Hindawi Publishing Corporation
Journal of Climatology
Volume 2013, Article ID 390945, 15 pages
http://dx.doi.org/10.1155/2013/390945

Research Article

Efficiencies of Inhomogeneity-Detection Algorithms: Comparison of Different Detection Methods and Efficiency Measures

Peter Domonkos

Centre for Climate Change, University of Rovira i Virgili, Campus Terres de l'Ebre, Avenue Remolins 13-15, 43500 Tortosa Tarragona, Spain

Correspondence should be addressed to Peter Domonkos; peter.domonkos@urv.cat

Received 18 March 2013; Accepted 2 September 2013

Academic Editors: S. Feng, L. Makra, and A. P. Trishchenko

Efficiency evaluations for change point Detection methods used in nine major Objective Homogenization Methods (DOHMs) are presented. The evaluations are conducted using ten different simulated datasets and four efficiency measures: detection skill, skill of linear trend estimation, sum of squared error, and a combined efficiency measure. Test datasets applied have a diverse set of inhomogeneity (IH) characteristics and include one dataset that is similar to the monthly benchmark temperature dataset of the European benchmarking effort known by the acronym COST HOME. The performance of DOHMs is highly dependent on the characteristics of test datasets and efficiency measures. Measures of skills differ markedly according to the frequency and mean duration of inhomogeneities and vary with the ratio of IH-magnitudes and background noise. The study focuses on cases when high quality relative time series (i.e., the difference between a candidate and reference series) can be created, but the frequency and intensity of inhomogeneities are high. Results show that in these cases the Caussinus-Mestre method is the most effective, although appreciably good results can also be achieved by the use of several other DOHMs, such as the Multiple Analysis of Series for Homogenisation, Bayes method, Multiple Linear Regression, and the Standard Normal Homogeneity Test.

1. Introduction

The underlying climate signal in observed in situ climatic data is often masked either by changes in observational practices, exposure, and instrumentation or by local changes in the environment where the observations are taken. If these changes (called inhomogeneities (IH)) have a significant impact on the statistical characteristics of the observed data, then the time series are inhomogeneous, and their usefulness is limited in assessments of observed climate change. Given that almost any long climate series is potentially inhomogeneous, various techniques have been developed to detect and adjust series where necessary (see, among others, the seminar series of Homogenisation and Quality Control in Climatological Databases, WMO-HMS [1–6]).

There are several options to eliminate the IHs from observed time series. The timing and cause of many potential IHs are documented in network management documents (so-called metadata). Good metadata information greatly facilitates the development of appropriate corrections for inhomogeneous time series, so that the time series can be made more suitable for climate studies. However, metadata is generally incomplete ([7–12], etc.) or at least cannot be assumed to be comprehensive. Thus, the benefit of metadata [13] cannot always be fully exploited in practice, and "even with the best possible metadata, some statistical inhomogeneity detection is advised" [14].

The need for efficient homogenization methods has encouraged researchers to create and use various statistical tools. As a result, nearly twenty statistical homogenization methods are in use. Taking into consideration the number of options available in how these methods are used and the parametric choices within a particular DOHM, the diversity of the applied DOHMs is even greater in practice.

Although several reviewing papers about homogenization methods have been published in the recent years

([13–19], etc.), a realistic evaluation of the advantages and possible disadvantages of homogenization procedures remains elusive. Sometimes even the principles are questioned. For example, in an investigation of real and simulated time series of radiosonde data with some arbitrary selected DOHMs, [20] found that the rate of false detections was usually higher than that of the correct IH detections. While this example is relevant only to the special properties of radiosonde data, results like this may lower confidence more generally in homogenization procedures.

Most homogenization studies focus on the detection and correction of the shifts in the monthly, seasonal, and annual means, since the correction of biases in section means is the most important for the reliable estimation of climate trends and low frequency climate variability. The homogenization of daily data is an even more complex problem as the homogenization on longer time scales, but the number of daily homogenization methods has been growing fast recently ([21–23], etc.), from which the more accurate estimation of extreme value statistics is expected. In this study the background noise of the simulated time series is white noise, and the inhomogeneities in them are changes (biases) of the section means by definition; thus, the results and conclusions are primarily applicable for monthly and annual homogenization.

Detection of IHs is usually conducted via relative homogeneity testing (e.g., the differences between two series are examined for breaks). In that case, detected IHs are supposed to belong to the so-called candidate series, although the chance that they at least partly belong to some reference series usually cannot be ruled out. To keep this risk low, application of statistical homogenization can be recommended only when several time series of the same geographical-climatic region are available [7, 24–27], and many of the spatial correlations are higher than 0.7 [13, 18, 28–32]. These conditions are generally true for the surface air temperature datasets in the extratropical land areas, and in most cases also for precipitation datasets. Monthly and annual temperature time series in Europe and the USA usually have dense networks, and the spatial correlations are often around or above 0.9 for them [12, 33]. In these networks the use of relative time series is advantageous since there is no comparable alternative for eliminating the impacts of local IHs from observed datasets.

Here, we examine the efficiencies of DOHMs as applied to series simulated to represent the test series in a relative homogenization approach. Only objective methods are tested that can be applied in fully automated way. The study does not address the mechanics of creating relative time series (differences or ratios) from raw climate observations nor do we consider the implications of using iteration in detection and correction even though these aspects may significantly impact the ultimate efficiency in practice. Rather, the rationale is that the change point detection components of DOHMs should be analyzed separately from the complete homogenization procedures, (i) because the most efficient detection parts can then likely be paired with spatial comparison and iteration segments of any other homogenization methods, and (ii) considering only the complete homogenization procedure without addressing specific components tends towards a "black box" approach whereby the advantages and disadvantages of particular elements are difficult to identify.

Although there have been made some comparative examinations aiming to reveal the capability of detecting IHs by different homogenization methods [16, 30, 34–45], all these studies examine some arbitrary selections of DOHMs, and the test datasets used mostly do not have realistic statistical properties. One of the fundamental open questions is the role of similarities and dissimilarities between the statistical properties of real and simulated time series, a topic which has only been discussed by Menne and Williams Jr. [16, 40], Domonkos [41, 42], Titchner et al. [45], and Venema et al. [46]. This topic is intensively discussed in this study relying on the empirical efficiencies from various simulated datasets.

The methodology of the present study mostly follows the rules introduced in [42]. The new lines of investigations in the present study in revealing the performances of DOHMs are as follows: (i) comparisons between test series with randomly positioned change points on the one hand and those with short term, platform shaped biases on the other hand; (ii) comparisons between test series with relatively large shift sizes (as they are in [46]) on the one hand, and those with mostly moderate shift sizes (as which were found empirically in [42]) on the other hand; (iii) the role of empirical autocorrelation in relative time series; (iv) experiments with moving signal-to-noise ratio. After presenting the results, the reality of the test datasets used will be discussed along with some peculiarities of the results, and we will make some comparisons between the blind test results of [46] and our results.

2. Methods

2.1. Concepts and Definitions. The efficiencies of DOHMs are quantified using ten test datasets as benchmarks. All the simulated time series that comprise the benchmarks were generated to mimic the properties of time series of differences derived by the comparison of one candidate series with some IHs and a reference series of "good quality" (i.e., without IHs). They are referred to as relative time series (**X**). Specifically, each dataset comprises N relative time series of n year length as follows:

$$\mathbf{X}_p = \left[x_{p,1}, x_{p,2}, \ldots x_{p,n}\right]^T, \quad p = 1, 2, \ldots N. \tag{1}$$

In this study $n = 100$ and $N = 10,000$. All the time series contain a standard white noise process (**W**) whose standard deviation equals 1, as well as a term for cumulated effects of IHs, named also station effect (**H**).

The origin of the noise is the natural fluctuation of spatial differences of climatic elements. The relative time series generally do not contain low frequency noise, because in a particular observing network of the same climatic region, the low frequency climatic changes tend to be common. However, this assumption is not exactly true; so two of the test datasets were simulated to have characteristics with slight deviations from this rule (see Section 2.3).

Each element of time series (the index p will not be in use hereafter) can be expressed as a sum of the interstation noise (w) and station-specific IH effects (h) as follows:

$$x_i = w_i + h_i, \quad i = [1, 2, \ldots n]. \tag{2}$$

X always represents raw time series (before homogenization), while **U** represents homogenized time series. If homogenization is perfect, then **U** ≡ **W**.

Always the first moment (section-average) of the time series is biased by the imposed IHs in the simulations. Three types of IHs are used, namely, (i) change point (sudden shift), (ii) trend (gradual change), and (iii) platform (pair of sudden shifts). These are defined in more detail as follows.

(i) Change point: if $h_{i+1} \neq h_i$ and the change is not a part of a gradual change (cf. (iii)), then a change point type IH exists at time i.

(ii) Trend: gradual change of h over a period $[j, k] (1 \leq j < k \leq n)$. The artificial trends are always linear ($h_{i+1} - h_i = h_i - h_{i-1}$ for each i ($i \in [j+1, k-1]$)), and their minimum duration is 5 years.

(iii) Platform: a pair of change points of the same magnitude, but with different signs. If $h_{i+1} \neq h_i$ and $\exists k$, ($k \in [i+1, n]$) for which $h_{k+1} - h_k = -(h_{i+1} - h_i)$, then a platform exists whose first year is $i+1$ and last year is k. When $k - i$ represents a relatively short time period (say, $k - i \leq 10$), platforms are also referred to as short-term IHs. From this point of view, outliers are also platforms with 1 year duration.

Note that although from the combination of IHs type (i) and type (ii), IH of any shape could be constructed; type (iii) is included as a distinct type in the simulation process, because an earlier study [42] showed that the shifts of successive change points often have the opposite signs.

The absolute value of an IH is considered to be its "magnitude", while the magnitude with the sign is considered to be its "size." During the simulation, IH magnitudes (m) and other statistical characteristics are expressed with their ratio to the standard deviation of the white noise (s_e), while detected magnitudes (m^*) are expressed with their ratio to the estimated standard deviation of the white noise (s_e^*) as follows:

$$\begin{aligned} s_e^* &= \sqrt{1 - R^2} \cdot s_T \quad \text{if } R > 0, \\ s_e^* &= s_T \qquad\qquad\quad \text{if } R \leq 0. \end{aligned} \tag{3}$$

In (3), R denotes 1-year lag autocorrelation, and s_T means the empirical standard deviation of the time series. The application of the unit s_e^* follows from the fact that during the detection process s_e is known only for simulated time series, while for relative time series from real observations this characteristic is unknown. In contrast, s_e^* can easily be calculated for any time series. s_e^* is usually higher than s_e but never higher than s_T. Thus s_e^* is a better estimation of s_e than s_T would be.

The detection processes are always paired with a standard adjustment procedure (SA) in this study. Let us suppose that Q change points have been detected with timings $t_1, t_2, \ldots t_Q$, as well as $t_0 = 0$ and $t_{Q+1} = n$ by definition. The segment between adjacent change points t_k and t_{k+1} is denoted by K and segment means with upper stroke. For the SA, $m^*(t_k)$ is calculated as the difference of the adjacent segment means around t_k. If there is a detected trend with e_K annual change for section K, it is taken into account in the calculation according to

$$\begin{aligned} m^*(t_k) = {} & \overline{x_K} - \overline{x_{K-1}} \\ & - 0.5\left(e_K\left(t_{k+1} - t_k\right) + e_{K-1}\left(t_k - t_{k-1}\right)\right). \end{aligned} \tag{4}$$

Then the adjustment of m^* is applied for all x_i of $i \leq t_k$. Derivation of m^* and SA is applied for trend IHs with the same logic, but taking into account the gradualness. Note that the estimation of m^* in the various homogenization approaches may differ from the SA used here; however, the uniformity of adjustment technique applied in this study is essential for testing the performance of detection parts separately from other properties of homogenization methods.

Finally, in the discussion below, f stands for the average frequency of IHs (i.e., the number of events in 100 years), while s equals the standard deviation of the IH sizes.

2.2. DOHMs Examined. The nine DOHMs that we examine (Table 1) are widely used in climatology. All they are objective methods, hence they can be applied automatically to find IHs in time series. Only one nonparametric method, the Wilcoxon Rank Sum test, was selected because the efficiencies of nonparametric methods are more limited (see, e.g., [37]).

The significance thresholds used attempt to ensure the 0.05 rate first type error (FTE) in pure white noise processes. The values are generally taken from the reference studies with some exceptions. For Bay, the Caussinus-Lyazrhi criterion [47] is applied, while for MLR and tts the thresholds are calculated with Monte-Carlo technique. In C-M and MAS some kind of joint detection of multiple IHs is applied, while in tts and in the second phase of E-P single change points are searched in subsections of the time series. The cutting algorithm [35] is applied when reference studies do not give other proposal for treating multiple IHs (i.e., for Bay, MLR, SNH, SNT, WRS, and in the first phase of E-P). In the cutting algorithm only one change point can be detected at a specific step, but cutting the time series into two parts at the timing of detected change points, multiple IHs can be detected. Subsection examination and cutting algorithm need the definition of the minimum length of subperiods for searching change points in them, it is 10 years in this study after Easterling and Peterson [35] and Moberg and Alexandersson [48]. The shortest period between two adjacent detected change points is 5 years for such methods. In SNT the detected IHs are always trends when the estimated duration of change is at least 5 years and always change points in the reverse case. The version of MLR used here differs in one more detail from the original description, that is, only 1-year-lag autocorrelations are considered in calculating FTE (instead of all lags between 1 year and 3 year). This modification has no substantial effect on the performance of MLR (not shown).

Table 1: DOHMs with their abbreviations used in the study.

DOHM	Abbreviation	Reference
Bayes method	Bay	Ducré-Robitaille et al. 2003 [37]
Caussinus-Mestre method (PRODIGE)	C-M	Caussinus and Mestre 2004 [53]
Easterling-Peterson method	E-P	Easterling and Peterson 1995 [35]
Multiple Analysis of Series for Homogenisation	MAS	Szentimrey 1999 [67]
Multiple Linear Regression	MLR	Vincent 1998 [68]
Standard Normal Homogeneity Test for shifts only	SNH	Alexandersson 1986 [69]
Standard Normal Homogeneity Test for shifts and trends	SNT	Alexandersson and Moberg 1997 [70]
t-test	tts	Ducré-Robitaille et al. 2003 [37]
Wilcoxon Rank Sum test	WRS	Wilcoxon 1945 [71]

A uniform prefiltering of outliers is applied before the use of any DOHM. Anomalies from the average of the time series are considered to be outliers if their absolute values are higher than 4 standard deviations of the time series elements. This threshold is often used in practice (e.g., [9, 25, 49]). Detected outliers are replaced with an anomaly value of zero.

2.3. Test Datasets. Obviously, the higher the resemblance between the simulated and real statistical properties, the higher the confidence that the assessed efficiencies based on simulated datasets are valid for real climatic datasets. Unfortunately, the exact statistical properties of IHs occurring in real climatic time series are not known. In the benchmark surrogated dataset of the COST HOME project (hereafter: Benchmark, [46]) the mean frequency of IHs is 5 per 100 years in artificial test datasets. It is based on the experience that the frequency of detected IHs in long climatic time series is approximately 5 per 100 years [13, 50], and considering that the IHs with low magnitudes are more frequent than those with high magnitudes [16, 40, 42], the IH sizes have normal distribution with 0 expected value. However, it is only a rough approach to the true properties of observational time series; see more discussion about this problem in Section 4.1 of this study and in [42], as well.

The mathematical description of ten test datasets is presented below, together with some ideas about the motivation of creating them. When no specification for IH-type is given, the type is always change point. Trends are included in three datasets only, but in those three their role is much greater (about 25% of long-term biases) than in the Benchmark (2% of all IHs).

(A) CH1B0 refers to 1 Big Change point without any restriction for R (0). In this dataset, exactly one IH is included in each time series. Its timing (j) is 40 or 60, and $m = 3$. In this simple case it is easy to demonstrate the time function of station effect as

$$h_i = 0, \quad \text{if } i \le j,$$
$$h_i = 3, \quad \text{if } i > j, \tag{5}$$
$$1 \le i < n, \quad j = 40 \quad \text{or} \quad j = 60.$$

(B) CH5B0: The intension of this dataset is to mimic the Benchmark (note that $f = 5$ and $s = 0.8^{\circ}$C are the key characteristics for the raw time series of the Benchmark). In the simulation process of this dataset, the probability of the introduction of a new change point was 0.05 at each year. Thus the average number of change points in time series is 5, and the mean frequency per time series has binomial distribution. The sizes are normally distributed with a mean of zero and s of 3.5. The value of s came from the estimation that 0.8°C is likely 3-4 times higher than the typical s_e for true relative time series in dense observing networks.

(C) PF5B0 was generated in the same way as CH5B0, but instead of individual change points platforms were introduced to the time series, again with probability of 0.05 at each year. The length of the platform has uniform distribution between 1 year and 10 years. The sizes of IHs have normal distribution with a mean of zero and s of 3.5.

(D) HUSTR: this dataset is the "Hungarian standard," because its characteristics were provided by an empirical procedure in which the statistical properties of the detected IHs in test datasets were approached via series of experiments to the same properties for true relative time series [42]. Those relative time series were constructed from observed temperature series in Hungary. The time series of HUSTR include rather complex structures of randomly distributed IHs of different types (change points, platforms, and trends) and magnitudes, as well as noise that differs from white noise. In this dataset the number of IHs is high, and short-term platforms are particularly frequent. Some IH-sizes are large, but the majority of them is small. $R \ge 0.4$ in each time series. The full description of its generation is presented in [42]. The change point frequency and the standard deviation of the change point magnitudes are $f = 31.1$ and $s = 1.20$.

(E) HUST0: like HUSTR, but without any restriction for R, $f = 30.3$ and $s = 1.02$.

For the following four datasets s was set to be similar to the s of HUSTR and HUST0.

(F) CH5S0: the same as CH5B0, but $s = 1$.

(G) PF5S0: the same as PF5B0, but $s = 1$.

(H) CH5SR: it was generated in the same way as CH5S0, but only time series with $R \geq 0.4$ were retained.

(I) PF5SR: it was generated in the same way as PF5S0, but only time series with $R \geq 0.4$ were retained.

(J) CHPF0: "compromise dataset." The term compromise is used because this dataset contains both long-term IHs and short-term platforms, as well as a few trend IHs. The frequency of IHs ($f = 10.4$) is higher than in the Benchmark, but much lower than in the Hungarian standard. During its simulation a new IH is introduced with 7% probability at each year, more specifically with 3, 3, and 1% probability for change points, platforms, and trends, respectively. Data were simulated from 50 years before the starting point of time series until 50 years after the end of time series for ensuring the temporal uniformity of occurrences of IHs within the examined 100 years. The duration of trends has even distribution between 5 and 99 years, while that of the platforms is the same as in PF5B0. This dataset is examined with moving *s*.

Table 2 summarizes the properties of the IHs in the ten test datasets. The diversity of the shown characteristics serves well the objective of the paper, that is, to find conclusions that are not related to the specific IH-properties of a given test dataset.

2.4. Measures of Efficiency. Four kinds of measures are examined: (a) detection skill, (b) skill of linear trend estimation, (c) sum of squared errors (SSE), and (d) combined maximal bias (CMB).

Let the sum of correct detections, that of false detections, and the total number of change points be denoted by S_R, S_F, and S, respectively. Although these concepts are clear in case of one or a few of fairly large IHs, their occurrences are not easy to be identified in complex structures. Therefore, the concepts "change point," "correct detection," and "false detection" must be defined for a quantitative and objective evaluation. The application of some arbitrary parameters is unavoidable for these definitions.

A *true change point* exists in time series **X** at year j ($3 \leq j \leq n-3$), if

$$\frac{1}{k}\left|\sum_{i=j-k+1}^{j} x_i - \sum_{i=j+1}^{j+k} x_i\right| \geq 2, \quad \text{for each } k \text{ of } k = \{1,2,3\}. \tag{6}$$

Equation (6) means that change points with magnitude (m^*) at least 2 are considered only, and the shift of this magnitude must be apparent comparing each symmetric half-window pairs, up to window width of 6 years.

Correct detection: there is a detected change point at year j with $m^* \geq 1.5$, and a true change point with a shift of the same sign as the detected IH has, really exists in section $[j-1, j+1]$ of **X**.

False detection: there is a detected change point at year j with $m^* \geq 1.5$, but no true change with the same direction occurs at all, taking into account any of the possible comparisons of section means for symmetric half windows around j up to window width of 6 years in **X**. There is no minimum threshold here for the magnitudes of true changes, only their signs are considered.

(a) *Detection skill* (E_D):

$$E_D = \frac{S_R - S_F}{S}. \tag{7}$$

Pieces of the detection result that do not meet with the conditions of either the correct detection or the false detection are not taken into account in the calculation of the detection skill. For perfect detection $E_D = 1$. In case of half of the detected change points are false, $E_D = 0$. Note that in datasets with very few change points (S is small) E_D can easily be negative.

For the following three measures error terms will be defined first, following them the way of their conversion to efficiency measures.

(b) *Error of linear trend estimation*: linear trends are fitted to the time series with minimizing the SSE between the trend line and the annual values of time series. The fitting is accomplished both for **U** and **W**, and the one for **W** is considered to be perfect. The differences between these two slopes are error terms. The procedure was accomplished for the whole (100 year long) time series, as well as for the last 50 years of the series. Thereafter the arithmetical average of these two errors is taken.

(c) *Sum of squared errors* (SSE):

$$\mathrm{SSE} = \sum_{i=1}^{n} (u_i - w_i)^2. \tag{8}$$

(d) *Combined maximal bias* (CMB): This measure evaluates the maximum difference between **U** and **W**, but in a way where detections with time-lapse error only are considered as partially right detections. When true IHs of **X** are detected right but with some time lapse, in CMB the detection is considered good, but a penalty term is applied for the time lapse. The penalization depends on the size of the time-lapse. Following this idea and comparing the annual values of **U** and **W**, the annual series of a combined error term **B** (the combination of size errors and time lapses) can be calculated. Naturally, when $u_i = w_i$, $b_i = 0$. The below formula shows the case, when $u_i > w_i$ (the reverse case is handled with the same logic rules):

$$b_i = \min_k \left(g_k + u_i - \min\left(u_i, \max_j\left(w_j\right)\right)\right), \tag{9}$$

where g_k is the penalty term of k-year lapse:

$$\begin{aligned} g_k &= \exp\left(c_1\left(k - c_2\right)\right) - c_3, \\ c_1 &= 0.369, \qquad c_2 = 3.297, \\ c_3 &= 0.2962; \quad j \in \left[j_1, j_2\right], \\ j_1 &= \max\left(1, i-k\right), \\ j_2 &= \min\left(i+k, n\right), \\ k &= [0,1,2,\ldots,15]. \end{aligned} \tag{10}$$

Table 2: Properties of the test datasets.

	f(CH)	f(trend)	f(PFa)	f(PFb)	f'	f	s	R
CH1B0	1.0	0	0	0	1.0	1.0	0	—
CH5B0	5.0	0	0	0	5.0	5.0	3.50	—
PF5B0	0	0	2.5	2.5	5.0	9.2	3.50	—
HUSTR	3.1	2.3	5.4	17.9	16.2	31.1	1.20	$R \geq 0.4$
HUST0	2.9	2.4	5.1	17.9	15.5	30.3	1.02	—
CH5S0	5.0	0	0	0	5.0	5.0	1.00	—
PF5S0	0	0	2.5	2.5	5.0	9.2	1.00	—
CH5SR	5.8	0	0	0	5.8	5.8	1.15	$R \geq 0.4$
PF5SR	0	0	3.9	2.7	7.8	12.2	1.34	$R \geq 0.4$
CHPF0	3.3	1.5	1.4	1.4	7.6	10.4	0.0. . .4.0	—

f(CH): frequency of change-points that are not parts of platforms, f(trend): frequency of trends, f(Pfa): frequency of platforms with longer than 5 years duration, f(Pfb): frequency of platforms with maximum 5 years duration, f': frequency of all but Pfb IHs, f: frequency of all IHs, s: standard deviation of IH-sizes, and R: restriction for the autocorrelation. Each frequency characteristic is shown as number per time series. Frequency characteristics show the frequency of introduced IHs during the generation, except for column f, where the empirical total frequencies are presented. (f is often slightly lower than the frequency of all introduced IHs, due to superposition of IHs or stretch-out over the end of the time series.)

In (9), the term $\max(w_j)$ indicates that each homogenized value (u_i) is compared with a true value for which the bias is minimal in the $2k$ wide window around i. This optimization is repeated applying different window width, but the penalty for time lapse exponentially increases with k. With the present parameterization the penalty for 3-year (4-year) lapse is 0.6 (1.0). After having b_i for each i, CMB is calculated as the difference between the extremes of b_i values as

$$\mathrm{CMB} = \max_{i,j} \left| b_i - b_j \right|, \quad i, j \in [1, 2, \ldots, n]. \tag{11}$$

The transformation of the error terms in (7), (8), and (9) to efficiency measures is as follows. Let the error terms and the efficiencies be denoted by q and E, respectively; then the general form of the connection between the error terms and efficiencies is given by

$$E = \frac{q(\mathbf{X}) - q(\mathbf{U})}{q(\mathbf{X})}. \tag{12}$$

In this way the maximal achievable value of E is always 1, and the sign of E shows whether a homogenization has resulted in quality improvement or not. Efficiency measures of trend detection, SSE, and CMB are denoted by E_T, E_S, and E_C, respectively.

The described four efficiency measures characterize DOHMs in different ways. Only E_D evaluates directly the detection results, while the other measures can be applied on adjusted time series. However, the usefulness of E_D is limited by the facts that (i) the calculation of E_D contains arbitrary parameters and (ii) the general purpose of homogenization is to achieve the highest reliability of trends and other characteristics of variability that are present in true time series [20], and not the identification of change points. Therefore the use of efficiency characteristics showing the performance in preserving or reconstructing the true climatic characteristics of time series such as E_T, E_S, and E_C is preferred rather than that of the detection skill [46, 51]. When E_T, E_S, and E_C are applied on evaluating DOHMs, the meanings of the obtained characteristics slightly differ from those for whole homogenization methods. The similarities and differences are characterized by the following peculiarities of time series homogenization. (i) The application of SA is optimal when the reference series is homogeneous. (ii) Although reference series are seldom homogeneous when real time series are homogenized, the bias in candidate series is often substantially larger than that in reference series. Therefore the good performance of DOHM+SA is a necessary, although not satisfactory, requirement from any homogenization method. (iii) The lack of the detection of IHs leaves the station effect to be inhomogeneous between two adjacent detected IHs on the one hand, while false detections shorten the sections between adjacent IHs unnecessarily reducing the sub samples for calculating sub section means on the other hand. The mentioned two problems reduce efficiencies, any kind of correction method is applied. The potential impact of these errors on the final homogenization results can be quantified by the remaining SSE after homogenization when the remaining SSE is influenced by detection errors only.

3. Results

3.1. Case Studies. Figure 1 presents the efficiencies showing the four kinds of characteristics in four distinct sections. The presentation starts with the characteristics of E_S followed by those of E_T, E_C, and E_D, in this order.

The simplest case is when only 1 change point is included in each time series. These time series can be treated most easily, but, unfortunately, the appearance of this case is not common in climatic datasets. Figure 1 shows that for CH1B0 all the DOHMs perform well, except for tts. The weakness of tts arises from its low detection power. As tts examines parts of time series separately, a relatively strict significance threshold has to be applied to keep the rate of FTE low, but it results in relatively poor detection power.

The results show that DOHMs usually perform well also for CH5B0. The largest IHs of CH5B0 can relatively easily be identified, owing to their characteristic magnitude

E_S (SSE)	Bay	C-M	E-P	MAS	MLR	SNH	SNT	tts	WRS
CH1B0	**98.6**	**97**	**95.8**	**98.6**	**98.9**	**98.7**	**98.6**	(*43.1*)	**98.8**
CH5B0	**97.3**	**97.4**	**94.3**	**97.2**	**94.2**	**97**	**95.2**	*74.8*	**94.5**
PF5B0	(*26.3*)	**60.3**	*31.6*	**58.7**	48.6	(*20.5*)	(*11.1*)	(*−24*)	(*−15*)
HUSTR	**76.7**	**78.8**	73.2	**77.9**	**76.8**	**76.4**	**75**	(*39.3*)	73.3
HUST0	**64.3**	**65**	**61.8**	**62.1**	**64.7**	**63.7**	**62.4**	53.1	**60.9**
CH5S0	**85**	**82.4**	73.4	**80**	78.4	**84.4**	**83.8**	(*18.1*)	**83.2**
PF5S0	**−51**	**−55**	−66	**−56**	**−56**	**−54**	**−56**	*−75*	−61
CH5SR	**91.9**	**90.2**	83.8	**89.8**	**88.5**	**91.4**	**90.9**	(*22.7*)	**90.7**
PF5SR	*5.3*	**23.5**	11.4	**19.5**	**20.3**	*1.6*	*−4*	(*−48*)	(*−17*)
E_T (Trend)	Bay	C-M	E-P	MAS	MLR	SNH	SNT	tts	WRS
CH1B0	**93**	**92.7**	86.7	**93.5**	**93.5**	**93.1**	**92.9**	(*39.1*)	**93.2**
CH5B0	**92.9**	**93.7**	**89.3**	**92.5**	**89.8**	**92.5**	**89.3**	(*57.3*)	**91**
PF5B0	*55.5*	**74.2**	63.1	**71.8**	**69.4**	*52.5*	*47.3*	(*23.5*)	(*38.8*)
HUSTR	**73.3**	**76.7**	69.4	**74**	**74.4**	**72.3**	69.8	(*22.9*)	71.3
HUST0	**59.2**	**62.7**	56.7	56.5	**59.6**	**58.9**	**57.8**	*44.1*	**57.7**
CH5S0	**73.7**	**75.2**	60.7	68.7	66.6	**72.9**	**71.2**	(*9.4*)	**72.5**
PF5S0	11	**20.3**	*3.2*	12.9	11.6	10.1	9.1	*−1*	7.3
CH5SR	**83.7**	**84.5**	73.7	**81.5**	**79.9**	**83.2**	**81.1**	(*14.4*)	**82.8**
PF5SR	*42.4*	**57.6**	48.8	**53.4**	**53.6**	*38.5*	*34.6*	(*8.1*)	*33.7*
E_C (CMB)	Bay	C-M	E-P	MAS	MLR	SNH	SNT	tts	WRS
CH1B0	**91.2**	**87.2**	*72.6*	86.4	**92**	**91.3**	**88.3**	(*38.5*)	**91.8**
CH5B0	**81.1**	**81.4**	75	**80**	72.8	**81.1**	72.4	(*41.6*)	**77.5**
PF5B0	20.9	**35.3**	24.5	**31.9**	24	*19.5*	*16.5*	(*−6*)	*5.9*
HUSTR	**45.5**	**44.8**	**42**	**44.2**	**45.2**	**45.5**	**45.3**	(*14.8*)	**43.8**
HUST0	**34.7**	**32.1**	**31**	26.1	**34.1**	**34.8**	**34.4**	26.2	**33.6**
CH5S0	**46.2**	40.5	*31.1*	35.6	40.3	**45.9**	**44.7**	(*6.7*)	**45.1**
PF5S0	**1.7**	**−2**	−9	−6	**1.2**	**1.2**	**1**	−4	**−1**
CH5SR	**59.1**	**55.1**	47.4	52.7	53.7	**58.7**	**56.7**	(*10.4*)	**57.8**
PF5SR	**22**	**26.7**	**25.8**	**26.5**	**26.7**	20.6	20.1	*1.2*	14
E_D (Detection)	Bay	C-M	E-P	MAS	MLR	SNH	SNT	tts	WRS
CH1B0	**96.1**	**91.6**	90.4	84	**96.2**	**96.2**	88.2	(*39*)	**94.7**
CH5B0	86.8	**92.7**	81.8	**88.8**	*67.8*	86.1	(*54.4*)	(*46.5*)	*74.9*
PF5B0	(*44.6*)	**92.6**	*62.6*	**88.2**	*62.7*	(*42.5*)	(*40.1*)	(*18.2*)	(*18.4*)
HUSTR	*54.8*	**79.3**	*53.1*	**77.1**	*58.7*	*52.1*	(*49.1*)	(*24*)	(*39.7*)
HUST0	(*25.8*)	**60**	*25.3*	**56.5**	*34.2*	(*29*)	(*27.2*)	(*9.5*)	(*20*)
CH5S0	*3.1*	(*−133*)	(*−109*)	(*−210*)	(*−13*)	*1.4*	**26.8**	13.5	*−1*
PF5S0	**17.9**	**18.7**	*−7*	(*−17*)	**21.1**	**18.6**	**18.2**	*5.8*	*5.5*
CH5SR	*23.5*	(*−47*)	(*−24*)	(*−85*)	(*6*)	*23.2*	**46.2**	*20.4*	(*11.8*)
PF5SR	*43.8*	**62.6**	56.6	**65.1**	55.5	*42.2*	*41.3*	(*12.9*)	(*17.3*)

Figure 1: Four efficiency characteristics for each DOHM with each test dataset. Bold and underlined: best DOHM of a particular test (BDOHM hereafter), bold: less than 5% lag behind BDOHM, italics: more than 15% lag behind BDOHM, and italics with brackets: more than 30% lag behind BDOHM. Lightness of cell background improves with growing efficiency.

and duration. Yet the average performance is usually lower for CH5B0 than for CH1B0. The highest performances are produced by C-M, but Bay, MAS, and SNH are nearly as good, and almost all DOHMs have an efficiency above 50%. E_S values are particularly high; they are usually well above 90%. The detection skill is relatively poor for SNT and MLR, its likely reason is that using these DOHMs sometimes trends are identified instead of two or more change points.

The only difference between PF5B0 and CH5B0 is that time series in PF5B0 include short-term platforms instead of long-term IHs. The frequency and the magnitude distribution of the events are unchanged (although the number of change points is doubled in this way, since a platform consists of two change points). This change of dataset properties has dramatic effect on the performances of DOHMs. For PF5B0 the majority of the efficiencies calculated are lower than 50%,

although they are still positive with few exceptions. C-M and MAS have considerably higher efficiencies than the other DOHMs have; MLR shows the third best performance, while tts and WRS produced the poorest results.

Examining the results of the Hungarian standard (HUSTR) one can see that the performances of the DOHMs are generally better than for PF5B0, but poorer than for CH5B0. C-M has the best results again, but the advantage of C-M and MAS is much smaller than for dataset PF5B0; moreover, Bay and MLR have similar or slightly better performance than MAS has for E_T and E_C. Considerably poorer results occur only with tts.

When the simulation method of the standard dataset is applied without limiting autocorrelation values (HUST0), the performances are poorer than for HUSTR, but the mean difference is moderated. The order of the skills does not change much either. In detection skill the C-M and MAS are much better than the other DOHMs, while for other skills the differences are small.

The results for CH5S0 show a spectacularly big difference in comparison with the previously analyzed results, namely, the detection skill of C-M and MAS is not the best in this case, but just the opposite relation can be seen. While most of the DOHMs perform near zero detection skill, those of MAS, C-M, and E-P are strongly negative. To understand this phenomenon, consider that (i) in CH5S0 the frequency of IHs of considerable size ($m^* > 2$) is very low, it is only 0.064 per 100 years on average; (ii) MAS, C-M, and E-P often have higher false alarm rate than the other DOHMs have (not shown). While in case of usual frequency of large IHs the high power of detection overbalances the drawback of relatively high false alarm rate; it does not operate for cases of rare IH occurrences. Interestingly, in spite of the large negative skills in E_D, the performances of C-M and MAS are still good for E_S, E_T, and E_C, and in trend detection skill C-M is the best.

Very different results were obtained using PF5S0. In this case most of the efficiencies scattered around zero, while SSE-reduction is not achieved by any of the DOHMs. This is the kind of results that would be good to avoid in the practical application of homogenization methods. Although the skills in preserving long-term linear trends are still slightly positive, the E_S results show the disruption of short-term variability. An interesting feature is the very big difference between the results of CH5S0 and PF5S0 (similarly, as between CH5B0 and PF5B0), since the magnitude distribution of the inserted IHs is the same for the two datasets. In PF5S0 the mean frequency of IHs with $m^* > 2$ is 0.166 per 100 years which is although significantly higher than that for CH5S0, it still indicates that most of the time series are free from IHs of large magnitudes. It can be seen that the change of the IH-form from solely shifts to short-term platforms resulted in the complete cessation of negative detection skills of C-M and MAS, but, on the other hand, dominant E_S values are dropped 140% approximately (from +80% to −60% in most of the methods).

The results with CH5SR and PF5SR prove that limiting the autocorrelation at 0.4 has a significant positive impact on the efficiencies. Although E_D values of MAS, C-M, and E-P are still negative, its importance seems to be minor relative to the high efficiencies in E_S and E_T. Large negative efficiencies related to platform type IHs of small size, as like with PF5S0, are not present with PF5SR except for tts, and most of the efficiencies are significantly positive.

The general differences between the performances of individual DOHMs when they are applied to the same test datasets in the nine case studies are as follows. Apart from tts the DOHMs tend to show similar efficiencies concerning E_T, E_S, and E_C, while the differences are generally larger in detection skill. C-M has generally the highest performance, except when there is only one IH in the time series or when the signal to noise ratio is very unfavorable. The skill of C-M is particularly good in preserving long-term climatic trends, which is given by the fact that the E_T of C-M is always positive, and it is always the highest in comparison with the E_T values of other DOHMs with the only exception of CH1B0. MAS is most often the second best DOHM, while the following three places in the rank order are for Bay, MLR, and SNH. Bay often, SNH several times produced better results than MLR, but the performance of MLR seems to be more uniform for different datasets than that of Bay and SNH. tts (WRS) performs markedly (slightly) poorer than the other DOHMs examined.

3.2. Sensitivity to IH Magnitudes. Experiments with compromise-form datasets (CHPF0) were performed applying increasing s from 0 to 4, and the remaining SSE after homogenization were calculated in relation to the background noise (s_e^*). In Figure 2 the results of tts are not included, because the SSE of tts are excessively large. For large signal to noise ratio ($s > 2$) the rank order is similar to what was dominant in the case studies: C-M has the best performance, followed by MAS, Bay, SNH, and MLR. For moderate signal to noise ratio ($1 < s < 2$) C-M is still the best; it is followed by Bay, but all the performances have little variation, that is, they hardly depend on the choice of the DOHM. When the signal to noise ratio is small ($s < 1$), the C-M is not the best, but SSE are generally small with any of the examined DOHMs, except with E-P they are slightly larger.

4. Discussion

4.1. Appropriateness of Test Datasets. The creation of test datasets with realistic statistical properties needs the knowledge of the relation between the characteristics of detected and true IHs. This relation is examined applying various DOHMs on two test datasets (Figure 3). Although the detected frequency (f^*) depends on the DOHM applied, the difference from the true frequency is always negative, because very small IHs cannot be detected by any of the DOHMs. This negative difference is even more striking when the IHs are short-term platforms (Figure 3(b)). From these results it is clear that the direct observation of the "best" IH-properties through the examination of observational datasets is not possible. The properties of Benchmark were set by expert decisions of experienced homogenizers, but when time series contain large number of hardly detectable IHs, experts' estimations might be biased. When HUSTR was constructed,

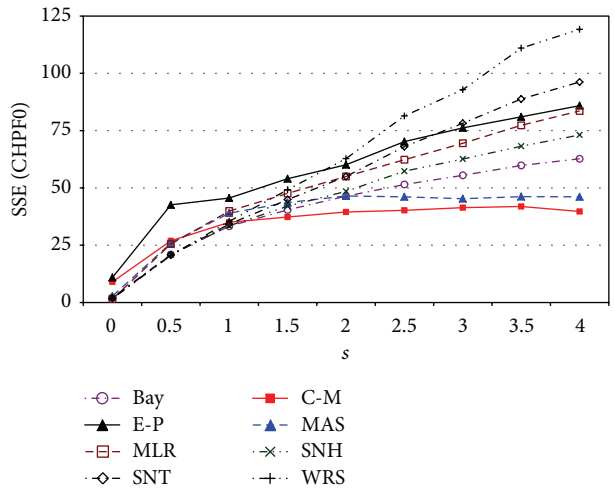

FIGURE 2: Remaining sum of squared errors (SSE) after the homogenization of CHPF0 with various DOHMs and using ideal reference series. The unit of SSE is the estimated standard deviation of background noise (s_e^*).

a distinct approach to the assessment of IH-properties was applied relative to the Benchmark. The main idea was to find IH-populations for which the detection results are similar to the detection results from real climatic time series. This approach could provide more reliable assessments of true IH properties than earlier studies, but further examination is probably needed because the very high frequency of detected short-term platforms in HUSTR may have origins other than IHs, for example, the natural variability of spatial temperature-gradients.

Figure 4 shows the magnitude distributions for change point type IHs in three datasets. It can be seen that the amount of small IHs is the highest in HUSTR, though it must be noted that there is no way to be assured about the reality of the amount of very small IHs ($m^* < 1$), because they have little impact on the detection results. For $m^* > 2$ the relation between CH5B0 and HUSTR is reversed, that is, CH5B0 contains more medium-size and large IHs than HUSTR does. Note that the differences are large, since the scale is logarithmic. Figure 4 illustrates also that the general failure with homogenizing PF5S0 is due to the lack of large IHs, beyond the short duration of IH caused biases in that dataset.

Relying on [42], the use of HUSTR-like test datasets can be favored. Taking into account that an unknown portion of IHs in HUSTR might have other sources than true IHs, datasets with smaller frequency of platforms than in HUSTR but with higher frequency of them than in the Benchmark could be closer to the reality, and CHPF0 is a realization of this idea. Anyhow, the author does not state that all kinds of climatic time series could be well represented with one or two test datasets of specific properties, but he suggests that the use of different kind test datasets, especially ones including large number of short-term IHs, is essentially important in testing efficiencies of IH detection algorithms.

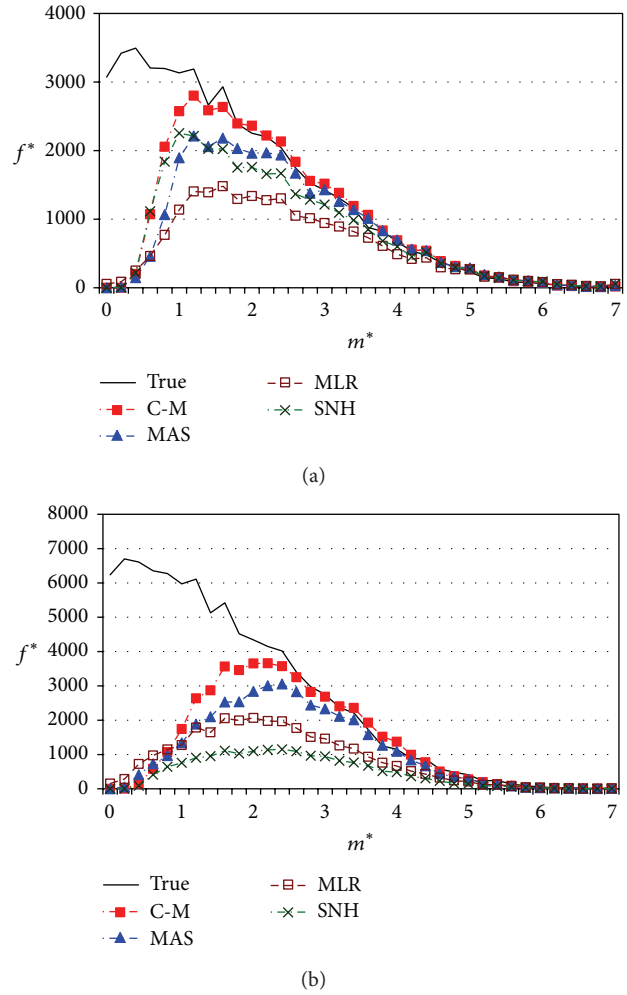

FIGURE 3: Magnitude distribution of true and detected IHs in test datasets, (a) (upper) CH5B0 and (b) (bottom) PF5B0. Frequencies (f^*) are shown using an arbitrary unit.

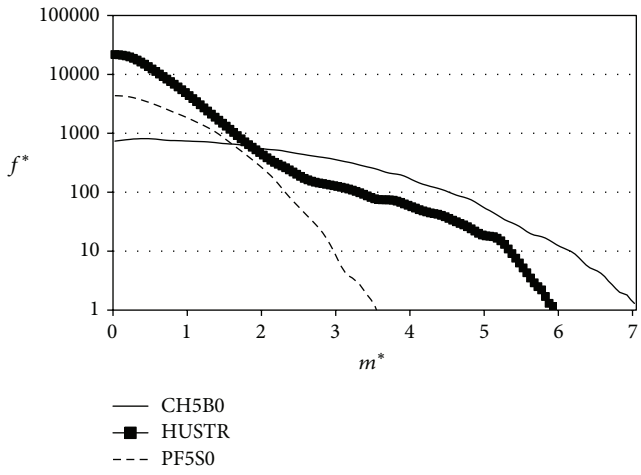

FIGURE 4: Magnitude distribution of IHs in three test datasets. Frequencies (f^*) are shown using an arbitrary unit.

4.2. Efficiency of DOHMs. The most important finding is that the performance of DOHMs strongly depends on the properties of the used test datasets, and thus the results are varied. (A similarly high diversity of results using simulated radiosonde datasets is described in [45].)

Nevertheless, some consistencies are evident in the results shown here. For example, efficiencies are always higher (a) for datasets with random sequences with only change points than for those with short-term platforms, (b) for datasets with time series of high autocorrelation than for datasets without restriction of R, and (c) for datasets with large IHs compared to those with low spread of IH-magnitudes. The apparent diversity of efficiencies is likely realistic; first because the frequency and intensity of IHs are diverse in practice, and second, because the networks of observed data have different spatial correlations causing strong diversity of signal/noise ratios.

In some earlier studies [41, 42] it was found that DOHMs with joint detection of multiple IHs perform better than other DOHMs. In this respect the present study confirms the earlier findings. The results presented here show that C-M is generally the best DOHM, but there are several other DOHMs whose performances are usually not much poorer. DOHMs of the best performances include joint detection of multiple IHs (C-M and MAS) or cutting algorithm (Bay, MLR, and SNH). By contrast, sequential detection of IHs using moving windows cannot be recommended which is proven by the fact that the performance of tts is always substantially poorer than using DOHMs with joint detection or cutting algorithm. The cause of the failure with tts is that the use of time windows results in distinct decisions about the significances of individual IHs, disregarding the data in other parts of the time series. Note that E-P also includes examinations for sections of limited length, and likely it does not effect positively its final performance. The traditional way of correcting inhomogeneous series is to apply the bias term on the candidate series with the opposite sign as it was detected in the relative time series. In examinations with limited window width the biases are usually assessed also with such windows [9, 35], and the adjustments are derived from such assessments. However, as the temporal coherence of such bias estimations is poor, it often results in particularly poor final performance in preserving the true climatic variability of time series, in spite of the fact that the detection skill of change points can be better than applying the SA (not shown). Examinations of arbitrary separated sections can be useful when the final objective is to make decisions about the significance of one or few potential IHs, but not for the task of general statistical homogenization. Among DOHMs with cutting algorithm the nonparametric WRS performed the poorest.

All the results show that the performance of MAS is very similar to that of C-M. Note that the performance of MAS is not better with its originally suggested correction technique than with SA (not shown), at least in the model task of this study (i.e., in case of homogeneous reference series).

An interesting feature of Figure 2 is that when time series are homogenized with C-M or MAS the increase of SSE with rising s stops around $s = 2$. It promises that applying one of these DOHMs, the remaining uncertainty of homogenized time series can be assessed, since that does not depend on the magnitudes of IHs in the raw dataset. However, it should be noted that that problem is still complex, since the remaining SSE depends on the frequency, shape, and temporal structure of IHs.

In an earlier study [41], moving parameter examinations were performed for DOHMs, in which significance thresholds and the shortest period allowed between two adjacent change points were varied in homogenizing HUSTR. The main findings of that study are that when time series are presumed to have large IHs, the optimal significance thresholds are lower than those which are generally applied, and the optimal shortest period is 2-3 years for C-M and MAS (1 year in this study) and 3–5 years (most often 4 years) for other DOHMs (5 years in this study). On the other hand, that study also showed that the optimization of parameters does not have robust effect on the performance of DOHMs. In this study we did not apply the parameters optimized on HUSTR by Domonkos [41] because the true optimums obviously depend on the properties of test datasets.

Most of the examined test datasets do not contain trend type IHs. MLR and SNT might be expected to perform better when time series contain more trends. However, the variation of trend frequency in HUSTR (not shown) indicated that raising the trend ratio within long-term IHs hardly influences the rank-order of efficiencies. Even with a 70% trend ratio, the MLR and SNT did not perform better than C-M, the only exception being that the E_C of SNT became to be the best among DOHMs. On the other hand, the performance of SNT is generally slightly poorer than that of C-M, MAS, Bay, MLR, and SNH.

Most DOHMs are applied with 0.05–010 FTE except E-P. In E-P the significance thresholds that are suggested by Easterling and Peterson [35] are applied. Lund and Reeves [52] presented the mathematical explanation why these thresholds are not restrictive enough for keeping FTE low. However, the application of stricter significance thresholds substantially worsens the performance of E-P in E_S, E_T, and E_C (not shown); therefore, we did not change the original parameterization.

Our results indicate that DOHMs with joint detection of multiple IHs, and particularly C-M, perform better than the other DOHMs except when the signal to noise ratio is too low. C-M is a maximum likelihood method, and the step function fitting in that aims at the minimization of the residual SSE [53]. The number of steps is determined by the Caussinus-Lyazrhi criterion, which is based on the information theory [47]. However, the methodological development of DOHMs has not been terminated with the creation of C-M. A further improvement of performance is expected from the network-wide joint detection of IHs [54] and from the harmonization of work on monthly and annual time scales [55], not mentioning here the problem of daily data homogenization which requires the development of distinct methods relative to the homogenization of annual and monthly data.

While in the COST HOME project full homogenization methods were tested, in this study only the detection parts are examined. The comparison of the results show that

the efficiencies of complete methods are usually much lower [46] than the efficiencies of detection methods with idealized reference series. This difference indicates that the time series comparison, the calculation of correction terms, as well as the treatment of data-gaps and outliers may also be substantial sources of homogenization errors. This finding confirms the need for further investigations into strategies for minimizing the risk of negative efficiencies. Optimally the best segments of homogenization methods should be put together (i.e., the best detection method with the best correction method, etc.), and the performance of the best methods should be repeatedly checked applying test datasets of varied properties.

4.3. Reduction of the Risk of Negative Efficiency. Given that almost all results indicate positive efficiency, the expectation that the application of DOHMs generally results in improvement in the quality of observed climatic time series is realistic. However, under certain conditions, alterations from the natural climatic characteristics are relatively frequent (e.g., with PF5S0). It is important to note that obviously no improvement can be achieved using any DOHM on a homogeneous time series; thus the fact that test datasets with negative efficiencies of DOHMs for them can be constructed is not a discouraging indication in itself. On the other hand, a general expectation from DOHMs is that their procedures should indicate with fairly high probability if there is no realistic chance to make quality improvement on time series. The application of DOHMs for time series with low signal to noise ratio is problematic because of the increasing proportion of false alarms and the possible disruption of the natural variability structure shown by raw observed time series. For time series comprising pure white noise, the traditional expectation of preserving time series without adjustments is 90–95%. Figure 5 shows the FTE values for the examined DOHMs. It can be seen that 100 year long series of white noise are not subjected to adjustments with 89–96% probability, except for E-P the ratio is 82% only. When time series contain IHs, but the signal to noise ratio is insufficient for their correct detection, these series should be treated in the same way as homogeneous series. Unfortunately, in case of PF5S0, 25–50% of the series are adjusted (the ratio depends on DOHM; see Figure 6), and their quality is worsened, since the detection results are generally poor due to the low signal-to-noise ratio. The signal to noise ratio can be low due to the very small size of IHs, short duration of biases, high level of background noise, or the existence of other noise than white noise (see also [20, 56]). The frequent adjustment of time series in low signal to noise conditions is a kind of over-homogenization. For instance, when the dataset properties are changed from CH5S0 to PF5S0, it resulted in the mean decline of E_S from +80% to −60%, which might seem to be shocking at the first glance. Notwithstanding, the degree of disruption by over-homogenization is usually much less serious in absolute scale. Examining more the differences between CH5S0 and PF5S0, the mean SSE of raw time series in s_e unit (which is common for all datasets) is 254.4 for CH5S0, but only 26.8 for PF5S0. Calculating with +80% E_S for CH5S0 and −60% E_S for PF5S0, the remaining mean SSE after adjustments is 50.9 (42.8) for

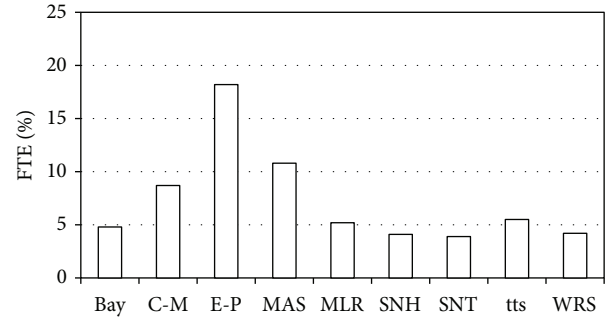

Figure 5: First type errors of DOHMs in pure white noise.

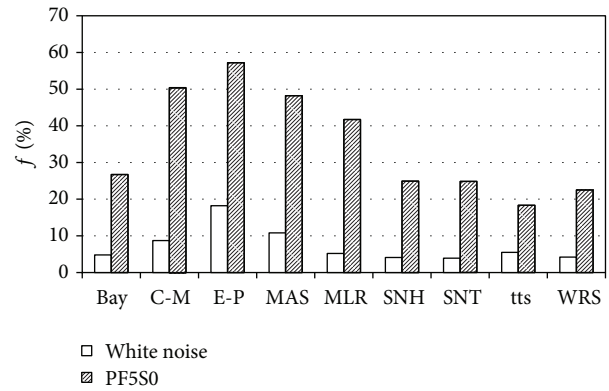

Figure 6: Ratios of qualification "nonhomogeneous" for time series of pure white noise and those for PF5S0.

CH5S0 (PF5S0). The latter results show that the possible over-homogenization usually has little true impact on the quality of observed time series, but one has to keep in mind that the undesired impacts can be great when the background noise of relative time series is high (e.g., due to relatively low spatial correlations in network), as well as additional error-terms should be added due to the imperfectness of reference time series, a problem that is not examined in this study.

A crucial question is that how often low signal to noise ratio occurs in relative time series of true observed datasets; because if its occurrence is frequent, the problem of over-homogenization can be serious. Unfortunately, short-term IHs are likely frequent in true observed time series, because they can easily be produced (i) when the cause of the IH is temporal, (ii) when technical problems are discovered with some delay, and the elimination of the problem is not paired with the backward correction of the data, and (iii) if two shifts of the same direction and significant magnitude are consecutives, the chance that the bias will be realized and corrected increases; therefore, the probability of two consecutive shifts with the opposite directions is higher than that with the same direction.

Another problem is that homogenizations are often step-by-step procedures, searching and correcting inhomogeneous time series from the ones with the highest biases proceeding towards the ones with smaller biases. Results of Figure 6 indicate that the chance that such homogenization procedures do not stop at the best stage is unfavorably high.

There are several options for minimizing the chance of over-homogenization. The intensive use of metadata information in the homogenization process is a widely applied alternative ([18, 29, 57–63], etc.). In spite of the fact that metadata usually do not provide quantitative information [58], the building of metadata information into the numerical evaluation process is one of the most promising ways of improving efficiency of homogenization methods [30, 40, 64–66]. Another option of possible developments is to think out again the parameterization of DOHMs. When some introductory pieces of information (e.g., autocorrelation) clearly indicate that a time series under examination is inhomogeneous, conventional significance thresholds often turn out to be too strict [41]. On the other hand, the problem of potential over-homogenization indicates that traditional significance thresholds may be too low for deciding the usefulness of applying adjustments. The results of this study show that the over-homogenization effects cease when 0.4 minimum threshold is applied for the autocorrelation of relative time series. The author does not suggest that it is an optimal solution, but he points to the fact that some problems exist around the basic mathematical model applied in homogenization procedures, that is, relative time series cannot be modeled well by one or several randomly situated IHs plus white noise. Some homogenization procedures apply a kind of moderation of correction terms to reduce the chance of over-homogenization. Examples for such treatments are the USHCN homogenization method in which adjustments are applied only when the data of at least three nearby stations concordantly indicate the existence and sign of a local shift in the candidate series [40], as well as Multiple Analysis of Series for Homogenization [67] in which always the lowest threshold of the confidence interval of shift-size is applied in particular adjustments. Naturally, a moderation of correction terms like these might sometimes cause under-homogenization (i.e., lack of adjustments application or too small adjustments). Thus the proper way of the minimization of over-homogenization is still an open question.

5. Conclusions

In this study the efficiency for detection algorithms of nine DOHMs is tested using ten different test datasets and four measures of efficiency. The statistical properties of the examined datasets are varied; one of them is similar to the benchmark surrogated dataset of COST HOME project (CH5B0), another is derived empirically to reproduce statistical characteristics of IH-detection results from real climatic time series (HUSTR), and a third one forms a compromise around the halfway between the Benchmark-like CH5B0 and the Hungarian standard (CHPF0). The diversity of test dataset properties and the different kinds of efficiency measures are necessary because the efficiency of homogenization strongly depends on the properties of real observed time series (which are truly diverse) and on the preferred purpose(s) of the homogenization. Our main findings are as follows.

(i) The application of DOHMs is generally beneficial for the quality of time series. When DOHMs are used for time series containing at least one large IH relative to the noise level, the bias of adjusted time series is usually smaller than half of that in raw time series. The mean error of linear trend estimation can be reduced by 75% in the Hungarian standard and by 90% in the Benchmark dataset with the application of DOHMs. Note that this estimation is valid only when all the errors out of the detection-segment (i.e., time series comparison, assessment of correction terms, and treatment of data gaps and outliers) are insignificantly small.

(ii) Short-term, platform-shaped IHs are much more difficult to detect precisely than randomly scattered solely IHs are, particularly when the IH-magnitudes are small. Differences between efficiencies for different kinds of datasets are often larger than those for different DOHMs.

(iii) Results of different efficiency measures often strongly differ, indicating different skills even by applying the same DOHM and for the same test dataset.

(iv) Efficiencies of individual DOHMs do not usually have the same rank order when different efficiency measures or results from different test datasets are compared, although several relationships seem to be stable. The results presented show that C-M is often the most effective DOHM, although the performances of further 4-5 DOHMs (MAS, Bay, MLR, and SNH) are still not much poorer. DOHMs capable of detecting both change points and trends (MLR and SNT) have no better performance than C-M has, even when time series contain trends. The efficiency of SNT is usually slightly lower than that of the earlier version of Standard Normal Homogeneity Test (SNH), and it is not among the best five. DOHMs have to treat multiple IHs either with cutting algorithm or (optimally) through the joint detection of all IHs. tts does not apply any of these techniques, and thus it is not capable of preserving real climatic characteristics of time series. Consequently, DOHMs with sequential tests cannot be recommended for homogenizing time series.

(v) Observed climatic time series cannot be modeled well with the composition of a white noise process plus one or several randomly scattered change points. Not considering the differences between such simple models and the true world may result in unnecessary disruption in the real climatic characteristics of time series. On the other hand, the results show that the degree of the potential disruption is small, at least when some conservative conditions of DOHM application are given, that is, when relative time series of low noise level can be built from the time series of networks of high spatial correlations.

Acknowledgments

The author thanks Matthew Menne, Manola Brunet, Phil Jones, and three anonymous reviewers for their useful comments in improving the clearness of the paper. The research was supported by the European projects COST ES0601 and EURO4M FP7-SPACE-2009-1/242093 and by the Spanish project "Tourism, Environment and Politics" ECO 2010-18158.

References

[1] WMO Hungarian Meteorological Service, *Proceedings of the First Seminar for Homogenization of Surface Climatological Data*, Edited by S. Szalai, Hungarian Meteorlogical Service, Budapest, Hungary, 144 pages, 1996.

[2] WMO Hungarian Meteorological Service, *Proceedings of the Second Seminar for Homogenization of Surface Climatological Data*, Edited by S. Szalai, T. Szentimrey, and Cs. Szinell, WCDMP-41, WMO-TD 932. WMO, Geneva, Switzerland, 214 pages, 1999.

[3] WMO Hungarian Meteorological Service, *Third Seminar for Homogenization and Quality Control in Climatological Databases*, Edited by S. Szalai, Hungarian Meteorological Service, 2001, http://www.met.hu/.

[4] WMO Hungarian Meteorological Service, *Fourth Seminar for Homogenization and Quality Control in Climatological Databases*, Edited by S. Szalai, WCDMP-56, WMO-TD, 1236, WMO, Geneva, Switzerland, 243 pages, 2004.

[5] WMO Hungarian Meteorological Service, *Proceedings of the Fifth Seminar for Homogenization and Quality Control in Climatological Databases*, Edited by M. Lakatos, T. Szentimrey, Z. Bihari, and S. Szalai, WCDMP-No. 71, WMO-TD, 1493, WMO, Geneva, Switzerland, 203 pages, 2008.

[6] WMO Hungarian Meteorological Service, *Proceedings of the Sixth Seminar for Homogenization and Quality Control in Climatological Databases*, Edited by M. Lakatos, T. Szentimrey, Z. Bihari, and S. Szalai, WCDMP-No. 76, WMO-TD., 1576, WMO, Geneva, Switzerland, 116 pages, 2010.

[7] M. Brunet, O. Saladié, P. Jones et al., "The development of a new dataset of Spanish daily adjusted temperature series (SDATS) (1850–2003)," *International Journal of Climatology*, vol. 26, no. 13, pp. 1777–1802, 2006.

[8] D. P. Rayner, "Wind run changes: the dominant factor affecting pan evaporation trends in Australia," *Journal of Climate*, vol. 20, no. 14, pp. 3379–3394, 2007.

[9] M. Staudt, M. J. Esteban-Parra, and Y. Castro-Díez, "Homogenization of long-term monthly Spanish temperature data," *International Journal of Climatology*, vol. 27, no. 13, pp. 1809–1823, 2007.

[10] M. Rusticucci and M. Renom, "Variability and trends in indices of quality-controlled daily temperature extremes in Uruguay," *International Journal of Climatology*, vol. 28, no. 8, pp. 1083–1095, 2008.

[11] S. C. Sherwood, C. L. Meyer, R. J. Allen, and H. A. Titchner, "Robust tropospheric warming revealed by iteratively homogenized radiosonde data," *Journal of Climate*, vol. 21, no. 20, pp. 5336–5350, 2008.

[12] M. J. Menne, C. N. Williams Jr., and R. S. Vose, "The U.S. Historical Climatology Network monthly temperature data, version 2," *Bulletin of the American Meteorological Society*, vol. 90, pp. 993–1007, 2009.

[13] I. Auer, R. Böhm, A. Jurković et al., "A new instrumental precipitation dataset for the greater Alpine region for the period 1800–2002," *International Journal of Climatology*, vol. 25, pp. 139–166, 2005.

[14] E. Aguilar, I. Auer, M. Brunet, T. C. Peterson, and J. Wieringa, "WMO Guidelines on climate metadata and homogenization," WCDMP 53, WMO, Geneva, Switzerland, 2003.

[15] T. C. Peterson, D. R. Easterling, and T. R. Karl, "Homogeneity adjustments of in situ atmospheric climate data: a review," *International Journal of Climatology*, vol. 18, pp. 1493–1517, 1998.

[16] M. J. Menne and C. N. Williams Jr., "Detection of undocumented changepoints using multiple test statistics and composite reference series," *Journal of Climate*, vol. 18, no. 20, pp. 4271–4286, 2005.

[17] J. Reeves, J. Chen, X. L. Wang, R. Lund, and X. Lu, "A review and comparison of change-point detection techniques for climate data," *Journal of Applied Meteorology and Climatology*, vol. 46, pp. 900–915, 2007.

[18] M. Brunet, O. Saladié, P. Jones et al., "A case-study/guidance on the development of long-term daily adjusted temperature datasets," Tech. Rep. WC-DMP-66/WMO-TD-1425, WMO, Geneva, Switzerland, 2008.

[19] B. Trewin, "A daily homogenized temperature data set for Australia," *International Journal of Climatology*, vol. 33, pp. 1510–1529, 2013.

[20] S. C. Sherwood, "Simultaneous detection of climate change and observing biases in a network with incomplete sampling," *Journal of Climate*, vol. 20, no. 15, pp. 4047–4062, 2007.

[21] P. M. Della-Marta and H. Wanner, "A method of homogenizing the extremes and mean of daily temperature measurements," *Journal of Climate*, vol. 19, no. 17, pp. 4179–4197, 2006.

[22] X. L. Wang, H. Chen, Y. Wu, Y. Feng, and Q. Pu, "New techniques for the detection and adjustment of shifts in daily precipitation data series," *Journal of Applied Meteorology and Climatology*, vol. 49, no. 12, pp. 2416–2436, 2010.

[23] Z. Li, Z. Yan, L. Cao, and P. Jones, "Adjusting inhomogeneous daily temperature variability using wavelet analysis," *International Journal of Climatology*, 2013.

[24] V. Alexandrov, M. Schneider, E. Koleva, and J.-M. Moisselin, "Climate variability and change in Bulgaria during the 20th century," *Theoretical and Applied Climatology*, vol. 79, no. 3-4, pp. 133–149, 2004.

[25] F. G. Kuglitsch, A. Toreti, E. Xoplaki, P. M. Della-Marta, J. Luterbacher, and H. Wanner, "Homogenization of daily maximum temperature series in the Mediterranean," *Journal of Geophysical Research D*, vol. 114, no. 15, Article ID D15108, 2009.

[26] A. C. Costa and A. Soares, "Trends in extreme precipitation indices derived from a daily rainfall database for the South of Portugal," *International Journal of Climatology*, vol. 29, no. 13, pp. 1956–1975, 2009.

[27] Z. Yan, Z. Li, Q. Li, and P. Jones, "Effects of site change and urbanisation in the Beijing temperature series 1977–2006," *International Journal of Climatology*, vol. 30, no. 8, pp. 1226–1234, 2010.

[28] T. C. Peterson and D. R. Easterling, "Creation of homogeneous composite climatological reference series," *International Journal of Climatology*, vol. 14, no. 6, pp. 671–679, 1994.

[29] M. Begert, T. Schlegel, and W. Kirchhofer, "Homogeneous temperature and precipitation series of Switzerland from 1864 to 2000," *International Journal of Climatology*, vol. 25, no. 1, pp. 65–80, 2005.

[30] A. T. DeGaetano, "Attributes of several methods for detecting discontinuities in mean temperature series," *Journal of Climate*, vol. 19, no. 5, pp. 838–853, 2006.

[31] J. C. Gonzalez-Hidalgo, J.-A. Lopez-Bustins, P. Štepánek, J. Martin-Vide, and M. de Luis, "Monthly precipitation trends on the Mediterranean fringe of the Iberian Peninsula during the second-half of the twentieth century (1951–2000)," *International Journal of Climatology*, vol. 29, no. 10, pp. 1415–1429, 2009.

[32] S. M. Vicente-Serrano, S. Beguería, J. I. López-Moreno, M. A. García-Vera, and P. Štěpánek, "A complete daily precipitation database for northeast Spain: reconstruction, quality control, and homogeneity," *International Journal of Climatology*, vol. 30, pp. 1146–1163, 2010.

[33] P. Domonkos and P. Štěpánek, "Statistical characteristics of detectable inhomogeneities in observed meteorological time series," *Studia Geophysica et Geodaetica*, vol. 53, pp. 239–260, 2009.

[34] T. A. Buishand, "Some methods for testing the homogeneity of rainfall records," *Journal of Hydrology*, vol. 58, no. 1-2, pp. 11–27, 1982.

[35] D. R. Easterling and T. C. Peterson, "A new method for detecting undocumented discontinuities in climatological time series," *International Journal of Climatology*, vol. 15, no. 4, pp. 369–377, 1995.

[36] J. R. Lanzante, "Resistant, robust and non-parametric techniques for the analysis of climate data: theory and examples, including applications to historical radiosonde station data," *International Journal of Climatology*, vol. 16, no. 11, pp. 1197–1226, 1996.

[37] J. F. Ducré-Robitaille, L. A. Vincent, and G. Boulet, "Comparison of techniques for detection of discontinuities in temperature series," *International Journal of Climatology*, vol. 23, pp. 1087–1101, 2003.

[38] M. Syrakova, "Homogeneity analysis of climatological time series—experiments and problems," *Időjárás*, vol. 107, pp. 31–48, 2003.

[39] G. Drogue, O. Mestre, L. Hoffmann, J.-F. Iffly, and L. Pfister, "Recent warming in a small region with semi-oceanic climate, 1949–1998: what is the ground truth?" *Theoretical and Applied Climatology*, vol. 81, no. 1-2, pp. 1–10, 2005.

[40] M. J. Menne and C. N. Williams Jr., "Homogenization of temperature series via pairwise comparisons," *Journal of Climate*, vol. 22, no. 7, pp. 1700–1717, 2009.

[41] P. Domonkos, "Testing of homogenisation methods: purposes, tools and problems of implementation," in *Proceedings of the 5th Seminar and Quality Control in Climatological Databases*, WCDMP-No. 71, WMO-TD, 1493, pp. 126–145, WMO, 2008.

[42] P. Domonkos, "Efficiency evaluation for detecting inhomogeneities by objective homogenisation methods," *Theoretical and Applied Climatology*, vol. 105, pp. 455–467, 2011.

[43] P. G. F. Gérard-Marchant, D. E. Stooksbury, and L. Seymour, "Methods for starting the detection of undocumented multiple changepoints," *Journal of Climate*, vol. 21, no. 18, pp. 4887–4899, 2008.

[44] C. Beaulieu, O. Seidou, T. B. M. J. Ouarda, X. Zhang, G. Boulet, and A. Yagouti, "Intercomparison of homogenization techniques for precipitation data," *Water Resources Research*, vol. 44, no. 2, Article ID W02425, 2008.

[45] H. A. Titchner, P. W. Thorne, M. P. McCarthy, S. F. B. Tett, L. Haimberger, and D. E. Parker, "Critically reassessing tropospheric temperature trends from radiosondes using realistic validation experiments," *Journal of Climate*, vol. 22, no. 3, pp. 465–485, 2009.

[46] V. Venema, O. Mestre, E. Aguilar et al., "Benchmarking monthly homogenization algorithms," *Climate of the Past*, vol. 8, pp. 89–115, 2012.

[47] H. Caussinus and F. Lyazrhi, "Choosing a linear model with a random number of change-points and outliers," *Annals of the Institute of Statistical Mathematics*, vol. 49, no. 4, pp. 761–775, 1997.

[48] A. Moberg and H. Alexandersson, "Homogenization of Swedish temperature data—part II: homogenized gridded air temperature compared with a subset of global gridded air temperature since 1861," *International Journal of Climatology*, vol. 17, no. 1, pp. 35–54, 1997.

[49] L. A. Vincent, T. C. Peterson, V. R. Barros et al., "Observed trends in indices of daily temperature extremes in South America 1960–2000," *Journal of Climate*, vol. 18, pp. 5011–5023, 2005.

[50] T. C. Peterson, "Assessment of urban versus rural in situ surface temperatures in the contiguous United States: no difference found," *Journal of Climate*, vol. 16, pp. 2941–2959, 2003.

[51] P. Domonkos, "Homogenising time series: beliefs, dogmas and facts," *Advances in Science and Research*, vol. 6, pp. 167–172, 2011.

[52] R. Lund and J. Reeves, "Detection of undocumented change-points: a revision of the two-phase regression model," *Journal of Climate*, vol. 15, no. 17, pp. 2547–2554, 2002.

[53] H. Caussinus and O. Mestre, "Detection and correction of artificial shifts in climate series," *Journal of the Royal Statistical Society C*, vol. 53, no. 3, pp. 405–425, 2004.

[54] F. Picard, E. Lebarbier, M. Hoebeke, G. Rigaill, B. Thiam, and S. Robin, "Joint segmentation, calling, and normalization of multiple CGH profiles," *Biostatistics*, vol. 12, no. 3, pp. 413–428, 2011.

[55] P. Domonkos, "Adapted caussinus-mestre algorithm for networks of temperature series (ACMANT)," *International Journal of Geosciences*, vol. 2, pp. 293–309, 2011.

[56] V. C. Slonosky and E. Graham, "Canadian pressure observations and circulation variability: links to air temperature," *International Journal of Climatology*, vol. 25, no. 11, pp. 1473–1492, 2005.

[57] E. Aguilar, T. C. Peterson, P. Ramírez et al., "Changes in precipitation and temperature extremes in Central America and northern South America, 1961–2003," *Journal of Geophysical Research*, vol. 110, Article ID D23107, 2005.

[58] M. Brunetti, M. Maugeri, F. Monti, and T. Nanni, "Temperature and precipitation variability in Italy in the last two centuries from homogenised instrumental time series," *International Journal of Climatology*, vol. 26, no. 3, pp. 345–381, 2006.

[59] D. Camuffo, C. Cocheo, and G. Sturaro, "Corrections of systematic errors, data homogenisation and climatic analysis of the Padova pressure series (1725–1999)," *Climatic Change*, vol. 78, no. 2-4, pp. 493–514, 2006.

[60] R. Brázdil, K. Chromá, P. Dobrovolný, and R. Tolasz, "Climate fluctuations in the Czech Republic during the period 1961–2005," *International Journal of Climatology*, vol. 29, pp. 223–242, 2009.

[61] Q. Li, H. Zhang, J. I. Chen, W. Li, X. Liu, and P. Jones, "A mainland china homogenized historical temperature dataset of 1951–2004," *Bulletin of the American Meteorological Society*, vol. 90, no. 8, pp. 1062–1065, 2009.

[62] M. Türkes, T. Koç, and F. Sariş, "Spatiotemporal variability of precipitation total series over Turkey," *International Journal of Climatology*, vol. 29, pp. 1056–1074, 2009.

[63] M. Syrakova and M. Stefanova, "Homogenization of Bulgarian temperature series," *International Journal of Climatology*, vol. 29, no. 12, pp. 1835–1849, 2009.

[64] J. R. Christy, W. B. Norris, K. Redmond, and K. P. Gallo, "Methodology and results of calculating central California surface temperature trends: evidence of human-induced climate change?" *Journal of Climate*, vol. 19, no. 4, pp. 548–563, 2006.

[65] L. Haimberger, "Homogenization of radiosonde temperature time series using innovation statistics," *Journal of Climate*, vol. 20, no. 7, pp. 1377–1403, 2007.

[66] M. P. McCarthy, H. A. Titchner, P. W. Thorne, S. F. B. Tett, L. Haimberger, and D. E. Parker, "Assessing bias and uncertainty in the HadAT-adjusted radiosonde climate record," *Journal of Climate*, vol. 21, no. 4, pp. 817–832, 2008.

[67] T. Szentimrey, "Multiple Analysis of Series for Homogenization (MASH)," in *Proceedings of the 2nd Seminar for Homogenization of Surface Climatological Data*, WCDMP 41, WMO-TD 962, pp. 27–46, WMO, 1999.

[68] L. A. Vincent, "A technique for the identification of inhomogeneities in Canadian temperature series," *Journal of Climate*, vol. 11, no. 5, pp. 1094–1104, 1998.

[69] H. Alexandersson, "A homogeneity test applied to precipitation data," *Journal of Climatology*, vol. 6, no. 6, pp. 661–675, 1986.

[70] H. Alexandersson and A. Moberg, "Homogemzation of Swedish temperature data—part I: homogeneity test for linear trends," *International Journal of Climatology*, vol. 17, no. 1, pp. 25–34, 1997.

[71] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometric Bulletin*, vol. 1, pp. 80–83, 1945.