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. Fratianni1[5], 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], T. Brandsma[23], K. Willett[24]

[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] *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*
[24] *Met Office Hadley Centre, Exeter, United Kingdom*

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. The algorithms were validated against a realistic benchmark dataset. Participants provided 25 separate homogenized contributions as part of the blind study as well as 22 additional solutions submitted after the details of the imposed inhomogeneities were revealed. These homogenized datasets were assessed by a number of performance metrics including i) the centered root mean square error relative to the true homogeneous values 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. Although relative homogenization algorithms typically improve the homogeneity of temperature data, only the best ones improve precipitation data. Moreover, state-of-the-art relative homogenization algorithms developed to work with an inhomogeneous reference are shown to perform best. The study showed that currently automatic algorithms can perform as well as manual ones.

Keywords: Surface climate network, instrumental climate records; monthly temperature records; monthly precipitation records; surface stations; homogenization, benchmarking; blind validation study; surrogate data.

INTRODUCTION

To study climate variability the original surface observations are indispensable, but these have to be treated with care. Long observational records always contain changes due to non-climatic factors as well. Such inhomogeneities can be either sudden jumps (breaks) or gradual trends in one station. In a recent paper [1], the methods to remove these factors were blindly tested on a benchmark temperature and precipitation dataset with inserted inhomogeneities. This abstract will present the main ideas from this study. All the most common and the most developed algorithms were tested.

Most surface stations are not operated for climatic purposes, but rather to meet the needs of weather forecasting, agriculture and hydrology [2]. Consequently, the average period between detected breaks in Western instrumental records is 15 to 20 years. The typical size of the breaks is of the same order as the climate change signal during the 20th century [3-7]. Specific inhomogeneities are typical for certain periods and common to many stations, these can collectively lead to artificial biases in climate trends across large regions. Inhomogeneities are thus a significant source of uncertainty for the estimation of secular trends and decadal-scale variability.

To the general public, the best known inhomogeneity is probably the urban heat island effect. The temperature in cities can be warmer than in the surrounding country side, especially during calm nights. As cities have grown they have encroached on the weather stations, raising the ambient temperature. It is not clear how important this effect is; the Austrian network even had a bias because the surrounding of the stations became less urban [8]. World wide the advent of aviation led to relocations of stations from cities to nearby, typically cooler, airports [9]. In general, relocations are an important cause of inhomogeneities.

Inhomogeneities caused by changes in the screens that protect the instruments for radiation and wetting are also common [10]. In 19th century Europe it was common to use a metal screen in front of a window on a North facing wall. However, the building may warm the screen leading to higher temperature measurements. When this problem was realized the cotton region shelter was introduced. Other typical causes of inhomogeneities are changes in the surrounding environment, e.g., land use change and building activity. In recent times, the most important inhomogeneity is the change over to automatic weather stations [11].

HOMOGENIZATION

Ideally, the date of a change of instruments, locations or observing practices would be documented and parallel measurements made with the original and the new set-up for several years [12]. By making parallel measurement with replicas of historical instruments, screens, etc., the influence of some historical inhomogeneities can still be studied today.

Because you are never sure that your metadata is complete, statistical homogenization is necessary as well. The most commonly used principle to detect and remove the artificial changes is relative homogenization [13]. This assumes that nearby stations are exposed to almost the same climate signal, but not any non-climatic changes. By looking at the difference between nearby stations, the year to year variability of the climate is removed, as well as the regional climatic trend. In such a difference time series, a clear and persistent jump can easily be detected and can only be due to changes in the measurement conditions.

If there is a jump (break) in a difference time series of a pair of stations, it is not yet clear which of the two stations it belongs to. Furthermore, time series typically have more than just one jump. These two features make statistical homogenization a challenging and beautiful statistical problem. Homogenization algorithms typically differ in how they solve these two fundamental problems.

Traditionally, this first fundamental problem is solved in relative homogenization by comparing a candidate series with a composite reference time series computed from its neighboring stations. This composite reference is assumed to be homogeneous due to averaging, which is only approximately true. The main research impetus for the last two decades has been the development of so-called direct homogenization algorithms that also function with an inhomogeneous reference time series.

Sometimes there are no other stations in the same climate region. In this case, sometimes absolute homogenization is applied and the inhomogeneities are detected in the time series of one station, i.e. without using a reference [14]. If there is a clear and large break at a certain date, such a break may be removed reasonably accurately, but smaller jumps and gradually occurring inhomogeneities (for instance due to the urban heat island or growing vegetation) cannot be distinguished from real natural variability and climate change. Data homogenized this way does not have the quality one may expect and should be used with much care.

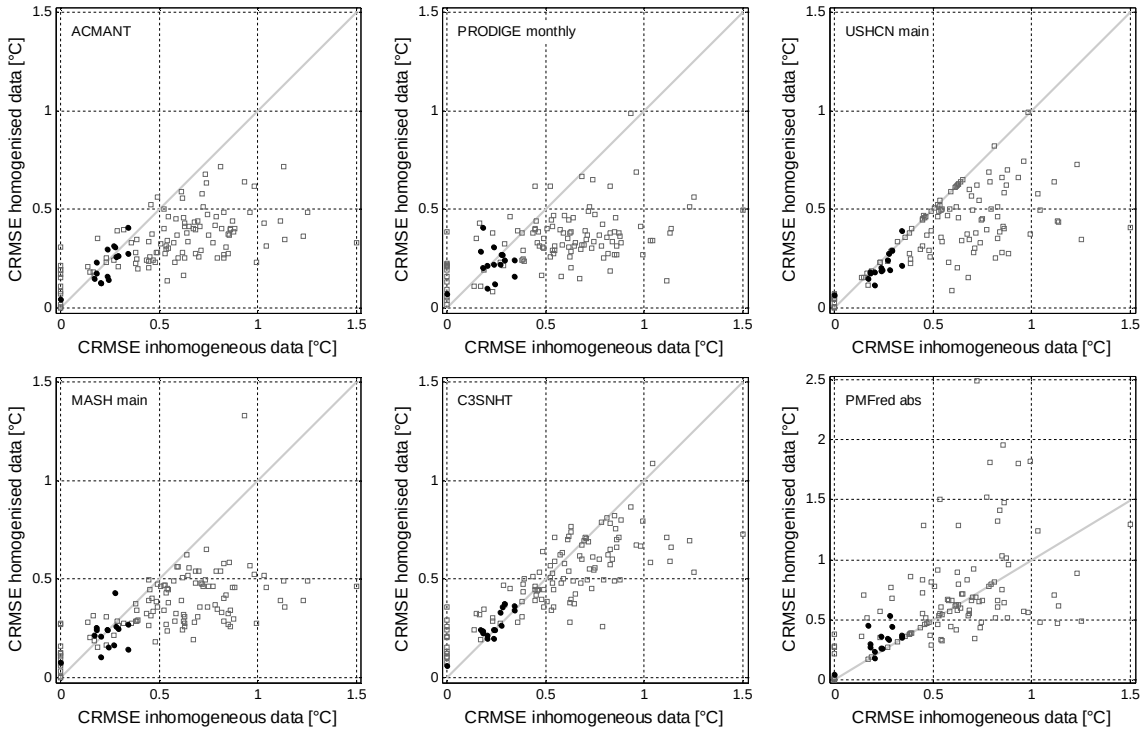

FIGURE 1. Scatterplot of the centered RMSE before and after homogenization for selected contributions. The squares display the errors of the stations; the dots show the errors of the network mean (regional climate) time series. Points on the bisect indicate no change, above the bisect the data is made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

BENCHMARKING

The benchmark dataset mimics station networks and their data problems with unprecedented realism. Homogeneous data was generated using the so-called surrogate data approach, which reproduces the cross- and auto-correlation functions, as well as the non-Gaussian distribution of climate observations [15, 16]. To this data random break-type inhomogeneities, as well as breaks that occur simultaneously in multiple stations, were added. Furthermore, local trends, which either continue at the end (to model for instance the urban heat island effect) or go back to baseline (to model growing vegetation that is cut back at the end) were inserted. The sizes of the breaks and local trends follow a normal distribution with a width of 0.8 °C for temperature [17], and 15 % for precipitation. Further, a stochastic nonlinear network-wide trend was added. The main novelty was that the test was blind. Furthermore, the benchmark was generated and the analysis of results was performed by independent researchers, who did not homogenize the data themselves. Everyone was invited to homogenize the data; 25 homogenized blind contributions were returned.

RESULTS

For a clear presentation of all results, the homogenization methods used to homogenize the 25 contributions would have to be explained, this is not possible in the limited space of this abstract.

To get a flavor of the results, see Figure 1. It shows the root mean square error of the centered monthly data. The time series are centered by subtracting the mean because homogenization aims to improve the temporal consistency of the data, not the absolute level. The first four contributions in Figure 1 — ACMANT (Spain) [18], PRODIGE monthly (Meteo France) [19], USHCN main (NOAA, USA) [4] and MASH main (Hungarian weather service) [20, 21] — are all direct homogenization algorithms and clearly perform better than the traditionally used SNHT method [22], here exemplified by CSNHT.

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) [23], which typically makes the benchmark data more inhomogeneous for both monthly and yearly mean values.

The mean centered RMSE for the inhomogeneous monthly temperature station data is 0.57 °C. The best complete contribution reduced this error to 0.34 °C. The same numbers for precipitation are 10.6 mm and 9.0 mm, respectively. On yearly and decadal scales the reduction in errors by homogenization is much larger. The CRMSE for the inhomogeneous annual temperature station data is 0.47 °C, whereas the best complete contribution reduced this error to 0.15 °C. The same numbers for precipitation are 7.3 mm and 4.7 mm.

The RMSE of the linear trend estimate for the inhomogeneous annual temperature station data is 1.19 °C/100yr. The best complete contribution reduced this error to 0.32 °C/100yr. The same numbers for precipitation are 15 mm/100yr and 7.5 mm/100yr, respectively.

The probability a break is detected in the right year is modest, which is also due to the large number of small breaks inherent in normal break size distribution. The probability of false detection is typically (well) below 5 percent. The Peirce and Heidke skill scores are positive, that is the results are better than random. Interestingly, the best contributions with respect to CRMSE and trends are not necessarily the ones with best scores for detection.

CONCLUSIONS AND OUTLOOK

The first main conclusion is that relative homogenization improves the temperature data; it reduces the root mean square error of the data and its linear trend coefficients and does not cause artificial climate trends. This conclusion can be stated with confidence because the test was blind and because of the realism of the data. The exceptions, where relative homogenization made the data more inhomogeneous, could mostly be explained by inexperienced users or be traced back to algorithms (or parts thereof) newly written for this exercise. This points to an important disadvantage of blind studies: mistakes discovered after the results are shared with participants cannot be corrected anymore. The results confirm the theoretical expectation that statistical absolute homogenization has the potential to make climate data more inhomogeneous; it should be noted that the benchmark data may have been more difficult for absolute methods as real data is due to the strong nonlinear trends added to the networks. In contrast to the results for temperature the positive results for precipitation are more mixed, still all but one relative method did improve the station trends. Many inhomogeneities are documented. Had such documentation been provided for part of the breaks in the benchmark data, the errors would have been lower.

The second main conclusion is that direct relative homogenization algorithms are clearly better than traditional ones. It needed a realistic benchmark dataset to see this difference with such clarity. 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.

The performance ranking of the homogenization methods depends on the error metric considered, on whether the root mean square error is computed on the monthly, yearly or decadal data and on whether it is computed on the station data or on the network mean climate signal. These rankings also do not correlate strongly with the error in the linear trend estimates (or break detection scores). In other words, it is difficult to compute one error metric that would signify the remaining error after homogenization for all climatic purposes. The computation and communication of the remaining uncertainties of homogenized data should be one of the research priorities for the coming years.

We feel that benchmarking has helped the homogenization community to mature [24]. The discussions on the properties of the benchmark, the nature of inhomogeneities in the various regions and on homogenization methods, as well as the joint work on the same dataset helped to bring scientists closer together in a way that writing individual papers cannot. The International Surface Temperature Initiative has started a follow-up benchmarking program for homogenization algorithms [25]. This benchmark will be global and be even more realistic, especially due to the inclusion of metadata, biased inhomogeneities and random missing data.

Everyone is invited to download and analyze the benchmark dataset. The homogeneous, inhomogeneous and homogenized datasets are published in the internet. Another offspring of the Action is HOMER, an open-source state-of-the-art homogenization package based on the best methods. The package is written in R and also performs basic quality control. Furthermore a mailing list for researchers working on homogenization has been started. All these resources can be accessed via the webpage of the Action, http://www.homogenisation.org, which will be kept running for the coming years and which also contains an extensive bibliography.

With advanced and well-validated statistical methods, the homogenization of annual and monthly station data is a mature field. The homogenization of daily data is still in its infancy, however. Daily data are es-

sential for studying extremes of weather and climate and therefore the basis for important political decisions with huge socio-economic consequences. For such studies the complete distribution needs to be homogenized. Looking at the physical causes of inhomogeneities, one would expect that many of them especially affect the tails of the distribution of the daily data. Likewise the IPCC AR4 report warns that changes in extremes are often more sensitive to inhomogeneous climate monitoring practices than changes in the mean [26]. This is of concern given that homogenization methods for daily data are often limited to adjustments on the mean of the distribution. Some correction algorithms for the distribution do exist, but these only reliably correct the first three moments and have currently only been applied to some networks and require highly correlated neighboring stations. A better understanding of the nature of daily inhomogeneities and better tools to correct them will be the main challenge for the coming years.

REFERENCES

1. Venema, V., 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. J., Lindau, R., Rasol, D., Rustemeier, E., Kolokythas, K., Marinova, T., Andresen, L., Acquaotta, F., Fratianni, S., Cheval, S., Klancar, M., Brunetti, M., Gruber, Ch., Prohom Duran, M., Likso, T., Esteban, and Brandsma, Th., Benchmarking homogenization algorithms for monthly data. *Climate of the Past*, **8**, pp. 89-115, doi: 10.5194/cp-8-89-2012 (2012).
2. Williams, C. N. jr., Menne, M. J., Thorne, P. W., Benchmarking the performance of pairwise homogenization of surface temperatures in the United States. *Journal of Geophysical Research-Atmospheres*, in press (2012).
3. Auer I., Böhm R., Jurkovic A., Lipa W., Orlik A., Potzmann R., et al., HISTALP – Historical Instrumental Climatological Surface Time Series of the Greater Alpine Region. *Int. J. Climatol.*, **27**, 17-46. doi: 10.1002/joc.1377 (2007).
4. Menne, M. J., Williams, C. N. jr., and Vose, R. S., The U.S. historical climatology network monthly temperature data, version 2. *Bull. Am. Meteorol. Soc.*, **90**, no.7, 993-1007, doi: 10.1175/2008BAMS2613.1 (2009).
5. 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. *International Journal of Climatology*, **26**, 345–381, (2006).
6. Caussinus, H. and Mestre, O., Detection and correction of artificial shifts in climate series. *Appl. Statist.*, **53**, part 3, 405-425 (2004).
7. Della-Marta, P. M., Collins, D., and Braganza, K., Updating Australia's high quality annual temperature dataset. *Austr. Meteor. Mag.*, **53**, 277-292 (2004).
8. 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. *International Journal of Climatology*, **21**, 1779–1801 (2001).
9. Trewin, B., Exposure, instrumentation, and observing practice effects on land temperature measurements. *WIREs Clim. Change*, **1**, 490–506, doi: 10.1002/wcc.46 (2010).
10. Meulen, van der, J. P., and Brandsma, T., Thermometer screen intercomparison in De Bilt (The Netherlands), part I: Understanding the weather-dependent temperature differences. *Int. J. Climatol.*, **28**, pp. 371-387 (2008).
11. 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).
12. 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, 55 p., 2003.
13. Conrad, V. and Pollak, C., *Methods in Climatology. Harvard University Press*, Cambridge, MA, 459 p., 1950.
14. 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).
15. Schreiber, T. and Schmitz, A., Improved surrogate data for nonlinearity tests. *Phys. Rev. Lett.*, **77**, 635–638 (1996).
16. Venema, V. K. C., Ament, F., and Simmer, C., A stochastic iterative amplitude adjusted Fourier transform algorithm with improved accuracy. *Nonlin. Proc. Geophys.*, **13**, no. 3, 247-363 (2006).
17. 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).
18. Domonkos, P., Adapted Caussinus-Mestre Algorithm for Networks of Temperature series (ACMANT). *Int. J. Geosci.*, **2**, 293-309, doi: 10.4236/ijg.2011.23032. (2011).
19. Caussinus, H. and Mestre, O., Detection and correction of artificial shifts in climate series. *Appl. Statist.*, **53**, part 3, 405-425 (2004).
20. Szentimrey, T., *Manual of homogenization software MASHv3.02*. Hungarian Meteorological Service, Budapest, 65 p. (2007).
21. Szentimrey, T., Development of MASH homogenization procedure for daily data. In *Proceedings of the fifth seminar for homogenization and quality control in climatological databases*. Budapest, Hungary, 2006; WCDMP-No. 71, 2008, 123-130.
22. Alexandersson, H. and Moberg, A., Homogenization of Swedish temperature data.1. Homogeneity test for linear trends. *Int. J. Climatol.*, **17**, 25-34 (1997).
23. 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**, no. 9, 2423-2444 (2008).
24. 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, 2003, ISBN: 0-7695-1877-X, 74-83.

25. Thorne, P. W., Willett, K. M., Allan, R. J., Bojinski, S., Christy, J. R., Fox, N., et al., Guiding the Creation of a Comprehensive Surface Temperature Resource for 21st Century Climate Science. *Bull. Am. Meteorol. Soc.*, Submitted, 2011.
26. Trenberth, K. E., Jones, P. D., Ambenje, P., Bojariu, R., Easterling, D., Klein Tank, A., Parker, D., Rahimzadeh, F., Renwick, J. A., Rusticucci, M., Soden, B., and Zhai, P., Observations: surface and atmospheric climate change. In: *Climate change 2007: the physical science basis. Contribution of working group I to the fourth assessment report of the Intergovernmental Panel on Climate Change*. Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor, and H.L. Miller (eds.). Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2007.