COMPARISON OF HOMOGENIZATION PACKAGES APPLIED TO MONTHLY SERIES OF TEMPERATURE AND PRECIPITATION: THE MULTITEST PROJECT

José A. Guijarro[1], José A. López[1], Enric Aguilar[2], Peter Domonkos, Victor K.C. Venema[3], Javier Sigró[2] and Manola Brunet[2]

[1]State Meteorological Agency (AEMET), Balearic Islands Office, Spain
<jguijarrop@aemet.es>
[2]Universitat Rovira i Virgili, Tarragona, Spain
[3]Meteorological Institute, University of Bonn, Germany

1. INTRODUCTION

It is well known that observational climatic series are exposed to unwanted alterations due to changes in the observational practices, instrumentation, relocations or changes in the surroundings of the stations. Many methods have been proposed to remove these perturbations from the series and leave the climate signal only, and the successful COST Action ES0601 "HOME" provided interesting inter-comparison results to understand the strengths and weaknesses of many of them (*Venema et al*., 2012).

However, some methods, implemented in computer packages, have been upgraded to new versions (in part as a result of the fruitful discussions maintained along the COST Action), and therefore new inter-comparisons are needed to evaluate their performance. Yet repeating such an effort, which involved the work of dozens of researchers along five years, is not foreseen in a near future. Hence, the only practical alternative is to implement a benchmarking system to test the homogenization packages in a completely automatic way. The drawbacks are that only packages able to be run in this mode can be tested, with default parameters in those that can be tuned to different climatic variables, and that the added value of manual homogenization is lost.. But on the other side, the methods can be tested on a high number of networks with varying characteristics, which enhance the knowledge about its applicability to different climatic zones.

The results of a preliminary automatic comparison on synthetic monthly temperature series can still be found at http://www.climatol.eu/DARE/testhomog.html, but the Spanish project MULTITEST (Multiple verification of automatic software homogenizing monthly temperature and precipitation series) aims at updating and improving those benchmarking experiments in various ways:

- More realistic temperature networks.
- Inclusion of precipitation networks with different climatic characteristics (Temperate, Mediterranean and Monsoonal).
- More realistic inhomogeneities.
- Comparison of more homogenization packages.

The details of the benchmarking implementation, the packages tested and the results obtained so far are explained in the following sections, ending with some conclusions and prospects of future work.

2. BENCHMARKING METHODOLOGY

2.1. General procedure

Several "master" networks were generated consisting in 100 series containing 720 monthly values (equivalent to 60 years of data). For each of them, for each of several different inhomogeneous settings (experiments), and for each chosen homogenization package, 100 tests were done by:

- Randomly sampling a subset of 10 series (true solution). (Some supplemental tests were done with 20, 40 and 80 series.)
- Inserting inhomogeneities into them (problem series).
- Homogenizing them with a backward adjustment (results).
- Comparing the results with the true solutions, computing Root Mean Squared Errors (RMSE), trend differences, and other metrics.

2.2. Homogeneous synthetic series

Six homogeneous master networks were generated, three for temperatures and three for precipitations. The temperature networks were built in the following way:

- Random locations were assigned to 100 points in a 4x3° lon-lat geographic domain.
- Mean monthly homogenized temperatures from Valladolid (Duero basin, Spain) were assigned to the first point, located near the center of the domain.
- Its closest point was assigned the same series plus white noise from a normal distribution with mean = 0 and standard deviation = 1.5, multiplied by a constant C. The rests of the series were computed with the same procedure, in order of minimum distance to any other already assigned series.
- Three different constant coefficients were used: C = 0.18, 0.30 and 0.65, yielding three master networks with decreasing correlation between stations, which were called Tm1, Tm2 and Tm3 (*Figure 1*, left column).
- All series were shifted to account for simulated elevation, a 2 °C/100yr trend was added, and their annual oscillation were varied in ±20%.

The three monthly master precipitation networks were generated taking as models real precipitation series from three different climatic regions from which, after their homogenization by Climatol 3.0, derive variograms, gamma coefficients and frequencies of zeros, which were used to compute synthetic series by means of the R package "gstat", preserving the spatial correlation structure. The names assigned to this networks, type of simulated climates and data used to model them were:

- PEir (Atlantic temperate): 198 Irish precipitations series (1941-2010).
- PMca (Mediterranean): 107 Majorcan precipitation series (1951-2015).
- PInd (Monsoonal): 64 SW India series from 0.5° resolution gridded monthly precipitations from the Global Precipitation Climate Center (Schneider, 2015).

Figure 1 (right column) shows the cross-correlations of these monthly data, computed on the first differences of the series.

2.3. Added inhomogeneities

In a first stage, inhomogeneities were applied to the synthetic homogeneous series. We used five different settings with increasing difficulty and realism:

- i) Big shifts in half of the series.
- ii) The same with a strong seasonality.
- iii) Short term platforms and local trends.
- iv) Random number of shifts with random size and location in all series.
- v) The same plus seasonality of random amplitude.

For setting 'i', the shifts have a size of 2°C; series 1 to 3 have fixed position and signs ("-,-", "+,+" and "-,+" respectively); series 4 and 5 have only one shift, but with random sign and position. For the case with seasonal cycle (ii) the size of the shifts is 1.5 or 2°C and the seasonal cycle is a sinusoidal function with an amplitude of 2°C. The size of the platforms and linear gradients used for setting 'iii' are 2°C, with random lengths within certain limits to avoid overlapping. The number of shifts in settings 'iv' and 'v' was taken from a Poisson distribution with a mean of 5 every 100 years. The shifts were applied as deviations from the baseline, additive for temperature, with an amplitude drawn from the standard normal distribution N(0, 1), and multiplicative for precipitation, with factors taken from N(1, 0.2) without any seasonal perturbation (only setting 'iv' was applied to this variable). Seasonality amplitudes were taken from N(0, 0.7). In all cases, the last 10 years are always kept untouched. Sample examples of these five settings can be seen in *Figure 2*.

2.4. Tested homogenization packages

Most commonly used homogenization programs that could be run in completely automatic mode were tested, namely:

- Climatol 3.0 (*Guijarro*, 2016), with constant and variable corrections.
- ACMANT 3.0 (*Domonkos*, 2015), versions for temperature and precipitation (sinusoidal and irregular seasonalities).
- MASH 3.03 (*Szentimrey*, 2007), with constant corrections.
- RHtestsV4 (*Wang & Feng*, 2013), absolute and relative (average series were given as references), with or without quantile adjustment.
- USHCN v52d (*Menne & Williams*, 2005), which makes constant corrections.
- HOMER 2.6 (*Mestre et al.,* 2013), with different iteration strategies.

Tests were run on a Linux PC by means of bash scripts. USHCN was compiled on a Linux computer and could be run natively. Climatol, RHtestsV4 and HOMER are implemented in R (R Core Team, 2015), and hence could also be easily run, although HOMER could not be automated by a simple redirection of the input from a bash script and the utility "expect" needed to be used to provide automatic responses to the questions of the program.

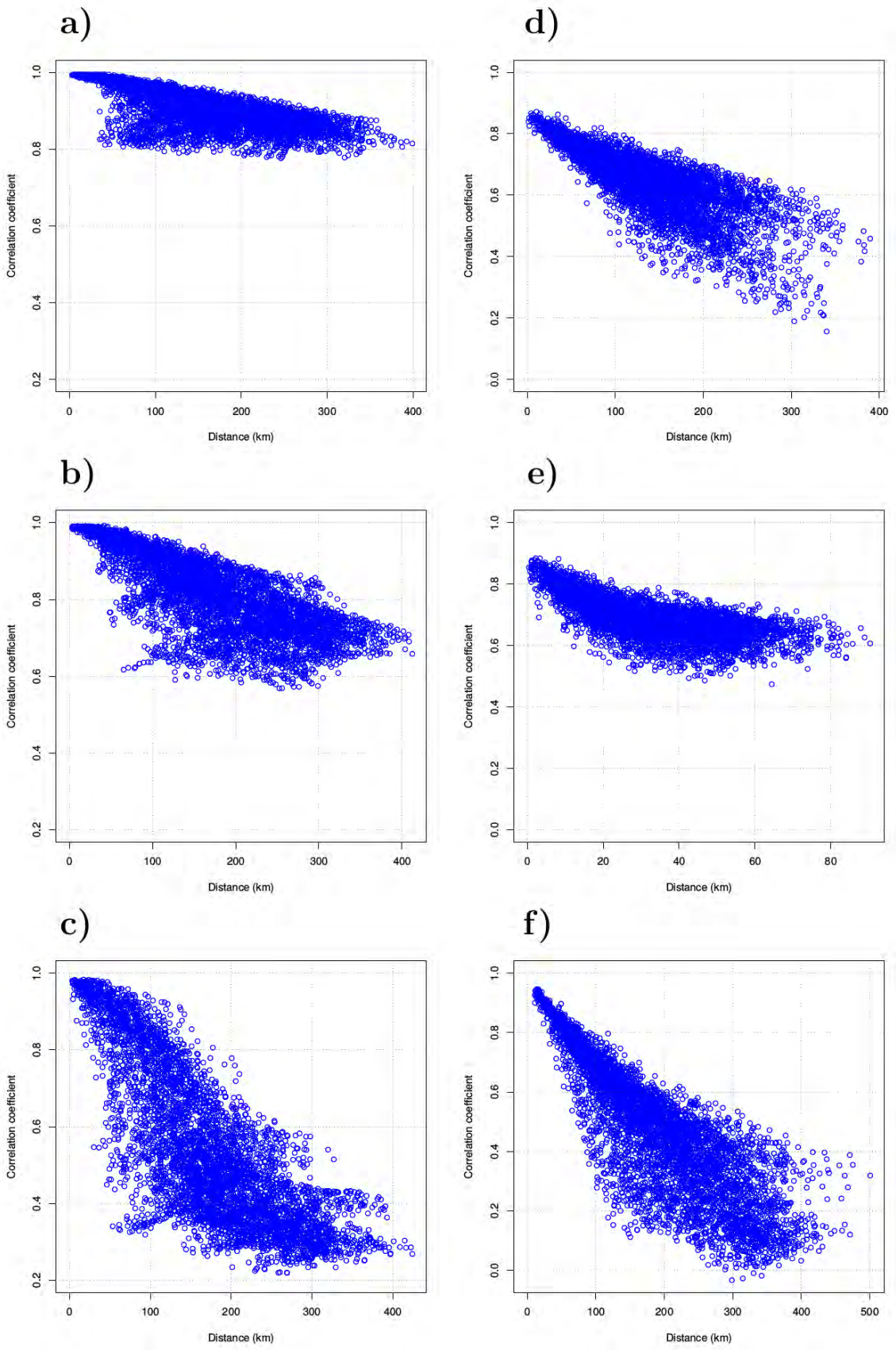

Fig. 1. Correlograms of the first differences of the master networks. Temperatures in the left column: a) Tm1, b) Tm2 and c) Tm3. Precipitations in the right column: d) PEir, e) Pmca, f) Pind. (Vertical axis ranges from 0.2 to 1.0 in the left column and from 0.0 to 1.0 in the right one).

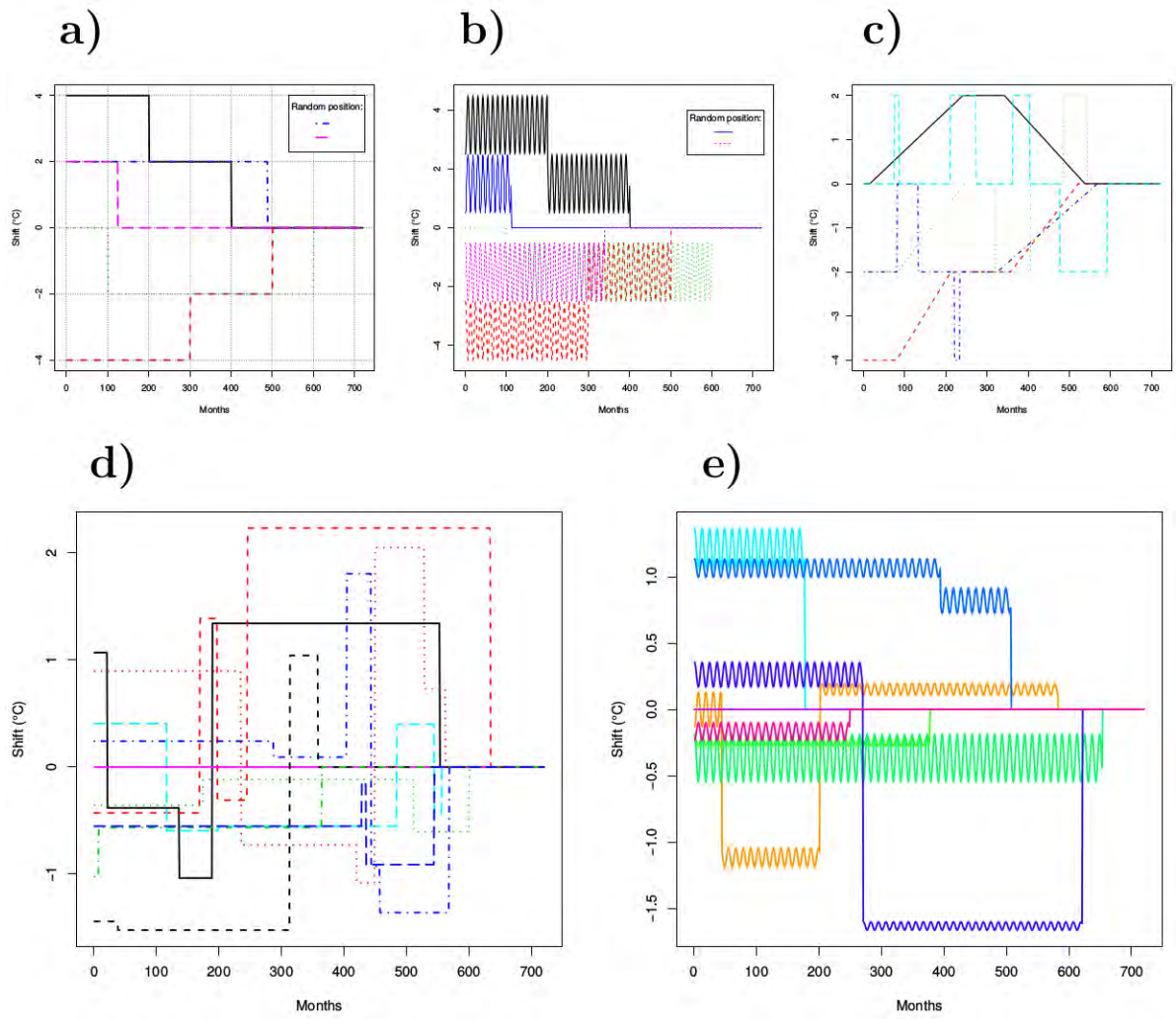

Fig. 2. Examples of inhomogeneities with increasing complexity introduced in 10 series sample networks: a) Small number of big constant shifts in five of the series. b) The same, but with a strong seasonal variation in the shifts. c) Local trends and short term "platforms". d) Random number of shifts with random constant amplitude can affect all 10 series. e) The same, but with a seasonal variation in the shifts of random amplitude. The last 10 years of the series are kept homogeneous in settings a) to c), and only the 5 last years in d) and e), to allow a reliable backward reconstruction of the homogenized solutions.

ACMANT and MASH are Windows executables, but could be run on Linux through "wine" (a Linux implementation of the Windows Application Interface). This was straightforward in the case of ACMANT, but MASH automatic procedures are implemented in DOS batch scripts that gave errors of incorrect file specification in wine when "*." were copied or moved between directories, so these DOS batch scripts had to be translated into Linux bash versions.

Whenever a method stop with an error condition (sometimes simply stating that the problem series was homogeneous, sometimes due to a limitation of the software), the problem series was taken as the solution provided by the tested method. This procedure allowed the unsupervised run of hundreds of tests in a continuous flow, but in some experiments HOMER gave more serious errors that aborted the process, giving incomplete solutions.

3. RESULTS AND DISCUSSION

The performance of the different methods in each experiment is evaluated mainly by looking at the Root Mean Squared Errors (RMSE) of the solutions returned by the different software packages, and also by comparing their trends with those of the original series. Box-plot figures allow an easy comparison of the performances of the methods between them and with respect to the problems (inhomogeneous networks) they had to solve.

3.1. Temperatures

Figure 3 shows the RMSE box-plots of the methods for the five different settings displayed in *Figure 2*, using the network Tm2 (intermediate level of cross-correlations). Horizontal axis labels are: "Inh" (Inhomogeneous, problem series), "cl1" (Climatol with constant correction, "Cl1" (Climatol with variable correction), "A3i" (ACMANT with irregular seasonality), "A3s" (ACMANT with sinusoidal seasonality), "MSH" (MASH), "RHa" and "RHA" (RHtestsV4 in absolute mode, without and with quantile matching correction), "RHr" and "RHR" (RHtestsV4 with reference series, without and with quantile matching correction), "US1" (USHCN), "Hoa", "Hob" and "Hom" (HOMER, with different iterative approaches).

It is clear that absolute homogenization without strong metadata support should be avoided. (An exception will be discussed later on). All other relative homogenization methods provide results clearly better than the inhomogeneous ("Inh") problem, although with a different performance degree. In particular, comparison of results from experiments "a" (where RHtestsV4 gives the best results) and "b" show a clear improvement in the methods that are able to correct a seasonally variable bias, the lowest RMSE being in this case achieved by ACMANT and HOMER. As to experiment "c" (short term platforms and local trends), ACMANT is also ranking the best, closely followed by MASH, Climatol and RHtestsV4 without quantile adjustment.

The lower row in *Figure 3* displays the results for the more realistic experiments. When the random inhomogeneities do not have a seasonal cycle ("d"), the lowest mean RMSE correspond to ACMANT sinusoidal, Climatol (with varying correction ranking better than the constant correction version!) and USHCN, followed not too far away by the other relative methods. And when more realism is added by imposing a sinusoidal seasonal cycle of the inhomogeneities ("e"), ACMANT (sinusoidal and irregular) is still ranking the best, followed by Climatol with variable correction and, more distantly, HOMER, USHCN, MASH and RHtestsV4 with quantile adjustment.

Figure 4 shows these last more realistic results in the middle left box-plots (4b), which can now be compared with the same results when the sample problems are drawn from the better (Tm1) and worse (Tm2) correlated master series (4a and 4c respectively). The performance of the methods decay with a decreasing level of cross-correlation in the networks as could be expected, but the ranking is quite constant in 4a and 4b. When the correlations between stations are worse (4c), ACMANT is still the method producing the lower RMSE, while the other methods give more similar results. It is worth mentioning that in this latter case the quantile adjustment worsens the performance of RHtestsV4, and that, even in this lower correlated scenario, all relative methods return series more homogeneous than the problem networks.

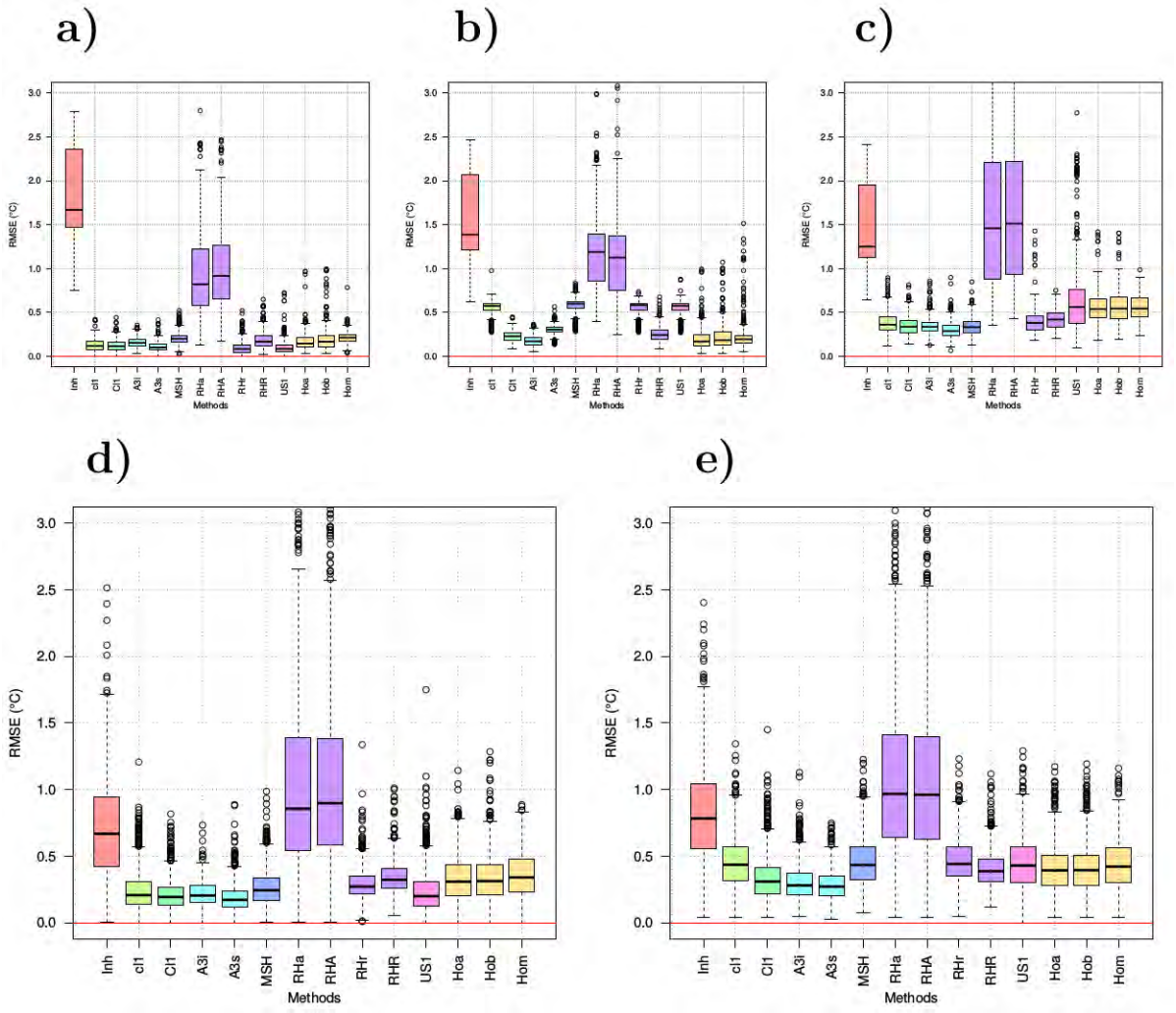

Fig. 3. RMSE of the solutions of the tested software packages for the five types of experiments displayed in Figure 2 using the master network with an intermediate level of cross-correlations (Tm2). Each box of the upper row contains results from 2000 series, and potentially up to 4000 series in the lower row, but in this case an undetermined number of homogeneous series have been excluded from the evaluation. (RMSE axis has been set constant from 0 to 3, so some outliers may lay out the figures).

The right column of *Figure 4* (d, e, f) show the box-plots of the trends errors, whose dispersions are lower than the original and unbiased, except for the absolute homogenization and for the HOMER results. These have also a lower trend dispersion than the problem series, but are negatively biased in all three temperature networks. The lower dispersion of trend errors is achieved by ACMANT, followed by USHCN, MASH and Climatol in the best correlated network (4d). ACMANT is still showing the best results with moderate correlations (4e), but in the worse correlated networks (4f) the dispersions are very similar between the methods. RHtestsV4 show dispersions very similar irrespectively of the correlation degree, while USHCN is the less robust, changing from one of the lower dispersions in 4d to one of the worse in 4f.

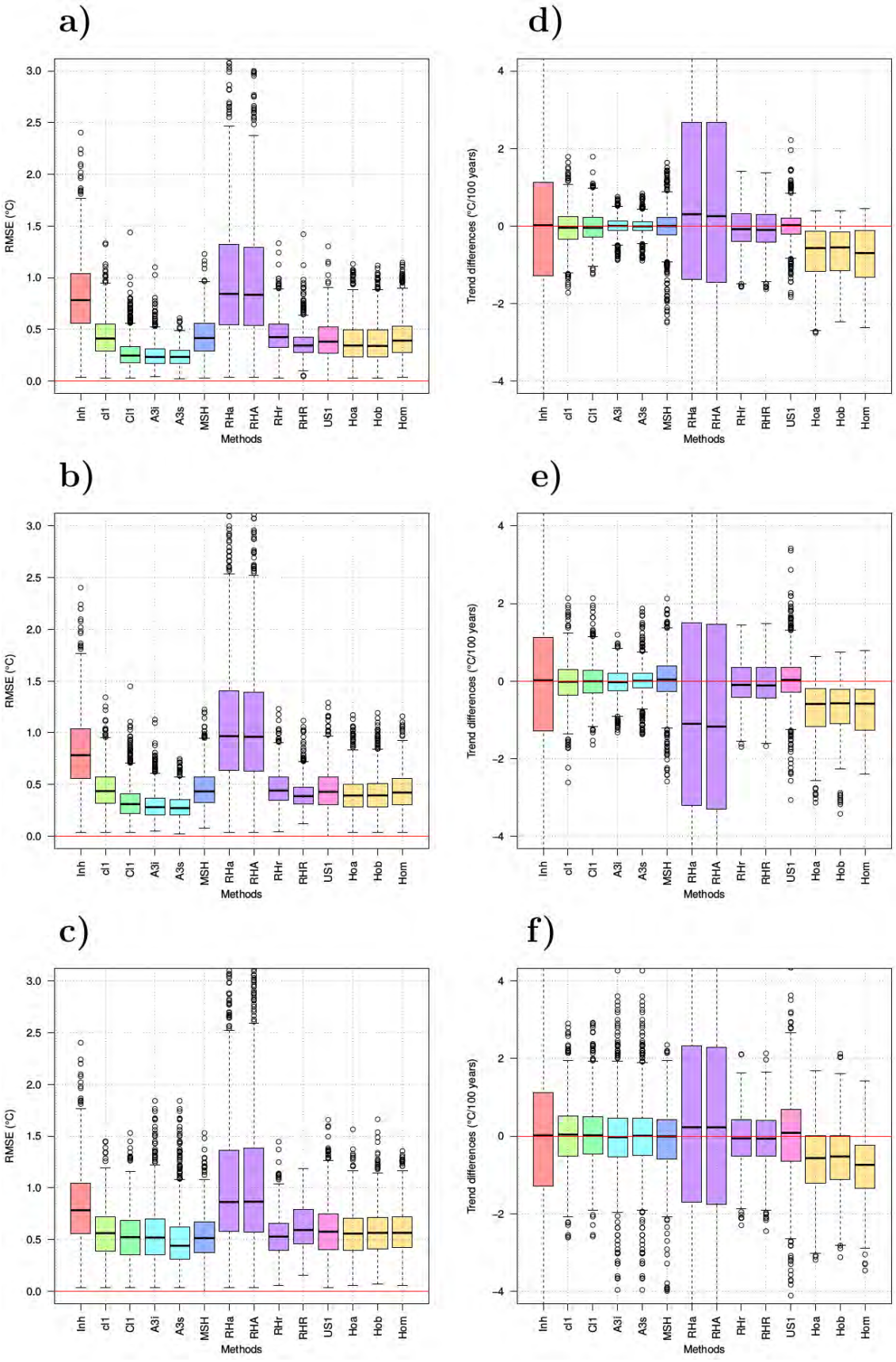

Fig. 4. The left column show the RMSE of the solutions provided by the methods for experiment v (the most realistic) in networks with decreasing level of correlations, a (Tm1), b (Tm2) and c (Tm3). Also for these three master networks, trend errors are shown in the right column (d, e, f).

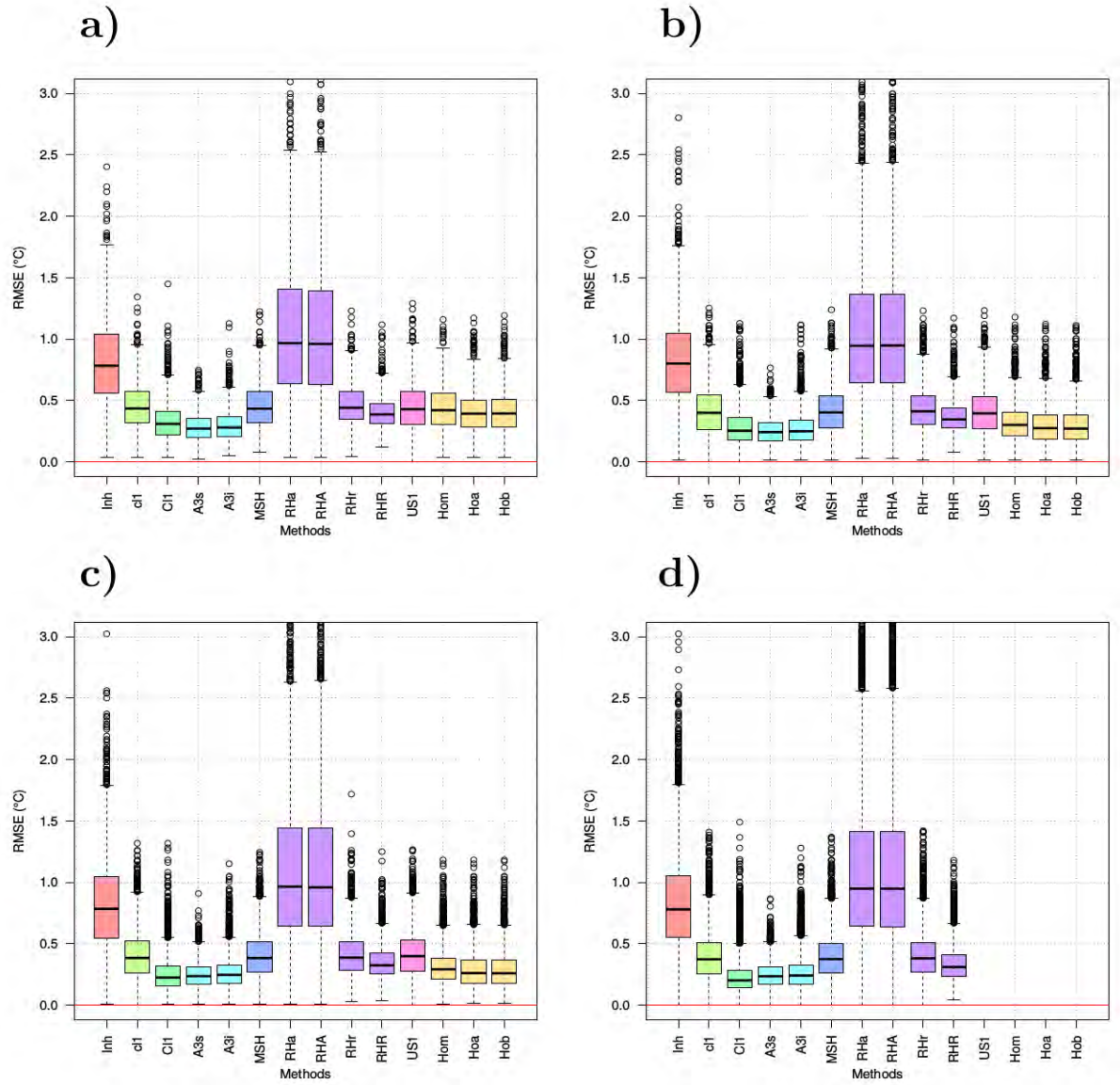

Fig. 5. RMSE of the methods increasing the number of series in each problem sample from 10 (a) to 20 (b), 40 (c) and 80 (c). (In this latter case we could not test USHCN because we had it compiled with a limit of 40 series, and HOMER aborted with errors and produced no results.)

The effect of increasing the number of series in the problem was tested on the master network with intermediate cross-correlations (Tm2) with the more realistic settings (v: random number of shifts of random amplitude with random seasonality) by increasing the sample size from 10 to 20, 40 and 80 series, although in this latter case no results could be obtained from HOMER due to unresolved errors, and USHCN could not be tested because our compilation was done with a limit of 40 series and could not be repeated because the only computer that allowed a successful compilation is no longer available. The box-plots of the RMSE obtained with this increasing number of series can be seen in *Figure 5*.

Results seem to improve when more series are available, especially in HOMER (although limited to the 40 series sample size). The lower RMSE are achieved in all cases by both versions of ACMANT and by Climatol with variable correction (the default); in this order when samples contain 10 or 20 series, and in the reverse order when 40 or 80 series are involved. The errors in the trends of the series also improve with an increased sample size (*Figure 6*), and in the case of HOMER, the worrying bias found in the 10 series samples vanishes as the number of series increases to 20 and 40.

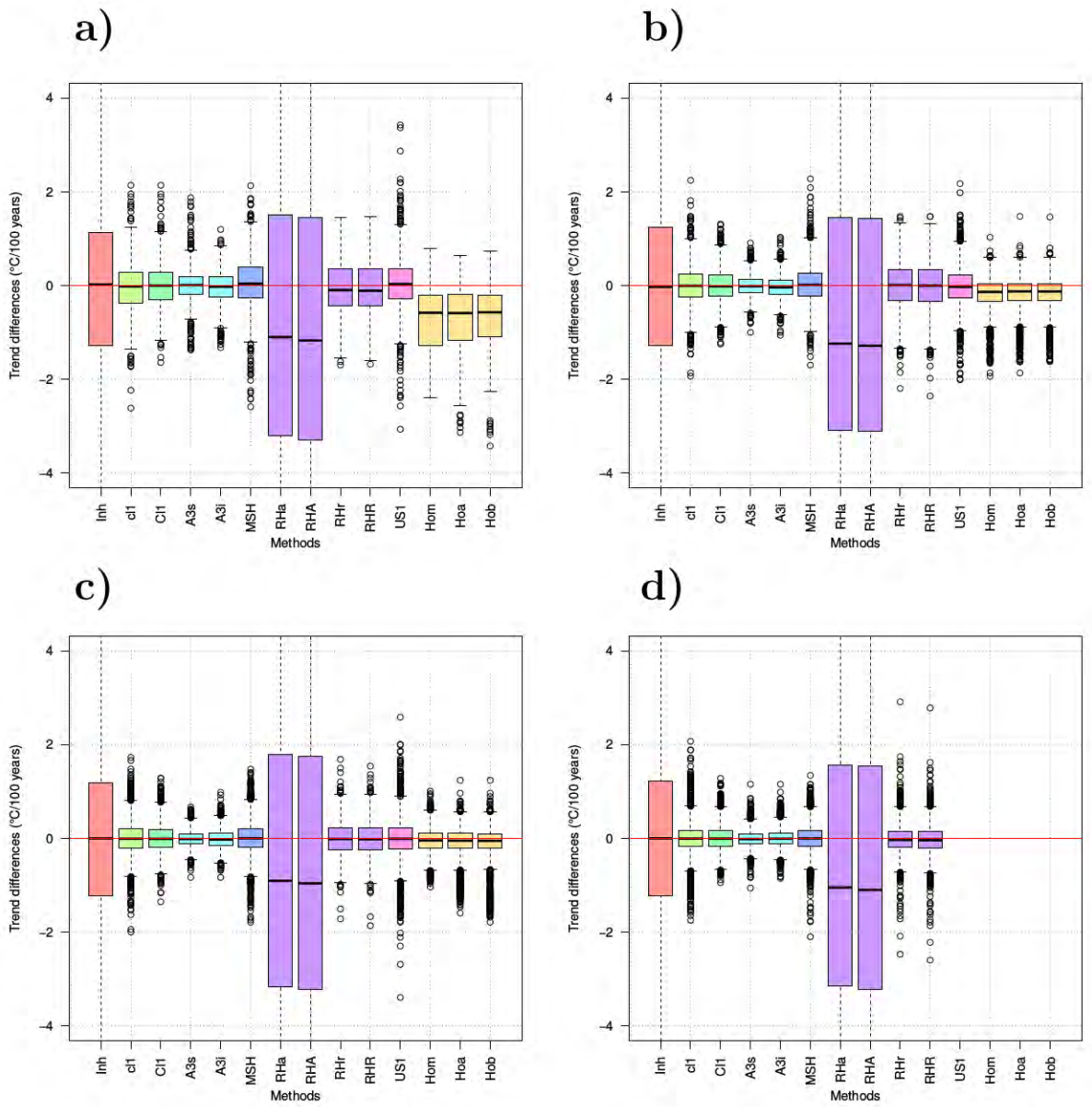

Fig. 6. Trend errors of the methods obtained with an increasing number of series in each problem sample from 10 (a) to 20 (b), 40 (c) and 80 (d). (USHCN and HOMER could not be run in the latter case.)

Another aspect worth to be tested is the performance of the methods when there is a simultaneous shift in a many of the series. It is expected that relative homogenization methods will fail to detect such inhomogeneities when they affect most or all of the series, since those changes will be attributed to real climate variability. Therefore, an additional very simple experiment was performed by introducing a simultaneous shift of 2°C in the middle of the series in 4, 7 and all 10 series of the problem samples. *Figure 7* shows the RMSE and trend errors when 40 and 70% of the series are affected (HOMER did not return any results due to unresolved errors). In the first case (upper row of the figure) all tested methods improved the problem series, although RHtestsV4 results are much worse than the others (*Figure 7a*). The higher reduction in RMSE correspond to USHCN, probably due to its pairwise detection strategy, followed closely by ACMANT. These methods also gave the best unbiased trend corrections (*Figure 7b*), although some outliers appear in the ACMANT trends. On the contrary, *Figures 7c* and *7d* show that when a majority of the series suffer a simultaneous shift (70% in this case), only RHtestsV4 makes a significant reduction on RMSE and, to a

lesser extent, in trend errors. (In this case the absolute homogenization over-corrects the trends and introduces an important negative bias.) When all series in the samples present this big simultaneous shift, results (not shown) are similar to those with 70% affected, except that RHtestsV4 relative is also unable to reduce the errors (as expected), and in this case quantile adjustments introduce huge errors. The big magnitude of the introduced simultaneous shifts has allowed absolute homogenization to give relative good results, but real simultaneous changes in the conditions of observation are not expected to produce such big shifts.

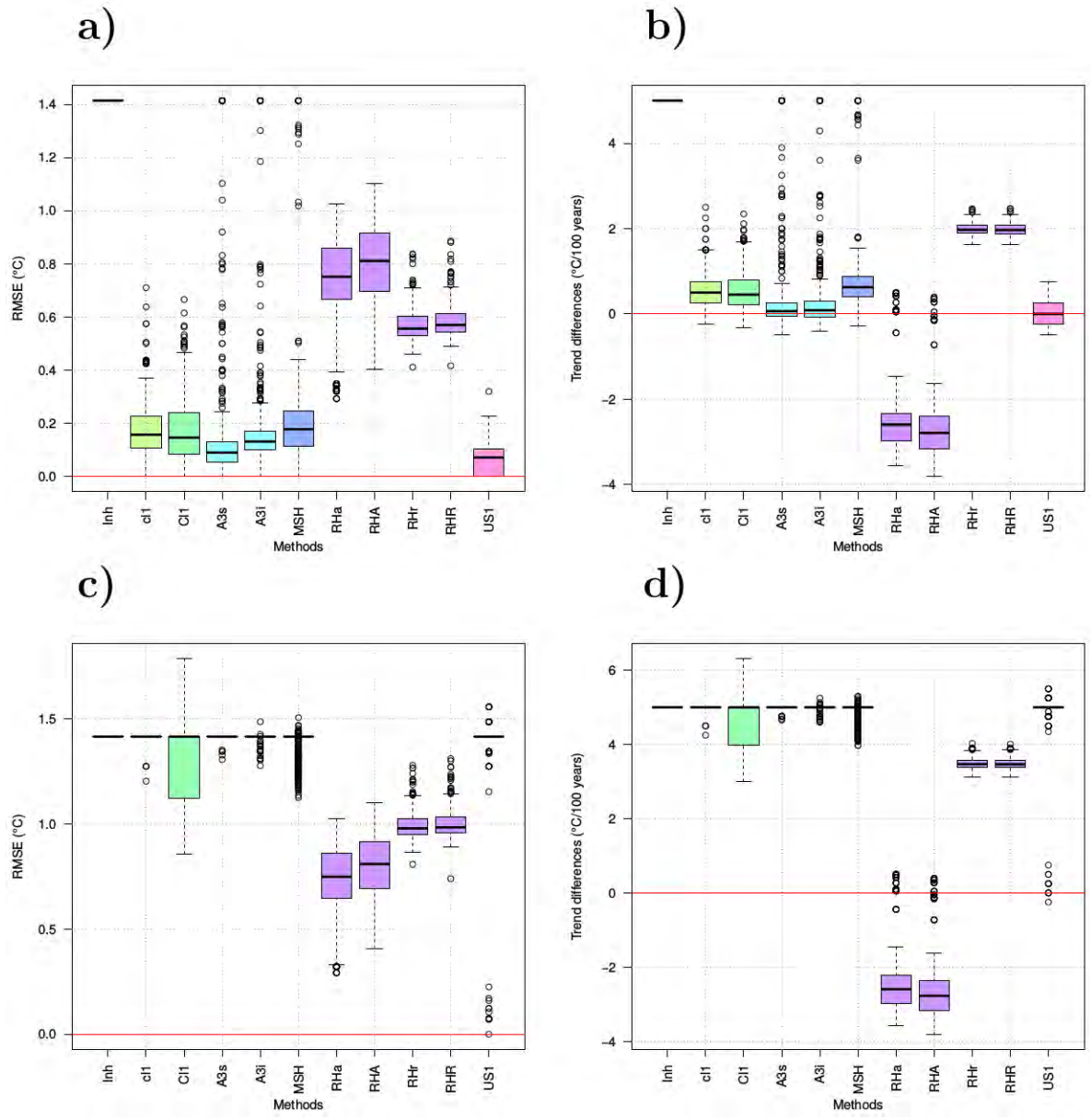

Fig. 7. Boxplots of RMSE (a and c) and trend errors (b and d) when a simultaneous shift of 2°C is introduced in the middle of the series in 40% (upper row) and 70% (lower row) of them. (HOMER gave errors in this experiment and did not produce any results).

3.2. Precipitations

Precipitation tests were performed only with random N(1, 0.2) factors affecting a random number of series at random locations (excluding the last five years of the series) on 10 series samples drawn from the three master networks PEir (Atlantic temperate), PMca (Mediterranean) and PInd (monsoonal). *Figure 8* displays box-plots of RMSE (left column) and trend errors (right column) showing that the homogenization of precipitation series is much more problematic than those of temperature, which can be attributed to their much higher variability, both in time and space, the presence of zeros and its biased probability distribution. There is a general reduction of RMSE in the Atlantic temperate precipitation case (*Figure 8a*), but much less in the Mediterranean (*8b*) and monsoonal (*8c*) climates, in which some methods even gave greater errors than the problem series (especially RHtestsV4 relative with quantile adjustment).

Without quantile adjustment, RHtestsV4 relative performed quite well in the Atlantic and Mediterranean climates, with error reductions similar to HOMER and Climatol, and not too far from ACMANP (the precipitation version of ACMANT), which gave the best results in all three climates. MASH results improve in the most difficult monsoonal climate (*Figure 7c*), where it performs as well as HOMER.

Trend errors (*Figures 8d, e, f*) are also reduced after the homogenization, except for the absolute mode. Biases are small in general, although the greatest (positive) deviations appear in the "easier" Atlantic precipitations with RHtestsV4 and HOMER. This latter method produced negative biases in all three Tm1, Tm2 and Tm3 temperature master networks (*Figures 4d, e, f*), but in this case biases are positive, less noticeable in the Mediterranean precipitation and inexistent in the monsoonal climate case.

Climatol was tested with three different settings: ratio normalization (cl1) and full standardization of cubic root (Cr1) or log(x+1) (Cl1) transformed data. These two last methods produced much worse results that the simpler first setting, probably due to an amplification of errors when undoing the transformation of the data. Therefore, it seems better to disregard the use of the transformation options included in this package, although many other packages use transformations when dealing with multiplicative model variables.

The poor performance found in these precipitation tests does not dismiss applying homogenization procedures to precipitation series, since the best methods did improve the problem series, and shifts greater than those drawn form a N(1, 0.2) may appear in real cases.

4. SUMMARY AND FUTURE WORK

Thousands of tests have been performed by applying six of the most used homogenization computer packages to hundreds of sample networks of monthly temperatures and precipitations with different characteristics and affected by a variety of prescribed inhomogeneities. Results have allowed a comparison of the performance of the tested methods under different circumstances, although only an automated procedure could be applied. Therefore, the packages were run with default parameterizations, and better results could perhaps be obtained by tuning them to every kind of network. Moreover, some methods were not devised to be operated automatically, but were included in this intercomparison project because of their high number of users. That was the case with RHtestsV4, which homogenizes the series one by one and it is the user who must provide a suitable homogeneous reference. Also HOMER is intended to be applied by experienced users who must take subjective decisions.

The simple unrealistic experiments help in detecting the strengths and weaknesses of the methods, while more realistic problem networks give results more useful to evaluate what can be expected when applied to real problems. The ranking of the methods is relatively consistent, but changes appear between different networks and types of inhomogeneities, hence the importance of showing results from networks that can be representative of different real climates and varying station densities. Overall results indicate that ACMANT and Climatol produced the most reliable results in this study, with MASH and USHCN following with also remarkable efficiency metrics, but user preferences may be driven by other characteristics of the software packages, as presented in comparative tables at http://www.climatol.eu/tt-hom/index.html. Moreover, as no significance testing of the differences between methods has been done, the rankings are informative and do not account for the uncertainty associated to the benchmarking process itself.

Automatic testing of homogenization computer packages is the only feasible way of updating the evaluation of the performance of new methods or new versions of existing packages, as far as they can be run in unattended mode. These intercomparisons will be valuable not only for users, but also to the developers of the tested packages, who can see how their algorithms behave under varied climate conditions.

Work under progress include tests with seasonalities other than sinusoidal, and with shifts concentrated over a short period for a high proportion of series. Also the unexpected behavior of RHtestsV4 and HOMER trend biases in some experiments must be investigated to understand their causes and discard possible flaws in the testing scripts.

Finally, tests will be performed on a longer and more realistic benchmark, with varying number of missing data along time, similar to that used in the COST Action ES0601, and biased inhomogeneities producing abnormal trends in the networks will also be considered.

At the end of the project, all results and scripts will be accessible in a web page to allow reproducibility. This benchmarking infrastructure will remain operative after the end of the project, allowing future updates of intercomparison exercises, both of the currently tested packages and any other willing to be added. The most straightforward way to widen the number of tested methods would be the involvement of the developers by providing suitable automatic scripts to apply their software to the problem networks. This way would not only facilitate the time consuming task of studying their way of operation, input and output file formats and other requirements, but their operation procedures would be directly designed by the developers, minimizing the possibilities of applying them incorrectly.

Acknowledgments

Project MULTITEST (CGL2014-52901-P) is funded by the Spanish Ministry of Economy and Competitiveness. Manola Brunet is also supported by the European Union-funded project "Uncertainties in Ensembles of Regional Reanalyses" (UERRA, FP7-SPACE-2013-1 project number 607193). Victor Venema is also supported by the DFG project Daily HUME (VE 366 - 8).

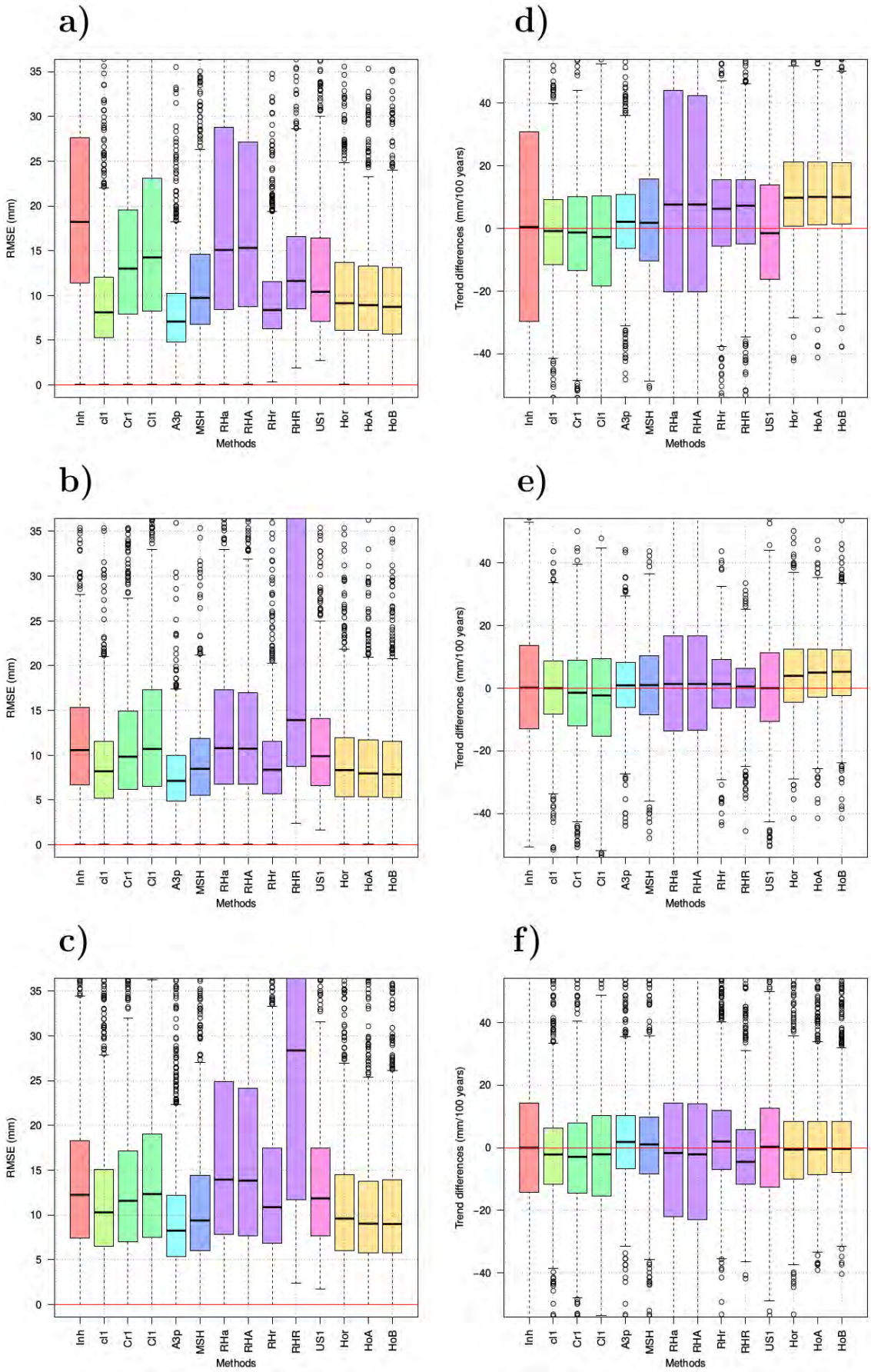

Fig. 8. RMSE (a, b, c) and trend errors (d, e, f) of the methods applied to Atlantic temperate (upper row), Mediterranean (middle row) and monsoonal (bottom row) simulated precipitation series.

Our thanks go to Met Éireann for for providing the Irish monthly precipitation series that served as model to synthesize the network of Atlantic Temperate precipitations. Mallorca monthly precipitations were taken from AEMET data bases, and monthly precipitations from SW India, gridded at 0.5° resolution, were obtained from the Global Precipitation Climate Center (GPCC).

References

Domonkos P (2015): Homogenization of precipitation time series with ACMANT. *Theor. Appl. Climatol.*, 122:303-314.

Guijarro JA (2016): Package 'climatol'. https://cran.r-project.org/web/packages/climatol/climatol.pdf (Accessed in June 2017).

López JA, Guijarro JA, Aguilar E, Domonkos P, Brunet M (2016): Una propuesta metodológica para la generación de redes de precipitación simuladas a partir de redes de precipitación observadas en el marco del proyecto MULTITEST. In Olcina J, Rico AM, Moltó E (eds.): "Clima, sociedad, riesgos y ordenación del territorio", Universidad de Alicante (Spain), *Asociación Española de Climatología*, ISBN 978-84-16724-19-2, pp. 183-194.

Menne MJ, Williams CN Jr (2005): Detection of undocumented changepoints using multiple test statistics and composite reference series. *J. Climate* 18:4271-4286.

Mestre O, Domonkos P, Picard F, Auer I, Robin S, Lebarbier E, Böhm R, Aguilar E, Guijarro J, Vertachnik G, Klancar M, Dubuisson B, Stepanek P (2013): HOMER: a homogenization software - methods and applications. *Időjárás*, 117:47-67.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/ (Accessed in June 2017).

Schneider U; Becker A, Finger P, Meyer-Christoffer A, Rudolf B, Ziese M (2015): GPCC Full Data Reanalysis Version 7.0 at 0.5°: Monthly Land-Surface Precipitation from Rain-Gauges built on GTS-based and Historic Data. DOI: 10.5676/DWD_GPCC/FD_M_V7_050.

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

Venema V, Mestre O, Aguilar E, Auer I, Guijarro JA, Domonkos P, Vertacnik G, Szentimrey T, Stepanek P, Zahradnicek P, Viarre J, Müller-Westermeier G, Lakatos M, Williams CN, 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 (2012): Benchmarking homogenization algorithms for monthly data. *Clim. Past*, 8:89-115.

Wang XL, Feng Y (2013): RHtestsV4 User Manual. http://etccdi.pacificclimate.org/software.shtml, 29 pp. (Accessed in June 2017).