined by a beta-distribution and also inserted trend segments in the candidate.

The annual dataset generated by Menne and Williams (2005) was more realistic than the previously mentioned studies. They also inserted breaks in the reference time series and did not enforce an artificial minimum period between breaks. Moreover, by studying the sizes of breaks known from metadata, they showed that these sizes follow a normal distribution; such breaks were thus implemented in their dataset. The consequence of such a distribution is that the dataset contains many small breaks that are hardly detectable; see also Domonkos and Štěpánek (2009). However, these small breaks are important for the detection of the climatologically more important detectable ones (Domonkos, 2011a) and likely for the correction as well (Easterling and Peterson, 1995). A recent validation study by Domonkos (2011a) directly generated artificial difference time series to compare eight different objective detection methods. The inserted inhomogeneities range from simple one-break cases to cases with a very complete and realistic description of the inhomogeneities, including platform-like inhomogeneities in which after the first break there is soon a second break in the opposite direction.

The large number of different monthly homogenization methods and the need for a realistic comparative study was the reason to start a coordinated European initiative, the COST Action HOME ES0601: advances in homogenization Methods of Climate Series: an integrated approach (HOME). Its main objective was to review and improve common homogenization methods, and to assess their impact on climate time series (HOME, 2011). As part of the Action a dataset was generated that serves as a benchmark (Sim et al., 2003) for comparing homogenization algorithms. This study analyses the results of this exercise. Based upon a survey among homogenization experts, the Action has chosen to focus on networks with monthly values for temperature and precipitation. Temperature and precipitation were selected because most participants consider these elements as most relevant. Furthermore, these elements represent two important types of statistical models (additive and multiplicative). For climate data aggregated to monthly scales, there is a large selection of possible homogenization algorithms. However, so far intercomparison studies have been based on annual data. Consequently, an intercomparison study is most needed for monthly data.

All studies before Domonkos (2008) have assessed the skill of homogenization algorithms based on the accuracy of the detection of breaks, which is a basic metric for a developer of homogenization algorithms. However, a climatologist may want to know to what degree decadal variability and trends in homogenized data may be due to remaining small inhomogeneities. To be able to answer such questions requires an evaluation of the output of full homogenization methods in terms of other statistical metrics, for instance the remaining error in linear trend estimates and the mean square error between the true time series and the homogenized ones (Domonkos, 2008; Domonkos et al., 2011). For these errors to be applicable to real datasets and to be able to perform a benchmarking of homogenization algorithms, the structure of the artificial data and its inserted inhomogeneities should be realistic.

Realistic climate data are generated with the surrogate data approach (Venema et al., 2006a), which is able to reproduce the cross-correlation structure of existing homogenized networks, as well as the auto-correlation functions of the stations and their difference time series. For comparison also Gaussian white noise is generated for the so-called synthetic data section of the benchmark dataset. In the homogeneous artificial datasets, known inhomogeneities are randomly inserted. Break inhomogeneities are modeled as an independent Poisson process and the sizes are normally distributed. Additionally, breaks are introduced that occur simultaneously in a multiple stations. Furthermore, outliers, missing data and local trends are inserted and a random global (network-wide) trend is added.

To be able to study how realistic the inserted inhomogeneities are, a third section of the benchmark contains real inhomogeneous data. This allows for a comparison of the statistical properties of the detected inhomogeneities in real and artificial data.

The organization of this study is different from previous works, being the first open – the dataset was published online and everyone was invited to homogenize it – as well as the first blind test – the truth was only revealed after all homogenized datasets were returned. Among the papers studying multiple algorithms, this study can be considered the most comprehensive one with 25 contributions based on 13 algorithms being returned by the participants, including contributions based on manual methods. For well-known algorithms – MASH, PRODIGE and SNHT – multiple contributions have been returned; see Sect. 4. This allows the study of the importance of the implementation of an algorithm or of the operator of the software.

This paper will focus on the properties of the benchmark dataset and provides a first analysis of the accuracy of the algorithms. It is intended as a reference for follow-up studies analyzing the results in more detail. In Sect. 2, the data and the methods are presented that are used to generate the three data sections (real, surrogate and synthetic data) of the benchmark. The surrogate and synthetic data are treated as real homogeneous climate data, to which inhomogeneities are added. Section 3 will explain how the inhomogeneities are introduced to the artificial dataset. Further details on the datasets and the types of breaks added can be found in the report by Venema et al. (2011). Section 4 provides a discussion of the homogenization principles and algorithms employed. The metrics used in the assessment are explained in Sect. 5. A general analysis of the submitted results is provided in Sect. 6. Some discussion and conclusions are offered in Sect. 7.

## 2 Data for benchmark dataset

The benchmark contains three data sections, one with observed, unhomogenized climate data (see Sect. 2.1) and two with artificial data. The main features of the real inhomogeneous data set and the generation of the homogeneous artificial data are summarized below.

While the general statistical properties of the artificial data and the inhomogeneities required to simulate real world observing networks were discussed and approved within the COST Action HOME management team, the dataset was generated solely by the first author. The true underlying homogeneous artificial data was therefore not known to other participants until after the deadline for submitting homogenized results. After the deadline, the truth and all homogenized contributions were made available to all contributors for analysis and are now freely available via HOME (2011).

The main type of artificial data, which most contributors homogenized, is the so-called surrogate data section; see Sect. 2.2. Surrogate data reproduce the distribution, power spectrum and cross spectra of a real homogenized dataset. The power spectrum is equivalent to the correlation function, thus the auto- and cross-correlation functions of the input data are also replicated.

For every surrogate network, a so-called synthetic network was also generated. The difference (or ratio) time series of the synthetic dataset is temporally uncorrelated Gaussian white noise. To generate pairs of surrogate and synthetic networks with a similar configuration, the cross-correlation matrix, mean and standard deviation of the synthetic networks mimic those of a corresponding surrogate network; see Sect. 2.3.

While the surrogate data is most realistic, the statistical properties of the synthetic data are those assumed by most statistical tests used for homogenization. A comparison of the results between these two types of artificial data can thus be used to study the influence of violations of these conditions. The benchmark dataset contained 20 surrogate and 20 synthetic networks for both temperature and for precipitation. During the analysis it was found that some of the input data was not homogenized well enough. Consequently, the (long-term) variability of some difference time series in these networks is artificially too strong. The algorithm used to produce the surrogate networks is able and has reproduced this (long-term) variability, which the homogenization algorithms may interpret as inhomogeneities. Consequently, these networks had to be removed and only the best 15 surrogate networks were used in the analysis. Selecting stronger did not change the validation metrics anymore. For the comparison of surrogate and synthetic data, a new dataset was generated using only well homogenized input networks; see Sect. 6.3.1.

### 2.1 Real data section

The real data section contains inhomogeneous datasets from various European climates and aims to contain examples of normal Europe datasets. The six precipitation datasets come from The Netherlands, France (Bourgogne), Norway (two regions in Western Norway), the Catalonian region (North-East Spain, Andorra and Southern France), and Romania. The six temperature datasets originate from The Netherlands, Norway (a coastal region and a group of light houses, both in the south), Romania, France (Brittany), and the Catalonian region. Most real datasets are about one century long, except for Romania and Brittany, which span about half a century.

### 2.2 Surrogate data section

Networks with 100 yr of data (1900 to 1999) with 5, 9 or 15 stations were generated. The statistical properties of the surrogate data are based on homogenized complete (or with

estimated values for missing data) temperature datasets from Austria, France (Brittany), and the Catalonian region, as well as such precipitation datasets from Austria and France (Bourgogne). These precipitation datasets did not contain zero values and were demeaned, detrended and cropped to one century. The temperature records were deseasonalised and detrended. After generating the surrogate, these means of the precipitation stations and the seasonal cycles of the temperature stations were added again. Some temperature datasets were shorter than 100 yr and were extended by mirroring them as often as needed and then cropping the dataset to 100 yr. To generate networks with different network configurations and a range of spatial correlations a different subset of stations was selected for each surrogate network.

The surrogate data was generated using the Iterative Amplitude Adjusted Fourier Transform Algorithm (IAAFT), developed by Schreiber and Schmitz (1996), with a small modification of the second iterative step as described in Venema et al. (2006b). The IAAFT algorithm tends to generate time series that are not very intermittent in the sense of the variance of the (small-scale) variance (Venema et al., 2006a). Thus, if the input data contains inhomogeneities, its large-scale variability will be reproduced in the surrogate (difference) time series and the intense small-scale variability of the jump will be spread over the full period.

To produce a new time series each time, the iterative IAAFT algorithm starts with white noise. The first iterative step adjusts the Fourier coefficients. The second step adjusts the (temperature or precipitation sum) distribution. The latter changes the Fourier spectrum somewhat, which necessitates several iterations. These Fourier spectra and distributions stem from an example homogenized dataset

### 2.3 Synthetic data section

Every surrogate network has a corresponding synthetic network. The generation of the synthetic data begins with computing a time series with the network mean precipitation or temperature. A difference (temperature) or ratio (precipitation) this mean is then computed to create each station series. This relative time series is converted to Gaussian white noise, which has the same mean, standard deviation and a similar spatial cross-correlation matrix, and added (or multiplied) to the network mean time series as described in Venema et al. (2011).

After the transformation to a Gaussian distribution, negative precipitation totals may occur; these values are explicitly set to zero. The cross-correlation matrix of the ratio time series of the synthetic data is close to that of the surrogate data, but after multiplying the ratio time series to network mean time series the cross-correlations are perturbed. For this reason, the cross-correlation between the precipitation stations within a network are biased by several percent points towards low correlations.

## 3 Inserted inhomogeneities

The artificial surrogate and synthetic data represent homogeneous climate data. To create the benchmarks, known inhomogeneities and other data disturbances are added: two types of break-type inhomogeneities and local trends, as well as outliers. Furthermore, two types of missing data are simulated and a global trend is added.

The two types of step-type breaks are random and clustered. Random breakpoints are inserted to the serial data at an average rate of five per hundred years. To vary the quality of the data on a station by station basis, the average break frequency for a station is first drawn from a uniform distribution between 2 and 8 %. The actual break events themselves are drawn with this frequency and are independent of each other (Poisson process). Breaks are thus on occasion inserted in missing data periods, in close succession or near the beginning or end of the series.

The size of the break points is based on a Gaussian distribution with a standard deviation of 0.8 °C for temperature and 15 % for rain. These mean break sizes have a seasonal cycle with standard deviation 0.4 °C and 7.5 %. The seasonal perturbations are computed by smoothing white noise and, if needed, shifting one of its extremes to the summer period. The break points are inserted by multiplying the precipitation with monthly factors or adding monthly constants to temperature.

To simulate network-wide changes, clustered breaks are also added in 30 % of the networks. In the affected networks, 30 % of the stations have a break point at the same time. The random numbers for the mean size and seasonal cycle of these breaks are drawn from the same distributions and have the properties as the random breaks. However, in this case the random numbers are not only drawn for every station, but additional once for all breaks. The random numbers are then averaged with a weight of 80 % for the random number for all breaks and a weight of 20 % for the station specific break.

In 10 % of the temperature stations a local linear trend is introduced. The station and beginning date of the trend were selected at random. The length of the trend has a uniform distribution between 30 and 60 yr. The beginning and the trend length were reselected as often as necessary to ensure that the local trend ended before the year 2000. The size of the trend at the end is randomly selected from a Gaussian distribution with a standard deviation of 0.8 °C. In half of these cases the perturbation due to the local trend continues at the end of the trend, e.g. to simulate urbanization, in the other half the station returns to its original value, e.g. to simulate a growing bush or tree that is cut at the end.

A small number of outliers was inserted to study the influence of imperfect quality control. The outliers are generated with a frequency of 1 per 100 yr per station. The outliers are added to the anomaly time series, i.e. without the annual

cycle for temperature. The value of the outliers is determined at random by a value from the tails of the distribution.

Two types of missing data are added. The earliest data is removed to simulate a gradual increase in the availability of data, which is common in real datasets. This is done by forcing a linear increase in the number of stations from a total of three with data in 1900 to all stations having data in 1925. In addition, a large part of the network is set to missing during the years covered by World War II, which is typical for European datasets. In this case, there is a 50 % chance that the data is missing in 1945. For the years preceding backward from 1944 to 1940, the stations with missing data have a probability of 50 % that the data for the previous year is also missing.

Finally, a global trend is added to every station in a network to simulate climate change. This trend is nonlinear given that homogenization should be independent of preconceived ideas about climate change. Furthermore, a different trend is stochastically modeled for every network because a known trend would allow for an improper validation of the results. The trend is generated as very smooth fractal Fourier "noise" with a power law power spectrum with an exponent of $-4$; only part of the signal is used to avoid the Fourier periodicity. This noise is normalized to a minimum of zero and a maximum of unity and then multiplied by a random Gaussian number. The width of this distribution is 1 °C or 10 %.

## 4 homogenization algorithms

This section describes the main characteristics of the homogenization methods. This paper will only list features used to homogenize the benchmark; many tools have additional possibilities. Most of the algorithms test for relative homogeneity, which implies that a candidate series is compared to some estimation of the regional climate ("comparison phase").

One absolute homogenization algorithm is employed, in this case only the station time series itself is used for homogenization.

Comparison may be performed using one composite reference series assumed homogeneous (e.g. SNHT), several ones, not assumed homogeneous (MASH), or via direct pairwise comparison (USHCN, PRODIGE); see Table 1. The comparison series are computed as the difference (in case of temperature) or ratio (precipitation) between the candidate and the reference. The time step of comparisons may be annual, seasonal or monthly. All four seasonal or twelve monthly time series may be analyzed independently in parallel or serially as one time series. When several comparisons are performed because multiple references are utilized or monthly data are analyzed in parallel, a synthesis phase is necessary, that may be automatic, semi-automatic, or manual.

The comparison series are tested for changes. Detection implies a statistical criterion to assess significance of changes, which may be based on a statistical test – Student's t, Fisher, Maximum Likelihood Ratio (MLR) test, etc. – or on criteria derived from information theory (penalized likelihood). Detection requires an optimization scheme, to find the most probable positions of the changes among all possibilities. Such a searching scheme may be exhaustive (MASH), based on semi-hierarchic binary splitting (HBS), stepwise, or moving windows (AnClim) or may use dynamic programming (DP).

The homogenization corrections, see Table 2, may be estimated directly from the comparison series (SNHT). When several references or pairwise estimates are available, a combination of those estimates is used, e.g. a mean or median. PRODIGE employs a decomposition of the signal into three parts: a common signal for all stations, a station dependent step function to model the inhomogeneities and random white noise. In some methods, raw monthly estimates are smoothed according to a seasonal variation.

Once a first correction has been performed, most methods perform a review; see Table 2. If inhomogeneities are still detected, corrections with additional breaks are implemented in the raw series (examination; raw data), except in MASH where the corrected series receive additional corrections, until no break is found (called "examination; cumulative" in Table 2).

The 25 submitted contributions, their operators and main purposes are listed in Table 3, where contributions denoted by "main" are the ones where the developer of the algorithm deployed them themselves with typical settings. Additional details on the contributions can be found in the report Venema et al. (2011).

## 5 Error metrics

A true benchmark would produce one or two numbers for every contribution for a ranking and this error metric would be fixed in advance. In case of homogenization this is not possible, different users have different requirements for the homogenized data and the ranking of the contributions depends on the chosen error metric. For this study the focus is on a number of error metrics related to the expectations of the users of homogenized data.

As the main aim of homogenization is not to improve the absolute values but rather the temporal consistency, the time series are centered by subtracting their mean values before computing the RMSE. The centered root mean square error (centered RMSE, CRMSE) of the time series themselves is thus used as a basic accuracy metric of the data at the highest available resolution (Sect. 6.1.1). This metric is similar to the standard deviation of the time series of the difference between the homogenized data and the truth. It is computed on single station data directly (station CRMSE), as well as on

**Table 1.** Comparison and detection methods of participating homogenization algorithms.

| | Comparison | | Detection | | References |
|---|---|---|---|---|---|
| Method | Comparison | Time step | Search | Criterion | |
| MASH | Multiple references | Annual, parallel monthly | Exhaustive | Statistical test (MLR) | Szentimrey (2007, 2008) |
| PRODIGE | Pairwise, human synthesis | Annual, parallel monthly | DP | Penalized Likelihood | Caussinus and Mestre (2004) |
| USHCN | Pairwise, automatic synthesis | Serial monthly | HBS | Statistical test (MLR) | Menne et al. (2009) |
| AnClim | Reference series | Annual, parallel monthly | HBS, moving window | Statistical test | Štepanek et al. (2009) |
| Craddock | Pairwise, human synthesis | Serial monthly | Visual | Visual | Craddock (1979); Brunetti et al. (2006) |
| RhtestV2 | Reference series or absolute | Serial monthly | Stepwise | Statistical test (modified Fisher) | Wang (2008) |
| SNHT | Reference series | Annual | HBS | Statistical test (MLR) | Alexandersson and Moberg (1997) |
| Climatol | Reference series | Parallel monthly | HBS, moving window | Statistical test | Guijarro (2011) |
| ACMANT | Reference series | Annual, joint seasonal | DP | Penalized Likelihood | Domonkos et al. (2011) |

DP = Dynamic programming (optimization method); HBS = (semi-)hierarchic binairy splitting; MLR = Maximum Likelihood Ratio test.

**Table 2.** Correction methods of participating homogenization algorithms.

| Method | Estimation | Review | Monthly correction |
|---|---|---|---|
| MASH | Smallest estimate from multiple comparisons | Examination; cumulative | Raw |
| PRODIGE | ANOVA | Examination; raw data | Raw |
| USHCN | Median of multiple comparisons | No review | Annual coefficients |
| AnClim | Estimated from comparison | Examination; raw data | Smoothed |
| Craddock | Mean of multiple comparisons | Examination; raw data | Smoothed |
| RhtestV2 | Estimated on comparison | No review | Annual coefficients |
| SNHT | Estimated on comparison | Examination; raw data | Raw |
| ACMANT | Estimated from comparison | No review | Smoothed |
| Climatol | Estimated from comparison | No review | Raw |

the average climate signal of all stations in one network (network CRMSE). When one or more of the stations is missing for a particular month, the network mean is not computed.

This metric is aggregated over all networks of each benchmark section in three different ways. The most direct way and important for a user is the arithmetic mean. However, because not all contributions homogenized all networks and some networks may be easier than others, the arithmetic mean may lead to a distorted judgment for the smaller contributions. Therefore, the mean of the CRMSE anomalies is also computed, where the anomalies are computed by subtracting the mean station or network CRMSE of a number of complete reference contributions (*MASH main*, *PRODIGE monthly*, *USHCN main*, *ACMANT* and *PMTred*). This anomaly is the best metric to compare (incomplete) contributions. Furthermore, to show the improvements after homogenization, the ratio between the mean CRMSE over all homogenized data with the mean CRMSE of the inhomogeneous data of the same cases is computed.

The same metrics are computed on yearly averages and results are presented in Sect. 6.1.2.

To assess the reproduction of decadal variability after homogenization, the yearly time series are first smoothed, after which the CRMSE is computed (Sect. 6.1.3). These smoothed time series or nonlinear trends are computed by a nonparametric regression method called locally weighted regression (LOESS; Cleveland and Devlin, 1998). For every year, the smoothed value is estimated by fitting a quadratic function using weighted regression on the nearest 25 % of the data points. The standard local weighting function described in Cleveland and Devlin (1998) is utilized. The effective smoothing period is about six years. An advantage of this method is that small-scale variability is strongly reduced. Furthermore, the method is robust to distortions at the edges of the time series. Nevertheless, the first and last five years were excluded from the computation of the CRMSE.

To study the remaining error in trend estimates after homogenization, the difference in the linear regression coefficient between the original data and the homogenized data is computed (for results see Sect. 6.1.4). The linear trend is estimated on the yearly time series using least squares regression and the standard RMSE of the trend coefficients over all stations (or networks) is computed as aggregated trend error metric.

**Table 3.** Names of contributions, contributors, and the main purpose of the contributions.

| Contribution | Operator | Main purpose |
|---|---|---|
| MASH main | Szentimrey & Lakatos | Main submission |
| MASH Marinova, Kolokythas | Marinova or Kolokythas | Two first-time users |
| MASH Basic, Light, Strict and No meta | Cheval | Experimental[1] |
| PRODIGE main | Mestre, Rasol & Rustemeier | Main submission[2] |
| PRODIGE monthly | Idem | Monthly detection |
| PRODIGE trendy | Idem | Local trends corrected |
| PRODIGE Acquaotta | Acquaotta & Fratianni | First-time users |
| USHCN main | Williams & Menne | Produced USHCNv2 dataset |
| USHCN 52x, cx8 | Idem | Alternatives for small networks |
| AnClim main | Stepanek | Main submission |
| AnClim SNHT | Andresen | SNHT alternative |
| AnClim Bivariate | Likso | Bivariate test in AnClim |
| iCraddock Vertacnik, Klancar | Vertacnik or Klancar | Two first-time users |
| PMTred rel | Viarre & Aguilar | PMTred test of RhTestV2 |
| PMFred abs | Viarre & Aguilar | PMFred test, absolute method |
| C3SNHT | Aguilar | SNHT alternative |
| SNHT DWD | Müller-Westermeier | SNHT alternative |
| Climatol | Guijarro | Main submission |
| ACMANT | Domonkos | Main submission |

[1] Experimental version that performs the four rules to combine yearly and monthly data separately, in stead of the standard consecutive way. [2] Detection: yearly; Correction: temperature monthly, precipitation yearly.

Since some methods do not fill data gaps, or do not handle outliers, data corresponding to missing data or outliers are not taken into account in the above computations. Thus while there is an influence of the outliers on the results of the homogenization algorithm, the outliers do not influence the error metrics themselves.

In Sect. 6.1.5 the accuracy of break detection will be investigated. An algorithm, which ranks high on detection, but is less good with respect to CRMSE or trends, may need to work on its correction methods. Thus even if in many (iterative) algorithms detection and correction cannot be fully separated, such a comparison does give qualitatively important information for the developer.

A comparison of detection scores among the contributions is impaired by the use of different methodologies. Most contributions aim at estimating the exact date a break physically happened, while others (*PRODIGE main*, *C3SNHT*) associate the break with the beginning or the ending of a year. Alternatively, all *MASH* contributions report the breaks in the monthly time series, but do not synthesize these breaks to one date; one true break may thus lead to up to 12 detected breaks. To mitigate this difference the data was analyzed at yearly resolution, i.e. every year containing a break is considered as break point, in both the tested contribution and the original time series. Nevertheless, the *MASH* contributions should be compared to the other contributions with care.

Four cases can be distinguished: true positives (hits, $a$), false positives (false alarms, $b$), false negatives (misses, $c$) and true negatives (no breaks present, nor predicted, $d$). Periods with missing data or with a local trend are ignored in this computation. Using this notation, the most basic skill scores using are the probability of detection, POD, and the probability of false detection, POFD, defined as:

$$\mathrm{POD} = \frac{a}{a+c} \tag{1}$$

$$\mathrm{POFD} = \frac{b}{b+d} \tag{2}$$

The Peirce Skill Score (or true skill score) is defined as POD minus POFD. In addition, the standard Heidke Skill Score (HSS) can be computed as:

$$\mathrm{HSS_{std}} = \frac{p - r_{\mathrm{std}}}{1 - r_{\mathrm{std}}} \tag{3}$$

where

$$r_{\mathrm{std}} = \frac{a+c}{n}\frac{a+b}{n} + \frac{b+d}{n}\frac{c+d}{n} \tag{4}$$

$p$=$(a+d)/n$ and $n = a+b+c+d$. The reference $r_{std}$ in Eq. (3) intends to correct for randomly correct results: for a random prediction the HSS is on average zero. The reference used within the standard HSS is equal to the proportion of random agreement for a given number of predicted breaks. It is independent from the fact whether this number of predicted breaks is actually realistic, i.e. whether it is comparable to the number of true breaks.

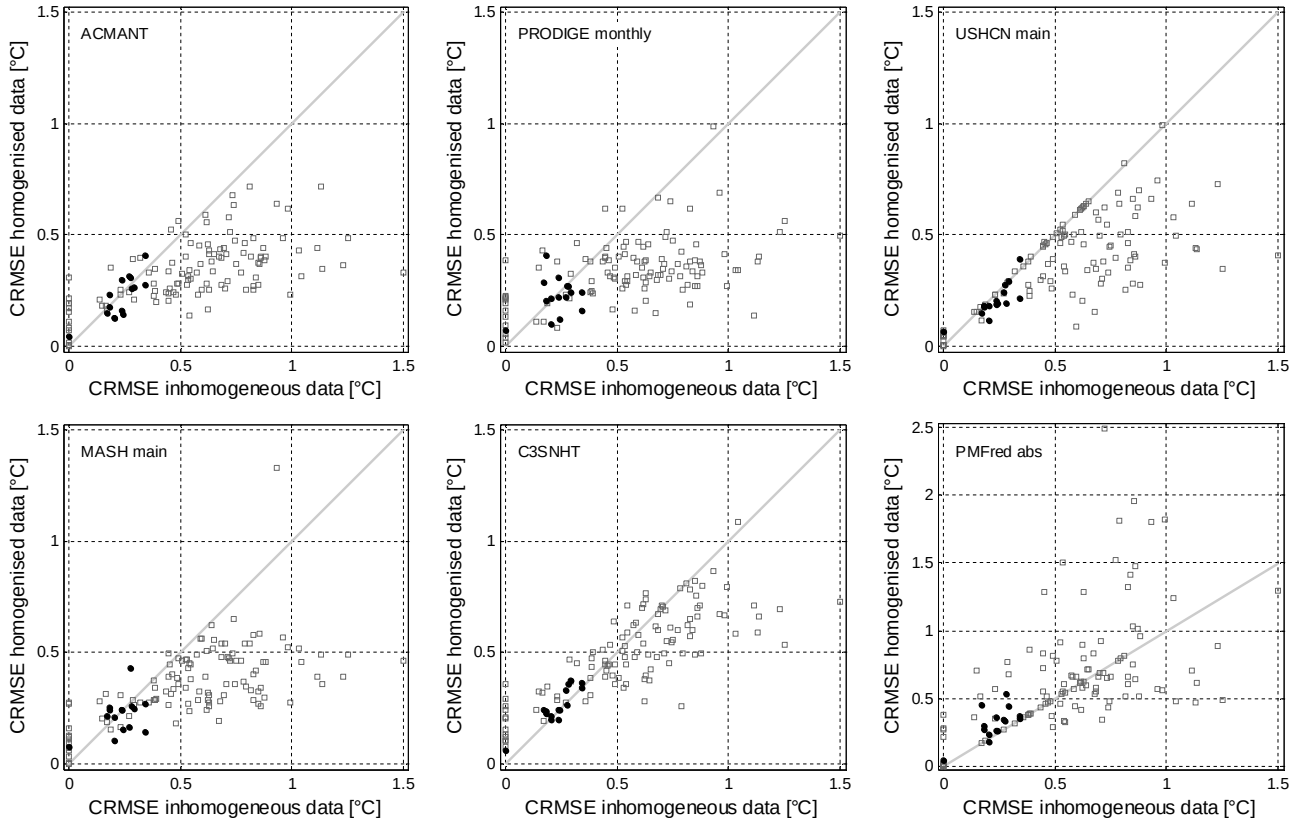


**Fig. 1.** Scatterplot of the centered RMSE before and after homogenization for selected contributions. The squares display the errors of the stations; the dots show the errors of the network mean (regional climate) time series. Points on the bisect indicate no change, above the bisect the data is made more inhomogeneous, while below the bisect homogenization improved the homogeneity of the data.

As an alternative Heidke special skill score, $HSS_{spc}$, is considered where the $r_{std}$ of Eq. (3) is substituted by $r_{spc}$ given by:

$$r_{spc} = \frac{a+c}{n} f + \frac{b+d}{n} (1-f) \quad (5)$$

with $f$, the mean frequency of true breaks as reference for the proportion of predicted positives and (1-$f$) the frequency for the predicted negatives. The special HSS becomes zero if the correct number of breaks is predicted and if this number were randomly inserted. Given that breaks are rarer than negatives, in essence this skill score mainly punishes false alarms stronger.

## 6 Results

This section starts with an analysis of the quality of the homogenized data for all blind contributions in Sect. 6.1. This analysis is largely mainly based on the surrogate data because these networks were homogenized most by the participants and are more realistic than synthetic. Furthermore, the focus is more on temperature than on precipitation because more contributions were submitted for this climatic element. The latter may be because homogenization of temperature is less challenging and because there is more interest in the homogeneity of temperature records.

Section 6.2 discusses some interesting contributions submitted after the deadline, when the break locations and magnitudes were known. In Sect. 6.3, the realism of the benchmark dataset is studied by comparing results obtained for surrogate and synthetic data, as well as by comparing the detected inhomogeneities of the artificial dataset with those of the real raw data section of the benchmark. This information is needed for the interpretation of the results in the discussion in Sect. 7.

### 6.1 Results for blind contributions

This section assesses the homogenized data based on a range of different error metrics. The analysis follows the temporal scale of the data: Sect. 6.1.1 discusses errors on monthly scales, Sect. 6.1.2 on yearly scales, Sect. 6.1.3 on decadal scales and Sect. 6.1.4 treats the errors in secular trends after homogenization. Finally in Sect. 6.1.5, contingency scores are computed to investigate the accuracy of the detection of break inhomogeneities.

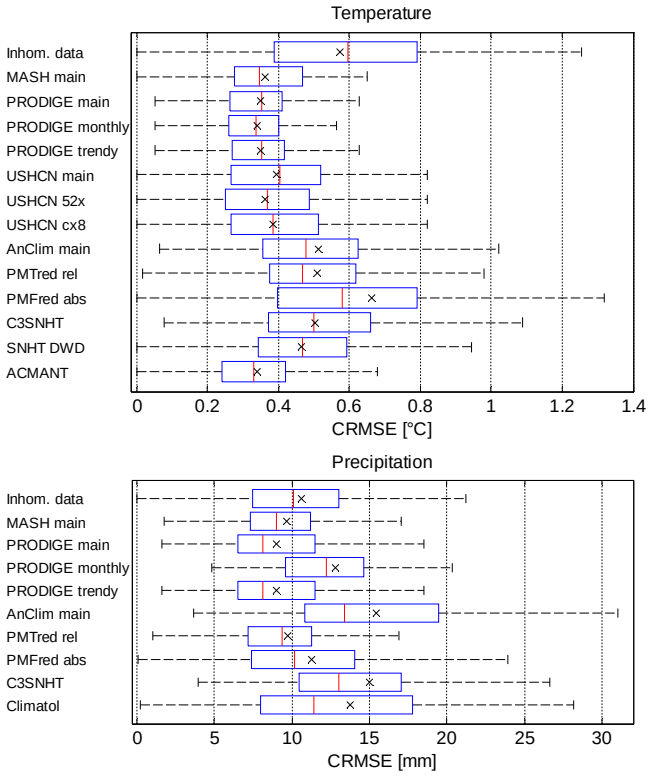


**Fig. 2.** Boxplot of the centered RMSE of the complete contributions, for temperature (top) and precipitation (bottom). For comparison the error metric for the inhomogeneous data is plotted at the top. The outliers are not displayed for legibility. The cross depicts the mean CRMSE, the vertical bar denotes the median; the box spans the interquartile range (the range of the 25 to the 75 percentile); the whiskers span the range of the data, but maximally span 1.5 times the interquartile range. Good homogenization algorithms should have low CRMSE values and little spread.

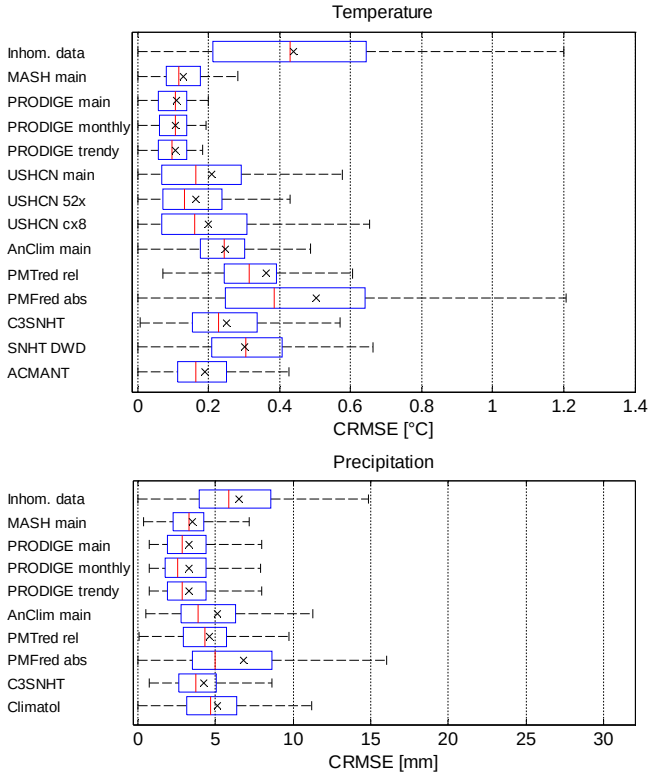


**Fig. 3.** Boxplot of the decadal CRMSE of the complete contributions, for temperature (top) and precipitation (bottom). For comparison the error metric for the inhomogeneous data is plotted at the top. The abscissa is the same as the one of Fig. 2 to emphasis the smaller errors and larger improvement over the inhomogeneous data for the decadal data. The conventions of the boxplots are explained in Fig. 2.

### 6.1.1 Errors on monthly scale

Figure 1 shows scatterplots of the centered RMSE before and after homogenization for monthly surrogate temperature data by six comprehensive contributions. Good results can be achieved either by improving the homogeneity on average or by never increasing the inhomogeneity of any station. *PRODIGE* seems to follow the former route, *USHCN* the latter, with the others making a compromise. The *USHCN* contribution is unique in that it has almost no stations with a higher error after homogenization, the contribution also has many values exactly on the bisect (no changes performed) and it made only small changes to the network without any inserted breaks (values on the ordinate). It should be noted that the same plots for yearly mean temperature show many fewer data points above the bisect for all contributions. The exception is absolute homogenization (*PMFred abs*), which typically decreases the homogeneity of the data for both monthly and yearly mean values.

For a more quantitative analysis of the monthly CRMSE, Fig. 2 shows boxplots for the complete blind contributions and Table 4 lists aggregated error metrics for all blind contributions for both temperature and precipitation. The boxplots show that the best contributions, with respect to the mean CRMSE of the temperature station data, are *PRODIGE*, *ACMANT*, *MASH main* and *USHCN 52x*; the CRMSE anomalies in the table reveal that the incomplete *iCraddock Vertacnik* contribution is actually the most accurate one for temperature. Five temperature contributions made the data more inhomogeneous, i.e. had an improvement quotient over the inhomogeneous data above one.

If all station series in a network are averaged to one network series representing the regional climate, the errors tend to become much smaller and results can be very different; see the last four columns in Table 4. For the network CRMSE the *USHCN 52x* performs best, followed by the best versions of *iCraddock*, *MASH* and *PRODIGE*. Interestingly, *ACMANT*, one of the best for the station CRMSE, performs much less well for the network CRMSE. Six contributions made the network average data more inhomogeneous.

**Table 4.** The centered RMSE of monthly data for all blind contributions.

| | Station | | | | Network | | | |
|---|---|---|---|---|---|---|---|---|
| Temperature | Number[1] | CRMSE[2] | CRMSE anomaly[3] | Impr.[4] | Number | CRMSE | CRMSE anomaly | Impr. |
| Inhomogeneous data | 111 | 0.57 | 0.18 | 1.00 | 15 | 0.23 | 0.01 | 1.00 |
| MASH main | 111 | 0.36 | −0.03 | 0.63 | 15 | 0.22 | 0.00 | 0.94 |
| MASH Marinova | 23 | 0.26 | −0.04 | 0.70 | 3 | 0.17 | 0.00 | 1.00 |
| MASH Kolokythas | 44 | 0.62 | 0.21 | 1.09 | 8 | 0.45 | 0.22 | 1.75 |
| MASH Basic | 20 | 0.35 | −0.02 | 0.54 | 2 | 0.20 | 0.01 | 0.81 |
| MASH Light | 20 | 0.35 | −0.02 | 0.54 | 2 | 0.20 | 0.02 | 0.83 |
| MASH Strict | 15 | 0.31 | -0.02 | 0.46 | 1 | 0.13 | 0.01 | 0.66 |
| MASH No meta | 20 | 0.35 | −0.01 | 0.55 | 2 | 0.20 | 0.02 | 0.83 |
| PRODIGE main | 111 | 0.35 | −0.04 | 0.61 | 15 | 0.23 | 0.01 | 0.98 |
| PRODIGE monthly | 111 | 0.34 | −0.05 | 0.59 | 15 | 0.22 | 0.01 | 0.96 |
| PRODIGE trendy | 111 | 0.35 | −0.04 | 0.61 | 15 | 0.23 | 0.01 | 0.99 |
| PRODIGE Acquaotta | 40 | 0.48 | 0.09 | 0.79 | 6 | 0.40 | 0.17 | 1.50 |
| USHCN main | 111 | 0.39 | 0.00 | 0.69 | 15 | 0.20 | −0.01 | 0.88 |
| USHCN 52x | 111 | 0.36 | −0.03 | 0.63 | 15 | 0.19 | −0.02 | 0.84 |
| USHCN cx8 | 111 | 0.39 | −0.01 | 0.67 | 15 | 0.20 | −0.02 | 0.86 |
| AnClim main | 111 | 0.51 | 0.12 | 0.89 | 15 | 0.29 | 0.07 | 1.26 |
| AnClim SNHT | 5 | 0.64 | 0.15 | 1.15 | 1 | 0.34 | 0.09 | 1.20 |
| AnClim Bivariate | 35 | 0.69 | 0.25 | 1.14 | 5 | 0.28 | 0.05 | 1.09 |
| iCraddock Vertacnik | 55 | 0.35 | −0.06 | 0.57 | 7 | 0.20 | −0.02 | 0.78 |
| iCraddock Klancar | 5 | 0.44 | −0.04 | 0.79 | 1 | 0.23 | −0.02 | 0.81 |
| PMTred rel | 111 | 0.51 | 0.12 | 0.89 | 15 | 0.22 | 0.00 | 0.95 |
| PMFred abs | 111 | 0.66 | 0.27 | 1.15 | 15 | 0.32 | 0.10 | 1.36 |
| C3SNHT | 111 | 0.50 | 0.11 | 0.88 | 15 | 0.26 | 0.04 | 1.12 |
| SNHT DWD | 111 | 0.46 | 0.07 | 0.81 | 15 | 0.23 | 0.01 | 1.00 |
| Climatol | 110 | 0.69 | 0.30 | 1.20 | 14 | 0.39 | 0.17 | 1.71 |
| ACMANT | 111 | 0.34 | −0.05 | 0.59 | 15 | 0.22 | 0.00 | 0.95 |
| Precipitation | | | | | | | | |
| Inhomogeneous data | 111 | 10.6 | 1.1 | 1.00 | 15 | 4.3 | −0.4 | 1.00 |
| MASH main | 111 | 9.7 | 0.2 | 0.91 | 15 | 4.9 | 0.2 | 1.13 |
| MASH Marinova | 14 | 8.5 | 0.1 | 0.84 | 2 | 3.8 | 0.3 | 1.03 |
| PRODIGE main | 111 | 9.0 | −0.5 | 0.85 | 15 | 5.0 | 0.3 | 1.16 |
| PRODIGE monthly | 111 | 12.8 | 3.3 | 1.20 | 15 | 7.0 | 2.3 | 1.63 |
| PRODIGE trendy | 111 | 9.0 | −0.5 | 0.85 | 15 | 5.0 | 0.3 | 1.16 |
| AnClim main | 111 | 15.4 | 5.9 | 1.45 | 15 | 6.2 | 1.4 | 1.43 |
| PMTred rel | 111 | 9.7 | 0.3 | 0.92 | 15 | 4.3 | −0.4 | 0.99 |
| PMFred abs | 111 | 11.3 | 1.8 | 1.06 | 15 | 4.9 | 0.2 | 1.15 |
| C3SNHT | 111 | 15.0 | 5.5 | 1.41 | 15 | 6.7 | 2.0 | 1.56 |
| SNHT DWD | 102 | 10.9 | 1.4 | 1.03 | 14 | 4.6 | −0.2 | 1.06 |
| Climatol | 111 | 13.7 | 4.3 | 1.30 | 15 | 7.6 | 2.9 | 1.76 |

[1] The number of homogenized stations or networks. [2] The mean CRMSE over all homogenized networks in °C or mm. [3] The mean anomaly of the CRMSE; anomalies are computed by subtracting the CRMSE of a number of complete reference contributions to be able to make a fair comparison for contributions that did not homogenize all networks, see Sect. 5. [4] The improvement over the inhomogeneous data is computed as the quotient of the mean CRMSE of the homogenized networks and the mean CRMSE of the same inhomogeneous networks.

For precipitation many fewer contributions were submitted. The best contribution regarding the monthly CRMSE anomaly of the station data is *PRODIGE main*, where monthly values are adjusted using a coefficient estimated on annual values. In contrast *PRODIGE monthly* made the data more inhomogeneous. The partial contribution *MASH Marinova* achieved the smallest CRMSE, but the larger mean CRMSE *anomaly* suggests that relatively easy networks were homogenized and that the contribution is actually second best. Over half of the contributions did not improve the

CRMSE of the station data and none of the algorithms improved the network CRMSE meaningfully.

### 6.1.2 Errors on yearly scale

The errors in the inhomogeneous yearly data are smaller than in the monthly data; see Table 5. The monthly station temperature error of the inhomogeneous monthly data is 0.57 °C, whereas at yearly scale the error is reduced to 0.47 °C. Notably, the reduction in error for the homogenized temperature data is typically much stronger; the average reduction factor over all contributions for monthly data is 77 %, whereas for yearly data it is 53 %. With some exceptions, the contributions with an improvement factor for monthly data of around 1.0, perform similarly for yearly data, whereas the better contributions for monthly data achieve an even better improvement factor for yearly data. For instance, where the best contributions improve the homogeneity of the monthly station data by about a factor 0.6, the improvement ratio of these contributions of the yearly data is around 0.3. As mentioned above, scatterplots of the CRMSE show that at yearly scales most contributions improve nearly all stations and networks individually.

For precipitation the yearly station-based results are more encouraging than the monthly results: only absolute homogenization increases the yearly CRMSE significantly. For the yearly CRMSE of precipitation *MASH main* is the most accurate algorithm. Network average precipitation data is not clearly improved by homogenization.

### 6.1.3 Errors on decadal scale

The errors in the inhomogeneous decadal data are again smaller than in the yearly data; see Table 6. Still, the intercomparison between the contributions are very similar for the CRMSE of yearly and decadal station data. The explained variance of a linear fit of the CRMSE at these two scales is 98 % (97 %) for temperature (precipitation). Therefore, only boxplots for the decadal CRMSE are shown in Fig. 3. Compared to the monthly data, the range of the results is larger because the errors of the best contributions decrease much more than for contributions that did not perform as well. At this scale *ACMANT* performs less well than the other contributions that were good with respect to the monthly CRMSE.

For the network mean signal there is a strong difference between yearly and decadal data as shown in Tables 5 and 6. The most evident difference is the typically much smaller error. In contrast to the yearly network CRMSE of precipitation, the decadal CRMSE is improved by homogenization. For the network mean precipitation there is almost no correlation between the yearly and decadal values. While in both cases *MASH main* is one of the best and absolute homogenization increases the inhomogeneity of the data, the ranking of most other contributions changes considerably.

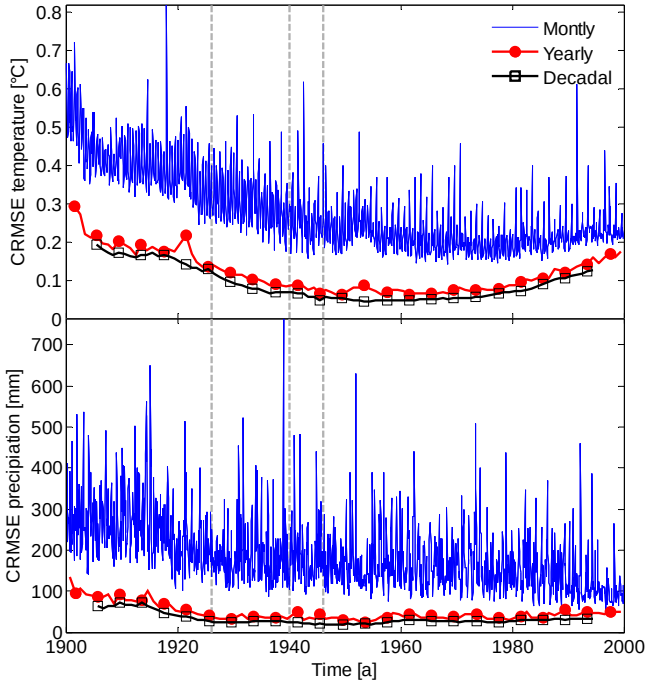


**Fig. 4.** The average temporal behavior over all contributions of the monthly, yearly and decadal CRMSE of the station data for temperature (top) and precipitation (bottom). The stripped vertical line at 1925 indicates the end of the period in which not all stations have started observations. The two stripped vertical lines at 1940 and 1945 indicate the period of the Second World War with much missing data.

### 6.1.4 Temporal behavior

For most contributions the CRMSE is lower near 2000 than in 1900. For example, the monthly station CRMSE of temperature (precipitation) is around 0.2 °C (100 mm) in 2000 and around 0.5 °C (250 mm) in 1900 averaged over all contributions; see Fig. 4. A clear feature of this figure is, furthermore, the u-shape of especially the yearly and decadal data. This is a natural consequence of using the centered time series to compute the errors in case of systematic deviations such as differences in slope.

The period with missing data during the WWII seems to be important. This is where the error often starts to grow more rapidly or even jumps higher. Another important period is the first quarter of the century where many stations do not yet have data. Therefore, the CRMSE of selected contributions are shown in Table 7 for the first and second quarter, as well as for the last half a century. The table shows that the error of the homogenized data in the first quarter is always higher or equal compared to the other two periods. For some contributions the errors in the second quarter are higher than for the last half of the century; this points to problems with the missing data in the middle of the time series after the Second World War. An exceptional contribution is

**Table 5.** The centered RMSE of yearly data for all blind contributions.

| | Station | | | | Network | | | |
|---|---|---|---|---|---|---|---|---|
| Temperature | Number[1] | CRMSE[2] | CRMSE anomaly[3] | Impr.[4] | Number | CRMSE | CRMSE anomaly | Impr. |
| Inhomogeneous data | 111 | 0.47 | 0.24 | 1.00 | 15 | 0.20 | 0.04 | 1.00 |
| MASH main | 111 | 0.16 | −0.07 | 0.35 | 15 | 0.13 | −0.02 | 0.67 |
| MASH Marinova | 23 | 0.12 | −0.06 | 0.40 | 3 | 0.10 | −0.03 | 0.62 |
| MASH Kolokythas | 44 | 0.28 | 0.02 | 0.60 | 8 | 0.19 | 0.02 | 0.86 |
| MASH Basic | 20 | 0.18 | −0.06 | 0.31 | 2 | 0.13 | −0.03 | 0.59 |
| MASH Light | 20 | 0.18 | −0.06 | 0.32 | 2 | 0.13 | −0.03 | 0.59 |
| MASH Strict | 15 | 0.17 | −0.02 | 0.29 | 1 | 0.11 | −0.01 | 0.59 |
| MASH No meta | 20 | 0.20 | −0.03 | 0.36 | 2 | 0.15 | −0.02 | 0.64 |
| PRODIGE main | 111 | 0.16 | −0.07 | 0.34 | 15 | 0.13 | −0.02 | 0.69 |
| PRODIGE monthly | 111 | 0.15 | −0.08 | 0.32 | 15 | 0.13 | −0.02 | 0.68 |
| PRODIGE trendy | 111 | 0.16 | −0.08 | 0.34 | 15 | 0.14 | −0.02 | 0.69 |
| PRODIGE Acquaotta | 40 | 0.19 | −0.06 | 0.37 | 6 | 0.19 | 0.02 | 0.83 |
| USHCN main | 111 | 0.25 | 0.01 | 0.52 | 15 | 0.17 | 0.01 | 0.86 |
| USHCN 52x | 111 | 0.20 | −0.03 | 0.43 | 15 | 0.16 | 0.00 | 0.80 |
| USHCN cx8 | 111 | 0.24 | 0.00 | 0.50 | 15 | 0.16 | 0.01 | 0.84 |
| AnClim main | 111 | 0.33 | 0.10 | 0.71 | 15 | 0.23 | 0.08 | 1.19 |
| AnClim SNHT | 5 | 0.52 | 0.15 | 1.02 | 1 | 0.31 | 0.09 | 1.16 |
| AnClim Bivariate | 35 | 0.45 | 0.19 | 0.93 | 5 | 0.20 | 0.04 | 0.95 |
| iCraddock Vertacnik | 55 | 0.15 | −0.10 | 0.29 | 7 | 0.13 | −0.03 | 0.59 |
| iCraddock Klancar | 5 | 0.17 | −0.19 | 0.34 | 1 | 0.10 | −0.11 | 0.39 |
| PMTred rel | 111 | 0.40 | 0.16 | 0.84 | 15 | 0.18 | 0.02 | 0.92 |
| PMFred abs | 111 | 0.56 | 0.33 | 1.19 | 15 | 0.29 | 0.13 | 1.48 |
| C3SNHT | 111 | 0.29 | 0.05 | 0.61 | 15 | 0.18 | 0.02 | 0.91 |
| SNHT DWD | 111 | 0.36 | 0.12 | 0.75 | 15 | 0.19 | 0.04 | 1.00 |
| Climatol | 110 | 0.28 | 0.05 | 0.60 | 14 | 0.18 | 0.03 | 0.95 |
| ACMANT | 111 | 0.21 | −0.02 | 0.45 | 15 | 0.17 | 0.01 | 0.85 |
| Precipitation | | | | | | | | |
| Inhomogeneous data | 111 | 7.3 | 2.4 | 1.00 | 15 | 3.1 | 0.0 | 1.00 |
| MASH main | 111 | 4.5 | −0.4 | 0.62 | 15 | 2.9 | −0.1 | 0.95 |
| MASH Marinova | 14 | 3.6 | −0.4 | 0.56 | 2 | 1.6 | −0.2 | 0.69 |
| PRODIGE main | 111 | 4.7 | −0.3 | 0.63 | 15 | 3.3 | 0.2 | 1.07 |
| PRODIGE monthly | 111 | 4.7 | −0.3 | 0.64 | 15 | 3.4 | 0.4 | 1.11 |
| PRODIGE trendy | 111 | 4.7 | −0.3 | 0.63 | 15 | 3.3 | 0.2 | 1.07 |
| AnClim main | 111 | 6.5 | 1.5 | 0.88 | 15 | 3.8 | 0.7 | 1.23 |
| PMTred rel | 111 | 5.7 | 0.8 | 0.78 | 15 | 3.0 | −0.1 | 0.97 |
| PMFred abs | 111 | 7.9 | 2.9 | 1.08 | 15 | 3.7 | 0.6 | 1.21 |
| C3SNHT | 111 | 5.8 | 0.8 | 0.79 | 15 | 3.0 | −0.1 | 0.98 |
| SNHT DWD | 102 | 6.7 | 1.6 | 0.90 | 14 | 3.1 | 0.0 | 1.01 |
| Climatol | 111 | 6.3 | 1.3 | 0.86 | 15 | 3.4 | 0.4 | 1.12 |

For footnotes see Table 4.

*Climatol*, which has the lowest monthly temperature errors around 1900, which grow slowly towards 2000; not shown.

### 6.1.5 Linear trends

More accurate trend estimation is a primary motivation to homogenize climate data. Figure 5 shows scatterplots of the station trends before and after homogenization for six selected contributions. Vertical lines start at the trend in the inhomogeneous data and end with a symbol at the trend estimate for the homogenized data. The figure illustrates the improvement of the temperature trend estimates and indicates that trend improvement was smaller for precipitation. Because all stations in one network have the same symbol, the figure also shows that all stations within one network tend to have a bias in the same direction, whereas for the networks overall there is no bias. *Climatol* is an exception in that it greatly decreases the magnitude of any trend in temperature.

**Table 6.** The centered RMSE of decadal data for all blind contributions.

| | Station | | | | Network | | | |
|---|---|---|---|---|---|---|---|---|
| Temperature | Number[1] | CRMSE[2] | CRMSE anomaly[3] | Impr.[4] | Number | CRMSE | CRMSE anomaly | Impr. |
| Inhomogeneous data | 111 | 0.44 | 0.24 | 1.00 | 15 | 0.18 | 0.06 | 1.00 |
| MASH main | 111 | 0.13 | −0.07 | 0.29 | 15 | 0.09 | −0.03 | 0.47 |
| MASH Marinova | 23 | 0.09 | −0.06 | 0.33 | 3 | 0.08 | −0.01 | 0.53 |
| MASH Kolokythas | 44 | 0.23 | 0.01 | 0.53 | 8 | 0.13 | −0.00 | 0.64 |
| MASH Basic | 20 | 0.15 | −0.04 | 0.29 | 2 | 0.12 | −0.00 | 0.54 |
| MASH Light | 20 | 0.15 | −0.04 | 0.29 | 2 | 0.12 | −0.00 | 0.55 |
| MASH Strict | 15 | 0.15 | −0.01 | 0.28 | 1 | 0.10 | 0.03 | 0.54 |
| MASH No meta | 20 | 0.17 | −0.02 | 0.33 | 2 | 0.13 | 0.01 | 0.60 |
| PRODIGE main | 111 | 0.11 | −0.09 | 0.25 | 15 | 0.06 | −0.06 | 0.35 |
| PRODIGE monthly | 111 | 0.11 | −0.09 | 0.24 | 15 | 0.07 | −0.05 | 0.35 |
| PRODIGE trendy | 111 | 0.11 | −0.09 | 0.24 | 15 | 0.06 | −0.06 | 0.35 |
| PRODIGE Acquaotta | 40 | 0.14 | −0.08 | 0.28 | 6 | 0.14 | 0.01 | 0.65 |
| USHCN main | 111 | 0.21 | 0.01 | 0.48 | 15 | 0.13 | 0.01 | 0.69 |
| USHCN 52x | 111 | 0.16 | −0.04 | 0.37 | 15 | 0.10 | −0.02 | 0.55 |
| USHCN cx8 | 111 | 0.20 | −0.00 | 0.45 | 15 | 0.12 | 0.00 | 0.66 |
| AnClim main | 111 | 0.25 | 0.05 | 0.56 | 15 | 0.18 | 0.06 | 1.00 |
| AnClim SNHT | 5 | 0.44 | 0.14 | 1.01 | 1 | 0.28 | 0.12 | 1.13 |
| AnClim Bivariate | 35 | 0.40 | 0.18 | 0.88 | 5 | 0.15 | 0.01 | 0.72 |
| iCraddock Vertacnik | 55 | 0.11 | −0.11 | 0.22 | 7 | 0.06 | −0.07 | 0.28 |
| iCraddock Klancar | 5 | 0.11 | −0.18 | 0.26 | 1 | 0.08 | −0.09 | 0.31 |
| PMTred rel | 111 | 0.36 | 0.16 | 0.82 | 15 | 0.16 | 0.04 | 0.90 |
| PMFred abs | 111 | 0.50 | 0.30 | 1.15 | 15 | 0.27 | 0.15 | 1.48 |
| C3SNHT | 111 | 0.25 | 0.05 | 0.57 | 15 | 0.16 | 0.04 | 0.90 |
| SNHT DWD | 111 | 0.30 | 0.10 | 0.69 | 15 | 0.18 | 0.06 | 0.97 |
| Climatol | 110 | 0.22 | 0.02 | 0.51 | 14 | 0.16 | 0.04 | 0.85 |
| ACMANT | 111 | 0.19 | −0.01 | 0.43 | 15 | 0.16 | 0.04 | 0.86 |
| Precipitation | | | | | | | | |
| Inhomogeneous data | 111 | 6.5 | 2.7 | 1.00 | 15 | 2.8 | 0.4 | 1.00 |
| MASH main | 111 | 3.5 | −0.3 | 0.54 | 15 | 2.3 | −0.2 | 0.81 |
| MASH Marinova | 14 | 2.7 | −0.1 | 0.51 | 2 | 1.4 | 0.1 | 0.69 |
| PRODIGE main | 111 | 3.3 | −0.5 | 0.51 | 15 | 2.4 | −0.1 | 0.84 |
| PRODIGE monthly | 111 | 3.3 | −0.5 | 0.50 | 15 | 2.5 | 0.0 | 0.87 |
| PRODIGE trendy | 111 | 3.3 | −0.5 | 0.51 | 15 | 2.4 | −0.1 | 0.84 |
| AnClim main | 111 | 5.2 | 1.3 | 0.79 | 15 | 2.7 | 0.2 | 0.95 |
| PMTred rel | 111 | 4.6 | 0.8 | 0.71 | 15 | 2.7 | 0.2 | 0.95 |
| PMFred abs | 111 | 6.8 | 3.0 | 1.04 | 15 | 3.4 | 1.0 | 1.21 |
| C3SNHT | 111 | 4.2 | 0.4 | 0.65 | 15 | 2.6 | 0.2 | 0.93 |
| SNHT DWD | 102 | 5.7 | 1.8 | 0.87 | 14 | 2.9 | 0.3 | 0.99 |
| Climatol | 111 | 5.1 | 1.3 | 0.79 | 15 | 2.6 | 0.1 | 0.91 |

For footnotes see Table 4.

Figure 6 gives an overview of the differences between the trends in the homogenized station data and the original data for all complete contributions; the smaller the spread, the better the contribution. *MASH main* performs best for precipitation. For this selection *PRODIGE monthly* performs best for temperature.

Table 8 summarizes all contributions and metrics for both station and network trends. Overall, the incomplete *iCraddock* and *MASH Marinova* contributions performed even better for temperature station trends. With respect to the trends in station or network precipitation trends *MASH Marinova* is the most accurate contribution.

The correlation between the scores for the station-based and the network-based trends is again modest. A considerable number of contributions do not decrease the uncertainty of the trends of the network. For network

**Table 7.** The centered RMSE of monthly, yearly and decadal station data for selected contributions for three periods. The first period contains much missing data, the second ends with WWII, the last fifty years contain high quality data.

| | Monthly | | | Yearly | | | Decadal | | |
|---|---|---|---|---|---|---|---|---|---|
| Temperature | 1900–1925 | 1925–1950 | 1950–2000 | 1900–1925 | 1925–1950 | 1950–2000 | 1900–1925 | 1925–1950 | 1950–2000 |
| Inhomogeneous data | 0.62 | 0.62 | 0.49 | 0.47 | 0.48 | 0.42 | 0.44 | 0.44 | 0.39 |
| MASH main | 0.47 | 0.39 | 0.30 | 0.19 | 0.15 | 0.14 | 0.16 | 0.12 | 0.11 |
| PRODIGE main | 0.41 | 0.37 | 0.28 | 0.17 | 0.13 | 0.15 | 0.13 | 0.09 | 0.10 |
| PRODIGE monthly | 0.41 | 0.36 | 0.28 | 0.16 | 0.12 | 0.14 | 0.12 | 0.09 | 0.10 |
| PRODIGE trendy | 0.42 | 0.37 | 0.28 | 0.16 | 0.13 | 0.15 | 0.12 | 0.09 | 0.10 |
| USHCN main | 0.45 | 0.41 | 0.33 | 0.26 | 0.21 | 0.23 | 0.23 | 0.19 | 0.18 |
| USHCN 52x | 0.41 | 0.38 | 0.30 | 0.20 | 0.16 | 0.20 | 0.17 | 0.15 | 0.15 |
| USHCN cx8 | 0.44 | 0.41 | 0.32 | 0.25 | 0.20 | 0.22 | 0.23 | 0.18 | 0.18 |
| AnClim main | 0.63 | 0.54 | 0.44 | 0.34 | 0.32 | 0.31 | 0.28 | 0.25 | 0.22 |
| iCraddock Vertacnik* | 0.44 | 0.37 | 0.29 | 0.16 | 0.11 | 0.14 | 0.11 | 0.08 | 0.10 |
| PMTred rel | 0.73 | 0.50 | 0.39 | 0.58 | 0.34 | 0.31 | 0.53 | 0.31 | 0.29 |
| PMFred abs | 0.81 | 0.62 | 0.60 | 0.68 | 0.47 | 0.54 | 0.64 | 0.44 | 0.46 |
| C3SNHT | 0.58 | 0.56 | 0.42 | 0.31 | 0.29 | 0.26 | 0.27 | 0.25 | 0.22 |
| SNHT DWD | 0.54 | 0.48 | 0.40 | 0.38 | 0.34 | 0.32 | 0.35 | 0.30 | 0.26 |
| ACMANT | 0.43 | 0.34 | 0.28 | 0.29 | 0.18 | 0.18 | 0.26 | 0.16 | 0.16 |
| Precipitation | | | | | | | | | |
| Inhomogeneous data | 12.1 | 10.0 | 9.6 | 8.2 | 6.6 | 6.7 | 7.2 | 5.7 | 6.0 |
| MASH main | 12.0 | 9.9 | 8.4 | 5.2 | 4.0 | 4.2 | 4.0 | 3.1 | 3.3 |
| PRODIGE main | 10.1 | 8.3 | 8.6 | 4.9 | 3.7 | 4.6 | 3.5 | 2.6 | 3.2 |
| PRODIGE monthly | 15.8 | 13.2 | 11.1 | 5.0 | 4.0 | 4.6 | 3.6 | 2.8 | 3.1 |
| PRODIGE trendy | 10.1 | 8.3 | 8.6 | 4.9 | 3.7 | 4.6 | 3.5 | 2.6 | 3.2 |
| AnClim main | 20.5 | 16.9 | 11.8 | 7.6 | 6.0 | 5.7 | 6.2 | 4.3 | 4.7 |
| PMTred rel | 11.9 | 9.0 | 8.8 | 7.6 | 4.8 | 5.1 | 6.2 | 3.8 | 4.1 |
| PMFred abs | 13.9 | 9.9 | 10.4 | 10.1 | 6.2 | 7.4 | 8.9 | 5.2 | 6.4 |
| C3SNHT | 17.3 | 16.3 | 13.1 | 7.3 | 5.7 | 4.9 | 4.9 | 4.0 | 3.8 |
| Climatol | 12.8 | 12.0 | 14.2 | 7.1 | 5.5 | 5.9 | 5.9 | 4.4 | 4.7 |

* This contribution homogenized 55 stations, all other contributions are complete (contain 111 stations). If this contribution had been complete, its errors would have been slightly smaller.

averaged precipitation only three contributions improve the trends: *MASH Marinova*, *C3SNHT* and *AnClim main.* Absolute homogenization (*PMFred*) increases the uncertainty of the trends in the raw data by about a factor two for all four metrics in Table 8.

#### 6.1.6 Detection scores

A scatterplot with the probability of detection, POD, against the probability of false detection, POFD, for all complete contributions is presented in Fig. 7. As the Peirce Skill Score, PSS, is defined as POD minus POFD, the isolines of PSS can be indicated by slant lines in Fig. 7. Table 9 shows all contributions and more detection skill scores. Because these skill scores are computed on all networks simultaneously, anomalies could not be computed as before. Therefore comparisons with incomplete contributions have to be made with care.

The scatterplot shows that *MASH* is an outlier with respect to both detection scores. Because *MASH* reports breaks for multiple monthly time series, it naturally has more breaks than the other algorithms, which combine monthly results to one date per break. The scores are computed on the yearly scale to reduce this problem. However, because of the noise in the detected date, the larger number of detected monthly breaks for MASH still leads to an artificially larger number of annual breaks and thus false alarms. Thus intercomparisons of *MASH* with the other contributions remain difficult, especially for the POD and POFD. For both temperature and precipitation *MASH main* performs best according to the Peirce skill score, while it has the lowest Heidke special score.

Most remarkable is that most other algorithms have a probability of false detection well below the target 5 % level. *C3SNHT PMTred rel* and *AnClim main* are close to this target level. The *USHCN* contributions have the lowest POFD. With respect to the POD and the Heidke skill scores the incomplete *iCraddock* contributions stand out and the three *USHCN* contributions perform very well. *ACMANT*,

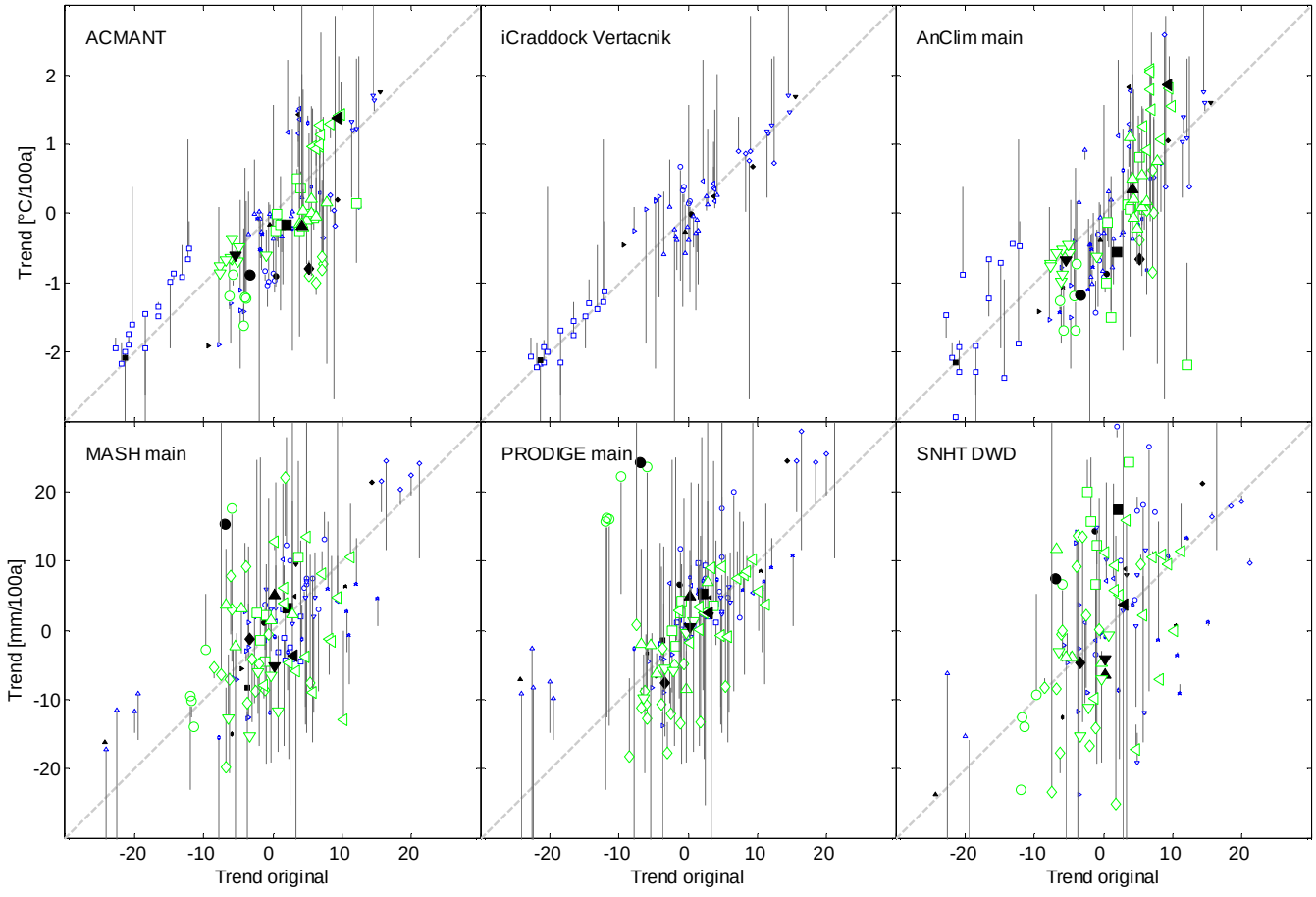


**Fig. 5.** Trends in the original data versus the trends in the inhomogeneous or homogenized data. The top row shows trends for selected temperature contributions, the bottom row for precipitation. The open symbols denote the trends of homogenized stations, the closed black symbols the trend of the homogenized regional network averaged trend; every network has its own symbol, which shows that station trend errors are correlated. The vertical grey lines run from the trend in the inhomogeneous data to the trend in the homogenized data.

*PMTred rel*, and *Climatol* perform well, especially in contrast to the previous error metrics; *Climatol* is even the best precipitation contribution with respect to the Heidke special score. All *SNHT* and *AnClim* contributions as well as *PMFred abs* are characterized by relatively low skill scores, mostly due to low probabilities of detection. The correlations between the various probability of detection and skill scores is modest, even between the two Heidke scores.

Figure 8 shows the temporal behavior of the number of true and predicted breaks (top panel), as well as the POD and the POFD (bottom) averaged over all complete surrogate temperature contributions. In the middle of the period, between about 1925 and 1975, a high correlation between true and predicted data is found in the top panel. However, there is a surplus of predicted breaks of 1 to 2 percentage points in this period.

The POD and POFD are reduced markedly at the edges of the time series, especially in the beginning of the century. The reason for this is a decrease in the total number of predicted breaks. This is presumptive due to a combination of a large uncertainty in the means needed to find a break and the smaller number of stations in the beginning. *PMFred abs* and *PMTred rel* are designed to compensate for the former problem. *PMFred abs* shows a reasonably constant POFD around the 2 percent level. On the other hand, *PMTred rel* shows a strong decline in POFD from 8 % in 1925 to 1 % in 1900, likely in response to the missing data.

### 6.2 Late contributions

This section describes contributions submitted after the deadline at which the truth was revealed to the participants. Some of these contributions aim to mend problems discovered by the results for the blind contributions. While the results found for these late contributions are interesting, their performance should be interpreted with care as these updated contributions are by definition benefiting from knowing the truth.

**Table 8.** The RMSE of linear trend estimates for all blind contributions.

| | Station | | | | Network | | | |
|---|---|---|---|---|---|---|---|---|
| Temperature | Number[1] | RMSE[2] | RMSE anomaly[3] | Impr.[4] | Number | RMSE | RMSE anomaly | Impr. |
| Inhomogeneous data | 111 | 1.19 | 0.57 | 1.00 | 15 | 0.53 | 0.08 | 1.00 |
| MASH main | 111 | 0.35 | −0.27 | 0.29 | 15 | 0.33 | −0.12 | 0.63 |
| MASH Marinova | 23 | 0.26 | −0.36 | 0.22 | 3 | 0.24 | −0.22 | 0.45 |
| MASH Kolokythas | 44 | 0.57 | −0.05 | 0.48 | 8 | 0.39 | −0.06 | 0.74 |
| MASH Basic | 20 | 0.31 | −0.30 | 0.26 | 2 | 0.22 | −0.24 | 0.41 |
| MASH Light | 20 | 0.31 | −0.30 | 0.26 | 2 | 0.20 | −0.25 | 0.39 |
| MASH Strict | 15 | 0.37 | −0.25 | 0.31 | 1 | 0.12 | −0.33 | 0.23 |
| MASH No meta | 20 | 0.40 | −0.22 | 0.33 | 2 | 0.23 | −0.22 | 0.44 |
| PRODIGE main | 111 | 0.33 | −0.29 | 0.28 | 15 | 0.26 | −0.19 | 0.50 |
| PRODIGE monthly | 111 | 0.32 | −0.30 | 0.27 | 15 | 0.27 | −0.18 | 0.52 |
| PRODIGE trendy | 111 | 0.32 | −0.29 | 0.27 | 15 | 0.25 | −0.20 | 0.48 |
| PRODIGE Acquaotta | 40 | 0.42 | −0.20 | 0.35 | 6 | 0.54 | 0.09 | 1.03 |
| USHCN main | 111 | 0.69 | 0.07 | 0.58 | 15 | 0.48 | 0.03 | 0.92 |
| USHCN 52x | 111 | 0.61 | −0.01 | 0.51 | 15 | 0.46 | 0.01 | 0.88 |
| USHCN cx8 | 111 | 0.64 | 0.02 | 0.54 | 15 | 0.43 | −0.02 | 0.82 |
| AnClim main | 111 | 0.77 | 0.15 | 0.65 | 15 | 0.70 | 0.25 | 1.32 |
| AnClim SNHT | 5 | 0.98 | 0.36 | 0.83 | 1 | 1.09 | 0.63 | 2.06 |
| AnClim Bivariate | 35 | 1.13 | 0.51 | 0.95 | 5 | 0.38 | −0.07 | 0.73 |
| iCraddock Vertacnik | 55 | 0.30 | −0.32 | 0.25 | 7 | 0.24 | −0.22 | 0.45 |
| iCraddock Klancar | 5 | 0.10 | −0.52 | 0.08 | 1 | 0.22 | −0.23 | 0.42 |
| PMTred rel | 111 | 1.09 | 0.47 | 0.92 | 15 | 0.52 | 0.07 | 0.99 |
| PMFred abs | 111 | 2.52 | 1.90 | 2.12 | 15 | 1.15 | 0.69 | 2.17 |
| C3SNHT | 111 | 0.66 | 0.04 | 0.56 | 15 | 0.57 | 0.12 | 1.08 |
| SNHT DWD | 111 | 0.73 | 0.11 | 0.61 | 15 | 0.52 | 0.07 | 0.99 |
| Climatol | 110 | 0.72 | 0.10 | 0.61 | 14 | 0.69 | 0.24 | 1.31 |
| ACMANT | 111 | 0.63 | 0.01 | 0.53 | 15 | 0.66 | 0.20 | 1.24 |
| Precipitation | | | | | | | | |
| Inhomogeneous data | 111 | 15.0 | 6.2 | 1.00 | 15 | 7.4 | −1.2 | 1.00 |
| MASH main | 111 | 7.5 | −1.3 | 0.50 | 15 | 7.7 | −0.9 | 1.04 |
| MASH Marinova | 14 | 7.1 | −1.7 | 0.47 | 2 | 5.8 | −2.8 | 0.78 |
| PRODIGE main | 111 | 8.8 | 0.0 | 0.59 | 15 | 10.0 | 1.4 | 1.36 |
| PRODIGE monthly | 111 | 9.3 | 0.5 | 0.62 | 15 | 10.3 | 1.7 | 1.39 |
| PRODIGE trendy | 111 | 8.8 | 0.0 | 0.59 | 15 | 10.0 | 1.4 | 1.36 |
| AnClim main | 111 | 16.0 | 7.2 | 1.07 | 15 | 7.3 | −1.3 | 0.99 |
| PMTred rel | 111 | 10.1 | 1.3 | 0.67 | 15 | 8.1 | −0.5 | 1.10 |
| PMFred abs | 111 | 27.8 | 19.0 | 1.86 | 15 | 17.4 | 8.8 | 2.36 |
| C3SNHT | 111 | 9.2 | 0.4 | 0.61 | 15 | 7.1 | −1.4 | 0.97 |
| SNHT DWD | 102 | 12.9 | 4.2 | 0.86 | 14 | 8.5 | −0.1 | 1.15 |
| Climatol | 111 | 12.3 | 3.6 | 0.82 | 15 | 8.0 | −0.6 | 1.08 |

[1] The number of homogenized stations or networks. [2] The mean RMSE over all homogenized networks in °C/100 yr or mm/100 yr. [3] The mean anomaly of the RMSE; anomalies are computed by subtracting the RMSE of a number of complete reference contributions to be able to make a fair comparison for contributions that did not homogenize all networks, see Sect. 5. [4] The improvement over the inhomogeneous data is computed as the quotient of the mean RMSE of the homogenized networks and the mean RMSE of the same inhomogeneous networks.

### 6.2.1 ACMANT late

*ACMANT late* has been generated with an improved version of ACMANT (Domonkos, 2011b). The main changes of *ACMANT late* compared to *ACMANT* are as follows. Using station B as one of the references for station A and later station A as one of the references for station B can lead to biased results. Therefore, a pre-homogenization for large breaks is applied in which this is forbidden. Furthermore, *ACMANT late* applies the decomposition model of PRODIGE for the final adjustment.

**Table 9.** A number of skill scores to compare the ability to detect breaks. The acronyms are utilized in Fig. 7.

| Temperature | Acronym | Number | POD | POFD | Peirce | Heidke Standard | Heidke Special |
|---|---|---|---|---|---|---|---|
| MASH main | M1 | 111 | 0.63 | 0.09 | 0.53 | 0.31 | −0.20 |
| MASH Marinova | M2 | 23 | 0.71 | 0.08 | 0.63 | 0.33 | −0.12 |
| MASH Kolokythas | M3 | 44 | 0.43 | 0.07 | 0.36 | 0.27 | 0.00 |
| MASH Basic | M4 | 20 | 0.77 | 0.18 | 0.59 | 0.26 | −0.83 |
| MASH Light | M5 | 20 | 0.77 | 0.19 | 0.58 | 0.24 | −0.99 |
| MASH Strict | M6 | 15 | 0.82 | 0.20 | 0.62 | 0.25 | −1.06 |
| MASH No meta | M7 | 20 | 0.81 | 0.26 | 0.55 | 0.19 | −1.59 |
| PRODIGE main | P1 | 111 | 0.35 | 0.02 | 0.33 | 0.35 | 0.41 |
| PRODIGE monthly | P2 | 111 | 0.39 | 0.02 | 0.37 | 0.40 | 0.44 |
| PRODIGE trendy | P3 | 111 | 0.35 | 0.02 | 0.32 | 0.35 | 0.41 |
| PRODIGE Acquaotta | P4 | 40 | 0.34 | 0.04 | 0.31 | 0.30 | 0.28 |
| USHCN main | U1 | 111 | 0.34 | 0.00 | 0.33 | 0.46 | 0.61 |
| USHCN 52x | U2 | 111 | 0.40 | 0.01 | 0.39 | 0.51 | 0.62 |
| USHCN cx8 | U3 | 111 | 0.35 | 0.01 | 0.35 | 0.47 | 0.61 |
| AnClim main | A1 | 111 | 0.18 | 0.03 | 0.15 | 0.16 | 0.20 |
| AnClim SNHT | A2 | 5 | 0.14 | 0.04 | 0.10 | 0.12 | 0.12 |
| AnClim Bivariate | A3 | 35 | 0.44 | 0.12 | 0.32 | 0.17 | −0.56 |
| iCraddock Vertacnik | C1 | 55 | 0.60 | 0.03 | 0.57 | 0.54 | 0.49 |
| iCraddock Klancar | C2 | 5 | 0.61 | 0.01 | 0.60 | 0.68 | 0.68 |
| PMTred rel | PT | 111 | 0.41 | 0.04 | 0.37 | 0.34 | 0.27 |
| PMFred abs | PF | 111 | 0.21 | 0.01 | 0.20 | 0.27 | 0.46 |
| C3SNHT | C3 | 111 | 0.23 | 0.05 | 0.18 | 0.16 | 0.04 |
| SNHT DWD | SN | 111 | 0.12 | 0.01 | 0.11 | 0.15 | 0.40 |
| Climatol | CL | 111 | 0.38 | 0.01 | 0.37 | 0.45 | 0.55 |
| ACMANT | AC | 111 | 0.50 | 0.03 | 0.47 | 0.44 | 0.41 |
| Precipitation | | | | | | | |
| MASH main | M1 | 111 | 0.26 | 0.04 | 0.22 | 0.21 | 0.19 |
| MASH Marinova | M2 | 14 | 0.23 | 0.03 | 0.20 | 0.22 | 0.27 |
| PRODIGE main | P1 | 111 | 0.19 | 0.03 | 0.16 | 0.19 | 0.29 |
| PRODIGE monthly | P2 | 111 | 0.20 | 0.03 | 0.17 | 0.19 | 0.27 |
| PRODIGE trendy | P3 | 111 | 0.19 | 0.03 | 0.16 | 0.19 | 0.29 |
| AnClim main | A1 | 111 | 0.14 | 0.02 | 0.12 | 0.16 | 0.34 |
| PMTred rel | PT | 111 | 0.23 | 0.02 | 0.21 | 0.25 | 0.37 |
| PMFred abs | PF | 111 | 0.08 | 0.01 | 0.08 | 0.13 | 0.46 |
| C3SNHT | C3 | 111 | 0.19 | 0.05 | 0.15 | 0.14 | 0.11 |
| SNHT DWD | SN | 102 | 0.04 | 0.00 | 0.04 | 0.06 | 0.47 |
| Climatol | CL | 111 | 0.12 | 0.00 | 0.11 | 0.18 | 0.50 |

*ACMANT late* would have been the most accurate contribution with respect to the CRMSE of the station (0.27 °C) and network average (0.13 °C) data, as well as the station (0.23 °C/100 yr) and network average linear trends (0.19 °C/100 yr). Especially, the performance up to 1930 has improved considerably. However, *ACMANT late* is optimized based on the benchmark data itself. It is thus not clear how much of this performance would be realized in an application to a real dataset.

### 6.2.2 Craddock late

After the deadline a contribution by Michele Brunetti, who is an experienced Craddock user, was solicited. This contribution, *Craddock late*, with four networks is about as accurate as the blind *Craddock* contributions. For instance, the monthly CRMSE of the stations is 0.34 °C and of the network average data is 0.16 °C. The linear trend estimate shows an error of 0.26 °C/100 yr (station) or 0.21 °C/100 yr (network).

Notable is that the CRMSE is almost constant as a function of time. *Craddock late* is consequently more accurate in the first half of the century, but less accurate than *iCraddock*

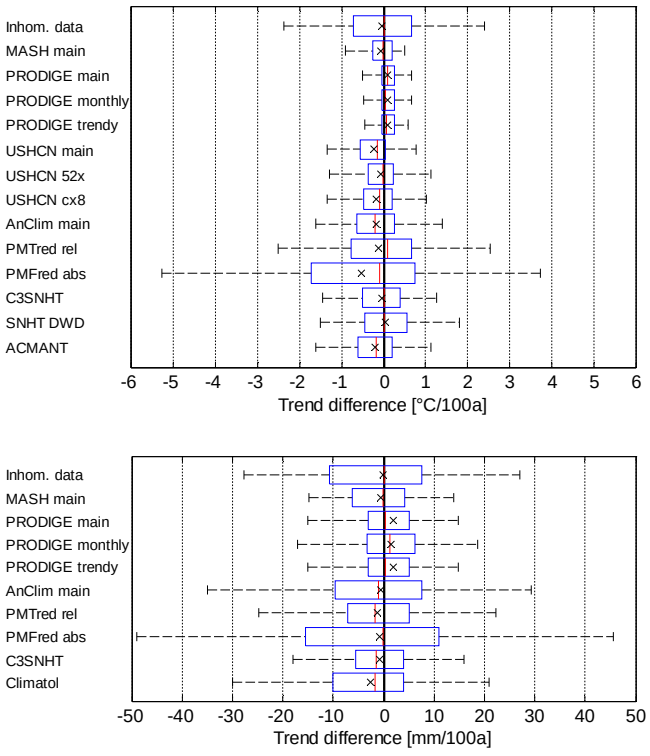


**Fig. 6.** Boxplot of the differences in the linear trends of the complete contributions, for the surrogate temperature section (top) and the surrogate precipitation section (bottom). Good homogenization algorithms should have little spread. The outliers are not displayed for legibility. The cross depicts the mean RMSE, the vertical bar denotes the median; the box spans the interquartile range (the range of the 25 to the 75 percentile); the whiskers span the range of the data, but maximally span 1.5 times the interquartile range.

*Vertacnik* or *Klanar* in the second half. This may be due to four strategies. Firstly, the most relevant pairs of stations are selected not only based on correlation, but for climatological similarity, e.g. exposure. Secondly, often only a part of the homogeneous subperiod is used for correction. Thirdly, also breaks that are not clearly evident are corrected. Finally, depending on the strength of the seasonal cycle of the break, the operator selects annual or monthly corrections.

#### 6.2.3 Climatol2.1a

Climatol's blind contribution showed good results for detection, but strongly reduced the trends. After the deadline a new *Climatol2.1a* contribution was submitted. The important changes are as follows. The main change is in the normalization of the series by the mean. As series are often incomplete, the means of the whole period are unknown, and therefore the normalization must be computed iteratively until getting stable values. The new stopping criterion for the iterations is stricter. Furthermore, the test of the squared relative mean difference was replaced by the SNHT test.

The late contribution shows a clear improvement over the blind contribution. With respect to all CRMSE metrics *Climatol2.1a* is the most accurate SNHT version; except for precipitation on decadal scales for which C3SNHT is more accurate. More importantly, *Climatol2.1a* no longer shows the reduction in the linear trends and the RMSE of the station temperature trends decreased from 0.72 °C/100 yr to 0.55 °C/100 yr, for the trends in the network means from 0.69 °C/100 yr to 0.55 °C/100 yr.

#### 6.2.4 PRODIGE automatic

This late contribution is similar to *PRODIGE main*, but the synthesis of the change points is performed automatically. It computes a weighted mean number of breaks per year, based on the cross-correlations between the stations. The decision to accept a break depends on thresholds, which were found by training on the first two precipitation networks.

For monthly precipitation, this automatic version is more accurate than *PRODIGE main*, whereas on larger averaging scales the error is larger. For linear trends in the precipitation, the RMSE of *PRODIGE automatic* for station (network) data is 9.9 mm/100 yr (12.52 mm/100 yr), respectively. Because this contribution was trained on a part of the benchmark dataset, these errors may not be representative.

#### 6.2.5 RhTestV3

After the deadline 16 surrogate temperature contributions similar to *PMTred rel* and *PMFred abs* were produced, but with the detection and correction functions from the new software package RhTestV3. After the deadline the outliers were known. Consequently in half of these late contributions the outliers could be removed to study their influence. Furthermore, half of the contributions corrected monthly and the other half yearly values; half did so correcting the mean values, half with quantile matching.

Comparing the contributions with and without outliers did not show a clear influence of outliers on the CRMSE at different averaging scales and periods, nor on the RMSE of the linear trends. All contributions corrected using quantile matching or absolute homogenization made the station data more inhomogeneous. All contributions made the network data more inhomogeneous. The results for the comparable late contributions are similar to the blind ones.

### 6.3 Benchmark properties

#### 6.3.1 Surrogate versus synthetic

To answer the question whether there are differences between the surrogate and the synthetic data, an additional large dataset with 200 networks for each data section was generated. This dataset was homogenized with a newer version of ACMANT; see also Sect. 6.2.1. The analysis of

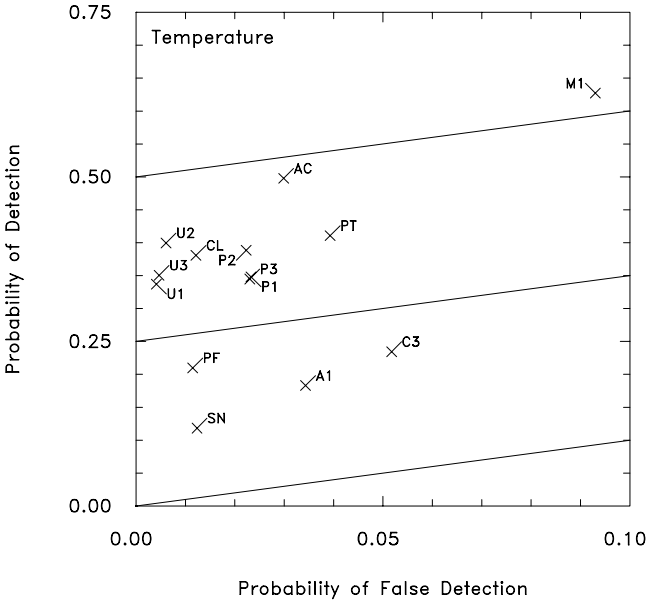


**Fig. 7.** Scatterplot of the probability of false detection against the probability of detection for the surrogate temperature dataset. The slant lines represent the Peirce (true) skill score. The crosses are the mean detection values of all complete surrogate temperature contributions. For the abbreviations see Table 9.

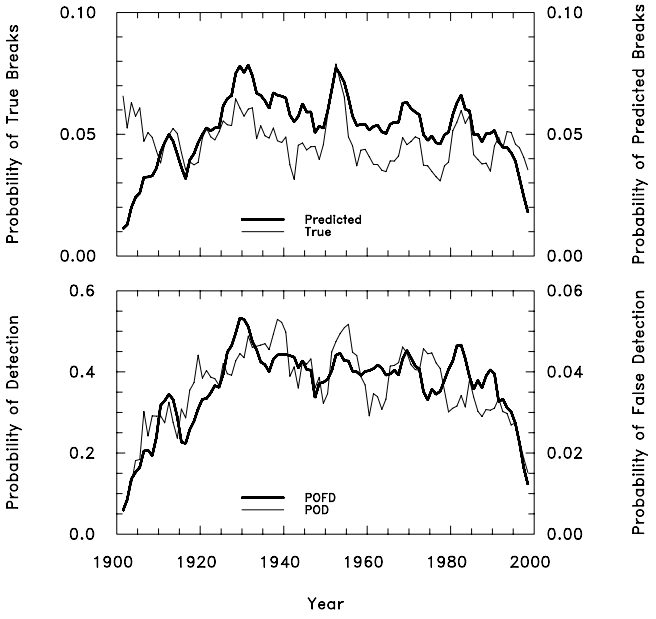


**Fig. 8.** The temporal behavior of the true breaks and predicted breaks (in top panel) and the probability of (false) detection (bottom panel) based on the homogeneous surrogate temperature stations and all complete contributions.

the homogenized data showed that the remaining error after homogenization, in terms of the monthly CRMSE, is 7 % smaller for the synthetic data. The standard deviation of the trend differences is 15 % smaller for the synthetic data compared to the surrogate data. All differences between surrogate and synthetic data are statistically highly significant. Thus synthetic data is easier to homogenize than the more realistic surrogate data.

### 6.3.2 Artificial inhomogeneities

To investigate how realistic the inserted inhomogeneities are, the detected breaks in the artificial data (surrogate and synthetic) are compared to those of the real data section of the benchmark. Only *USHCN*, *Climatol*, *Acmant*, and *AnClim main* have homogenized all real temperature networks. From the three *USHCN* contributions, *USHCN main* was selected to obtain independent data. *Climatol* was omitted as it showed problems with temperature trends. For precipitation, only *AnClim main* is available for analysis.

In the comparison below between the real and artificial networks of the properties of the detected breaks, also the power of detection should be taken into account and is analyzed first. The length of the record of the artificial data is set at 100 yr, whereas the real temperature (precipitation) data has a lower average record length of 87 yr (95 yr). The real temperature data has more missing data (on average about 20 yr) and it is more interspersed than in the artificial data, which on average has only 10 yr of missing data. The precipitation in all data sections contains about 90 yr of data. The average cross-correlation of the best correlating temperature pairs is higher for the real data (94 %) than for the artificial data (90 %). For precipitation these cross-correlations are 86, 81, 72 percent for real, surrogate and synthetic data, respectively.

The average annual break size in all data sections is not statistically different from zero. The magnitude of the artificial temperature breaks is larger: the average standard deviation of the annual detected break size distribution is 0.94 °C in the artificial data, whereas in the real data it is only 0.72 °C. For comparison, the average magnitude of all inserted breaks was 0.8 °C. The artificial annual precipitation break sizes are larger than the real ones: the standard deviation of the detected real breaks is 9.5 mm (10 %), but of the artificial breaks 15 mm (19 %). For comparison: the size of the inserted breaks is 15 %. Partially the smaller mean break size may be due to the stronger spatial correlations in the real precipitation dataset, which allows for the detection of smaller breaks.

The frequency of the artificial temperature breaks is lower: average frequency of detected breaks is 4.0 % and 4.7 % in the artificial and real data, respectively. More breaks are detected in the artificial precipitation data: 2.3 %, against 1.0 % in the real data.

Taken together the statistical properties of the networks and the nature of the breaks discovered do not differ greatly among the three data sections. Thus the differences discussed below are probably due to real differences in the statistical

properties of the inhomogeneities and not due to differences in the accuracy of homogenization.

If the perturbations applied at a break were independent, the perturbation time series would be a random walk. In the benchmark the perturbations are modeled as random noise, as a deviation from a baseline signal, which means that after a large break up (down) the probability of a break down (up) is increased. Defining a platform as a pair of breaks with opposite sign, this means that modeling the breaks as a random noise produces more than 50 % platform pairs. The percentage of platforms in the real temperature data section is 59 ($n = 742$), in the surrogate data 64 ($n = 1360$), and in the synthetic data 62 ($n = 1267$). The artificial temperature data thus contains more platforms; the real data is more like a random walk. This percentage of platforms and the difference between real and artificial data become larger if only pairs of breaks with a minimum magnitude are considered. In the precipitation data, the percentage of platforms is also above 50 %, but the values for the real and artificial data are similar. The perturbations in precipitation may thus be modeled as random noise, but more data and algorithms would be needed for firm conclusions.

Another important parameter is the seasonal cycle of the inhomogeneities. First the monthly anomalies are computed by subtracting the yearly means. Consecutively, the homogenization perturbations are computed from these anomalies. The size of the seasonal cycle of a break is parameterized as the change in the standard deviation of these perturbations before and after a break. The distribution of the break sizes of this seasonal cycle has a standard deviation of 0.19 (0.23) °C in the real (artificial) data. The seasonal cycle of the breaks in the artificial data is thus larger than in the real data and the homogenization algorithms underestimate the size of the seasonal cycle of the breaks (the average seasonal cycle of the breaks inserted into the benchmark is 0.4 °C). *USHCN* does not introduce a seasonal cycle and was omitted. *ACMANT* found stronger seasonal cycles in the breaks than *AnClim main*, but the difference between real and artificial data is about the same. In the precipitation data, the seasonal cycle of the breaks is 12 % in the real data and 19 % in the artificial data.

#### 6.3.3 Global biases and inhomogeneities

If inhomogeneities have a tendency to be in one direction during a certain period, they may have an influence on the network average signal, even for large networks. This could, for instance, happen in case new technologies or measurement procedures are introduced; see introduction. This effect can be studied in the cross-correlations between stations of the homogenization adjustments implemented and can be best seen in smoothed data.

Therefore, the perturbations were computed by comparing the inhomogeneous with the homogenized data and smoothing these perturbations in the same way as for the computation for the decadal CRMSE (Sect. 5). Consecutively, the average cross-correlation between all pairs of stations in a network was computed, after which this correlation was averaged over all networks in one of the three data sections of the benchmark. The same contributions were analyzed as in Sect. 6.3.2.

For the real, surrogate and synthetic data the cross-correlations are 9.1, −4.3 and 3.5 percent, respectively. Surprisingly, the cross-correlation for the surrogate data is negative. For the real and surrogate data these correlations are significant and they are also significantly different from each other. The values depend strongly on the homogenization method. Therefore only complete contributions have been used. However, when additionally including incomplete contributions the above inferences stay the same.

For precipitation, only *AnClim main* is available for analysis. The same inferences as for temperature may be made, but the difference between real and surrogate data is only significant at the $p = 7$ % level.

The raw datasets studied here are relatively recent. Records from the early instrumental records typically show artificial trends in all stations, because all stations made similar measurement errors (Böhm et al., 2001; Auer et al., 2005). The bias effect studied here may thus be stronger in older data.

## 7 Discussion

The discussion is divided into two parts. The lessons learned about homogenization of climate records will be discussed in Sect. 7.1, while Sect. 7.2 will deal with the benchmarking itself.

### 7.1 homogenization

Before discussing the performance of the algorithms it should be stated that the results for individual contributions should not be compared in too much detail for three reasons. First of all, the errors are non-Gaussian and dependent within one network. Especially in case of networks with multiple breaks that happen in multiple stations simultaneously, basically neutral changes in the algorithms can make the difference between solving a combinatorial problem or not. Therefore, the number of 15 networks is still quite limited and especially results for partial contributions should be interpreted with care. Secondly, there are uncertainties due to the limited realism of the benchmark data. While Sect. 6.3 showed that the average properties of the breaks in the temperature stations are reasonable in general, some deviations were found. The annual cycle of the breaks is somewhat exaggerated, which unfairly benefits the detection of breaks by *ACMANT*. Moreover, the perturbations due to inhomogeneities in the stations are more strongly cross-correlated in real data, which leads to larger perturbations in the network

mean signal. As a consequence, the errors in the network mean signals of the benchmark are small and harder to improve than in reality. See Sect. 7.2 for more details. Thirdly, results depend on the error metric analyzed, not only between the CRMSE of the time series, the RMSE of the linear trends and the detection scores, but also for the different averaging scale at which the CRMSE is computed and the period under consideration. Moreover, different treatments of the data particularly with respect to the missing data and the annual cycle, which are all reasonable, lead to differences in the errors found. In this context it should be noted that while many contributed to the analysis, the final pre-processing and analysis was performed by authors who did not submit homogenized data to avoid unfair biases.

The all-over best blind contributions are homogenized by *Craddock*, *MASH*, and *PRODIGE*. The blind *ACMANT* contribution had some problems with the network mean signal and trends, but the updated *ACMANT late* contribution suggests that ACMANT is currently the most accurate method available. *USHCN*, while less proficient than the four best ones, is nonetheless the best for the monthly network mean CRMSE and achieves its performance with a very low false alarm rate and without correcting the seasonal cycle.

All of these best methods have been designed to work with an inhomogeneous reference, either by using pairs or testing multiple reference time series for their suitability. Algorithms that circumvent the inhomogeneous-reference problem by first detecting the largest breaks are clearly less accurate. In praxis, the choice of a homogenization algorithm will also depend on the degree of automation desired or needed, which is related to the size of the network, and the access to expertise. Expertise and training is important; contributions using good algorithms by first time users often produced suboptimal results.

Some contributions result in data that is more inhomogeneous. In case of relative homogenization of temperature data, these cases could mostly be traced back to operating or programming errors. The latter are often related to the way iterations are performed. Algorithms using iterations have to be validated with extra care. Implicitly, this is connected to the advice "to always start homogenization from the beginning, assuming all series contain potential breaks and ignoring any previous homogeneity work undertaken for any of the series" (Auer et al., 2005).

Unfortunately only one contribution utilized absolute homogenization. This contribution produced much more inhomogeneous data, both for temperature as well as for precipitation. Absolute homogenization should thus be used with care and always accompanied by metadata. A more detailed study using multiple absolute homogenization methods (Reeves et al., 2007) would be worthwhile. The performance of absolute homogenization may have been reduced by the sometimes strong nonlinear global trends added to the data; see Sect. 3.

Precipitation data is expected to be more difficult to homogenize due to lower cross-correlations. The lower correlations should, however, only lead to less improvement of the data. The increases in inhomogeneity, found especially for the network average signals, are worrisome and warrant more research into the homogenization of precipitation. Given that the break detection score were positive, the problem probably lies in the noisy correction of precipitation data, especially for monthly correction. This is also suggested by the considerable difference for precipitation between *PRODIGE monthly*, which experimentally performed monthly corrections, and *PRODIGE main*, which applied more stable yearly corrections and was more accurate. Annual corrections are thus currently recommended for homogenization of typical precipitation networks.

The improvements achieved in CRMSE were much larger for yearly and decadal data than for monthly data. This is mainly related to the much smaller signal to noise ratio in the ratio time series due to the high spatial variability of precipitation, but may also be related to the fact that previous validation studies were limited to annual data. Monthly correction methods warrant more study. The correlations between the error metrics based on the time series themselves and break detection scores are modest (Sherwood et al., 2009), as well as for the detection scores amongst each other. The use of detection scores as sole performance criterion should thus be discouraged.

Most, but not all contributions, showed much larger errors in the beginning quarter or half of the century. Partially this is unavoidable due to the sparser density of the networks for the earlier periods. Consequently, detection of the changes is less precise, consequently also the corrections. These errors may also point to possibilities for developers of homogenization algorithms to improve the handling of missing data and of networks with few stations. Another reason may be that most algorithms perform no corrections for the more recent period and compute break sizes from one homogeneous subperiod to the next, which may lead to an accumulation of errors.

Some contributions applied algorithms that did not remove outliers themselves. The late surrogate temperature contributions applying the tests *PMTred* and *PMFred* did not show an influence of outliers on the results. Probably the results for the other temperature contributions without outlier removal are thus representative.

The contribution *PRODIGE trendy* that corrected local trends did not perform better than the versions that only corrected breaks, but trends were also only implemented in ten series. It should be studied whether improvements are more evident in those stations where local trends were present.

### 7.2 Benchmark

The synthetic data is apparently easier to homogenize than surrogate data. Especially the about 15 % smaller error in the linear trend estimation is climatologically relevant when

interpreting results based on homogenized data. As many validation studies did take into account the lag-one autocorrelation, it would be interesting to study in more detail whether this aspect of the surrogate data made it harder to homogenize. Alternative explanations could be the variability on large temporal scales (the correlations for all lags), or maybe the non-Gaussian nature of the distributions.

In software engineering it has been observed that a benchmark can help a field of science to mature, both due to social as well as technical factors (Sim et al., 2003). Also in the COST Action, the definition of the properties of the benchmark and the joint work on the same dataset helped to bring scientists closer together. The benchmarking also led to technical improvements, ranging from finding bugs, to improved understanding, and to an upcoming open-source state-of-the-art homogenization package.

Sim et al. (2003) state that benchmarking is more than providing a problem, but that it should also be announced in advance how the solutions will be judged. In this respect, the homogenization effort did not constitute a true benchmark. In case of homogenization, it is difficult, and may even be impossible, to boil down the results to one or two accuracy metrics. The contributions have been judged with respect to how well they reconstruct the temporal climatic variability, which is the most common reason to homogenize data. The data could also have been judged on how well the cross-correlations are reproduced or even the absolute values of the measured elements. With such an aim, another benchmark should have been produced, one in which observations performed at different locations are not merged to one long record. With the current experience, it is possible to communicate how the contributions will be judged in more detail for a future benchmarking exercise.

It is planned to redo the exercise every few years to monitor improvements in homogenization. As typical for a benchmarking project, also this benchmark will likely evolve. Updates will be implemented to avoid tuning and based on lessons from this study, see Sect. 6.3. Correlations in the perturbation applied to stations are important to increase the perturbations in network average data to realistic values.

A few remaining outliers were found to have little influence on homogenization; a future dataset could do without outliers. The periods with much missing data clearly made homogenization more difficult; in future also inserting random missing data may thus be interesting and enhance the realism of the benchmark.

The best contributions and especially *ACMANT late* perform very well. A future benchmark dataset should thus be more challenging, for instance by reducing the density of the networks.

The participants were requested to focus on homogenizing the surrogate data section. In retrospect more emphasis on the importance of the real data section should have been given and the real and surrogate data should be based on similar datasets for better comparison. While the surrogate data provides an estimate of the accuracy of the homogenization algorithms, the comparison of the results for the surrogate and the real data is needed to interpret the differences between the contributions. Furthermore, this comparison is important for the development of more realistic future benchmarks.

## 8 General conclusions and recommendations

The main research impetus for the last two decades has been the development of homogenization algorithms that also function with an inhomogeneous reference time series. This effort has paid off. There is a clear split in performance on the benchmark data between these direct algorithms and the ones, which evade the inhomogeneous-reference problem using older concepts such as stepwise or semi-hierarchical splitting, as well as detection on moving windows. With mathematical argumentation, climatological reasoning and the benchmark metrics all pointing in the same direction, we thus strongly recommend the use of direct homogenization algorithms. Such participating algorithms are: ACMANT, Craddock, MASH, PRODIGE and USHCN. ACMANT, MASH and PRODIGE also tackle the multiple break-point problem directly, which is also important for their performance.

Almost all relative homogenization algorithms improved the homogeneity of the temperature data. The exceptions could mostly be explained by inexperienced users or be traced back to algorithms (or parts thereof) newly written for this exercise. The results illustrate that statistical absolute homogenization has the potential to make the data even more inhomogeneous. Some contributions created with the best algorithms were much less accurate than the contributions by the developers. This indicates that training of the operator is very important and that developers should invest more effort into making their software easy to use and give out relevant warnings.

We feel that this blind test of homogenization algorithms has benefited the homogenization community, see Sect. 7.2, and advocate to repeat the exercise in future. One follow-up is the surface temperature initiative, which is working on a global homogenized surface temperature dataset and has started a benchmarking initiative for its homogenization algorithms (Thorne et al., 2011). Due to its sheer size, such a benchmark would only be of interest to automatic homogenization algorithms. There may thus be room for additional initiatives studying other climate variables and utilizing smaller networks for comparison with manual methods.

Benchmarking is not only useful to study the performance of the homogenization algorithms. The definition of the properties of the benchmark, the work on the same dataset and the joint analysis of the results has strengthened the community. The benchmarking has also let to technical

improvements, ranging from finding bugs, to improved understanding, and to the recommendations for an upcoming open-source state-of-the-art homogenization package.

### 8.1 Recommendations

Benchmarking officially requires agreeing on the error metrics in advance. For homogenization there is not one clearly preferred metric, however. With the current experience, it should be possible, though, to define the initial analysis in more detail for a future benchmark. The results showed only modest correlations between the break detection scores, which developers of homogenization methods tend to focus on, and the other error metrics, which are close to the needs of climatologists. It is thus recommended to use both types of error metrics in future validation studies.

In retrospect, too little emphasis was given to the homogenization of the real data section, which provides a validation of the statistical properties of the inserted inhomogeneities. For future benchmarking exercises, more studies on the statistical characteristics of inhomogeneities for various climate elements would be important. The size distribution of temperature inhomogeneities in Western countries is studied reasonably well, but for other regions and climatic variables more information would be valuable. Too little studied and quantified are cross-correlations of the breaks between stations, see Sect. 6.3.2. Especially periods in which breaks are biased in one direction lead to a much stronger perturbation of the regional climate signal (average over multiple stations), as the random breaks used in this study, and should be included in any future benchmark dataset.

Furthermore, the breaks in the benchmark are modeled as deviations from the baseline values, i.e. as random noise. An alternative way to model breaks would be relative to the previous values, i.e. as a random walk. The random noise model was found to be reasonable, but for the temperature records a mixed model with a small random-walk component may be even more realistic.

Irrespective of the above mentioned advantages of benchmarking and the reliability of the blind results, there are also disadvantages to benchmarking and alternative validation methodologies should also be used. An important disadvantage is that the blind test does not allow for the correction of problems discovered during the analysis. Consequently, not all methods could deliver their optimal performance. The interpretation is furthermore hampered by differences in experience and effort of the participants. Finally, because of its competitive character, it is paramount that the statistical properties of the data and the inhomogeneities are realistic. Otherwise it would be possible to tailor the algorithms to the benchmark and perform better on the benchmark than on real data. Therefore, benchmarking does not allow for systematic studies aimed at understanding the algorithms, for instance by systematically testing varieties of an algorithm, and for testing the limits of the methods with unrealistic easy or difficult data – the latter being the strength of standard intercomparison studies and mathematical analysis. Another valuable validation strategy is the testing of the methods on real data with good metadata. Given that every methodology has its own advantages and disadvantages, we expect that progress is best served by a diversity of methodologies. Benchmarking is important for its ability to obtain reliable accuracy metrics, due to the blind testing of the contributions and the realism of the data.

The use of metadata and reconstructions of past observation methodologies is preferred over statistical homogenization, especially to designate the dates of the breaks more precisely. To find additional not documented breaks, statistical homogenization should always be used as well. In future, more homogenization algorithms should implement the automatic use of metadata, so that a future benchmark can also include simulated metadata. National Meteorological Services should intensify their work on the digitization of metadata (Brunet and Jones, 2011) and the formulation of a standard machine-readable format for metadata.

The intelligent use of metadata is an advantage of manual methods over automatic ones, yet automatic methods may tempt people to rely less on metadata. Further advantages of manual methods are the climatological knowledge of the operator on how much variability is allowed in the difference time series, which accordingly allows for an intelligent selection of similar stations. Furthermore, humans are good at solving combinatorial problems, which explains the quality of the Craddock and PRODIGE contributions. Strengths of automatic methods are their objectivity and reproducibility. Furthermore, automatic methods can be easily applied to large datasets and thus also lend themselves better to validation and benchmarking, which aids their refinement. The study showed that currently automatic and semi-automatic algorithms (ACMANT, MASH, USHCN) can perform as well as manual ones.

A considerable difference in improvement of the data by homogenization was found between annual and monthly data. Furthermore, the break detection scores are only modestly related to the remaining centered root mean square error. Both findings suggest that more work on the correction algorithms could be fruitful. The benchmark dataset could be used to study the performance of various correction methods.

The results for precipitation were not as good as for temperature. This may well be due to the more difficult estimation of the correction factors. This is suggested by the positive performance for detection and the higher accuracy of the PRODIGE contribution with annual factors compared with the version with monthly factors. The operators also have more experience with temperature and the algorithms are better validated for temperature. It should also be noted that the properties of the benchmark data may have been less good for precipitation, as less is known about the statistical properties of breaks in precipitation and too little homogenized real

data was available for a stringent validation of the benchmark. Given these results and the importance of precipitation for climate impact research, the homogenization of precipitation should be given priority. It may be worthwhile to generate a dedicated benchmark for precipitation.

Many evidently interesting questions are not yet answered and will hopefully be studied in subsequent articles. For instance, in networks without breaks, homogenization algorithms should change as little as possible; this can be studied in the network without inserted inhomogeneities. How well the gradual local trends are removed by homogenization would warrant a dedicated study, as well. This analysis was mainly based on statistical metrics of interest to many users of the homogenized data. With the benchmark dataset being available, any climatologist can now study the influence of remaining inhomogeneities on a specific analysis. Users may, for instance, be interested in the annual cycle, the cross-correlations between stations, as well as secular trends for individual months, (interannual) variability, intermittence and long range dependence (Rust et al., 2008).

Based upon the results on the benchmark and theoretical consideration, the Action is currently working on providing a free software package with recommended homogenization tools, which will be published on the HOME homepage (HOME, 2011).

*Acknowledgements.* This study has been performed with support of the European Union, through the COST Action ES0601 – Advances in Homogenisation Methods of Climate Series: an Integrated Approach (HOME), as well as the project Large Scale Climate Changes and their Environmental Relevance funded by the North Rhine-Westphalia Academy of Science. The contribution of VV was supported by the surrogate cloud project (VE 366/3), the one of RL by the Daily Stew project (VE366/5), both sponsored by the German Science Foundation (DFG). The contribution of EA was sponsored by the "Cambios en la Frecuencia, Intensidad y Duración de eventos Extremos en la Península Ibérica", code number: CGL2007-65546-C03-02. This dataset would have been impossible without contributed climate records. Our thanks thus go to Meteo France, Ecole Nationale de la Meteorologie, Toulouse, France, the Zentralanstalt für Meteorologie und Geodynamik, Wien, Austria, the Center on Climate Change (C3), Tarragona, Spain, the Centre d'Estudis de la Neu i de la Muntanya d'Andorra, Andorra, the Royal Netherlands Meteorological Institute, De Bilt, The Netherlands, the National Institute of Meteorology and Hydrology – BAS, Sofia, Bulgaria, and the National Meteorological Administration, Bucharest, Romania. Furthermore, we would like to thank Lucie Vincent and Mário Pereira for their comments.

Edited by: P. Brohan

## References

Aguilar, E., Auer, I., Brunet, M., Peterson, T. C., and Wieringa, J.: Guidelines on climate metadata and homogenization. World Meteorological Organization, WMO-TD No. 1186, WCDMP No. 53, Geneva, Switzerland, p. 55, 2003.

Alexandersson, A.: A homogeneity test applied to precipitation data, J. Climatol., 6, 661–675, 1986.

Alexandersson, H. and Moberg, A.: homogenization of Swedish temperature data.1, Homogeneity test for linear trends, Int. J. Climatol., 17, 25–34, 1997.

Auer, I., Böhm, R., Jurkovic, A., Orlik, A., Potzmann, R., Schöner, W., Ungersböck, M., Brunetti, M., Nanni, T., Maugeri, M., Briffa, K., Jones, P., Efthymiadis, D., Mestre, O., Moisselin, J.-M., Begert, M., Brazdil, R., Bochnicek, O., Cegnar, T., Gajic-Capka, M., Zaninovic, K., Majstorovicp, Z., Szalai, S., Szentimrey, T., and Mercalli, L.: A new instrumental precipitation dataset for the Greater Alpine Region for the period 1800–2002, Int. J. Climatol., 25, 139–166, 2005.

Auer, I., Böhm, R., Jurkovic, A., Lipa, W., Orlik, A., Potzmann, R., Schöner, W., Ungersböck, M., Matulla, C., Briffa, K., Jones, P., Efthymiadis, D., Brunetti, M., Nanni, T., Maugeri, M., Mercalli, L., Mestre, O., Moisselin, J.-M., Begert, M., Müller-Westermeier, G., Kveton, V., Bochnicek, O., Stastny, P., Lapin, M., Szalai, S., Szentimrey, T., Cegnar, T., Dolinar, M., Gajic-Capka, M., Zaninovic, K., Majstorovicp, Z., and Nieplovaq, E.: HISTALP – Historical Instrumental Climatological Surface Time Series of the Greater Alpine Region, Int. J. Climatol., 27, 17–46, doi:10.1002/joc.1377, 2007.

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

Begert, M., Schlegel, T., and Kirchhofer, W.: Homogeneous temperature and precipitation series of Switzerland from 1864 to 2000, Int. J. Climatol., 25, 65–80, 2005.

Böhm, R., Auer, I., Brunetti, M., Maugeri, M., Nanni, T., and Schöner, W.: Regional temperature variability in the European Alps 1760–1998 from homogenized instrumental time series, Int. J. Climatol., 21, 1779–1801, 2001.

Brunet, M. and Jones, P.: Data rescue initiatives: bringing historical climate data into the 21st century, Clim. Res., 47, 29–40, 2011.

Brunet, M., Asin, J., Sigró, J., Banón, M., García, F., Aguilar, E., Esteban Palenzuela, J., Peterson, T. C., and Jones, P.: The minimization of the screen bias from ancient Western Mediterranean air temperature records: an exploratory statistical analysis, Int. J. Climatol., 31, 1879–1895, 2011.

Brunetti, M., Maugeri, M., Monti, F., and Nanni, T.: Temperature and precipitation variability in Italy in the last two centuries from homogenized instrumental time series, Int. J. Climatol., 26, 345–381, 2006.

Buishand, T. A.: Some methods for testing the homogeneity of rainfall records, J. Hydrol., 58, 11–27, 1982.

Caussinus, H. and Mestre, O.: Detection and correction of artificial shifts in climate series, Appl. Statist., 53, 405–425, 2004.

Cleveland, W. S. and Devlin, S. J.: Locally weighted regression: An approach to regression analysis by local fitting, J. Am. Stat. Assoc., 83, 596–610, 1998.

Conrad, V.: Homogenitätsbestimmung meteorologischer Beobachtungsreihen, Meteorol. Z., 42, 482–485, 1925.

Conrad, V.: Methods in climatology, Havard University Press, p. 228, 1944.

Conrad, V. and Pollak, C.: Methods in climatology, Harvard University Press, Cambridge, MA, p. 459, 1950.

Craddock, J. M.: Methods of comparing annual rainfall records for

climatic purposes, Weather, 34, 332–346, 1979.

DeGaetano, A. T.: Attributes of several methods for detecting discontinuities in mean temperature series, J. Climate, 19, 838–853, 2006.

Della-Marta, P. M., Collins, D., and Braganza, K.: Updating Australia's high quality annual temperature dataset, Austr. Meteor. Mag., 53, 277–292, 2004.

Domonkos, P.: Testing of homogenisation methods: purposes, tools and problems of implementation. Proceedings of the 5th Seminar and Quality Control in Climatological Databases, edited by: Lakatos, M., Szentimrey, T., Bihari, Z. and Szalai, S., WCDMP-No. 71, WMO/TD-NO. 1493, 126–145, 2008.

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

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

Domonkos, P. and Štepánek, P.: Statistical characteristics of detectable inhomogeneities in observed meteorological time series, Stud. Geophys. Geod., 53, 239–260, 2009.

Domonkos, P., Poza, R., and Efthymiadis, D.: Newest developments of ACMANT, Adv. Sci. Res., 6, 7–11, doi:10.5194/asr-6-7-2011, 2011.

Ducré-Robitaille, J.-F., Vincent, L. A., and Boulet, G.: Comparison of techniques for detection of discontinuities in temperature series, Int. J. Climatol., 23, 1087–1101, 2003.

Easterling, D. R. and Peterson, T. C.: A new method for detecting undocumented discontinuities in climatological time series, Int. J. Climatol., 15, 369–377, 1995.

Guijarro, J. A.: User's guide to climatol. An R contributed package for homogenization of climatological series, Report, State Meteorological Agency, Balearic Islands Office, Spain, available at: http://webs.ono.com/climatol/climatol.html, 2011.

Gullett, D. W., Vincent, L., and Sajecki, P. J. F.: Testing for homogeneity in temperature time series at Canadian climate stations. CCC Report No. 90-4, Atmospheric Environment Service, Downsview, Ontario, p. 43, 1990.

Hansen, J., Ruedy, R., Sato, M., Imhoff, M., Lawrence, W., Easterling, D. R., Peterson, T. C., and Karl, T. R.: A closer look at United States and global surface temperature change, J. Geophys. Res., 106, 23947–23963, 2001.

Heidke, P.: Quantitative Begriffsbestimmung homogener Temperatur- und Niederschlagsreihen, Meteorol. Z., 40, 114–115, 1923.

Helmert, F. R.: Die Ausgleichrechnung nach der Methode der kleinsten Quadrate, 2. Auflage, Teubner Verlag, 1907.

HOME: Homepage of the COST Action ES0601 – Advances in Homogenisation Methods of Climate Series: an Integrated Approach (HOME), available at: http://www.homogenisation.org, last access: 22 December 2011, 2011.

HMS: Hungarian Meteorological Service. Proceedings of the First Seminar for homogenization of Surface Climatological Data, Budapest, Hungary, 6–12 October 1996, p. 44, 1996.

IPCC: Climate Change 2007: The physical science basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K. B., Tignor, M., and Miller, H. L., Cambridge University Press, Cambridge, United Kingdom, p. 996, 2007.

Kohler, M. A.: Double-mass analysis for testing the consistency of records and for making adjustments, B. Am. Meteorol. Soc., 30, 188–189, 1949.

Kreil, K.: Mehrjährige Beobachtungen in Wien vom Jahre 1775 bis 1850, Jahrbücher der k.k. Central-Anstalt für Meteorologie und Erdmagnetismus, I. Band – Jg. 1848 und 1849, 35–74, 1854a.

Kreil, K.: Mehrjährige Beobachtungen in Mailand vom Jahre 1763 bis 1850, Jahrbücher der k.k. Central-Anstalt für Meteorologie und Erdmagnetismus, I. Band – Jg. 1848 und 1849, 75–114, 1854b.

Menne, M. J. and Williams, C. N. Jr.: Detection of undocumented changepoints using multiple test statistics and composite reference series, J. Climate, 18, 4271–4286, 2005.

Menne, M. J., Williams, C. N. Jr., and Vose, R. S.: The U.S. historical climatology network monthly temperature data, version 2, B. Am. Meteorol. Soc., 90, 993–1007, doi:10.1175/2008BAMS2613.1, 2009.

Menne, M. J., Williams, C. N. Jr., and Palecki M. A.: On the reliability of the US surface temperature record, J. Geophys. Res. Atmos., 115, D11108, doi:10.1029/2009JD013094, 2010.

Mestre, O.: Step-by-step procedures for choosing a model with change-points. In Proceedings of the second seminar for homogenisation of surface climatological data, Budapest, Hungary, WCDMP-No.41, WMO-TD No. 962, 15–26, 1999.

OMSZ: Third Seminar for homogenization and Quality Control in climatological Databases, Budapest, available at: http://omsz.met.hu/omsz.php?almenu_id=omsz&pid=seminars&pri=16, last access: 22 December 2011, 2001.

Perreault, L., Bernier, J., Bobée, B., and Parent, E.: Bayesian change-point analysis in hydrometeorological time series, Part. 1. The Normal model revisited, J. Hydrol., 235, 221–241, 2000.

Peterson, T. C.: Assessment of urban versus rural in situ surface temperatures in the contiguous United States: No difference found, J. Climate, 16, 2941–2959, 2003.

Peterson, T. C., Easterling, D. R., Karl, T. R., Groisman, P., Nicholls, N., Plummer, N., Torok, S., Auer, I., Boehm, R., Gullett, D., Vincent, L., Heino, R., Tuomenvirta, H., Mestre, O., Szentimrey, T., Salingeri, J., Førland, E. J., Hanssen-Bauer, I., Alexandersson, H., Jones, P., and Parker, D.: Homogeneity adjustments of in situ atmospheric climate data: A review, Int. J. Climatol., 18, 1493–1517, 1998.

Reeves, J., Chen, J., Wang, X. L., Lund, R., and Lu, Q.: A review and comparison of changepoint detection techniques for climate data, J. Appl. Meteorol. Climatol., 46, 900–915, 2007.

Rust, H. W., Mestre, O., and Venema, V. K. C.: Less jumps, less memory: homogenized temperature records and long memory, J. Geophys. Res. Atmos., 113, D19110, doi:10.1029/2008JD009919, 2008.

Schreiber, T. and Schmitz, A.: Improved surrogate data for nonlinearity tests, Phys. Rev. Lett., 77, 635–638, 1996.

Sherwood, S. C., Titchner, H. A., Thorne, P. W., and McCarthy, M. P.: How do we tell which estimates of past climate change are correct?, Int. J. Climatol., 29, 1520–1523, 2009.

Sim, S. E., Easterbrook, S., and Holt, R. C.: Using benchmarking to advance research: A challenge to software engineering. Proceedings of the 25th International Conference on Software Engineering ICSE '03, IEEE Computer Society Washington, DC, USA, ISBN: 0-7695-1877-X, 74-83, 2003.

Štěpánek, P., Zahradníček, P., and Skalák, P.: Data quality control and homogenization of the air temperature and precipitation series in the Czech Republic in the period 1961–2007, Adv. Sci. Res., 3, 23–26, 2009.

Szentimrey, T.: Multiple Analysis of Series for homogenization (MASH). Proceedings of the second seminar for homogenization of surface climatological data, Budapest, Hungary, WMO, WCDMP-No. 41, 27–46, 1999.

Szentimrey, T.: Manual of homogenization software MASHv3.02, Hungarian Meteorological Service, p. 65, 2007.

Szentimrey, T.: Development of MASH homogenization procedure for daily data. Proceedings of the fifth seminar for homogenization and quality control in climatological databases, Budapest, Hungary, 2006, WCDMP-No. 71, 123–130, 2008.

Thorne, P. W., Willett, K. M., Allan, R. J., Bojinski, S., Christy, J. R., Fox, N., Gilbert, S., Jolliffe, I., Kennedy, J. J., Kent, E., Klein Tank, A., Lawrimore, J., Parker, D. E., Rayner, N., Simmons, A., Song, L., Stott, P. A., and Trewin, B.: Guiding the creation of a comprehensive surface temperature resource for 21st century climate science, B. Am. Meteorol. Soc., 92, ES40–ES47, doi:10.1175/2011BAMS3124.1, 2011.

Trewin, B.: Exposure, instrumentation, and observing practice effects on land temperature measurements, WIREs Clim. Change, 1, 490–506, doi:10.1002/wcc.46, 2010.

Venema, V. K. C., Bachner, S., Rust. H. W., and Simmer, C.: Statistical characteristics of surrogate data based on geophysical measurements, Nonlin. Proc. Geophys., 13, 449–466, 2006a.

Venema, V. K. C., Ament, F., and Simmer, C.: A stochastic iterative amplitude adjusted Fourier transform algorithm with improved accuracy, Nonlin. Proc. Geophys., 13, 247–363, 2006b.

Venema, V. K. C., Mestre, O., Aguilar, E., Auer, I., Guijarro, J. A., Domonkos, P., Vertacnik, G., Szentimrey, T., Stepanek, P., Zahradnicek, P., Viarre, J., Müller-Westermeier, G., Lakatos, M., Williams, C. N., Menne, M., Lindau, R., Rasol, D., Rustemeier, E., Kolokythas, K., Marinova, T., Andresen, L., Acquaotta, F., Fratianni, S., Cheval, S., Klancar, M., Brunetti, M., Gruber, C., Prohom Duran, M., Likso, T., Esteban, P., and Brandsma, T.: Description of the COST-HOME monthly benchmark dataset and the submitted homogenized contributions. Report, Meteorological Institute, University of Bonn, Germany, available at: http://www2.meteo.uni-bonn.de/venema/articles/2011/report_home.pdf, 2011.

Vincent, L. A.: A technique for the identification of inhomogeneities in Canadian temperature series, J. Climate, 11, 1094–1104, 1998.

Wang, X. L. L.: 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, 2008.

WMO: Proceedings of the Second Seminar for homogenization of Surface Climatological Data. Budapest, Hungary, 9–13 November 1998, p. 214, 1999.

WMO: Fourth Seminar for homogenization and Quality Control in Climatological Databases. Budapest, Hungary, 6–10 October 2003, WCDMP-No 56, WMO-TD No. 1236, p. 243, 2004.

WMO: Proceedings of the Fifth Seminar for homogenization and Quality Control in Climatological Databases. Budapest, Hungary, 29 May–2 June 2006. Climate Data and Monitoring WCDMP- No 71, WMO/TD- No. 1493, 2006.

WMO: Proceedings of the Sixth Seminar for homogenization and Quality Control in Climatological Databases. Budapest, Hungary, 26–30 2008, edited by: Lakatos, M., Szentimrey, T., Bihari, Z., and Szalai, S., WCDMP-No. 76, WMO/TD-NO. 1576, 2011.