INTERNATIONAL JOURNAL OF CLIMATOLOGY
Int. J. Climatol. **34**: 3528–3545 (2014)
Published online 18 February 2014 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/joc.3925

RMetS
Royal Meteorological Society

An automatized homogenization procedure via pairwise comparisons with application to Argentinean temperature series

Alexis Hannart,[a]* Olivier Mestre[b] and Philippe Naveau[c]
[a] *IFAECI, CNRS/CONICET/UBA, Buenos Aires, Argentina*
[b] *MÉTÉO FRANCE, Toulouse, France*
[c] *LSCE, CNRS/CEA, Saclay, France*

ABSTRACT: We describe a fully automatized homogenization procedure and illustrate it on Argentinean weather station temperature series. The procedure relies on multiple pairwise comparisons between a candidate station and its surrounding stations. The main advantage of this approach is to get around the difficulty of defining a reliable reference series; its main drawback is to often require visual attribution and grouping of shifts resulting in too high a cost in human time for implementation on large datasets. Here, we fully automatize these two steps by using a probabilistic metric of similarity between shifts which is leveraged within two optimized clustering schemes. Simulation results show performance improvements *versus* both visual inspection and the automatized procedure of Menne MJ, Williams CN, Jr. 2009. Homogenization of temperature series via pairwise comparisons. *J. Clim*. 22: 1700–1717. Implementation on Argentinean temperature series results in the identification and removal of numerous inhomogeneities; corrected series reveal stronger and spatially smoother warming trends.

KEY WORDS homogenization of climatic time series; statistical methods; temperature record

Received 7 February 2013; Revised 19 December 2013; Accepted 20 December 2013

1. Introduction

Long instrumental climatic records are often affected by artificial discontinuities due to changes in measurement conditions, for instance instrument modernization or maintenance, station relocation, modification of station surroundings, change in measurement practices among others. Because they have similar magnitude, these artificial shifts can wrongly modify the analysis of natural climate variations (Abarca-Del-Rio and Mestre, 2006). Removing these shifts is thus widely considered an essential step preliminary to climatic series analysis. This pre-processing is referred to as homogenization and may rely on metadata documenting changes when available. Otherwise, to cope with the common situation of undocumented changes, statistical homogenization procedures have been developed to detect and remove inhomogeneities, for a detailed review the reader is referred to Beaulieu *et al.* (2007).

Statistical procedures usually rely on the well-verified *relative homogeneity principle* (Conrad and Pollack, 1962), i.e. on the assumption that neighbouring series exhibit the same climate variations except when artificial changes perturb observations of one of the series. The statistical model of Equation (1) describes the relative homogeneity principle quite well. This model assumes that each series x_i of observation at a given station i is the sum of a time-varying climate effect s_t which is constant in space across the entire neighbourhood, a station instrumental effect $\delta_{i,t}$ which is piecewise constant in time between two consecutive breakpoints, and a Gaussian white noise $\epsilon_{i,t}$ which is independent in both space and time and represent the station-specific component of climatic variability. Denoting x the set of series to be homogenized, we have

$$x_{i,t} = s_t + \delta_{i,t} + \epsilon_{i,t} \quad (1)$$

where indice $i = 1,..,p$ corresponds to the station and $t = 1,...,n$ to the time step. In this framework, homogenization essentially consists in the statistical inference and removal of the piecewise constant instrumental effect $\delta_{i,t}$, which is equivalent to the estimation of the number, position and amplitude of breakpoints in each station i of the network. As the combination of those characteristics (number and position of breakpoints) quickly explodes when p and n increase, the straightforward resolution of this model faces a so-called combinatorial wall, which makes it technically infeasible. For this reason, homogenization procedures attempt first to remove the regional climate signal s_t from the series and second to detect breakpoints, thus making the inference lighter. So far, there are two strategies to remove the regional climate

* Correspondence to: A. Hannart, IFAECI, UBA/FCEN, Pab. II 2° Piso, Ciudad Universitaria, (1428) Buenos Aires, Argentina. E-mail: alexis.hannart@cima.fcen.uba.ar

signal s_t. Initial procedures (Easterling and Peterson, 1995; Moberg and Alexandersson, 1997) typically propose to estimate the climatic signal by means of a regional reference series obtained by averaging all series, e.g.

$$\widehat{s_t} = \sum_{i=1}^{p} \alpha_i . x_{i,t} = s_t + \sum_{i=1}^{p} \alpha_i . \delta_{i,t} + \sum_{i=1}^{p} \alpha_i . \epsilon_{i,t} \quad (2)$$

and the regional series $\widehat{s_t}$ is subtracted from the candidate series before detection. However, the non-homogeneity of such a regional reference series which is introduced by the term $\sum_{i=1}^{p} \alpha_i . \delta_{i,t}$ in Equation (2) can be considered a drawback. To circumvent this issue, more recent procedures (Alexandrov *et al.*, 2004; Caussinus and Mestre, 2004; Drogue *et al.*, 2005; Kuglitsch *et al.*, 2009; Menne and Williams, 2009) eliminate component s_t without estimating it in the first place, simply by computing series of pairwise differences, e.g.

$$x_{ij,t} = x_{i,t} - x_{j,t} = \delta_{ij,t} + \epsilon_{ij,t} \quad (3)$$

where i and j are referred to as the candidate series and the reference series, $\delta_{ij,t} = \delta_{i,t} - \delta_{j,t}$ and $\epsilon_{ij,t} = \epsilon_{i,t} - \epsilon_{j,t}$, respectively. Pairwise difference series x_{ij} can now be viewed as the sum of a compounded instrumental effect δ_{ij} which cumulates breakpoints from both series, and a noise $\epsilon_{ij,t}$ which is also Gaussian iid. Thus, difference series x_{ij} match assumptions commonly used in methods described in the abundant literature on the topic of the detection of multiple change-points in the mean of a Gaussian series (Caussinus and Lyazrhi, 1997; Lavielle and Lebarbier, 2001; Davis *et al.*, 2006; Fearnhead, 2006; Reeves *et al.*, 2007; Hannart and Naveau, 2009, and references therein for a review of most usual methods in climatology). The combined instrumental effect δ_{ij} can then be estimated from x_{ij} by mean of one such detection method.

Such an approach efficiently mitigates the non-homogeneity issue associated to regional reference series, but it raises several new issues. First, the set of series to be treated for breakpoints detection is larger; second, each breakpoint detected on a pairwise difference series may be caused by any of the two series and must hence be attributed to the so-called culprit series; third, multiple breakpoint locations and signs estimated on several difference series must be grouped and reconciled into a unique break to be used for adjusting the candidate series. This results in two additional steps referred to as attribution and grouping in this article. Most descriptions of pairwise procedures found in the literature rely on manual processing to complete these two steps based on a general approach first described in Caussinus and Mestre (2004), hereinafter referred to as CM04. This method relies on the visual inspection of a two-dimensional (2D) graph displaying the position of the breakpoints detected (horizontally) *versus* the difference series (vertically). Attribution and grouping are then performed by visually identifying alignments in the 2D graph. Indeed, if a breakpoint detected in a given difference series would belong to the candidate series, it should be detected in all, or at least several, other difference series, which becomes apparent in the graph via an alignment. By contrast, if a breakpoint would belong to the reference series, it should be detected in none, or at most few, other difference series and the graph should show no alignment. This results in a time-consuming and tedious process, especially for large datasets. Also, there are no objectivized rules for this manual review which remain arbitrary to a large extent, making the expertise of the manual reviewer critical. These two issues are important limitations of the pairwise difference approach.

To our knowledge, only one study (Menne and Williams, 2009, hereinafter referred to as MW09) has attempted to address those limitations so far. In this study, authors propose to formalize and automatize the attribution and grouping steps. The attribution step is based on the comparison of a straightforward counting of detected breaks by station and date between the candidate and the neighbour series, and the grouping step is based on overlapping confidence intervals on the dates of breaks attributed to the same candidate series. While this study provides a valuable contribution to the aforementioned issues, this automatized approach may be further improved by proposing further optimized decision rules.

The main objective of this study is to further the methodological research line initiated by MW09. We develop an alternative to their initial procedure which aims at improving the performance of the automatized attribution and grouping steps based on more advanced, optimized metrics. To achieve this goal, our strategy consists in defining a metric that measures more adequately the degree of similarity between two breakpoints obtained from two pairwise difference series, and to use this metric in a more formally optimized way. Our metric intends to take advantage of available Bayesian methods of change-point detection by leveraging the additional information provided by posterior distributions of breakpoints position τ, hereinafter denoted $p(\tau)$. This approach enables to deal explicitly with the uncertainty which affects point estimates more commonly used in existing metrics. As in MW09, attribution of detected breakpoints to a culprit series is based on counting, but based on an optimized criterion derived from our metric. Then, grouping of attributed breakpoints is based on a clustering procedure in which this metric plays a key role both for generating clusters and for determining their number. The first and last steps of our homogenization procedure, detection and correction, are based on existing procedures and do not make any use of this metric. These steps are recalled briefly in the text for the sake of self-containedness.

The remainder of this paper is organized as follows. Section 2 describes the procedure extensively and Section 3 evaluates the gain in performance based on simulated series. Section 4 describes an application of the procedure to series of temperatures in Argentina, presents results for 67 stations and analyses most obvious implications on

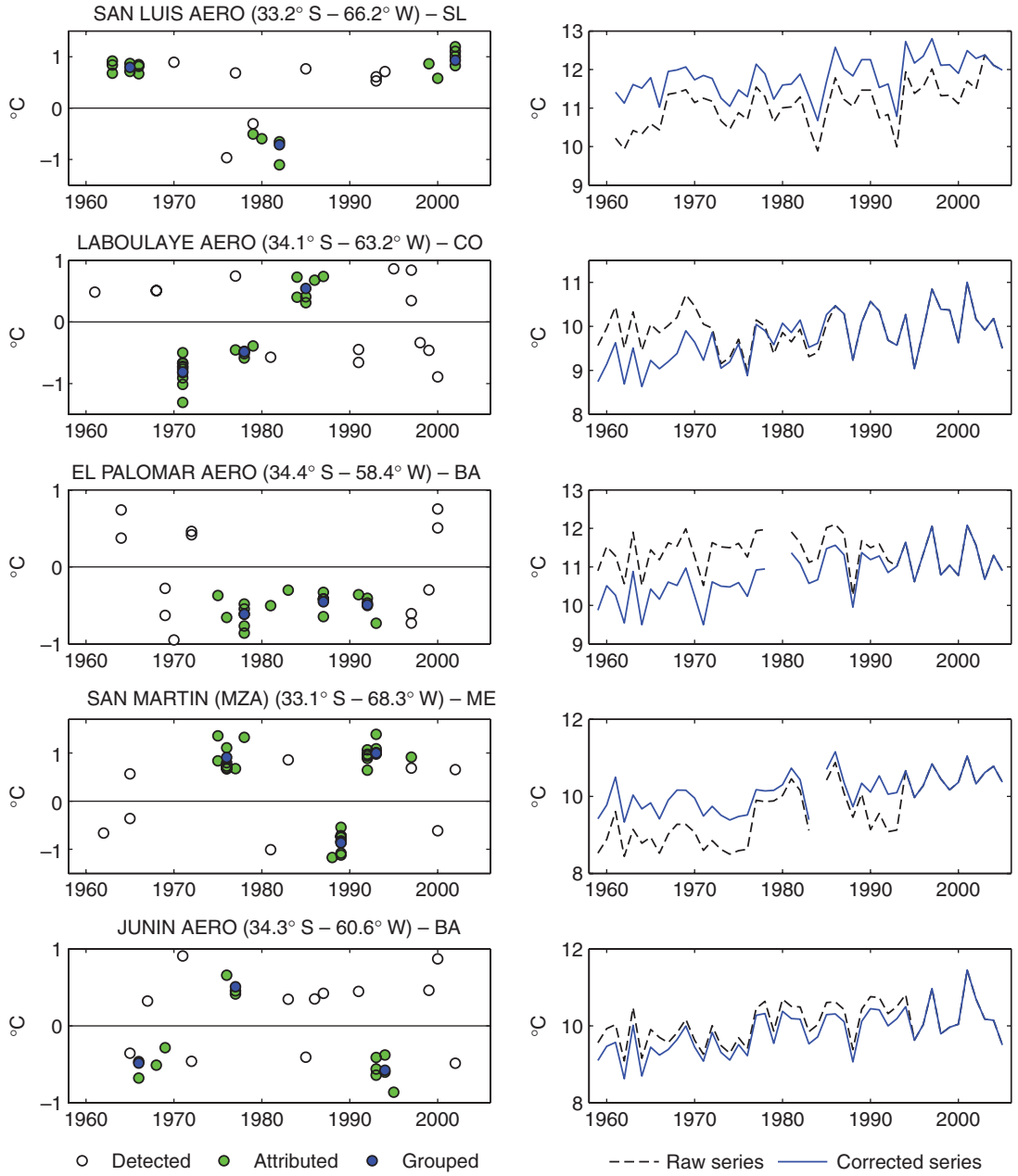

Figure 1. Illustration of the method on five series of minimal temperature. Right panels: raw series (dashed dark line) and corrected series (thick blue line). Left panels: position × amplitude scatterplot of detected breakpoints (white circles), attributed breakpoints (green circles) and grouped breakpoints (blue circles).

measured temperature trends. Sections 5 discusses results and concludes.

2. Method description

This section first introduces some notation and defines a similarity metric. The four steps of the procedure, e.g. detection, attribution, grouping and correction, are then described in detail. The procedure is illustrated graphically on the basis of actual series of temperatures – those series are further described in Section 4. Figure 1 gives an overall illustration of the four steps of the procedure, based on five actual series.

2.1. Notation

Our procedure requires intensive processing of the information on breakpoints detected on pairwise difference series, involving inter-comparison of breakpoints characteristics and compilation of the resulting outcomes. To describe this processing clearer, it is thus useful to introduce some specific notations regarding breakpoints and breakpoints sets:

$$\theta = (i, j, \widehat{a}, \widehat{\tau}, p(\tau)) \qquad \widetilde{\theta} = (i, j, a, \tau) \tag{4}$$

$$\Theta_i = \{\theta | i_\theta = i\} \qquad \Theta = \cup_{i=1}^{p} \Theta_i \qquad \Theta_i = \cup_{g=1}^{k} \Theta_{ig} \tag{5}$$

$$N_i = \text{Card}\,\{\Theta_i\} \quad N = \text{Card}\,\{\Theta\} \quad N_{ig} = \text{Card}\,\left\{\Theta_{ig}\right\} \tag{6}$$

where a breakpoint detected on a difference series x_{ij} is denoted by $\theta = (i, j, \widehat{a}, \widehat{\tau}, p(\tau))$, with i and j the index of the candidate and the reference series respectively; $\widehat{a}$ and $\widehat{\tau}$ the estimators of breakpoint sign and location respectively and $p(\tau)$ the posterior distribution of breakpoint position. The computation of those characteristics at the detection step is described further in this section. For a given detected breakpoint θ, $\widetilde{\theta} = (i, j, a, \tau)$ denotes the true breakpoint present in series x_{ij} having actual position τ and actual sign a, e.g. the detected breakpoint θ is an imperfect representation of the actual breakpoint $\widetilde{\theta}$. When needed, we refer to the characteristics of a particular detected breakpoint θ by using an indice, thus denoting its characteristics i_θ, j_θ, $\widehat{a}_\theta$, $\widehat{\tau}_\theta$ and $p_\theta(\tau)$. The set of all breakpoints θ detected on all difference series is denoted by Θ, and Θ_i denotes the subset of all breakpoints obtained for a particular candidate series i. The subsets $(\Theta_i)_{i=1}^{p}$ are thus a partition of Θ. When a subset Θ_i is further partitioned into k clusters, we will denote $\left(\Theta_{ig}\right)_{g=1}^{k}$ such a partition.

2.2. Similarity metric

We now introduce a metric of similarity between any two breakpoints θ and θ' in Θ, which we denote $\delta(\theta,\theta')$:

$$\delta\left(\theta, \theta'\right) = \left\{\sum_{\tau=1}^{n} p_\theta(\tau) . p_{\theta'}(\tau)\right\} . \mathbf{1}_{\widehat{a}_\theta \widehat{a}_{\theta'} > 0} \tag{7}$$

Figure 2 illustrates Equation (7) on actual breakpoints. The metric δ has a simple, natural meaning. It is an approximation of the probability that the true breakpoints $\widetilde{\theta}$ and $\widetilde{\theta}'$ present in the series are actually identical, conditional on the available information and having defined identicality between two breakpoints as sharing identical position and sign. Rephrased in the Bayesian decisioning framework, the metric $\delta(\theta,\theta')$ can also be viewed as the posterior cost associated to deciding whether the two actual breakpoints are identical, based on a simple $0-1$ cost function. This can be formally written as

$$\delta\left(\theta, \theta'\right) \sim \mathbb{P}\left\{\widetilde{\theta} \equiv \widetilde{\theta}'\right\}$$
$$\text{with :} \quad \mathbb{P}\left\{\widetilde{\theta} \equiv \widetilde{\theta}'\right\} =_{\text{bydef.}} \mathbb{P}\left\{\tau_{\widetilde{\theta}} = \tau_{\widetilde{\theta}'} \text{ and } a_{\widetilde{\theta}} a_{\widetilde{\theta}'} > 0\right\} \tag{8}$$

The justification of Equation (8) – both for the definition component and for the approximation component – is detailed in the Appendix. The metric δ thus differs from similarity metrics usually found in the literature which are straightforward functions of point estimators, in that it is probabilistic and takes into account the uncertainty associated to those estimators. It may thus provide a more grounded quantification of the amount of evidence in favour of identicality.

2.3. Detection

The detection step of the procedure consists in inferring the number of breakpoints as well as estimators of their position $\widehat{\tau}$ and sign $\widehat{a}$ from each difference series x_{ij}. In addition, the detection step should also provide the posterior distribution $p(\tau)$ of the position τ for each breakpoint θ. The information set Θ is then obtained

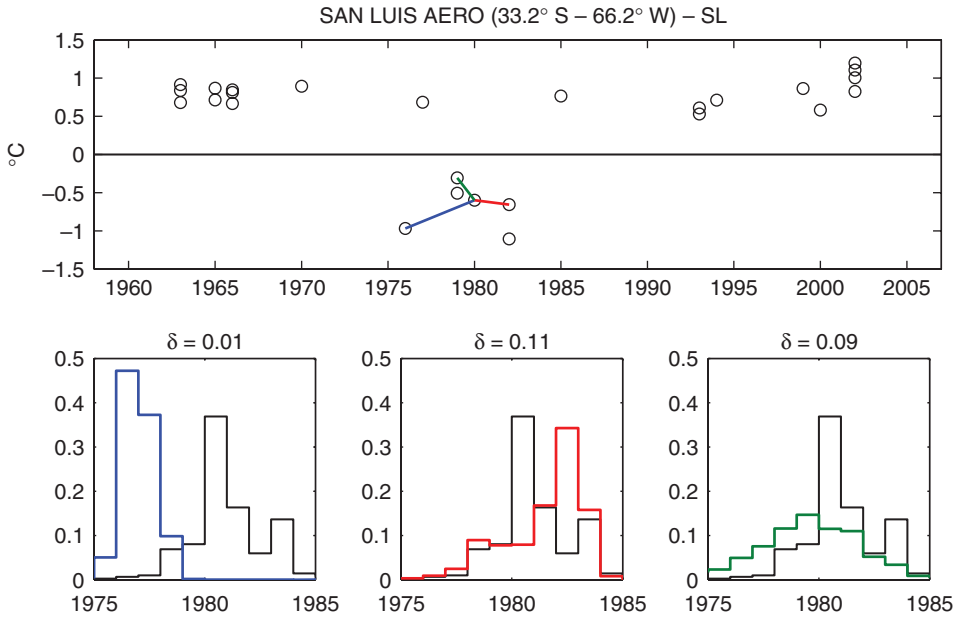

Figure 2. Upper panel: position × amplitude scatterplot of detected breakpoints for candidate station San Luis. Thick coloured lines highlight a central breakpoint and three nearby breakpoints. Lower panels: posterior distributions of the central breakpoint (light black line) compared to that of each nearby breakpoint (thick coloured line). Metric δ is indicated above each chart. Based on threshold $\delta^* = 0.05$, the lower right and middle panels correspond to similar breakpoints.

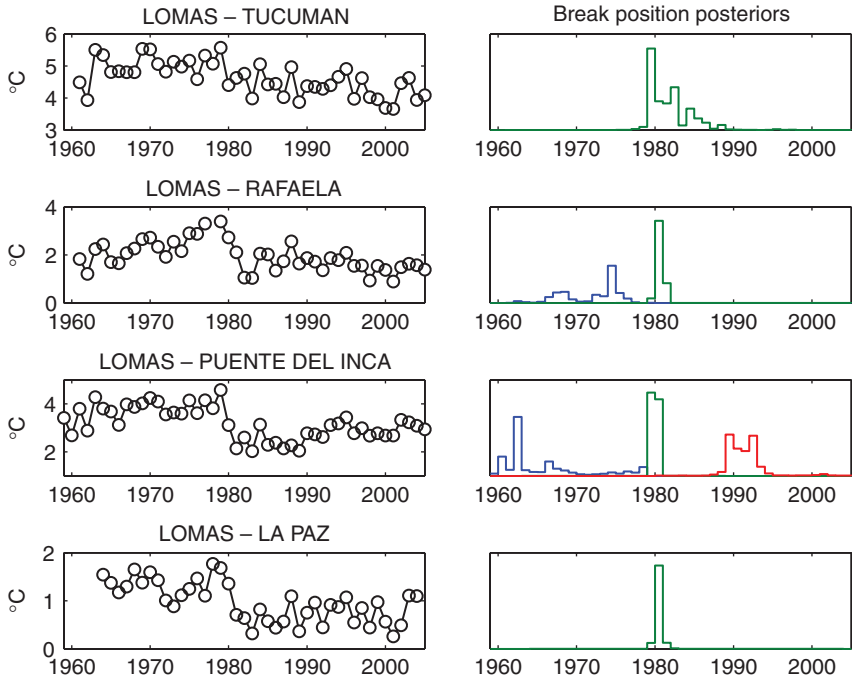

Figure 3. Illustration of the detection step on four pairwise difference series. Left panels: difference series between candidate station Lomas and four neighbouring stations. Right panels: posterior distributions of detected breakpoints.

by compiling breakpoints. For this purpose, we used the Bayesian method of Hannart and Naveau (2009). The method is adequately designed for the detection of an unknown number of multiple change-points in the mean of a Gaussian, independent series. Its specificity lies in the segmentation of the time series into subsequences corresponding to the episodes that contain a unique jump, which are found by applying Bayesian decision theory. This method is particularly fit for the computation of the posterior distribution $p(\tau)$ of the breakpoints position because it is Bayesian, and we used it here specifically for this reason. However, our procedure can be applied no matter the detection method used, and even if it is not Bayesian – in particular it can be applied to the detection method used in CM04 and MW09. In this case, we obtain breakpoints θ together with their estimators $\widehat{\tau}$ and $\widehat{a}$, but without the distribution $p(\tau)$ which is required for the sake of computing metric δ. The problem is thus to derive an approximation of this distribution. To obtain it, we isolate each breakpoint within a single breakpoint subsequence y. That is, for a series x_{ij} with r detected breakpoints at positions $0 < \widehat{\tau}_1 < \widehat{\tau}_2 < \ldots < \widehat{\tau}_r < n$, the subsequence corresponding to the lth breakpoint would be defined by $y = \left[\widehat{\tau}_{l-1} + 1, \widehat{\tau}_{l+1}\right]$ with the convention $\widehat{\tau}_0 = 0$ and $\widehat{\tau}_{r+1} = n$ and $l = 1,...,r$. Then, we apply the expression given in Lee and Heghinian (1977) to the consecutive segments y:

$$p(\tau) \propto \lambda_\tau^{-\frac{1}{2}} \left\{ 1 - \frac{\lambda_\tau \Delta\overline{y}_\tau^2}{\widehat{\sigma}_y^2} \right\}^{-\frac{n_y-2}{2}} \tag{9}$$

where $\widehat{\sigma}_y^2$ is the empirical variance of y obtained after subtraction of the empirical means on each two subsegments; n_y is the length of y; $\lambda_\tau = \frac{\tau}{n_y}\left(1 - \frac{\tau}{n_y}\right)$ is a weighting factor and $\Delta\overline{y}_\tau = \overline{y}_{\tau+1:n} - \overline{y}_{1:\tau}$ is the difference in partial means at time τ. It should be noted that $\Delta\overline{y}_\tau$ corresponds to the maximum-likelihood estimation (MLE) estimator of break amplitude when the break is assumed to occur at τ, and that the factor λ_τ appears in the expression of the variance of this estimator $V\left(\Delta\overline{y}_\tau\right) = \widehat{\sigma}_y^2/\lambda_\tau n_y$. Distributions of breakpoint positions are shown in Figure 3.

2.4. Attribution

The attribution step consists in identifying for each breakpoint $\theta \in \Theta$, which among series i_θ and j_θ is the culprit series. The general idea for this attribution is to compute the number of breakpoints in Θ_{i_θ} and Θ_{j_θ} that are similar to θ, and then to decide which is the culprit series based on the values of these numbers. For this purpose, we therefore need to define an accounting rule of similarity between two breakpoints and a rule of attribution based on similarity counts. For the accounting rule, we define two breakpoints θ and θ' to be similar whenever $\delta(\theta,\theta')$ is greater than a threshold δ. The number of breakpoints that are similar to θ in series i_θ are then obtained by

$$\gamma(\theta, i_\theta|\delta) = \sum_{\theta' \in \Theta_{i_\theta}} \mathbf{1}_{\delta\left(\theta,\theta'\right)>\delta} \tag{10}$$

In case i_θ is very clearly the culprit, it is expected that in each of the $p-1$ difference series $\left(x_{i_\theta j}\right)_{j \neq i_\theta}$, a

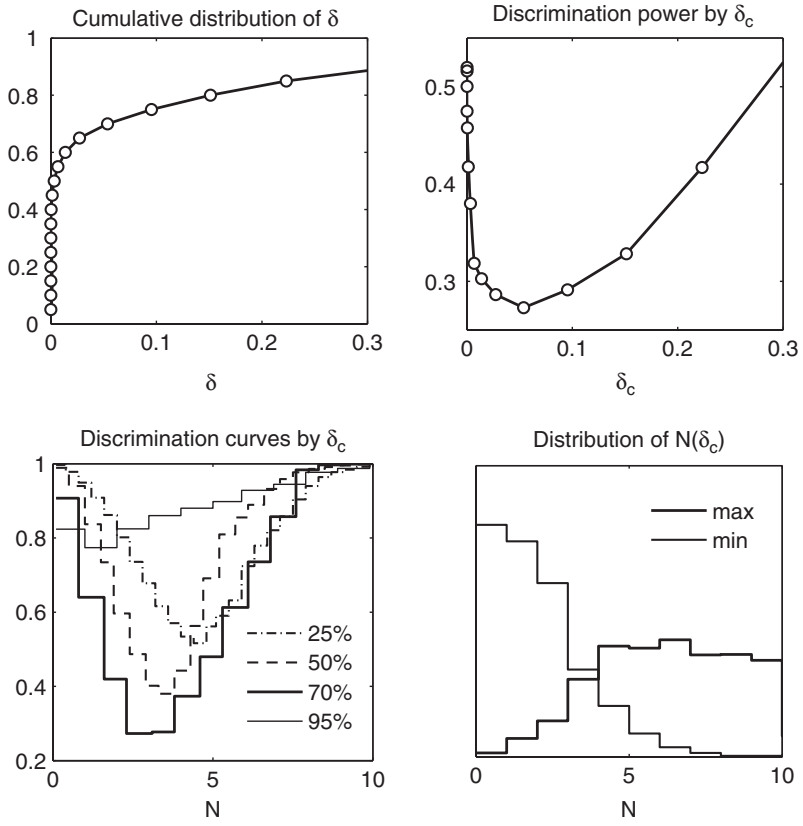

Figure 4. Upper left panel: empirical cumulative distribution of $\delta(\theta,\theta')$ for $(\theta,\theta') \in \Theta^2$, showing a concentration point in $\delta = 0$. Upper right panel: Discrimination power as a function of threshold δ, reaching its minimum in $\delta^* = 0.05$. Lower right panel: empirical distribution of γ_X and γ_N for $\delta = \delta^*$, e.g. for the best possible contrast. Lower left panel: Overlap between distributions of γ_X and γ_N for several values of δ, as a function of threshold γ.

breakpoint similar to θ should be found, hence γ should approach $p-1$. Conversely, in case i_θ is very clearly not the culprit, no similar breakpoint should be found, hence γ should approach 0. The choice of δ is greatly affecting the result of this procedure: for δ closed to 0 (resp. 1), all breakpoints tend to be spotted as similar to (resp. different from) any other breakpoint and γ tends to be always equal to $p-1$ (resp. 0), thus limiting its discriminative power. But for an intermediate value of δ, which makes a good job of establishing a clear difference between culprit and a non-culprit series, it is expected that the two values $\gamma(\theta, i_\theta|\delta)$ and $\gamma(\theta, j_\theta|\delta)$ are in general very different. As a consequence, for a suitable value δ, the empirical distributions of the maximum and minimum values of γ are well separated. On the basis of these considerations, our strategy for selecting an optimal value of δ consists in maximizing the contrast between these two distributions, e.g.

$$\delta^* = \arg\min_{\delta\in[0,1]} \left\{ \min_{\gamma} \{1 + F_X(\gamma|\delta) - F_N(\gamma|\delta)\} \right\} \quad (11)$$

where $F_X(\gamma|\delta) = \frac{1}{N}\sum_{\theta\in\Theta} \mathbf{1}_{\gamma_X(\theta|\delta)<\gamma}$ denotes the cumulative empirical distribution of the maximum of the two values obtained for a similarity threshold δ, denoted $\gamma_X(\theta|\delta) = \max_{i\in\{i_\theta, j_\theta\}} \gamma(\theta, i|\delta)$. The same notation applies for F_N and γ_N with respect to the minimum of the two values. With these notations, the quantity $\{1 + F_X(\gamma|\delta) - F_N(\gamma|\delta)\}$ that appears in Equation (11) measures the overlap between F_X and F_N for a given level γ. The overlap is thus a function of γ and δ. When the overlap is minimized in γ, the quantity $\min_\gamma\{1 + F_X(\gamma|\delta) - F_N(\gamma|\delta)\}$ measures the contrast between the two distributions F_X and F_N. The contrast is thus a function of δ. It is equal to 0 when the distributions have disjoint support, e.g. when there exists a level γ that fully separates both distributions. It is equal to 1 when distributions are identical. Figure 4 shows plots of the distribution of γ_N and γ_X, of the overlap function and the contrast function.

After obtaining δ^*, we are able to compute $\gamma(\theta, i_\theta|\delta^*)$ and $\gamma(\theta, j_\theta|\delta^*)$ for every θ. We must then establish an attribution rule based on the comparison of these two values. A straightforward rule is to attribute θ to the series i^* that has the largest of the two counts.

$$i^* = \arg\max_{i\in\{i_\theta, j_\theta\}} \gamma\left(\theta, i|\delta^*\right) \quad (12)$$

This natural rule can be slightly improved by allowing the attribution of γ to both series in case both values are high, and to no series in case both values are low. Such a rule therefore requires the definition of a threshold for γ, and we choose as a natural option the threshold γ^*

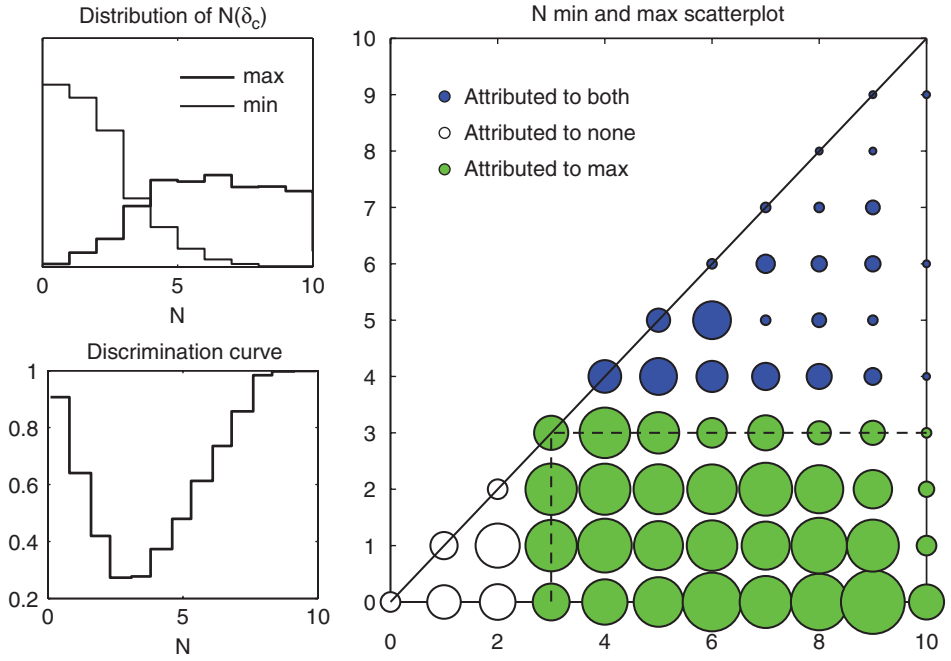

Figure 5. Upper left panel: empirical distribution of γ_X and γ_N for $\delta = \delta^*$, e.g. for the best possible contrast. Lower left panel: overlap between distributions of γ_X and γ_N for $\delta = \delta^*$, as a function of threshold γ. Right panel: empirical joint distribution of (γ_X, γ_N) (radius of circles indicate frequency) and graphical representation of the attribution rule.

which maximizes the separation between the distribution of $\gamma_X(\theta|\delta^*)$ and $\gamma_N(\theta|\delta^*)$:

$$\gamma^* = \arg \min_{\gamma \in [0,1]} \{1 + F_X(\gamma|\delta) - F_N(\gamma|\delta)\} \tag{13}$$

The attribution rule is illustrated in Figure 5. We perform attribution iteratively. At each step, the attribution rule is applied only to the breakpoint θ^* with best contrast, e.g.

$$\theta^* = \arg \max_{\theta \in \Theta} \left\{ \gamma_X\left(\theta|\delta^*\right) - \gamma_N\left(\theta|\delta^*\right) \right\} \tag{14}$$

Then, the set Θ is updated by removing the attributed breakpoint θ^* from the non-culprit series (if any) and by recomputing all the values of γ. Iterations are applied until no further attribution can be done.

2.5. Grouping

As a result of the above-described attribution procedure, a breakpoint may be attributed to the candidate series, to the reference series, or in some unfrequent instances to both. The result of this sorting is, for each candidate i, a subset Θ_i of all breakpoints that were identified as caused by i. Because they were detected on several difference series, there are several attributed *versions* of the same actual breakpoint, but of course, the candidate series must be corrected only once for this breakpoint hence those versions must be grouped. We approach this grouping as a clustering problem, and we deal with it using a classical, hierarchical ascending clustering procedure as described for instance in Jain and Dubes (1988). In order to apply such a procedure, we need to define a metric of similarity between two subsets of breakpoints Θ_{ig} and $\Theta_{ig'}$. For this purpose, we use the average of pairwise distances between elements of both subsets:

$$\delta\left(\Theta_{ig}, \Theta_{ig'}\right) = \frac{1}{N_{ig} N_{ig'}} \sum_{\theta \in \Theta_{ig}} \sum_{\theta' \in \Theta_{ig'}} \delta\left(\theta, \theta'\right) \tag{15}$$

Such an 'average distance' choice is a commonplace option in such clustering procedures, yet further justification for this choice in the present context can be found in the Appendix. This distance metric enables to run the clustering algorithm and to obtain the k clusters for each k in 1,...,N_i. At present, determining the natural number of clusters from those results is a distinct problem which has been extensively studied and also requires to choose an appropriate criterion. For this purpose, we define the following criterion:

$$\mathcal{B}(k|i) = -\sum_{g=1}^{k} \sum_{\theta \in \Theta_{ig}} \log \left\{ \frac{1}{N_{ig}} \sum_{\theta' \in \Theta_{ig}} \delta\left(\theta, \theta'\right) \right\} + \frac{1}{2} k \log N_i \tag{16}$$

The criterion $\mathcal{B}$ can be viewed as a pseudo-Bayesian Information Criterion (BIC) (Schwarz, 1978). Indeed, for reasons that are detailed in the Appendix, the first term of Equation (16) consists of a pseudo-log likelihood function which measures the quality of the fit provided by a particular cluster, and thus gets smaller as the number of clusters k increases. On the other hand, the second term $\frac{1}{2} k \log N_i$ is the classic penalty term of the BIC which increases with k to account for increasing complexity. The criterion $\mathcal{B}$ reaches a minimum in a value of k which

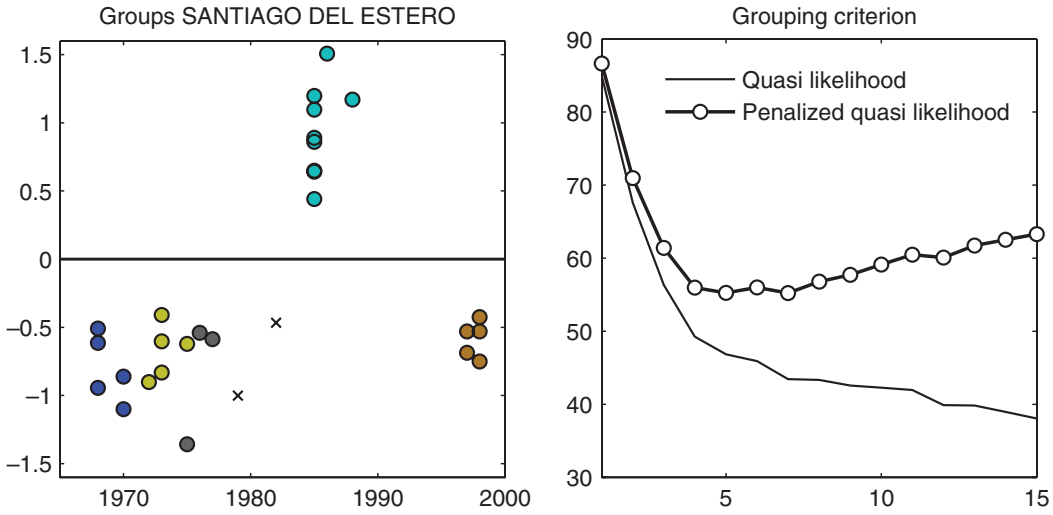

Figure 6. Illustration of the grouping step on breakpoints attributed to Santiago del Estero. Right panel: pseudo-BIC criterion and pseudo-likelihood as a function of the number of groups k, minimum is reached in $k^* = 7$. Left panel: representation of the seven groups obtained: coloured circles correspond to non-singleton groups, black crosses to singletons.

is defined as the optimal number of clusters.

$$k_i^* = \arg\min_k \mathcal{B}(k|i) \tag{17}$$

Figure 6 shows a plot of the pseudo-BIC $\mathcal{B}(k|i)$ and the resulting optimum clustering of attributed breakpoints obtained for $k = k_i^*$. Finally, a point estimator of position for each of the k_i^* subsets is obtained with

$$\widehat{\tau}_{ig} = \arg\max_\tau \sum_{\theta \in \Theta_{ig}} p_\theta(\tau) \tag{18}$$

which is the mode of the pdf obtained by direct intra-group averaging of the individual pdfs p_θ.

2.6. Correction

The above-described steps result in the determination of the number and position of breakpoints for each series. For correction, we choose to apply the approach CM04, which consists in resolving the standard model of Equation (1) using the final breakpoint positions obtained for each candidate series after grouping. While this model is technically inextricable with unknown breakpoint positions, resolution is actually straightforward when breakpoint positions are known using a direct application of a standard balanced two-way analysis of variance (2D ANOVA). Such a procedure leads to an estimate $\widehat{s}_t$ of the climate effect and an estimate $\widehat{\delta}_{i,t}$ of the piecewise constant station effect. By subtracting the latter estimate from the raw series $x_{i,t}$, the corrected series $\widehat{x}_{i,t}$ is obtained.

Finally, as in the case of CM04, the above-described four steps − detection, attribution, grouping and correction − are applied iteratively. That is, the corrected series obtained after application of the four-step procedure are then reprocessed using the exact same procedure. These iterations are conducted until no extra inhomogeneity is found. Series are then declared to be homogeneous. Note that such an iterative treatment is frequent but not systematic in homogenization literature. For instance, the MW09 procedure as described in Menne and Williams (2009) is a one-off, non-iterative procedure. However, in Section 3, for the sake of a like-for-like performance comparison with our procedure, we implement MW09 both in a one-off and in an iterative modality. Such an iterative application of the original MW09 procedure is straightforward and presents neither theoretical nor practical difficulty.

3. Simulation results

The purpose of this section is to assess the performance of the procedure described in Section 2, hereinafter referred to as HMN13, by implementing it on simulated series, and compare it to the performance achieved by the two available alternative procedures CM04 (manual) and MW09 (automatized). We conduct this performance evaluation on a simulated station network, and we assess performance based on two different metrics. All three procedures are applied iteratively, and performance metrics are derived at each iterations.

3.1. Simulations

We attempt to simulate realistic series of yearly temperature observations obtained on a representative network of stations. Simulations are based on the assumptions given in the standard homogenization model of Equation (1), recalled here:

$$x_{i,t} = s_t + \delta_{i,t} + \epsilon_{i,t} \quad i = 1, ..., p \quad t = 1, ..., n$$

We choose the number of stations in the network $p = 20$ and the length of the series $n = 100$. The common climate signal s_t is simulated as an AR(1) with autocorrelation 0.6 and standard deviation 0.5 °C. For

the station effect $\delta_{i,t}$, we first simulate the inter-event times, assumed to be iid with a Gamma distribution having mean 15 years and standard deviation 11 years. The simulated inter-event times then drive the number and position of breakpoints, leading to an average of 8 breakpoints by series. The piecewise constant values of $\delta_{i,t}$ on the successive segments are then simulated, assuming them to be iid with a Gaussian centred distribution having standard deviation 0.6 °C. Finally, the station noise $\epsilon_{i,t}$ is simulated assuming it to be Gaussian iid with independence in time and space, and having standard deviation 0.2 °C. All the above numerical values were chosen as realistic based on the results obtained on Argentinean series, which are detailed in Section 4.

It should be noted that the above simulation testbed is very similar to, and is largely inspired by, the COST synthetic data simulation testbed described in Venema *et al.* (2011), although a few simplifications were taken here. There are three main simplifications: (1) inhomogeneities do not have a seasonal cycle as we deal with yearly time series; (2) series do not have missing values nor local trends; (3) the common climate signal is an autoregressive noise that does not have a trend nor elaborated spectral features. For the purpose of the present discussion, which is merely to qualitatively highlight the potential benefit of our procedure as a starting point, rather than performing an in-depth simulation analysis as comprehensive as that of Venema *et al.* (2011), we argue that our simplifications do maintain the most important characteristics of the statistical problem under study and as such are acceptable for the aforementioned purpose.

3.2. Performance metrics

We define two distinct performance metrics denoted r_1 and r_2. Because the estimation of the piecewise constant term δ_{it} can be seen as the end goal of the procedure, we define our first performance metric r_1 as the mean squared error (MSE) of obtained estimators:

$$r_1 = \frac{1}{p}\sum_{i=1}^{p} \frac{\left\{\sum_{t=1}^{n}\left(\delta_{i,t}-\widehat{\delta}_{i,t}\right)^2\right\}^{\frac{1}{2}}}{\left\{\sum_{t=1}^{n}\delta_{i,t}^2\right\}^{\frac{1}{2}}} \tag{19}$$

It has the property to be an increasing function of the estimation error attached to the number, position and sign of breakpoints separately, and it conveniently integrates them together. Second, we use a metric r_2 which focuses more specifically on the performance in correctly spotting the existence of the breakpoints. The metric is described in detail in Hannart and Naveau (2009) and is defined by

$$r_2 = \frac{1}{p}\sum_{i=1}^{p}\left\{\frac{N_i^{(fp)}}{n/5-K_i} - \frac{N_i^{(tp)}}{K_i} + 1\right\} \tag{20}$$

Metric r_2 is based on a trade-off between the number of true positives $N_i^{(tp)}$ and the number of false positives $N_i^{(fp)}$ found in each series i. A detected breakpoint θ is considered here as a true positive whenever its position estimation error $\widehat{\tau}-\tau$ is lower than 2 years and as a false positive otherwise, so that $N_i^{(tp)}+N_i^{(fp)}=N_i$. The intercept 1 and the weights $(n/5-K_i)^{-1}$ and $-K_i^{-1}$ of the affine terms averaged in Equation (20), where K_i denotes the true number of breakpoints present in series i, are set in such a way that r_2 is equal to zero when detection is perfect and to one when detection is random.

3.3. Results

Procedures MW09 and HMN13 (resp. CM04) required four (resp. three) iterations until convergence was reached (i.e. no additional inhomogeneity is found) – the cumulated number of detected inhomogeneities for each procedure is shown in Figure 7(a). The average MSE metric r_1 obtained after the last iteration of the automatized procedure MW09, of the manual procedure CM04 and

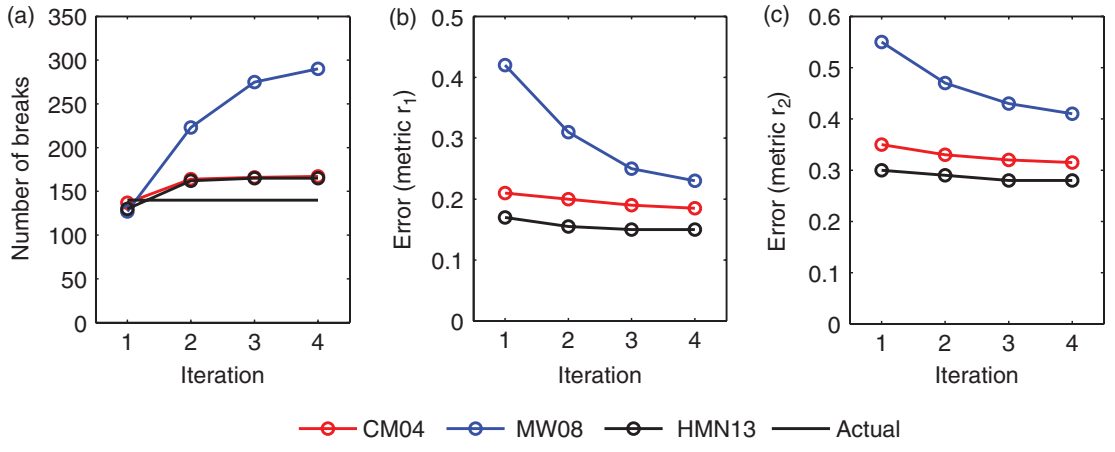

Figure 7. Results of the simulation study for each tested procedure, over a sample of 20 series. Left panel: cumulated total number of detected breakpoints at each iteration (bold lines and circles) and actual total number of breakpoints (thin dark line). Middle panel: r_1 error at each iteration. Right panel: r_2 error at each iteration.

of our procedure HMN13 are respectively 23 ± 1.5%, 18.5 ± 1.5% and 15 ± 1% (Figure 7(b)), and the average metric r_2 obtained after the last iteration are respectively 41 ± 3.5%, 31.5 ± 3% and 28 ± 2% (Figure 7(c)). The performance achieved by our procedure is thus superior to the performance achieved by both alternative procedures that were tested, based on both performance metrics. The performance gain *versus* the automatized procedure is significant, with a roughly 35% error reduction on both r_1 and r_2. The performance gain *versus* the manual procedure is small with a roughly 15% reduction on both r_1 and r_2. However, what matters in the latter case is not this small gain in performance, it is rather the gain in human time, as well as the possibility to obtain some results and a performance level that can be reproduced. These figures represent averages obtained for a reduced set of 20 simulated series and are therefore subject to some sampling error, yet both performance gaps are statistically significant here ($p = 0$ and $p = 0.03$ respectively).

As a side finding, simulation results also show that an iterative implementation is beneficial for the three tested procedures, but that the magnitude of the benefits is strongly contrasted across procedures. Indeed, on the one hand the performance of MW09 is greatly improved by iterative implementation, whereas on the other hand the performance gain is very limited for our procedure as well as for CM04. For the latter two procedures, we find that when measuring performance separately series by series, iterations sometimes do actually yield a deterioration instead of an improvement (not shown) – even though the sample average is found to be slightly improved overall. These qualitative findings are robust across both performance metrics. For MW09, the benefit of performing iterations is more important when measured with r_1 than with r_2; i.e. a 45 and 25% error reduction respectively. This is a consequence of the fact that the incrementally detected breakpoints at each iteration tend to have a much higher proportion of false positives, which is penalized more heavily by r_2 which explicitly accounts for this, than by r_1 which focuses solely on MSE. Overall the consolidated performance gap between the original, one-off version of MW09 and our proposed iterative procedure HMN13 is roughly twofold, although this comparison can not be considered to be like-for-like.

4. Application to Argentinean temperature series

The purpose of this section is twofold. First, it aims at illustrating the method by applying it to real data, consisting here of temperature series obtained from the Argentinean meteorological station network. Second, the final outcome of this illustrative application of the procedure, consisting of corrected Argentinean series, is at least as important a purpose because this network has never been previously homogenized to our knowledge, despite the fact that these series have been used for numerous climate studies in the region.

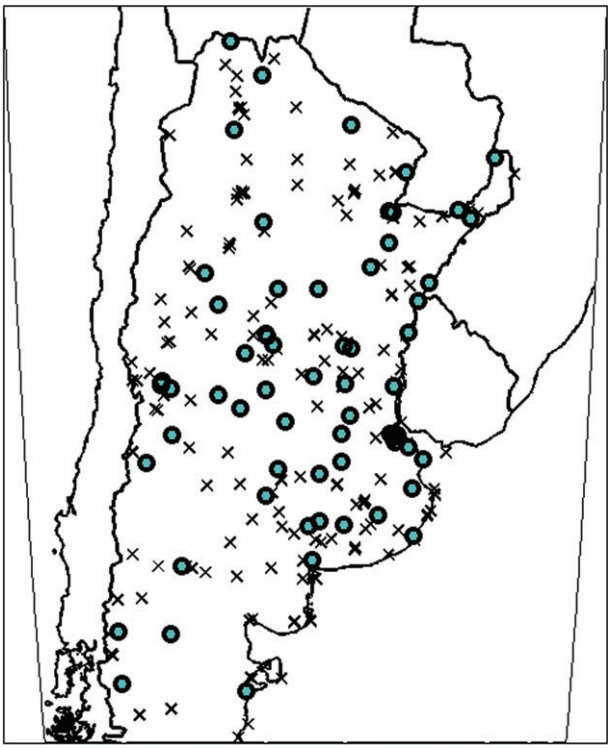

Figure 8. Map of the Argentinean National Meteorological Service (SMN) network. Green circles represent stations selected for homogenization. Black crosses represent stations that were not selected.

4.1. Pre-processing

We used a dataset of daily minimum and maximum temperatures (denoted TN and TX respectively hereinafter) from the station network of the Argentinean weather service, covering the period 1959–2006. The network has a total of 250 active stations heterogeneously spread over the Argentinean territory between latitude 55° and 22°S and longitude 72° and 54°W (Figure 8), with variable data availability across stations and time. To cope with missing data issues, we dismissed months with more than 4 missing days, years with more than 2 missing months and stations with less than 40 years of record. These suppressions resulted in a sample of 67 stations that were the same for TN and TX, which are listed in Table 1 and are highlighted in Figure 8. For these stations, missing years appear to be concentrated at the beginning of the recording period and becomes less of an issue after 1970. Series were quality checked based on the methodology of Vincent *et al.* (2005). Yearly series were derived from daily series by straightforward intra-annual averaging. As the Argentinean network is widespread over several climatic regions with contrasted features, the assumption of the standard model of Equation (1) according to which all stations share a common climatic signal is not realistic. As is usually done in such a situation, we defined station-specific climatic neighbourhoods on which the assumption holds, and then derived pairwise difference series from the stations belonging to the candidate's neighbourhood. Neighbours were selected according to a correlation threshold of 0.6. Correlations were computed

Table 1. Station identification numbers and locality names.

id	Name	id	Name
1	–	11	Las Lomitas
12	Salta	62	Santiago Del Estero
67	Bella Vista INTA	77	La Rioja
81	Ceres	82	Villa Maria Del Rio Seco
87	Monte Caseros	100	Cordoba
105	Cordoba Observatorio	111	Pilar Observatorio
113	Parana	117	Villa Dolores
131	Mendoza	132	Mendoza Observatorio
133	Rosario	134	Gualeguaychu
138	Rio Cuarto	139	San Luis
145	Pergamino INTA	148	Laboulaye
154	San Miguel	156	Buenos Aires
159	El Palomar	166	Ezeiza
450	Punta Indio	177	Nueve De Julio
178	Malargüe	179	Dolores
190	Santa Rosa	192	Coronel Suarez
204	Laprida	205	Pigüe
206	Mar Del Plata	210	Bahia Blanca
221	Neuquen	227	Maquinchao
248	Trelew	258	Comodoro Rivadavia
270	Rio Gallegos	293	Esquel
303	Reconquista	309	Tandil
311	Bariloche	323	San Rafael
325	Buenos Aires Aeroparque	332	General Pico
334	Villa Reynolds	335	Oran
339	Paso De Los Libres	346	Iguazu
353	Castelar INTA	358	Posadas
362	Marcos Juarez	369	San Martin
370	Cerro Azul INTA	423	La Plata Dique
451	Sauce Viejo	453	Junin
454	Chacras De Coria	456	Pehuajo
470	Corrientes	476	Chamical
477	Concordia	483	Formosa
487	Colonia Benitez INTA	–	–

Identification numbers are from the national weather service (SMN).

on the differentiated series to mitigate the correlation bias caused by inhomogeneities. When more than ten neighbours were found – as was most often the case – only the ten most correlated were selected. Accounting for both TN and TX, a total of $2\times10\times67=1340$ pairwise series was so obtained and processed.

4.2. Procedure results

The total number of inhomogeneities finally identified on the 67 stations is 252 for TN (resp. 182 for TX). The dates of these inhomogeneities and their estimated amplitudes used for correction are listed in Table 2. This means an inhomogeneity occurrence frequency of 8%, e.g. a return period of 12 years for TN (resp. 6% and 16 years for TX) – those figures match well with usual values of 10 to 15 years found in previous studies (CM04; Della-Marta *et al.*, 2004; Auer *et al.*, 2006). As illustrated in Figure 9, the time series of the total number of inhomogeneities detected in the network shows some strong variability with three peaks around 1972, 1983 and 1997, three lows around 1964, 1974, 1994, and a pronounced decrease from year 2000 onwards. The number of inhomogeneities on TN and TX show substantial correlation in time (equal to 0.6). Obtaining these inhomogeneities required five iterative applications of the entire procedure until no additional inhomogeneity was found. The five iterations obtained successively 196, 42, 7, 4 and 3 inhomogeneities for TN (resp. 92, 49, 23, 16 and 2 for TX).

Going further, we focus on the results of the first iteration. The detection step was applied to the 670 pairwise difference series for TN (resp. 670 for TX) resulting in a total of 1645 (resp. 748) breakpoints detected. This corresponds to 2.5 breakpoints (resp. 1.1) detected by series. With difference series on average 40 years long and breakpoint return period in candidate series expected to be around 15 years, there should be around 5 inhomogeneities on average in difference series. The large discrepancy between our results, and this expected value may be due to the fact that a large proportion of inhomogeneities are too small in amplitude to be detected, some inhomogeneities occurring nearly simultaneously in both the candidate and reference series may cancel out. Hence, the algorithm does not detect all breakpoints and only captures those that are sufficiently isolated in the series to stand out clearly. While this may be seen as a drawback of using pairwise difference series, we find that this is in fact not a problem in practice. Indeed, as highlighted below, the average number of breakpoints eventually detected on candidate series after attribution and grouping is much higher than the average number of breakpoints initially detected on difference series. This is owing to the diversity of difference series obtained for each candidate which gives more chance for each of its inhomogeneities to become apparent within a sufficiently long subsequence in one or several difference series. The optimal threshold δ^* was found to be equal to 0.05. The attribution step resulted in attributing 677 breakpoints (resp. 246) to the candidate series and 345 breakpoints (resp. 154) to both. The grouping step resulted in 196 (resp. 92) groups with an average of 6.8 (resp. 5.3) breakpoints per group. The total 196 (resp. 92) obtained inhomogeneities are unevenly distributed on the 67 series which have between 0 and 7 inhomogeneities, with an average of 3 (resp 1.5). For both, the distribution of breakpoint amplitudes is found to match well other studies results (CM04; Della-Marta *et al.*, 2004; Auer *et al.*, 2006), with a standard deviation of 0.6 °C (Figure 10) and a characteristic bimodal aspect around zero. It is interesting to note that, in spite of the latter feature, the distribution is nonetheless not exactly symmetric: the adjustments are slightly more often positive than negative.

4.3. Implications on trends and spatial coherence

We computed the linear trend coefficient for each series before and after homogenization. We find that for most stations, the trend is significantly affected by homogenization. When analysing those homogenization adjustments overall for the 67 stations, there are three main findings that stand out, and apply equally well for both TN and TX (Figures 11 and 12). First, the homogenized trend coefficients are on average significantly higher than

Table 2. Detected inhomogeneities: station id, breakpoint years (bold text) and °C amplitudes (italic when < 0) for TN and TX.

id	TN	TX
1	–	**93** 0.4, **97** 0.5
11	**62** 0.6, **80** −0.4, **82** −0.2, **97** −0.2	–
12	**65** −0.7, **92** −0.4, **97** 0.6	–
62	**68** −0.1, **70** −0.6, **75** −0.4, **85** 0.8, **98** −0.4	–
67	**65** 0.5, **86** −0.3, **00** 0.4	–
77	**62** −0.5, **66** 0.8, **73** −0.6, **78** −0.9, **82** 1.1, **94** 0.4, **99** −0.5	**73** −0.4, **74** −0.3, **86** 0.2
81	**67** −0.8, **73** 0.4, **92** −1, **93** −0.3, **96** 0.5, **00** 0.3	**70** −0.4, **82** 0.1, **89** −0.6, **94** 0.5
82	**64** −1.8, **66** 1.7, **96** −0.8	**61** 0.7, **66** 0.6, **71** −0.8, **95** −0.9
87	**80** −0.1, **99** −0.2	**67** −0.7, **69** 0.5, **75** 0.4, **90** 0.1
100	**71** −0.2, **76** −0.8, **82** 0.5, **04** 0.3	**71** −0.3, **87** −1.4, **88** 1.5, **98** −0.5, **01** 0.4
105	**63** 0.4, **71** −0.2, **86** 0.4, **92** −0.1	**73** 0.3, **83** −0.3, **98** −0.1
111	**71** −0.3, **87** 0.4, **92** −0.3, **97**−0.5, **00** 0.6	**66** −0.5, **81** 0.3, **90** −0.3, **98** −0.3
113	**65** −0.4, **73** −0.9, **80** 1.1, **86** 0.2, **99** 0.1	**61** 1.1, **69** −0.7, **72** −0.1
117	**61** 0.4, **71** −0.6, **79** −0.4, **85** 0.6, **88** −0.5, **93** −0.1, **02** 0.4	–
131	**70** 0.4, **80** 0.1, **91** −0.2, **95** 0.5	**78** −0.5, **83** 1.2, **84** −1.4, **87** 0.5
132	**71** −0.8, **93** −0.1, **96** −0.1, **99** −0.3	–
133	**61** −0.5, **63** 0.5, **80** 0.1, **95** 0.3	**63** 0.4, **70** −0.2, **80** −0.1, **89** 0.1, **97** 0.1
134	–	**62** −1, **78** 0.2, **88** 0.2, **93** 0.1, **97** −0.6
138	**75** −0.2, **85** −0.1, **92** −0.7, **93** 0.9, **01** −0.4	**77** −0.3, **86** 1.8, **88** −1.8, **90** −0.7, **95** 0.5
139	**66** 0.7, **71** −0.6, **82** −0.7, **85** 0.5, **93** 0.3, **02** 0.8	**83** 0.2, **87** −0.4
145	**61** 0.1, **64** −0.6, **71** 1, **78** −0.2, **95** −0.9	**64** −0.4, **71** −0.1
148	**71** −0.8, **78** −0.3, **85** 0.3, **97** 0.5, **99** −0.5	**71** −0.7, **81** −0.4, **97** −0.3
154	**61** 0.1, **69** 0.3, **83** 0.2, **98** 0.5, **00** −0.5	**72** 0.2, **00** −0.1
156	**83** −0.3, **87** 0.3, **98** −0.1	**85** 0.1, **92** 0.1
159	**69** −0.1, **76** −0.1, **78** −0.3, **87** −0.2, **92**−0.3	**68** 0.2, **88** −0.3, **91** 0.3, **96** −0.4, **00** 0.4
166	**67** −0.3, **80** −0.5, **83** 0.4, **86** 0.2, **98** 0.2	**72** 0.1, **83** 0.1, **93** −0.3
450	**77** −0.1, **90** −0.3, **99** 0.3	–
177	**71** 0.5, **84** 1.5, **85**−2.1, **99** 0.6	**76** −0.5, **85** −1.6, **86** 1.8, **00** −0.4
178	**63** 0.5, **69** −1.2, **72** 2.1, **73** −1, **88** −0.8, **99** 0.7	**64** 0.3, **69** 0.2, **83** −0.4
179	–	**72** 0.3, **83** 0.3, **85** −0.7, **88** 0.5, **00** 0.4
190	–	**64** 0.4, **85** −0.4
192	**76** 0.5, **85** 0.4, **89** 0.3, **99** −0.5, **02** 0.6	**65** −0.4, **70** −0.1, **83** −0.4
204	–	–
205	**59** 0.9, **66** −1.1, **78** −2.9, **81** 1.8, **93** 0.6, **99** −0.4	**69** 0.3, **90** 0.3, **99** −0.2
206	**83** 0.1, **89** 0.1, **93** −0.1	**72** 0.3, **90** −0.2, **95** 1.6, **98** −1.3
210	**66** −0.7, **70** 0.7, **84** −1.1, **87** 0.4, **93** −0.4, **97** 0.4	**72** 0.3, **80** 0.6, **84** −0.7, **87** −0.1, **92** −0.2
221	**72** −0.5, **75** −0.3, **80** 0.5, **82** −0.4, **86** 0.8, **02** 0.5	**64** −0.4, **84** 0.2, **91** 0.4
227	**63** 1.2, **68** −0.2, **86** 0.7, **00** −0.6, **02** 0.7	–
248	**66** −1.5, **69** 1, **75** 0.7, **81** −1, **89** −1.7, **92** 2.6	–
258	**79** −0.4, **82** −0.3, **92** −1.2, **02** 0.4	**69** −0.5, **77** 0.7, **79** −0.1
270	**71** −0.4, **89** 0.1, **93** −0.5	–
293	**60** 1, **77** 0.4, **90** −1.2	**73** 0.4, **83** 0.5, **89** −0.5
303	**64** −0.6, **73** 0.6	–
309	**66** −0.4, **74** −0.5, **81** 1.2, **84** −0.8, **89** 0.4	**71** −0.5, **76** −0.5, **81** −1.3, **83** 0.9, **85** 0.9
311	**84** −0.5, **93** −0.9	–
323	**64** −0.2, **65** −0.6, **92** 0.2	**75** 0.4, **77** 0.5, **94** 0.2, **97** 0.2
325	**62** −0.2, **67** −0.7, **71** 0.4, **97** −0.4	**68** −1.1, **70** 1.3, **84** −0.1
332	**78** 0.2, **83** −0.2, **99** −0.3	**71** 0.6, **75** −0.4, **90** −0.2
334	**72** 0.1, **78** 0.1, **97** 0.4, **02** −0.1	**62** −1.7, **64** 1.4, **83** −0.5
335	**75** 0.7, **77** −0.3, **84** −1.1, **91** 0.3, **95** 0.1	**62** −0.1, **69** 0.3, **77** −0.3, **83** −0.5, **88** 0.1, **02** 0.3
339	–	**79** −0.7, **94** 0.4
346	**67** −0.4, **77** −0.2, **84** 0.3	**72** 0.4, **74** −0.3, **80** 0.3, **88** 0.1, **93** −0.5
353	**68** 0.8, **76** 0.6	**61** −0.7, **73** −0.6, **82** 0.3
358	**59** −2.3, **89** −0.4	**59** −3, **84** −0.2, **97** −0.3
362	**85** 0.2, **97** 0.2	**71** 0.2, **76** 0.3, **86** −0.2
369	**78** 0.2, **82** −0.4, **90** −0.1, **98** −0.3	**66** −0.4, **77** 0.5, **86** −0.6, **95** 0.2
370	**71** −0.4, **76** 0.8, **89** −0.6, **93** 0.9	**67** 0.2, **69** −0.3, **82** −0.9, **83** 0.6, **87** 0.3
423	**67** 0.6, **92** 0.2, **95** −1.1	**81** 0.5, **91** −0.4
451	**79** 0.8, **87** 0.4, **88** −0.6, **90** −0.4	**70** −0.3, **80** 0.4, **89** −0.7, **98** 0.2
453	**66** −0.4, **77** 0.1, **94** −0.3	**72** −0.3, **88** −0.1, **96** 0.2
454	**67** −0.7, **90** 0.1, **97** 0.4	**66** −0.5, **72** 0.4, **77** −0.1, **96** 2.3, **99** −2.2
456	**72** 1, **75** −1, **85** 1.3, **89** −2.1, **97** 1.4	**72** −0.3, **83** −0.8, **87** 0.6, **96** −0.7
470	**61** −3.4, **76** −0.6, **83** 0.5, **97** −0.2	**61** −3.7, **70** 0.2, **92** −0.2, **98** 0.2
476	**62** −7.5, **70** −0.4, **87** 0.4, **91** −0.4, **98** −0.3	–
477	**62** −2.8, **77** −0.2, **80** −0.5, **88** 0.2	–
483	**62** −4, **68** −0.5, **81** 0.3, **00** −0.2	–
487	**72** 0.1, **80** 0.3, **84** −0.4	**67** 0.4, **78** −0.6, **96** 0.3

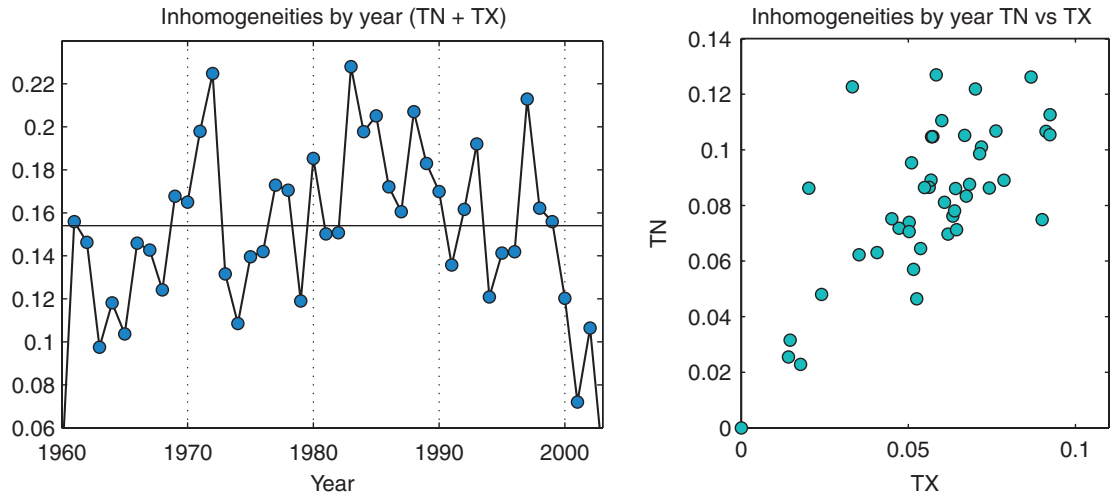

Figure 9. Left panel: time series of the total number of inhomogeneities by year (TN+TX). Right panel: scatterplot of the number of inhomogeneities by year on TX *versus* on TN (correlation = 0.6). Values are normalized by the number of stations available on each particular year.

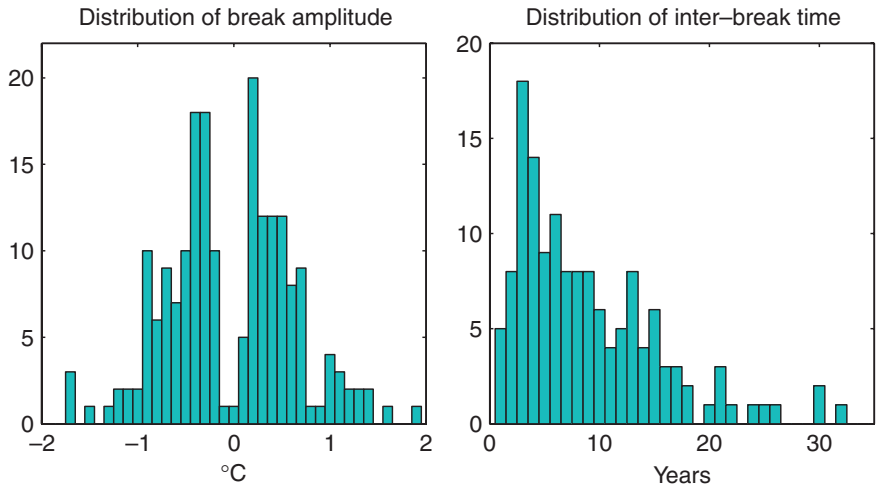

Figure 10. Empirical distributions of the characteristics of breakpoints on TN. Left panel: breakpoint amplitude. Right panel: time between two consecutive breakpoints.

the non-homogenized trend coefficients, in part owing to the aforementioned asymmetry in adjustments. For instance, the average trend on minimal temperature series is $+1.7$ °C per century before homogenization and it is increased to $+2.5$ °C per century after homogenization. Similarly, the average trend is on maximal temperature series is $+0.1$ °C per century before homogenization and it is increased to $+0.8$ °C per century after homogenization. Second, the adjustment result in a distribution of trend coefficients that is much more concentrated around its average value, e.g. trends after homogenization are much less spatially variable. For instance, the standard deviation of trends on minimal temperature series is $+1.5$ °C per century before homogenization and it is decreased to $+0.4$ °C per century after homogenization. Similarly, the standard deviation of trends on maximal temperature series is $+1.2$ °C per century before homogenization and it is decreased to $+0.5$ °C per century after homogenization. If one defines signal as the regional average trend and noise as its intra-regional spatial variability, homogenization thus translates into a strong enhancement of the signal-to-noise ratio on trends. For TN, it moves from strong (1.1) to very strong (6.3) and for TX, it moves from non-significant (0.08) to strong (1.6). Finally, the spatial coherence of trends after homogenization is much higher. As Figures 11 and 12 show, trend coefficients obtained on raw series are erratic and show no spatial coherence, whereas trend coefficients obtained on homogenized series show a smooth spatial aspect with some obvious large-scale features, such as a zonal gradient. The increased spatial coherence can also be visualized by considering the scatterplot of correlation between every pair of stations as a function of their lag distance (Figure 13). The scatterplot obtained after homogenization is considerably more concentrated around a decreasing autocorrelation line, than is the scatterplot before homogenization. A large part of this effect can be explained by the enhanced spatial coherence of

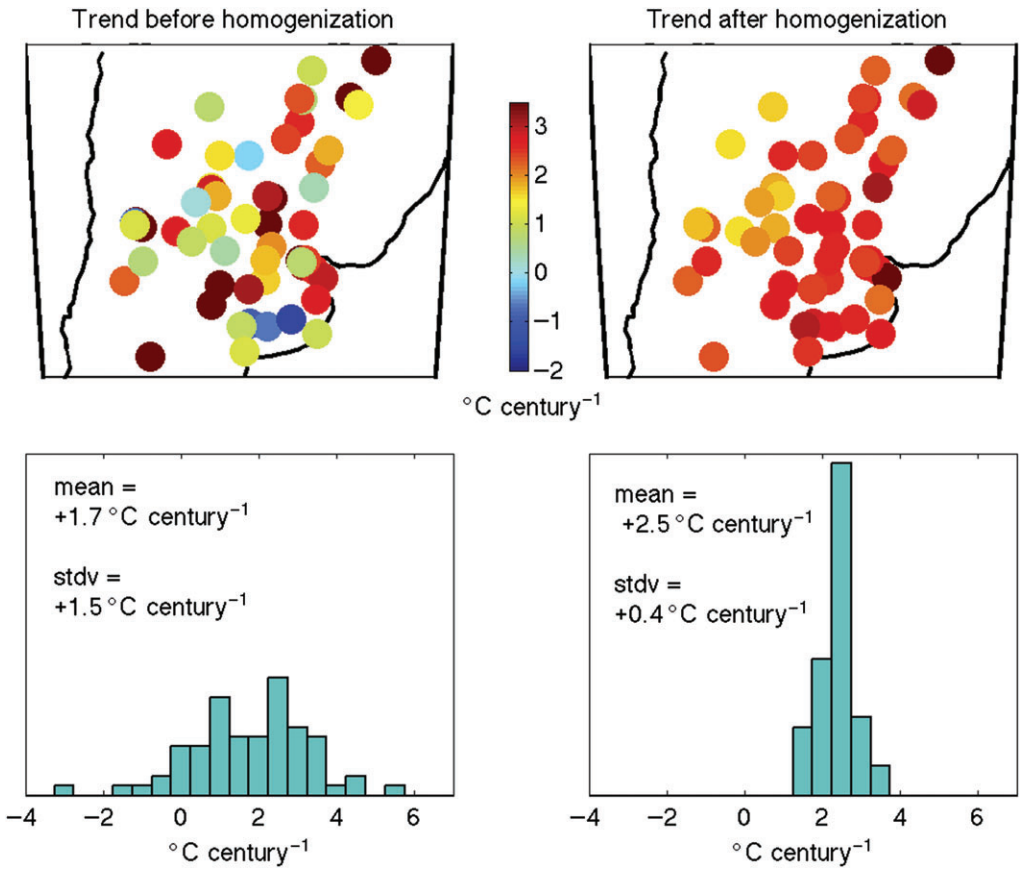

Figure 11. Implications of results for trends and spatial coherence. Upper panels: linear trend coefficients on TN by station, before and after homogenization. Lower panels: empirical distribution of trend coefficients, before and after homogenization.

trend alone. Nonetheless, this effect is still clearly visible even after trend removal, hence the enhancement of spatial coherence appears to apply similarly to variability components that are not related to the trend, such as periodic ones. Therefore, to summarize, homogenized series reveal overall a stronger warming signal and a much more pronounced spatial coherence and structure, than raw series do.

5. Conclusion

This section discusses implications and concludes on the improvements achieved methodologically for homogenization, and on the results obtained on Argentinean temperature series.

5.1. Methodological improvements

We have proposed a new homogenization procedure based on pairwise comparison of neighbouring series. The procedure is fully automatized and consequently avoids the tedious and lengthy visual inspection required by the approach CM04 and alike methods. Instead, it runs in a matter of minutes on a desktop computer and yet is able to maintain similar or better performance than a visual review of the series by a trained expert. To automatize, we followed a strategy similar in its bottom line to that of the only fully automatized method available so far to our knowledge, MW09. Indeed, as in this method, the automatization of attribution and grouping – the two manual, time-consuming steps of CM04 – relies in our procedure on a counting criterion to attribute breakpoints to a culprit series, and a distance criterion to group attributed breakpoints. The main methodological difference of our procedure with MW09 mainly lies in three aspects. On the one hand, a new metric is defined to measure similarity between breakpoints based on the posterior pdf of their position and sign. On the other hand, this metric is used in an optimized way to derive counting and clustering criteria used for attribution and grouping respectively. This optimization consists in the automatized calibration of an adaptive parameter. Finally, our procedure is implemented iteratively whereas MW09 was introduced as a one-off method, an aspect which is discussed further in this section. Those three differences are sufficient to materialize in a substantial performance gain with respect to MW09, measured by two different performance metrics, on a set of simulated series.

By achieving high performance at low cost in human time, our proposal thus resolves the main difficulty faced when homogenizing series based on pairwise differences: fully automatizing at a high cost in performance, *versus* visually reviewing series at a high cost in human time.

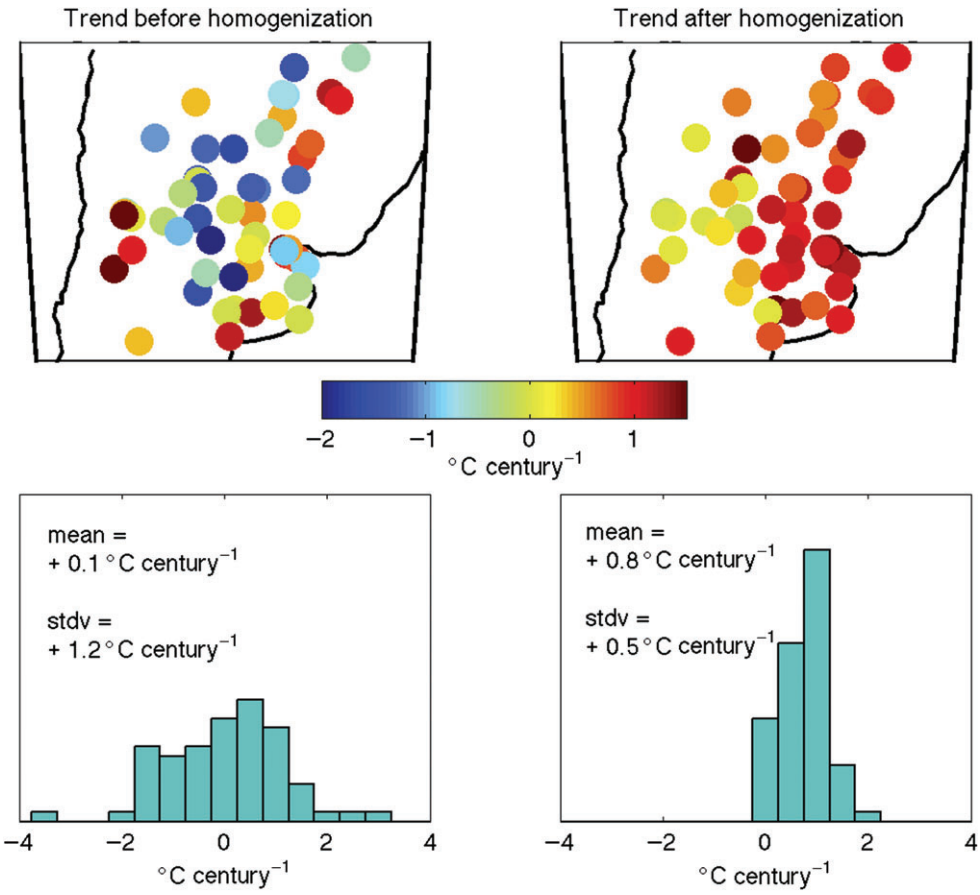

Figure 12. Implications of results for trends and spatial coherence. Upper panels: linear trend coefficients on TX by station, before and after homogenization. Lower panels: empirical distribution of trend coefficients, before and after homogenization.

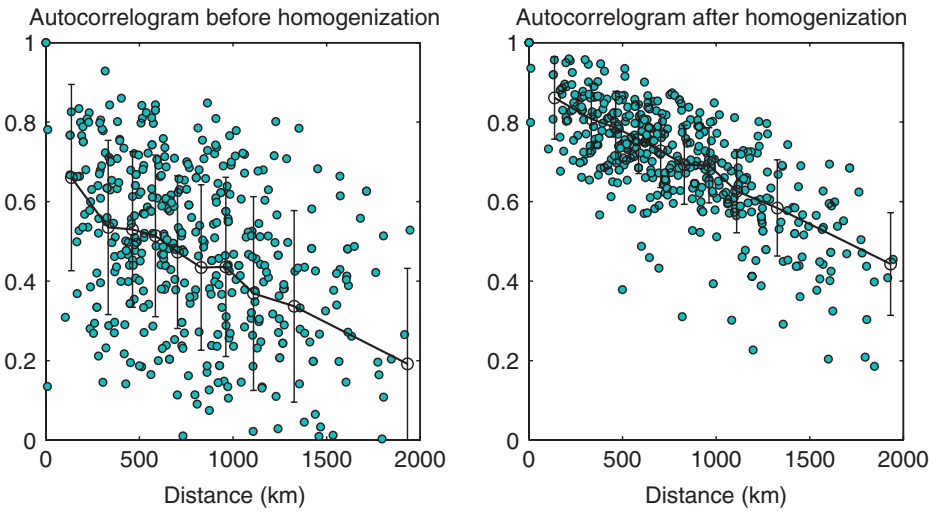

Figure 13. Implications of results for spatial coherence. Scatterplot of pairwise correlations on TN series × lag distances, before and after homogenization. Each circle represent a pair of stations (i,j).

The resolution of this difficulty may facilitate the popularization of the pairwise difference series approach instead of the regional reference series when homogenizing, even for very large data sets. In this respect, an interest of the procedure is that the attribution and grouping steps are independent of the detection and correction steps. Therefore, the former steps can easily be combined with or inserted within other alternative procedures based on different detection and correction strategies, thus further facilitating the diffusion of the approach. The increase in usage of pairwise difference methods resulting from such a possible popularization may lead to improvements in homogenization performance in general. In particular, as our procedure reduces operational constraints on

the number of neighbouring series to be used for difference, it would therefore be of interest to take advantage of the extra information provided by additional pairwise difference series and to gauge the resulting performance benefit.

The proposed procedure was implemented here on a relatively small meteorological network as a starting point. But, as a consequence of automatization, scalability to larger datasets does not appear to be a major issue here – which in our view is the most attractive feature of the proposed procedure. Indeed, the scalability problem is a purely computational one and thus relates mainly to the complexity of the algorithm, which is linear in the network size p, quadratic in the number of neighbouring series, and linear in series length n. Note that the latter linearity in n is a feature of the detection method of Hannart and Naveau (2009), by contrast for instance with the quadratic complexity in n of the detection step of the procedure CM04 associated to its dynamic optimization scheme. Running the entire procedure on the Argentinean network (i.e. with $n = 47$, $p = 67$ and 10 neighbours) took 90 s on a desktop computer. As a first raw estimate of the computational capacity required for instance for the homogenization of the International Surface Temperature Initiative databank stations (Lawrimore *et al.*, 2013), characterized by a typical series length $n = 130$ years and $p = 32\,000$ stations, we can merely scale linearly this amount of time in n and p, thereby assuming the same number of neighbouring series. This yields an estimated computational time of 45 h – before likely improvements in code performance and computer power for such a huge initiative – which is affordable.

The question of whether the proposed homogenization procedure, and more generally any homogenization procedure, should be applied iteratively or in a one-off mode was briefly addressed in Section 3. Even if our simulation results suggest some slight degree of benefit, they do not provide any obvious, general answer to this question. As discussed for instance in the radiosonde series homogenization study of Thorne *et al.* (2011), the answer may arguably depend, among other things, on the detailed features of the series and inhomogeneities under study, on the criterion used to stop iterations and on the metric used to measure performance. In any case, this emphasizes the need for further analysis on this matter. In particular, the problem of choosing a correct criterion to decide at each step whether further iterations should be performed, appears to us as especially relevant.

Finally, the proposed procedure still has a number of limitations, which leaves room for improvement and further extension. In particular, the method addresses restrictively abrupt shifts in the series mean, when inhomogeneities are known to also induce other types of changes (e.g. local trends). Further, the method focuses restrictively on yearly time series, when most recent procedures typically focus on monthly time series allowing for detection and correction of inhomogeneities presenting a seasonal cycle in their amplitude.

5.2. Argentinean temperature series

In the proposed application, yearly temperature series of the Argentinean weather service were homogenized for the first time to our knowledge. After correction, series obtained showed a strong enhancement of warming trends and a more pronounced spatial coherence of the network overall. Those two features – enhanced warming signal and higher spatial coherence – can be regarded as positive indications concerning the quality and reliability of our homogenization results. From a general standpoint, this brings an extra amount of evidence in favour of the case for homogenization as a necessary preliminary step before a detailed characterization of climate variability and change from observations.

From the standpoint of climatic studies in Argentina and South East South America, these results may have implications for several previous studies (Vincent *et al.*, 2005; Rusticucci and Barrucand, 2004; Rusticucci *et al.*, 2009) whose conclusions are based on the unhomogenized temperature record. It may be relevant to investigate whether those conclusions are affected when the analysis is based on homogenized series. For instance, a recent study (Camilloni and Barrucand, 2011) highlighted a negative trend in the nocturnal Urban Heat Island (UHI) intensity in Buenos Aires, Argentina, based on the temperature series difference between the Buenos Aires station (urban area) and the Ezeiza station (nearby rural area). For these two stations, our results show five inhomogeneities in the minimal temperature series at Ezeiza (two being particularly marked in 1980 and 1983) and three at the Buenos Aires station. After correction, the warming trends of homogenized series are found to be affected, gauging the importance of this effect would hence be relevant for this particular study. In this particular example, as in other studies, homogenized series may in some case mitigate some earlier conclusions. Nonetheless, it is also quite plausible that those conclusions would on the contrary be enhanced instead of mitigated, owing to the fact that homogenization tends to strengthen the climatic signal by removing what can be referred to as instrumental noise.

Finally, the enhancement of the climatic signal in homogenized series may create an opportunity for new, original studies. In particular, it may justify studies focusing on the detection and attribution of climate change over Argentina and South East South America. Indeed, the more pronounced warming trend of homogenized series should make detection easier. In addition, one may speculate that the enhanced spatial pattern of homogenized series would also facilitate an attribution study. Indeed, attribution requires a so-called fingerprint, i.e. a spatial pattern found in observations that allow for non-equivocal discrimination of the involved forcings, which may be enhanced after homogenization.

Acknowledgements

The European COST action HOME is acknowledged for financial support and for enabling valuable discussions and access to helpful insights from European experts on the topic. The ANR AssimilEx and FP7 ACQWA projects are also acknowledged by Philippe Naveau. The authors would like to thank the Centre National de la Recherche Scientifique (CNRS) and the CONICET for their support in this collaboration. Finally, we are grateful to the Editor and to two anonymous referees for their comments and suggestions which enabled to significantly improve this article.

Appendix

First, let us introduce some group characteristics $p_\Theta(\tau)$ and $\widehat{a}_\Theta$ associated to group Θ, defined by direct averaging of the characteristics of the breakpoints belonging to the Θ, e.g. $p_\Theta(\tau) = \frac{1}{N_\Theta}\sum_{\theta\in\Theta} p_\theta(\tau)$ and $\widehat{a}_\Theta = \frac{1}{N_\Theta}\sum_{\theta\in\Theta}\widehat{a}_\theta$. We can extend the definition of the distance $\delta(\theta,\theta')$ between two breakpoints, to the distance $\delta(\Theta,\Theta')$ between two groups of breakpoints.

$$\delta\left(\Theta,\Theta'\right) = \sum_{\tau=1}^{n} p_\Theta(\tau)\, p_{\Theta'}(\tau)\, \mathbf{1}_{\widehat{a}_\Theta \widehat{a}_{\Theta'}>0} \tag{A1}$$

Based on these definitions, we have

$$\delta\left(\Theta,\Theta'\right) = \frac{1}{N.N'}\sum_{\tau=1}^{n}\sum_{\theta\in\Theta}\sum_{\theta'\in\Theta'} p_\theta(\tau)\, p_{\theta'}(\tau)\, \mathbf{1}_{\widehat{a}_\theta \widehat{a}_{\theta'}>0} \tag{A2}$$

Indeed, we have that for any $\theta\in\Theta$ and $\theta'\in\Theta'$, $\mathbf{1}_{\widehat{a}_\theta \widehat{a}_{\theta'}>0} = \mathbf{1}_{\widehat{a}_\Theta \widehat{a}_{\Theta'}>0}$ because the clustering criterion guarantees that all the breakpoints within a given group have the same sign. By changing the order of the summation in Equation (A1) we obtain $\delta\left(\Theta,\Theta'\right) = \sum_{\theta\in\Theta}\sum_{\theta'\in\Theta'}\delta\left(\theta,\theta'\right)/N.N'$, i.e. the expression found in Equation (15).

Second, let us assume for a moment that each breakpoint $\theta\in\Theta$ has a unique deterministic position τ_θ. Then, it would be possible and straightforward to derive a likelihood associated to a given group Θ by forming the product $\Pi_{\theta\in\Theta}\,\ell(\theta)$, where the contribution of each θ to this product would simply be $\ell(\theta) = p_\Theta(\tau_\theta)$. That, because p_Θ is the pdf estimated on the subset Θ to which θ belongs. Nonetheless, θ has a probabilistic position hence we cannot use this expression directly. Instead, we can use its expected value $\mathbb{E}\left(p_\Theta(\tau)\,|\theta\right)$, instead of $p_\Theta(\tau_\theta)$, in order to obtain a pseudo-likelihood $\ell(\theta) = \mathbb{E}\left(p_\Theta(\tau)\,|\theta\right) = \sum_{\tau=1}^{n} p_\Theta(\tau)\, p_\theta(\tau)$. The log of the likelihood associated to the group Θ is then obtained as

$$\begin{aligned}\log\ell(\Theta) &= \log\Pi_{\theta\in\Theta}\,\ell(\theta) = \sum_{\theta\in\Theta}\log\ell(\theta)\\ &= \sum_{\theta\in\Theta}\log\left\{\frac{1}{N_\Theta}\sum_{\theta'\in\Theta}\delta\left(\theta,\theta'\right)\right\}\end{aligned} \tag{A3}$$

from which the pseudo-log likelihood of Equation (16) is obtained.

References

Abarca-Del-Rio R, Mestre O. 2006. Decadal to secular time scales variability in temperature measurements over France. *Geophys. Res. Lett.* **33**: L13705, DOI: 10.1029/2006GL026019.

Alexandrov V, Schneider M, Koleva E, Moisselin J-M. 2004. Climate variability and change in Bulgaria during the 20th century. *Theor. Appl. Climatol.* **79**: 133–149.

Auer I, Bohm R, Jurkovic A, Lipa W, Orlik A, Potzmann R, Schšner W, Ungersbšck M, Matulla C, Briffa K, Jones PD, Efthymiadis D, Brunetti M, Nanni T, Maugeri M, Mercalli L, Mestre O, Moisselin JM, Begert M, Mÿller-Westermeier G, Kveton V, Bochnicek O, Stastny P, Lapin M, Szalai S, Szentimrey T, Cegnar T, Dolinar M, Gajic-Capka M, Zaninovic K, Majstorovic Z, Nieplova E. 2006. HISTALP - Historical instrumental climatological surface time series of the Greater Alpine Region. *Int. J. Climatol.* **25**: 139–166, DOI: 10.1002/joc.1377.

Beaulieu C, Ouarda TBMJ, Seidou O. 2007. A review of homogenization techniques for precipitation data and their applicability to precipitation series (in French). *Hydrol. Sci. J.* **52**(1): 18–37.

Camilloni I, Barrucand M. 2011. Temporal variability of the Buenos Aires, Argentina, urban heat island. *Theor. Appl. Climatol.* **52**(1): 18–37.

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

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

Conrad V, Pollack LW. 1962. *Methods in Climatology*. Harvard University Press: Cambridge, MA.

Davis AR, Lee TCM, Rodriguez-Yam GA. 2006. Structural break estimation for nonstationary time series models. *J. Am. Stat. Assoc.* **101**: 473.

Della-Marta PM, Collins D, Braganza K. 2004. Updating Australia's high quality annual temperature dataset. *Aust. Meteor. Mag.* **53**: 277–292.

Drogue G, Mestre O, Hoffmann L, Iffly J-F, Pfister L. 2005. Recent warming in a small region with semi-oceanic climate, 1949–1998: what is the ground truth? *Theor. Appl. Climatol.* **81**: 1–10.

Easterling DR, Peterson TC. 1995. A new method for detecting undocumented discontinuities in climatological time series. *Int. J. Climatol.* **15**: 369–377.

Fearnhead P. 2006. Exact and efficient Bayesian inference for multiple changepoint. *Stat. Comput.* **16**: 203–213.

Hannart A, Naveau P. 2009. Bayesian multiple change-points and segmentation: application to homogenization of climatic series. *Water Resour. Res*. **45**(10): 1944–1973.

Jain A, Dubes R. 1988. *Algorithms for Clustering Data*. Prentice-Hall: Englewood Cliffs, NJ.

Kuglitsch FG, Toreti A, Xoplaki E, Della-Marta PM, Luterbacher J, Wanner H. 2009. Homogenization of daily maximum temperature series in the Mediterranean. *J. Geophys. Res.* **114**: D15108.

Lavielle M, Lebarbier E. 2001. An application of MCMC methods for the multiple change-points problem. *Signal Process.* **81**: 39–53.

Lawrimore J, Rennie J, Thorne P. 2013. Responding to the need for better global temperature data. *Eos, Trans. Am. Geophys. Union* **94–6**: 61–62.

Lee ASF, Heghinian SM. 1977. A shift of the mean level in a sequence of independent normal random variables – a Bayesian approach. *Technometrics* **19**: 503–506.

Menne MJ, Williams CN Jr. 2009. Homogenization of temperature series via pairwise comparisons. *J. Clim.* **22**: 1700–1717.

Moberg A, Alexandersson H. 1997. Homogenization of Swedish temperature data. Part II: homogenized gridded air temperature compared with a subset of global gridded air temperature since 1861. *Int. J. Climatol.* **14**: 35–34.

Reeves J, Chen J, Wang XL, Lund R, Lu Q. 2007. A review and comparison of change-point detection techniques for climate data. *J. Appl. Meteor. Climatol.* **46**(2): 900–915.

Rusticucci M, Barrucand M. 2004. Observed trends and changes in temperature extremes over Argentina. *J. Clim.* **17**(20): 4099–4107.

Rusticucci M, Marengo J, Penalba O, Renom M. 2009. Observed temperature and precipitation extreme indices and modeled by the IPCC AR4 models in South America. *Clim. Change* **98**(3–4): 493–508.

Schwarz G. 1978. Estimating the dimension of a model. *Ann. Stat.* **6**: 461–464.

Thorne PW *et al.* 2011. A quantification of the uncertainty in historical tropical tropospheric temperature trends from radiosondes. *J. Geophys. Res.* **116**: D12116.

Venema VO, Mestre EA, 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, Brandsma T. 2011. Description of the COST-HOME monthly benchmark dataset and the submitted homogenized contributions. Report, Meteorological Institute, University of Bonn, Germany, http://www2.meteo.uni-bonn.de/venema/articles/2011/reporthome.pdf

Vincent LA, Peterson TC, Barros VR, Marino MB, Rusticucci M, Carrasco G, Ramirez E, Alves LM, Ambrizzi T, Berlato MA, Grimm AM, Marengo JA, Molion L, Moncunill DF, Rebello E, Anunciação YMT, Quintana J, Santos JL, Baez J, Coronel G, Garcia J, Trebejo I, Bidegain M, Haylock MR, Karoly D. 2005. Observed trends in indices of daily temperature extremes in South America 1960–2000. *J. Clim.* **18**(23): 5011–5023.

