EFFICIENCIES OF HOMOGENISATION METHODS: OUR PRESENT KNOWLEDGE AND ITS LIMITATION

Peter Domonkos[1], **Victor Venema**[2] and **Olivier Mestre**[3]
[1]Center for Climate Change, Univ. Rovira i Virgili, Campus Terres de
l'Ebre, Tortosa, Spain, e-mail: peter.domonkos@urv.cat
[2]Meteorological institute of the University of Bonn, Germany
[3]Meteo France, Ecole Nationale de la Meteorologie, Toulouse, France

Abstract

In the recent years new methodological tools have been developed for measuring the efficiencies of homogenisation methods. This paper describes these new tools and presents the efficiency of some widely applied homogenisation methods focusing on the homogenisation of monthly temperature datasets. Conclusions are based on the synthesis of different testing methods. The results show that the best homogenisation methods are the PRODIGE, MASH, ACMANT and the Craddock test.

1. INTRODUCTION

A large range of homogenisation methods (HM) is available for climatologists to improve the quality of observational time series. However, selecting the best performing HM is not trivial. The core of the problem is that the efficiency depends on the properties of the time series the method is applied to. These properties of real networks are diverse, thus it is hard to provide generally valid settings for validation datasets. In addition, the practical efficiency depends on the purpose of homogenisation. For instance, the accuracy of linear trend estimation is usually more important in climatology than the accuracy of change-point estimations. Until recently, efficiency examination of homogenisation methods typically meant the computation of the detection skill in test datasets, composed of an arbitrary number of change-points and white noise. Although such examinations may serve important information for the method developers, the detection skills obtained in this way is not a relevant metric for the users of homogenised climatic data.

This study presents the latest results of the methodological development in measuring efficiencies of HMs, together with observed efficiencies for some widely used HMs. Three kinds of examinations are presented: i) Tests of detection parts only, in sets of relative time series (i.e. the difference between candidate series and reference series), ii) Blind tests of full homogenisation methods in the COST ES0601 (COST HOME) action with simulated datasets of networks, iii) The performance of ANOVA correction method with sets of detection results from various HMs. By comparing the results of these three kinds of examinations, the quality of the main components of HMs can be studied.

2. METHODS

In the first group of examinations (G1), the simulated test series are considered to be relative time series and the imaginary reference series are considered to be perfect. Consequently, there is no error due to imperfections in the reference time series used for detection and correction. Iterations or supplementary elements (e.g. metadata-use) are excluded. In this way the only source of the errors is from the detection part of HMs. In the second group of examinations (G2) the blind test results with the COST HOME benchmark dataset (hereafter: Benchmark) are evaluated. In these experiments full networks are simulated and complete HMs are tested. In the third group of examinations (G3) the performance of ANOVA correction method is shown when it is coupled with the change-point detection results from various HMs.

The properties of test datasets for G1 are very different from those of G2, while for G3 the input field does not include test series. The examined HMs also differ for G1 and G2, partly because only detection parts can be tested with G1, and partly because not all HMS tested in G1 also participated in the openly announced benchmarking exercise (G2).

2.1. Input data

In G1 the simulated series are relative time series, and the inhomogeneities in them fully belong to the candidate series. The length of time series is always 100 years, the time-resolution is annual, and the background noise is white noise. The frequency, shape, and size-distribution of inhomogeneities differ in the three versions used in this paper. The mean shift-magnitude (m) is expressed as a ratio to the standard deviation of the white noise.

- Dataset A: One change-point is included in each time series. Its timing is year = 40 or year = 60. The shift-size is constant, $m = 3$.
- Dataset B: Five change-points per 100 years are included on average, but the actual number of change-points varies in the series of the dataset. The break sizes are normally distributed with a mean of zero, $m = 3.5$.
- Dataset C: This dataset contains a rather complex structure with randomly distributed inhomogeneities (IHs) of different types (change-points, platform-shaped changes and trends) and magnitudes, possessing similar statistical properties to those of detected IHs in relative time series derived from observed temperature time series in Hungary (Domonkos, 2011a). In this dataset the number of IHs is high, and short-term platforms are particularly frequent. This dataset was derived in a way that the statistical properties of detected inhomogeneities from simulated datasets made to be similar to those from true observed temperature datasets in Hungary through an iterative development of test datasets. Due to the way of its derivation the inhomogeneity-properties of this dataset are likely the closest to the real-world properties. The mean frequency of change-points is 31.1 per 100 years and $m = 1.2$.
- In G2 the Benchmark is used. In this dataset monthly temperatures of complete observational networks are simulated. The statistical properties of the temperature data in the Benchmark (moments, spatial correlations, seasonality, low frequency fluctuations) mimic the properties of true European observational temperature networks (Venema et al., 2011). The frequency, size-distribution and seasonality of inhomogeneities were set by an expert-team of the COST Action HOME, and finally the properties are close to those of dataset B in most respects (the mean frequency is 5 per 100 years and $m = 0.8$°C).
- In G3 the input data are not test series, but the list of timings of detected change-points and outliers in the public homogenisation of the Benchmark. These data are available for the public (Venema, 2011).

2.2. Homogenisation methods

In G1 nine HMs are examined. All of them are widely used, objective methods that can be applied automatically. In this paper the abbreviations of the HMs for G1 have always three letters, thus they may deviate somewhat from their widely used acronyms. The Bayes method (Bay, Ducré-Robitaille et al., 2003), Caussinus - Mestre method (C-M, also known as

PRODIGE, Caussinus and Mestre, 2004), Easterling – Peterson method (E-P, also known as FTP, Easterling and Peterson, 1995), Multiple Analysis of Series for Homogenisation (MAS, Szentimrey, 1999), Multiple Linear Regression (MLR, Vincent, 1998), Standard Normal Homogeneity Test for shifts only (SNH, Alexandersson, 1986), Standard Normal Homogeneity Test for shifts and trends (SNT, Alexandersson and Moberg, 1997), t-test (tts, Ducré-Robitaille et al., 2003) and Wilcoxon Rank Sum test (WRS, Wilcoxon, 1945) are examined in G1.

Searching-algorithms for detecting multiple IHs, as well as parameterization for selecting significant IHs are usually the same as in the referred sources, but there are some deviations. The significance thresholds recommended by the constructors to ensure the 0.05 rate first type error in pure white noise processes are applied generally, but with some exceptions: The Caussinus – Lyazrhi criterion (Caussinus and Lyazrhi, 1997) is applied for Bay, while for MLR and tts the thresholds are based on the authors' Monte-Carlo simulations. In SNT the detected IHs are always trends when the estimated duration of change is at least 5 years, and always change-points in the reverse case. The version of MLR used in this study slightly differs from the original one, but it has no considerable effect on the detection results (not shown). Partly deviating from the original content of HMs, the cutting algorithm (Easterling and Peterson, 1995) is included in Bay, MLR, SNH, SNT and WRS.

A uniform pre-filtering of outliers is applied before the use of any HM, and the correction is also uniform, ensuring that any difference between the results is due to the different skills in the detection process.

In G2 and G3 the true contributions of the public homogenization of the Benchmark are examined. In this study only complete contributions with 15 homogenised networks from the surrogated temperature dataset are considered, except for the Craddock-test, which homogenised seven networks. The AnClim contribution was complete, but unfortunately had to be removed from the G3 examination because of discrepancies between applied and reported breaks. MASH detection has not been examined with ANOVA either, because it homogenizes every month separately (Venema et al. 2011).

From different versions of the same HM often the one with the best performance is selected only. The following HMs are analysed here: PRODIGE monthly (PROD), ACMANT late (ACMA, Domonkos, 2011b), MASH main (MASH), Craddock test by Vertacnik (Crad, Craddock, 1979), USHCN 52x (USHC, Menne and Williams, 2009), C3SNHT (SNHT), Climatol 2.1a (Clim, Guijarro, 2011), PMTred rel (PMT, Wang et al., 2007).

2.3. Efficiency measures

In this study efficiency (E) of homogenisation is usually expressed as the percentage of ceased root mean squared error (RMSE) relative to the RMSE in raw data (eq 1).

$$E = \frac{\mathrm{RMSE}_{\mathrm{raw}} - \mathrm{RMSE}_{\mathrm{hom}}}{\mathrm{RMSE}_{\mathrm{raw}}} \cdot 100\% \qquad (1)$$

In some examinations centred RMSE (CRMSE) is applied. It is the RMSE of the anomalies from the mean. Let true values (homogenisation estimations) be denoted by t (x) in n-year long time series, then the CRMSE is formulated by (2).

$$\mathrm{CRMSE} = \sqrt{\frac{1}{n}\left(x_i - t_i - \frac{1}{n}\sum_{j=1}^{n}\left(x_j - t_j\right)\right)^2} \qquad (2)$$

The advantage of CRMSE is that RMSE depends on the chosen fix point of the homogenisation, i.e. on a value or section of the time series that is considered to be free of errors, hence all the other values are adjusted to that. The drawback of CRMSE is that distortions of true spatial coherence in networks are not considered by that.

The CRMSE is chosen for evaluating E of annual value estimations in the Benchmark homogenisation (denotation: E_{CA}), because there was no fix point accepted by all contributors. The RMSE is applied for calculating the E for i) annual values in G1 (E_A), ii) trend-slopes for individual annual time series (E_T), iii) network-mean trend-slopes for the whole Benchmark-period (1900-1999, E_{TNL}), iv) network-mean trend-slopes for the second half of the Benchmark-period (1950-1999, E_{TNS}).

In the early part of Benchmark the ratio of missing data is high, therefore it has to be taken into account that the expected value of subset-means deviates from the expected value of the mean of all time series in network. Let the number of available stations be denoted by K for year j, then the subset-mean (denoted by X with upper stroke) is calculated by (3).

$$\overline{X_{j,K}} = \frac{1}{K}\sum_{k=1}^{K} x_{j,k} \tag{3}$$

When K equals the total number of stations the following notation is used:

$$\overline{X_j} = \overline{X_{j,k}} \tag{4}$$

Then the unbiased estimation of network-mean (Y) for year j is shown by (5).

$$Y_j = \overline{X_{j,k}} + \frac{1}{m}\sum_{i=1}^{m}\left(\overline{X_i} - \overline{X_{i,k}}\right) \tag{5}$$

In eq. 5 m denotes the number of years with available data in each station.

Detection skill is also applied. Let the number of right detections, that of false detections and that of all change-points be denoted by S_R, S_F and S, respectively. Then the detection skill (E_D) is calculated by (6).

$$E_D = \frac{S_R - S_F}{S} \tag{6}$$

The parameterisation of S_R, S_F and S is the same as in Domonkos (2011a).

2.4. ANOVA

The ANOVA is based on the minimisation of variance of homogenised data under the following criterions: i) The climate signal is the same for each time series, ii) the station-effect is always constant between two adjacent change-points of a time series. The minimum variance can be calculated by an equation system in which the time-dependent climate effects and the site-effects are the variables (Caussinus and Mestre, 2004).

The ANOVA provides the optimal solution of homogenisation-task when the input data meets with the written criterions. In practice, there is hardly any problem with the uniformity of climate signal, since networks are expected to be formed for a region of the same climate. The second criterion is more problematic, because the detected timings of change-points by HMs are usually neither complete nor accurate.

3. RESULTS

3.1. Tests of detection parts (G1)

Fig. 1 shows the E_A and E_T for Dataset A. Throughout this paper, the colour blue means C-M or C-M based HM, red means MAS and green means SNH or SNH based HM. It can be seen that in case of 1 change-point per time series the change-points are detected well with high certainty and with an ideally good reference series the homogenisation could be almost perfect, i.e. most efficiencies are above 90% and for E_A the values are often above 95%. E_A is always higher than E_T. E-P (tts) has slightly (markedly) poorer performance than the other HMs. The differences among the performances of other HMs are very small.

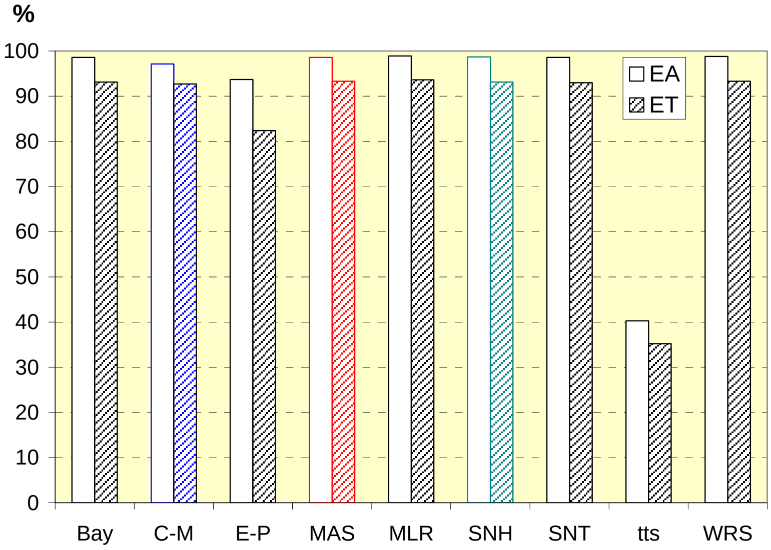

Fig. 1. Efficiencies of nine detection methods in RMSE of annual values (E_A) and individual trends (E_T) for Dataset A.

Fig. 2 presents the same kind results for Dataset B. In case of five change-points per time series on average, the homogenisation results are slightly poorer than in case of 1 change-point only, but the performances are generally still very good, the efficiencies are mostly above 90%. E_A is always higher than E_T again. Similarly as with Dataset A, E-P and tts show clearly poorer performance than the other HMs. Considering the mean of E_A and E_T, the rank order of the best HMs is C-M, Bay, MAS and SNH, but with insignificant differences in the performances (95.4%, 95.1%, 94.8% and 94.7%, respectively).

Fig. 3 presents the results for Dataset C. These results show that the presence of large number of small-size and platform-like inhomogeneities raises the uncertainty of inhomogeneity detection. The E-P and tts have very poor performances, while the other HMs produce rather similar results, around 70-75% efficiency. In this experiment E_T is sometimes

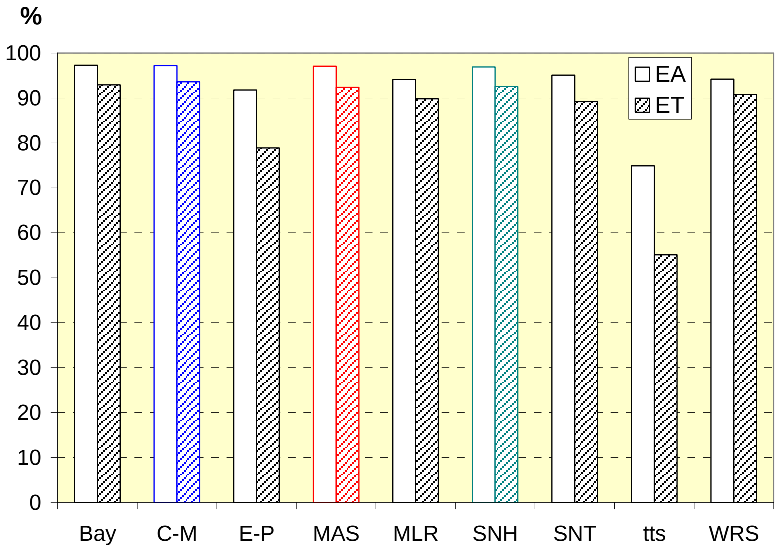

Fig. 2. The same as Fig. 1, but for Dataset B.

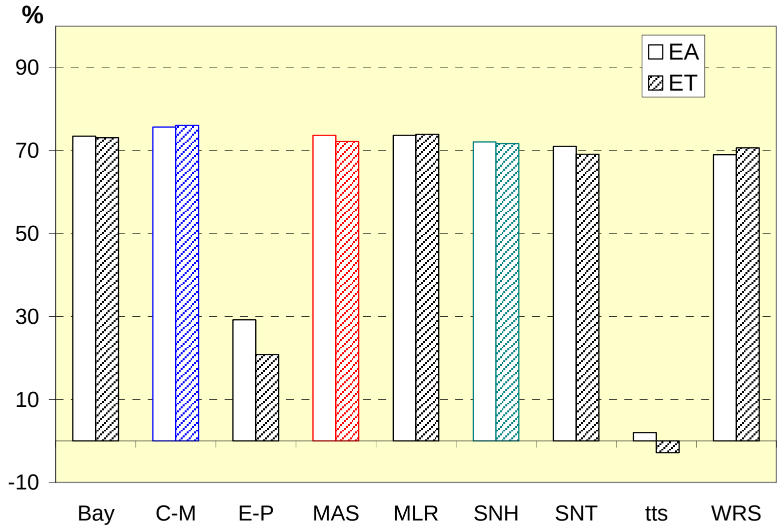

Fig. 3. The same as Fig. 1, but for Dataset C.

higher than E_A. Focusing on the rank order of the performances of the best methods, two important differences appear relative to the results of Dataset B: i) The lead of C-M is slightly larger, ii) The MLR has the second best performance here. The rank order: C-M (75.9%), MLR (73.8%), Bay (73.3%), MAS (73.0%), SNH (71.9%). For Dataset C the detection skills are also calculated (Fig. 4). Here the C-M is the best again, but otherwise the picture is quite different compared to Fig. 3. The main differences are: i) Only the MAS has comparably good performance with C-M, ii) The performance of E-P and tts is not markedly poorer than that of the other HMs, moreover, counting with E_D the E-P would be the third best HM, iii) The most popular HMs (SNH, SNT) have markedly poorer performances than the best methods.

With the joint evaluation of E_A, E_T and E_D we can prove that the traditional detection skill examinations could not give a reliable picture about the true performance of HMs. The E-P seems to be one of the best methods when only the E_D is considered, but the reconstruction of true annual means and trends is more important in climatology than the ratio

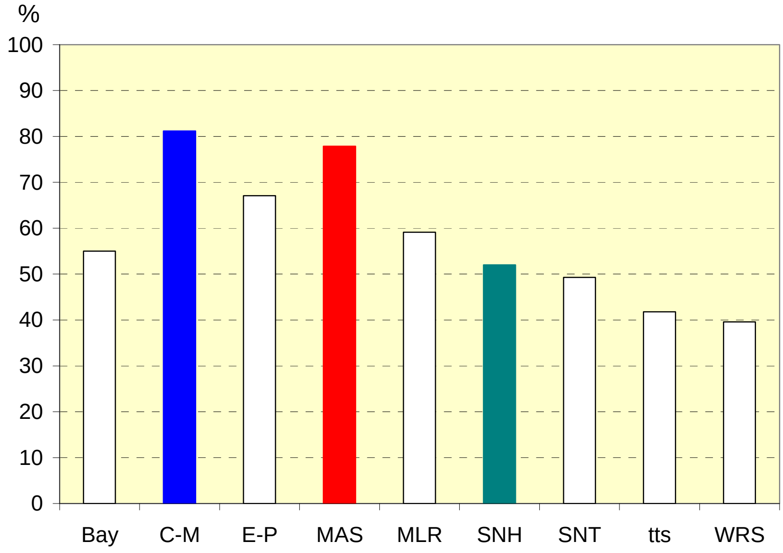

Fig. 4. Detection skills (E_D) of nine homogenisation methods for Dataset C.

of accurately detected change-points. On the other hand, when HMs of similar performances in E_A and E_T are examined, the differences in E_D is a valuable piece of information because in contrast with the simplified examinations in G1, the full homogenisation procedures often contain step-by-step elements, and for this reason the detection-errors in a certain step may affect the accuracy of estimations in later steps. All in all, the results of G1 show that C-M is the best detection method, followed by MAS. The other detection tools are less powerful, although in many cases the deficiencies from the best HMs are small.

3.2. Blind test results of the COST Action HOME (G2)

Fig. 5 presents the E_{CA} and E_T results for eight contributions which homogenised the Benchmark. In constructing this figure, HMs that are based on C-M, MAS or SNH detection are selected, as well as Crad that is included due to its outstanding performance. The main findings are as follows: i) The performances are much poorer than in

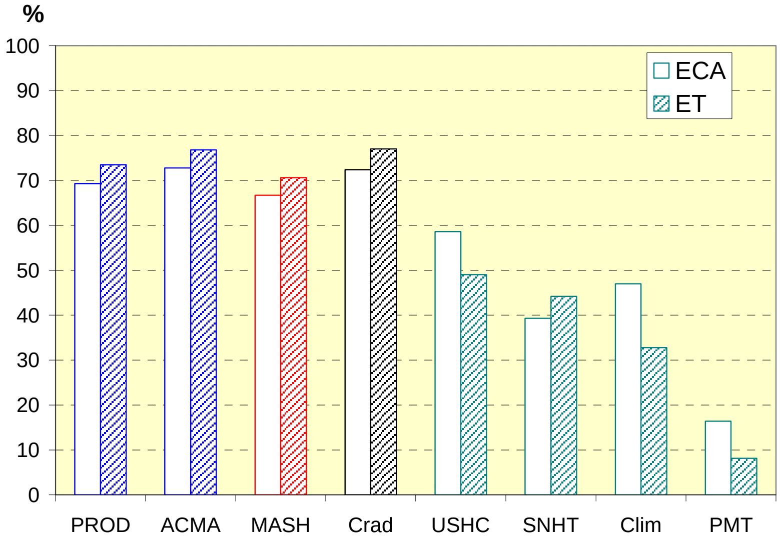

Fig. 5. Efficiencies of various homogenisation methods in the Benchmark experiments in the CRMSE of annual values (E_{CA}) and RMSE of individual trends (E_T).

the G1 experiment with Dataset B, which indicates that most homogenisation errors are not due to detection errors. ii) The efficiencies are positive what proves that the homogenised time series have better quality than before homogenisation. iii) The best four methods (including PROD and MASH) have a markedly better performance than SNH-based HMs. iv) ACMA and Crad have even better performance than PROD and MASH.

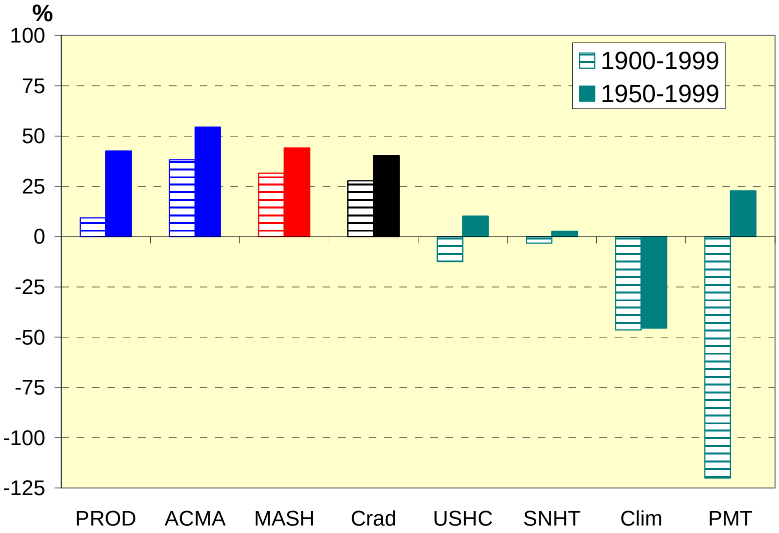

Fig. 6. The same as Fig. 5, but for the RMSE of network-mean trends between 1900-1999 (E_{NL}) and between 1950-1999 (E_{NS}).

Fig. 6 presents the performances in network-mean trend estimations. The mean bias of network mean trends in raw data is 0.39°C/100yr (0.69°C/100yr) for 1900-1999 (1950-1999). The results show that it is a difficult task to reduce these trend errors by homogenisation. Particularly E_{NL} is generally low or often even negative. Its likely explanation is that in the early section of the time series the ratio of missing data is very high. The rank order of the best HMs is somewhat different here than in Fig. 5. The best is the ACMA, which is followed by the MASH, Crad and PROD. Note that due to the small sample-size the stochastic noise is very high in these characteristics.

3.3. Unified evaluation of G1 and G2

First it has to be clarified that the possibility of reliable comparisons between the results of G1 and G2 is limited, because the full detection process of HMs is often different from the basic method. For instance, the detection part of ACMA is a modified version of C-M-detection. In addition, ACMA is applicable only for monthly data, and it is recommendable only when station-effects have monthly cycle (series of monthly temperatures). Similarly, the detection parts of USHC, Clim and PMT are not exactly identical with SNH, there are differences in the selection of significant change-points and sometimes also in other details. Crad, being subjective, has no counterpart in the G1 examinations. Note that objective versions of accumulated anomaly based detection methods, similar to that of Crad, exist (Buishand, 1982) and they have been tested in G1-like examinations (Domonkos, 2008). According to those examinations the accumulated anomaly based detection methods have similar performance to that of WRS, which means that its deficiency from the best methods is moderate. Note also that the detection part of PMT was also tested (Domonkos, 2011a) and its performance was almost exactly the same as that of the SNH.

One interesting finding is that for G2, the performance of PROD and MASH relative to the SNH-based HMs is much stronger than in G1 results. Its explanation is the consequent mathematical structures in the whole homogenisation processes of PROD and MASH. While in G1 results the differences between performances are often very small, in G2 examinations only two HMs have comparable or better results than PROD and MASH. These two methods are the ACMA and Crad. ACMA is an improved version of PROD, and the results show that the improvement has been successful. The good performance of Crad has a very different explanation. The Crad incorporates the unique features of human intelligence that are, according to our results, difficult to convert perfectly into automatic or semi-automatic HMs.

3.4. ANOVA with various detection results

This section examines the performance of ANOVA correction method, when it is applied to the detection results of various HMs that originally did not apply ANOVA for correction. The results are presented in Fig. 7 and 8. Fig. 7a (7b) shows the impact of ANOVA on E_{CA} (E_T). It

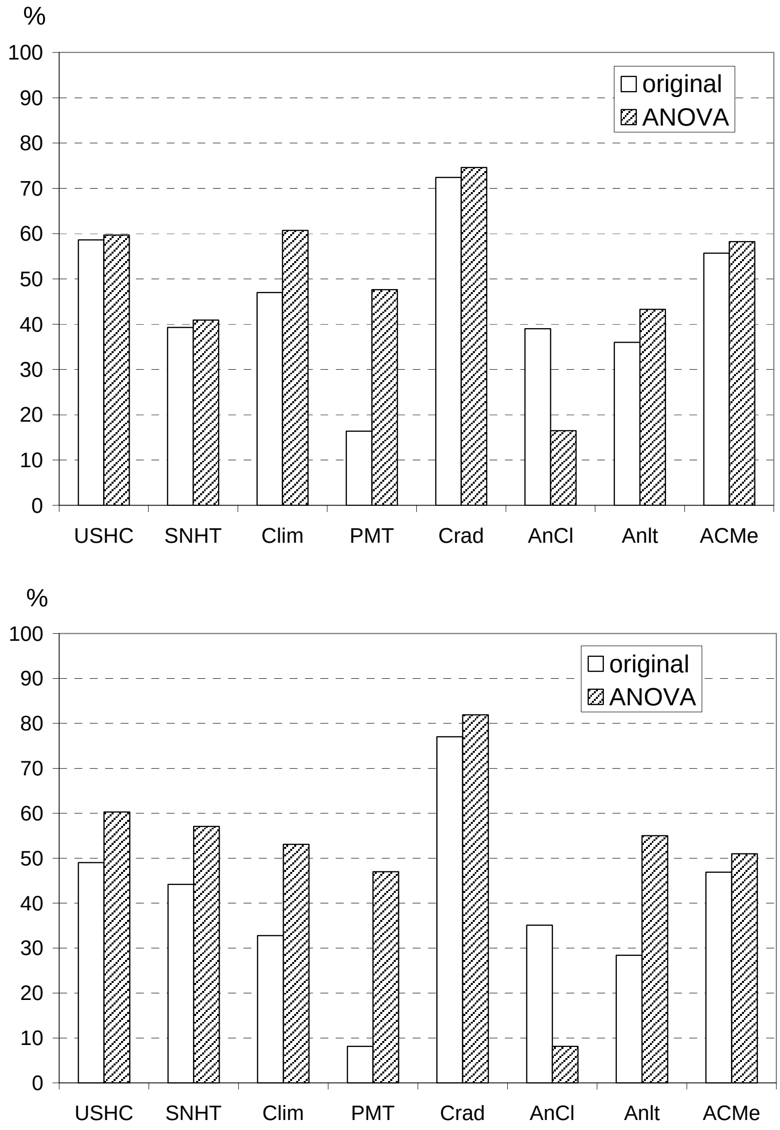

Fig. 7. Efficiencies of various homogenisation methods in the Benchmark experiments with and without ANOVA application. a) (upper) CRMSE of annual values (E_{CA}) b) (bottom) RMSE of individual trends (E_T).

can be seen that the application of ANOVA generally raises the efficiencies. The improvement is bigger in E_T than in E_{CA}. Interestingly, the ANOVA is even beneficially for the already very accurate contribution Crad, and this combination shows the highest efficiencies (except for E_{NS}, see later) for all the HMs examined in this study.

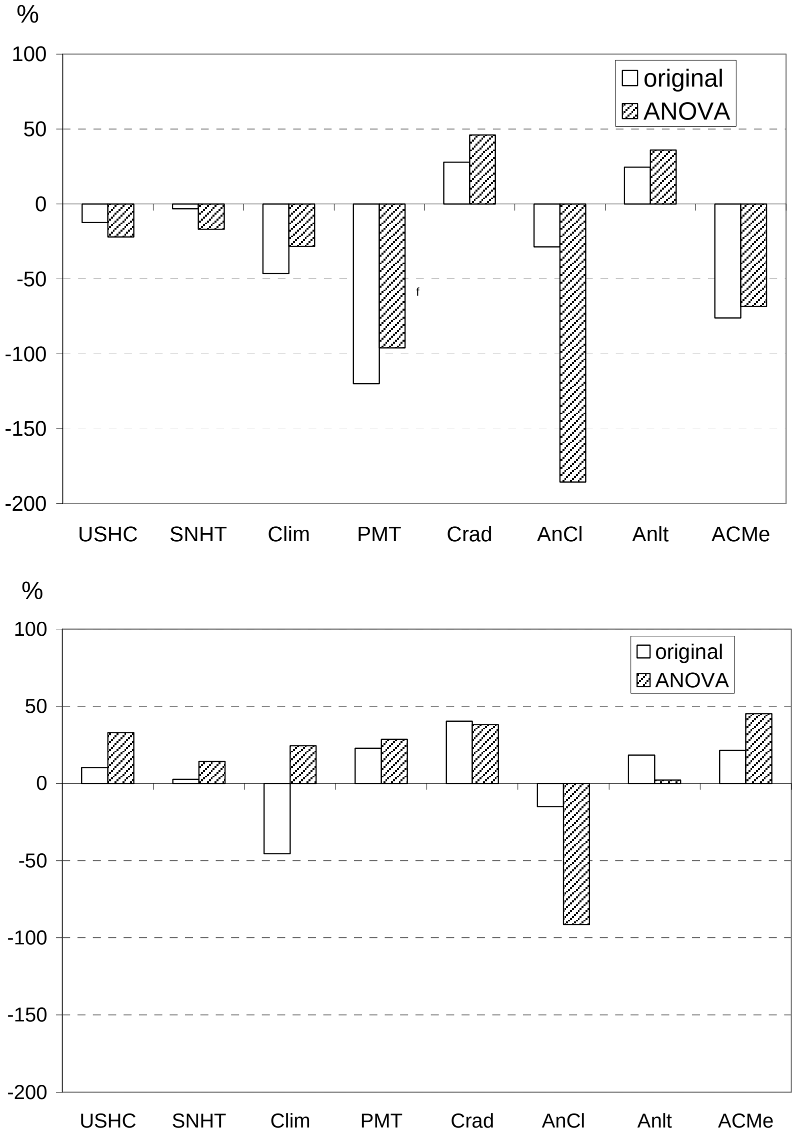

Fig. 8. The same as fig. 7, but for RMSE of network-mean trends a) (upper) 1900-1999 (E_{NL}), b) (bottom) 1950-1999 (E_{NS})

The general picture for network-mean trends is very different than for annual CRMSE and individual trends (Fig. 8), i.e. the effect of ANOVA is sometimes negative, particularly for E_{NL}. Note that the E_{NL} values are often negative both without ANOVA and together with that. These results tend to show that when original HMs yield negative efficiencies, the ANOVA might not be able to do improvement, perhaps due to the detection errors on which the network-mean trends are more sensitive than the annual CRMSE and individual trends.

4. DISCUSSION AND CONCLUSIONS

One important finding of this study is that apart from some poor detection methods the main source of the inaccuracy in homogenised time series is often not the detection part of HMs. In a complex homogenisation procedure, the time series comparison, the treatment of missing data and outliers, and the calculation of adjustment factors all together influence the final performance. For long the central question in homogenisation was the search for more accurate detection methods, while relatively little attention was paid to the influence of other parts of HMs. These results suggest that simple detection examinations, such as G1 are not suitable to reveal the true properties of complete homogenisation methods.

Another notable finding is that efficiencies are more often negative than it was thought earlier (Peterson et al. 1998, Auer et al. 2005, etc.). That overestimation was based on the fact that in the homogenisation results of real datasets the common bias of trends in networks does not appear at all, it can be seen only when the evaluation is made on simulated data with exactly known properties of artificially set inhomogeneities (Venema et al., 2011). On the other hand we note that in the true world the use of metadata may substantially help in achieving effective homogenisation results, while metadata was not simulated for the Benchmark homogenisation.

The reconstruction of network-mean trends, that is a crucially important task in climatology, seems to be more sensitive to the weak points of HMs. In the examination of the Benchmark homogenisation the network-mean trends for 1900-1999 often have larger errors after homogenisation than in the raw data. Note that the lengths of time series in the Benchmark are different, and only 3 time series start from 1900 in each network, while the other time series start later, and the low number of comparable time series significantly increases the uncertainty of the homogenisation results. One might think that to reconstruct the network-mean trends with these conditions is an unfairly hard task, but in fact, climatologists often have to face with similar problems, so the experience of negative efficiencies should be taken seriously. Other characteristics than the E_{NL} rarely show negative efficiencies in this study, and for the best HMs (ACMA, Crad, PROD, MASH) all the efficiencies are positive. However, one has to take into account at this point that real datasets could be less favourably for achieving good performances than the Benchmark.

The difference between Dataset B and Dataset C indicates that the Benchmark is likely easier for HMs than the true datasets (see more discussions about this in Domonkos, 2011a). Even if the Benchmark represents well a true class of real datasets, there are surely many others for which the homogenisation task is an even bigger challenge. For keeping low the chance of negative efficiency and its effect on climate variability analyses we have two main recommendations. First, moderately effective but popular HMs should be replaced with HMs of the best performance in climatological studies. Second, more benchmarking studies are needed, and ones with particular attention to the search of threshold conditions until which the statistical homogenisation is beneficial would be essentially useful.

Not correcting some detected breaks may well sometimes lead to more accurate data. To explain more the last idea, we mention that an indication of inhomogeneity e.g. by a homogeneity test does not mean necessarily that homogenisation-adjustments will provide more reliable and more accurate data in comparison with the raw data. Note that this idea is not new, and in the USHCN it has already been applied (Menne and Williams, 2009).

The examinations of ANOVA with the detection results of various HMs show that the ANOVA often produces higher efficiencies than the original results of the examined HMs. To understand the success of ANOVA it has to be noted that in ANOVA all inhomogeneity-caused biases and their mutual influences are treated together in one equation system. By contrast, when the correction-factors are computed for every break individually, errors, for instance due to undetected IH, may accumulate.

A tentative conclusion of G3 experiments could be that the ANOVA should be included in HMs that have not included ANOVA until now. However, some examples show that the performance of ANOVA is poor when the quality of detection results is insufficient. Therefore our final conclusion is that HMs with effective detection method and correction method together should be applied in the homogenisation of climatic time series. The SNHT and its modern versions (PMT, Climatol, USHCN) may function quite well in certain cases, but both their detection parts and correction parts are poorer than that of the best HMs. The same also refers to other HMs that are based on the single change-point detection and corrections to one or more reference series or reference sections. According to our present knowledge the best HMs are the ACMANT, Craddock, Craddock + ANOVA, PRODIGE and MASH. Note that the performance of Craddock is strongly user-dependent, and this HM cannot be tested in large datasets. Our conclusions refer primarily to the homogenisation of temperature datasets.

Acknowledgements: The research was supported by the European projects COST ES0601, EURO4M FP7-SPACE-2009-1 /242093 and by the Spanish project "Tourism and Environment" ECO 2010-18158.

REFERENCES

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

Alexandersson, H. and Moberg, A. 1997: Homogenization of Swedish temperature data. Part I: Homogeneity test for linear trends. Int. J. Climatol. 17, 25-34.

Auer, I., Böhm, R. and 23 coauthors, 2005: A new instrumental precipitation dataset for the greater Alpine region for the period 1800-2002. Int. J. Climatol. 25, 139-166. DOI: 10.1002/joc.1135.

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

Caussinus, H, and Lyazrhi, F. 1997: Choosing a linear model with a random number of change-points and outliers. Ann. Inst. Statist. Math. 49/4, 761-775.

Caussinus, H. and Mestre, O. 2004: Detection and correction of artificial shifts in climate series. J. Roy. Stat. Soc. Series C 53, 405-425.

Craddock, J.M. 1979: Methods of comparing annual rainfall records for climatic purposes. Weather 34: 332-346.

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

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

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

Ducré-Robitaille, J-F., Vincent, L.A. and Boulet, G. 2003: Comparison of techniques for detection of discontinuities in temperature series. Int. J. Climatol. 23, 1087-1101. DOI: 10.1002/joc.924.

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

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

Menne, M.J. and Williams Jr, C.N. 2009: Homogenization of temperature series via pairwise comparisons. J. Climate, 22, 1700-1717. DOI: 10.1175/2008JCLI2263.1.

Peterson, T.C., Easterling, D.E. and 19 co-authors, 1998: Homogeneity adjustments of in situ atmospheric climate data: a review. Int. J. Climatol. 18, 1493-1517.

Štěpánek, P., Zahradniéek, P. and Skalák, P. 2009: 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.

Szentimrey, T. 1999: Multiple Analysis of Series for Homogenization (MASH). Second Seminar for Homogenization of Surface Climatological Data. WCDMP 41, WMO-TD 962, WMO, Geneva, 27-46.

Venema, V., Mestre, O. and the COST HOME team, 2012: Benchmarking monthly homogenization algorithms. Climate of the Past, in press.

Venema, V. 2011: ftp://ftp.meteo.uni-bonn.de/pub/victor/costhome/homogenized_monthly_benchmark.

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

Wang, X. L., Wen, Q.H. and Wu, Y. 2007: Penalized maximal *t* test for detecting undocumented mean change in climate data series. J. Appl. Meteor. Climatol. 46 (No. 6), 916-931. DOI:10.1175/JAM2504.1

Wilcoxon, F. 1945: Individual comparisons by ranking methods. Biometrics Bull., 1, 80-83.