Theor Appl Climatol
DOI 10.1007/s00704-011-0399-7

ORIGINAL PAPER

Efficiency evaluation for detecting inhomogeneities by objective homogenisation methods

Peter Domonkos

Received: 16 December 2009 / Accepted: 23 December 2010

Abstract Evaluation and comparison of efficiencies of widely used objective homogenisation methods (OHOMs) are presented relying on some test-datasets and efficiency measures. Problems related to the choice of efficiency measure, creation of appropriate test-datasets and use of OHOM parameterisation are discussed. The detection parts of the OHOMs are examined only. Power of detection, false alarm rate, detection skill and skill of linear trend estimation are calculated and compared for eight OHOMs and six test-datasets. Each test-dataset comprises 10,000 100 year-long artificially simulated time series. In the simplest test dataset, each time series contains one inhomogeneity (IH), while a structure of inhomogeneities that is similar to that of real central European temperature time series is included in the most complex simulated dataset. Distinct attention is given to OHOMs that contain (1) cutting algorithm, (2) semihierarchic algorithm, (3) direct detection of multiple IHs, (4) detection of change-point and trend-line shaped IHs. Results show that Caussinus–Mestre method and Multiple Analysis of Series for Homogenization are the most powerful tools in detecting and correcting IHs in climatic time series.

1 Introduction

Examination of climate change and climate variability requires climatic time series of high quality. Fortunately, a large number of long-observed time series are available in the world, owing to the early realisation of the importance of collecting observational data. Several climatic time series from Europe and North America are much longer than a century. However, the constancy of the measuring techniques, location and physical surrounding is hard to maintain even for a few decades. Thus, a portion of the changes in observed time series has no climatic origin, but they indicate some changes in the direct or indirect conditions of the observing process (Peterson et al. 1998; Aguilar et al. 2003; Auer et al. 2005; Brunet et al. 2006; etc.). This kind of change is called an inhomogeneity (IH), and a time series with IHs is said to be inhomogeneous. The typical form of IHs is a sudden shift (a so-called change-point) in the values of time series, because most of non-climatic changes occur at some distinct time. However, some factors may cause slow, gradually increasing change in time series (urban effect, growing trees, etc.), and in such cases, IHs can be modelled with a non-climatic linear trend in some section of the time series. Pairs of change-points indicating changes of the same size but in opposite directions are referred to as platforms in the study. IHs can be detected by the comparison of several data series from the same climatic region, though small IHs are usually indistinguishable from random noise. The success of detecting and correcting IHs depends on numerous factors, with the skill of the method for resolving the change-points being of primary importance.

The demand for achieving the best quality of observed time series has resulted in a great development of homogenisation methods in the recent decades. Nearly 20 homogenisation methods are widely used in climatology nowadays, and considering the differences in the details of practical application the diversity is even greater. Most (but not all) homogenisation methods cannot detect trend-type

P. Domonkos (✉)
University of Rovira i Virgili, Centre for Climate Change,
Campus Terres de l'Ebre,
CP: 43500 C\ Betánia 5,
Tortosa, Spain
e-mail: peter.domonkos@urv.cat

Published online: 29 January 2011

IHs directly; rather, they often parse the trend-type IH into a series of step changes. Moreover, it would seem that the challenge remains to evaluate which method or methods produce the highest quality of climatic time series. Obviously, only objective and reproducible homogenisation methods (OHOMs) can be quantitatively evaluated. Although there have been some comparative examinations aiming to reveal the capability of detecting IHs by OHOMs (Buishand 1982; Easterling and Peterson 1995; Lanzante 1996; Ducré-Robitaille et al. 2003; Syrakova 2003; Drogue et al. 2005; Menne and Williams 2005; DeGaetano 2006; Domonkos 2006a; Gérard-Marchant et al. 2008; Beaulieu et al. 2008; Titchner et al. 2009), in these efforts, only a small selection of the methods were evaluated, and test-datasets whose statistical properties are usually far from the reality were used. Simulated datasets for testing OHOMs are commonly generated from a white noise process to which some imposed change-points of fixed magnitude are added. Although these experiments provide useful information about the general properties of OHOMs, one cannot draw direct conclusion from them for the performance of OHOMs in real climatic datasets, since, in the real world, IH-sizes are not constant. The introduction of variable-size artificial IHs with peak-frequency around 0 size (Menne and Williams 2005, 2009) was a substantial step towards making efficiency tests more appropriate for producing realistic results. Here, a special way of creating realistic test-dataset is applied (presented first in Domonkos 2006a): the statistical characteristics of detected IHs from test dataset made similar to those from real climatic dataset through an iterative development of test-datasets.

Given that the statistical characteristics of climatic time series are diverse and that the practical meaning of "efficiency" depends on the objectives of the time series analysis, a thorough evaluation of efficiencies requires the evaluation of a variety of test-datasets and a number of different measures of efficiency. This paper presents several examples of efficiency calculations for cases of high practical importance but makes no claim to be comprehensive. The principles for selecting the examples were as follows:

1. The most relevant test-datasets must have similar statistical properties as real, observed climatic time series have. The similarity refers also to the frequency and magnitude–distribution of IHs.
2. Test-datasets of widely different properties are examined in parallel in order to reveal general relations among efficiencies for individual OHOMs.
3. Detection parts of OHOMs are examined only; see its definition in section 2.1.
4. The efficiency measures chosen are relatively simple, but they have substantial practical importance.
5. OHOMs selected are widely used, and they can be applied automatically in the examination of relative time series.
6. Skill in detecting multiple IHs is tested with the parallel examination of OHOMs with hierarchic, semihierarchic and direct algorithms.
7. Skill of OHOMs for detecting both shift- and trend-type IHs is compared with that of OHOMs detecting shifts only.
8. OHOMs are usually applied with the originally recommended parameterisation.

In the following section, some basic concepts and definitions are provided. The OHOMs selected for testing, the test-datasets used and the efficiency measures applied are also discussed. Results are presented in section 3, which focus specifically on the efficiency characteristics among different OHOMs and different test-datasets. A discussion of the results is offered in section 4, and concluding comments are provided in section 5.

2 Methods

2.1 Concepts and definitions

Evaluations of efficiency are usually quantified via simulated time series that contain artificial IHs. In reality, the statistical properties of IHs that occur in real climatic time series are not known precisely. For this reason, the frequency of IHs, as well as the distribution of IH magnitudes may be substantially different in simulated datasets in comparison with the statistical properties of relative time series derived from the differences between a candidate and reference series. Obviously, the higher the resemblance between the simulated and real statistical properties, the higher the probability that the estimated efficiencies based on simulated datasets are valid for real climatic datasets. Domonkos (2006a) applied a simulation method that was empirically developed, and the similarity of such datasets to real observations was tested. This simulation method is also applied in this study [see item (E) of section 2.3 and the detailed description of the simulation in Appendix I].

Efficiency of homogenisation methods is calculated using six test-datasets in the study. Each dataset comprises N time series of n year length (Eq. 1).

$$X_P = \left[X_{p,1}, X_{p,2}, \ldots X_{p,n}\right]^{\mathrm{T}}, p = 1, 2, \ldots \mathrm{N} \quad (1)$$

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 (**H**). So that each element of time series (the index p will not be in use hereafter) can be

expressed as a sum of the contemporary noise and the cumulated effects of IHs (Eq. 2).

$$x_j = w_j + h_j, j = [1, 2, \ldots n] \quad (2)$$

The statistical properties of IHs in the six datasets are different. In certain datasets, the structures of inhomogeneities are very complex (in E and F datasets, see section 2.3.), and the noise part contains an additional term ($\mathbf{W}^*$) beyond white noise.

Simulated series are always handled as relative time series in this study, i.e. they are considered to be the difference between candidate and reference series in an imaginary dataset. Note that, in raw climatic time series X contains one more component than how many is shown in formula (2), namely the time-varying climate-component. Climate signals may have similar shape as IHs, and it may complicate the identification of IHs. Therefore, the application of OHOMs is recommended primarily for relative time series. Problems related to the ways of creating relative time series (through building reference series or with pair-wise comparisons) are beyond the scope of this study. This paper is restricted to the examination of the detection part of OHOMs, i.e. evaluates the skill of statistical methods for identifying IHs in given time series. Iterative parts of OHOMs are not considered here. It is important to point out that, while the results here may differ somewhat from those obtained by applying a complete homogenisation method, we nevertheless maintain that the skill of the detection component of OHOMs deserves to be quantified in isolation, and such an assessment becomes complicated when the other components of homogenisation procedures are not controlled for in the comparison.

During the simulation, IH magnitudes (m) are expressed with their ratio to the standard deviation of the white noise (s_e). However, during the detection of IHs another unit, the estimated standard deviation of white noise (s_e*) is used for IH magnitudes (m*).

$$\begin{aligned} S_e^* &= \sqrt{1 - R^2} \cdot s_T \quad \text{if } R > 0 \\ S_e^* &= s_T \quad \text{if } R \le 0 \end{aligned} \quad (3)$$

In formula (3), R denotes the 1 year lag autocorrelation in relative time series, and s_T refers to the empirical standard deviation of the time series. The use of unit s_e* is justified by the fact that during the detection process s_e is known only for simulated time series while, for relative time series derived 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.

2.2 OHOMs examined

OHOMs usually detect one change-point only or a structure of trend plus one change-point in a particular step of the detection procedure, and multiple IHs of time series are searched by hierarchic or semihierarchic organisation of individual steps. The only exceptions are the Caussinus–Mestre method (C-M, Caussinus and Mestre 2004) and the Multiple Analysis of Series for Homogenisation (MASH, Szentimrey 1999) whose algorithms are capable of detecting multiple structures of IHs directly. Thus, considering the theoretical bases, the latter two OHOMs promise the homogenisation results of the highest quality.

Efficiencies of detection parts of seven OHOMs are examined in the study. They represent different classes of detection algorithms of great theoretical importance by methods widely used in climate studies. The seven methods are the: Multiple Linear Regression (MLR, Vincent 1998), Penalised Maximal t test (PMT, Wang et al. 2007), Standard Normal Homogeneity Test for shifts only (SNH, Alexandersson 1986), Standard Normal Homogeneity Test for shifts and trends (SNHT, Alexandersson and Moberg 1997), t test (tts, Ducré-Robitaille et al. 2003), C-M and MASH. Two versions of SNH are used: SNH1 includes the common cutting algorithm, while SNH2 is supplied with the semihierarchic algorithm recommended by Moberg and Alexandersson (1997). Thus, the final number of examined OHOMs is eight. The use of two different versions of the SNH allows the comparison of different algorithms for detecting multiple IHs. The abbreviation SAMA is used hereafter for the semihierarchic algorithm by Moberg and Alexandersson.

The simplest OHOM examined here is the sequential t test. It assesses the homogeneity of distinct sections of time series not considering the properties of time series beyond the selected section. All the other OHOMs examine whole time series or sections that are delimited by either an IH detected in an earlier step or some end of the time series.

MLR, PMT and SNH1 operate with cutting algorithm. In these methods, when an IH is detected, the time series is divided to two parts, and the sections derived in this way are examined further as long as the length of sections is sufficient for further examination. This threshold is 10 years in this study, in accordance with the usual recommendation (Easterling and Peterson 1995; Lanzante 1996, etc.). A further restriction here is that the minimum distance between two IHs or between an IH and one end of the time series is 5 years.

In SNH2 and SNHT, the SAMA is applied. In this method, after identifying the timings of potential IHs detected with the cutting algorithm in the first phase, each potential IH is retested in the second phase using a section that contains exactly one potential IH and that is delimited by either some IH, potential IH or by one end of the time series. This double-phase procedure aims to eliminate

potential interferences when more than one IH can be present in examined sections (in the first phase).

C-M and MASH are capable of detecting multiple IHs in a direct way. C-M fits an optimised step function to time series (Hawkins 1972), and the number of steps is set by the Caussinus–Lyazrhi criterion (Caussinus and Lyazrhi 1997). In MASH, each possible combination of IHs is considered and subjected to hypothesis test (Szentimrey 1999).

MLR and SNHT detect both change-points and trend-type IHs, while the other OHOMs can only approximate the treatment of trend-type IHs with a series of small shifts (series of change-points).

Considering the parameterisation for selecting significant IHs, the original parameters are used in C-M and PMT; it is set according to the original recommendations aiming 0.05 rate first type error (FTE) in pure white noise processes when the null-hypothesis is that the time series is homogeneous in MASH, SNH2 and SNHT, the same for SNH2 as for SNH1, while it was calculated by the author for MLR and tts, approaching 0.05 rate FTE in pure white noise processes. In SNHT, 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. In MLR, only the 1 year lag autocorrelations are controlled in the present version. This parameterisation for MLR appears to generally provide higher efficiencies than the original formulation (not shown).

A uniform pre-filtering of outliers is applied before the use of any OHOM. Anomalies from the average of the time series are considered to be outliers if their absolute values are higher than 4 standard deviation of the time series elements. These values are replaced with 0 anomalies.

2.3 Test-datasets

Six kinds of test-datasets (A,B,…F) are used in this paper. Each dataset comprises 10,000 time series. All the time series are of 100 years length (n=100).

The inhomogeneity-properties of the six datasets are as follows.

A. One IH is included in each time series. Its type is change-point, the timing (j) equals to 40 or 60, and m equals to 3. In this simple case, it is easy to explicitly write down the values taken by h as a function of time (Eq. 4).

$$h_i = 0, \text{ if } i \le j, \text{ and } h_i = 3, \text{ if } i > j, 1 \le i < n, \quad (4)$$

$$j = 40 \text{ or } j = 60$$

B. Five change-points are included, one with j=40 and m=3, while the others are with random timings but of a fixed magnitude, m=1.5. The minimum distance between adjacent IHs and from the endpoints of the series was set to be 4 years.

C. The mean frequency occurrence is one IH per 20 years, but IH-frequencies in individual time series may deviate from the average. All the IHs are change-points; their signs (positive or negative), timings and magnitudes are random. Magnitudes (m) are between 0 and 4; they are exponentially distributed for m>1, and equally distributed for m<1 (Eq. 5).

$$\begin{aligned} m &= e^{2.39\cdot(q-0.42)} \text{if } q \ge 0.42 \\ m &= \tfrac{1}{0.42}\cdot q \qquad \text{if } q < 0.42 \end{aligned} \quad (5)$$

q is a random variable with equal distribution between 0 and 1.

D. It is similar to item 3; only, some parameters differ. The mean frequency occurrence is one IH per decade. All the IHs are change-points; their signs (positive or negative), timings and magnitudes are random. Magnitudes (m) are between 0 and 6; they are exponentially distributed for m>1, and equally distributed for m<1 (Eq. 6).

$$\begin{aligned} m &= e^{2.8\cdot(q-0.36)} \quad \text{if } q \ge 0.36 \\ m &= \tfrac{1}{0.36}\cdot q \qquad \text{if } q < 0.36 \end{aligned} \quad (6)$$

E. The standard dataset. A rather complex structure of randomly distributed IHs of different types (change-points, platforms, trends) and magnitudes. When this dataset was created, the goal was to simulate the properties of relative time series from an observed Hungarian temperature dataset (Domonkos 2006b) as closely as possible, which is difficult because the configuration of IHs is essentially unknown. Ultimately, an empirical approach obtained following numerous iterations focused on minimising the differences between the simulations and the observed data.

Appendix I provides a description of the simulation method in detail. In brief, the simulation yields a set of synthetic relative time series. Each series is categorised by a lag-1 autocorrelation greater than 0.4 (which is itself an indicator of non-homogeneous character for relative time series; see, e.g. Sneyers 1997). The simulation method was developed empirically, without setting any preconception of IH-structures. However, beyond the empirical justification, the basic IH properties of the standard dataset can be reasoned as follows:

- Small IHs are much more frequent than large ones since potentially large deviations are more obvious and can be corrected by observers in routine-controls or through an earlier implemented homogenisation. Also, the causes of potentially large deviations could easily be identified and eliminated.
- Platform-like IHs are very frequent. The cause of these types of IHs is twofold: First, the causes of deviations are often eliminated with some delay such that the excursion to a new mean level is

temporary. Second, two adjacent shifts relatively rarely have the same sign. This is because accumulated deviations of large magnitudes may be noticed and corrected with higher probability than deviations fluctuating near zero.

- Short-term deviations (i.e. IHs of small duration) are more frequent than long-lasting IHs. From the empirical results, it seems that most of the IHs of considerable size are recognised and eliminated within few years of their occurrence. However, it cannot be ruled out that a certain percentage of IH-like signatures of very short duration may in fact have true climatic origins through the natural fluctuations of spatial climatic gradients, which may contribute to the seemingly high rate of short-term IHs. We have no concrete information regarding the relative distribution of macroclimatic and truly local small IHs, but an arbitrarily determined percentage of small IHs is considered to be of true climatic origin (see Appendix I).

Figure 1 illustrates the similarity between the IH properties from simulated and real climatic datasets. To generate this figure, 204 statistical properties were calculated both for the real and the artificial datasets that are relevant to the 15 OHOMs, each of which has four different parameterisations. The frequency of detected shifts, first and second moments of magnitude distributions, as well as properties of detected trends (when OHOM calculates also trends) were considered (Domonkos 2006a). In Fig. 1, the frequency distribution of differences between the real and simulated characteristics is presented. The resemblance is evident in that there are very few cases with differences over 10%. We note that an examination of a Moravian temperature dataset resulted in very similar statistical properties of detected IHs (Domonkos and Štěpánek 2009) despite large differences in the spatial density and correlation between station series.

F. A quasi-standard dataset with reduced frequency of large IHs. The simulation method is almost the same as for the standard dataset, but the frequency of persistent large IHs is much lower in this dataset (see Appendix II). The examination of reduced frequency of persistent large IHs has great practical importance, because the quality-controller may not be aware of all earlier corrections made in time series, so homogenised, partly corrected and seriously erroneous time series cannot always be separated well. Thus, during the quality controls, one can easily meet with time series whose statistical characteristics are very similar to typical unchecked series, but the frequency of persistent large IHs is markedly reduced.

Figure 2 presents the magnitude distributions for change-point type IHs in the C–F datasets. Magnitudes are expressed in m^*, while for frequency values, an arbitrary unit is used. The difference between the frequency of large IHs and that of small IHs is very high, therefore the y-scale is logarithmic. It can be seen that the amount of small IHs is larger in the standard dataset than in the datasets with exponential distribution of shift-magnitudes. It must be noted that the amount of very small IHs ($m^*<<1$) cannot be determined with certainty because they have little impact on the detection results. However, for the very same reason, their importance is also limited. On the other hand, above $m^*\approx 0.5$, the impact of small IHs on the efficiency of homogenisation increases, and experimental homogenization results do not reproduce the statistical characteristics of observed datasets without the inclusion of a number of small IHs. This observation indicates that the inclusion of small IHs in test-datasets is necessary to obtain reliable results in testing the efficiency of OHOMs.

Datasets C and D contain more moderately large IHs ($2<m^*<3$) than the standard dataset does, but considering

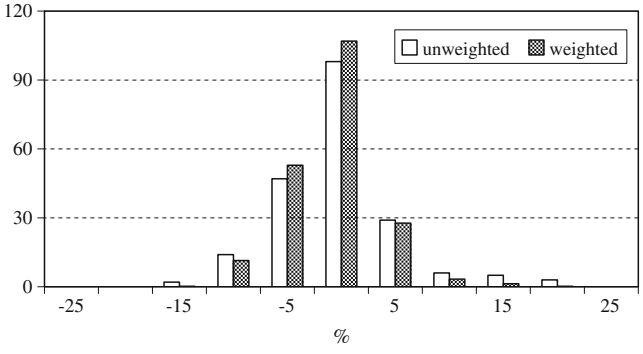

Fig. 1 Frequency distribution of differences between the same type statistical characteristics of detected inhomogeneities for observed temperature series in Hungary and for the standard test-dataset. *Filled columns* show results of weighting with sample size

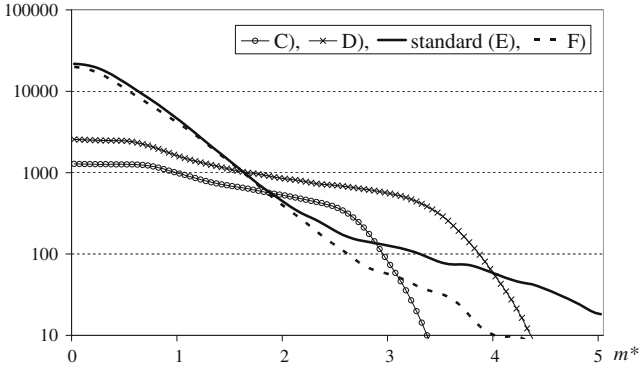

Fig. 2 Distribution of change-point magnitudes (m^*) in test-datasets *C…F*

very large IHs, the relation turns back again. Differences for datasets E and F are small below m^*=2.5, but more striking at higher magnitudes.

2.4 Measures of efficiency

Four measures are considered: (a) power of detection, (b) false alarm rate, (c) detection skill and (d) skill of linear trend estimation. Measures (a) to (c) test the skill of the identification of change-points, while (d) evaluates the reliability of linear trends in homogenised time series.

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. Measures (a) to (c) can be expressed with the combination of these simple characteristics. Although the concepts of correct and false detection are clear in case of one or a few of fairly large IHs, separating their occurrences is not as easy in complex IH-structures, such as those in datasets E and F.

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

$$\frac{1}{k}\left|\sum_{i=j-k+1}^{j} x_i - \sum_{i=j+1}^{j+k} x_i\right| \ge 2, \quad \text{for each } k \text{ of } k = \{1,2,3\} \tag{7}$$

Formula 7 shows that change-points with magnitude (m^*) at least 2.0 are considered. A shift of this magnitude must be detectable comparing each symmetric half-window pairs, up to window-width of 6 years. The mean frequencies of such change-points in datasets A–F are 1, 1, 0.66, 1.94, 0.64 and 0.37 per time series, respectively.

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

False detection The detection result at year j is a change-point with $m^* \ge 1.5$, but no 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. There is no minimum threshold here for the size of factual changes, only their signs are considered.

1. Power of detection (Pw):

$$Pw = \frac{S_R}{S} \tag{8}$$

2. False alarm rate (Fa):

$$Fa = \frac{S_F}{S_R + S_F} \tag{9}$$

3. Detection skill (E_D):

$$E_D = \frac{S_R - S_F}{S} \tag{10}$$

E_D presents a combined measure of the power and the rate of false detections. The highest possible value of E_D is 1, and E_D=1 means perfect identification of change-points. When half of the detected change-points are false, E_D=0. Note that in datasets including very few change-points (S is small), E_D can easily be negative.

4. Skill of linear trend estimation (E_T): The accuracy of slopes of linear trends for the whole (100 years long) time series and for the last 50 years of series is evaluated. Let the mean bias of trend estimations be denoted by f for homogenised time series and f_0 for time series without homogenisation.

$$E_T = \frac{f_0 - f}{f_0} \tag{11}$$

E_T indicates the reliability of trend estimations in homogenised time series. The maximum value of E_T is 1, and E_T= 1, only if all the trend estimations are perfect. E_T=0 means that neither any improvement, nor an increase of errors is typical for homogenised time series. If a homogenisation results in larger biases of trends than the biases for time series without homogenisation, E_T is negative.

3 Results

Figure 3 presents the E_D and E_T values of the investigated OHOMs for datasets A–F. When only one change-point is included (Fig. 3a), most of the OHOMs perform well and the efficiencies are near or above 90%, except for tts. When the same size change-point (m=3) is accompanied by four small shifts (Fig. 3b), the efficiencies are substantially lower. Nevertheless, although the detection skill has been dropped dramatically to 40–70%, the skill of trend estimation remained relatively high, between 80–90%, except for tts. While for "A" type series, SNH1 and MASH have the highest E_D and E_T values, respectively; for "B" series, the PMT and C-M are the best. However, the highest efficiency values are closely followed by the efficiencies of several other OHOMs in both of the experiments "A" and "B", thus positional differences do not indicate significantly different performances.

Results for dataset C (Fig. 3c) are very similar to those of dataset B. In this experiment, SNH2 has the highest E_D, and C-M has the highest E_T value.

Figure 3d–f show a substantially different distribution of E_D values than the previous ones. While in experiments A,

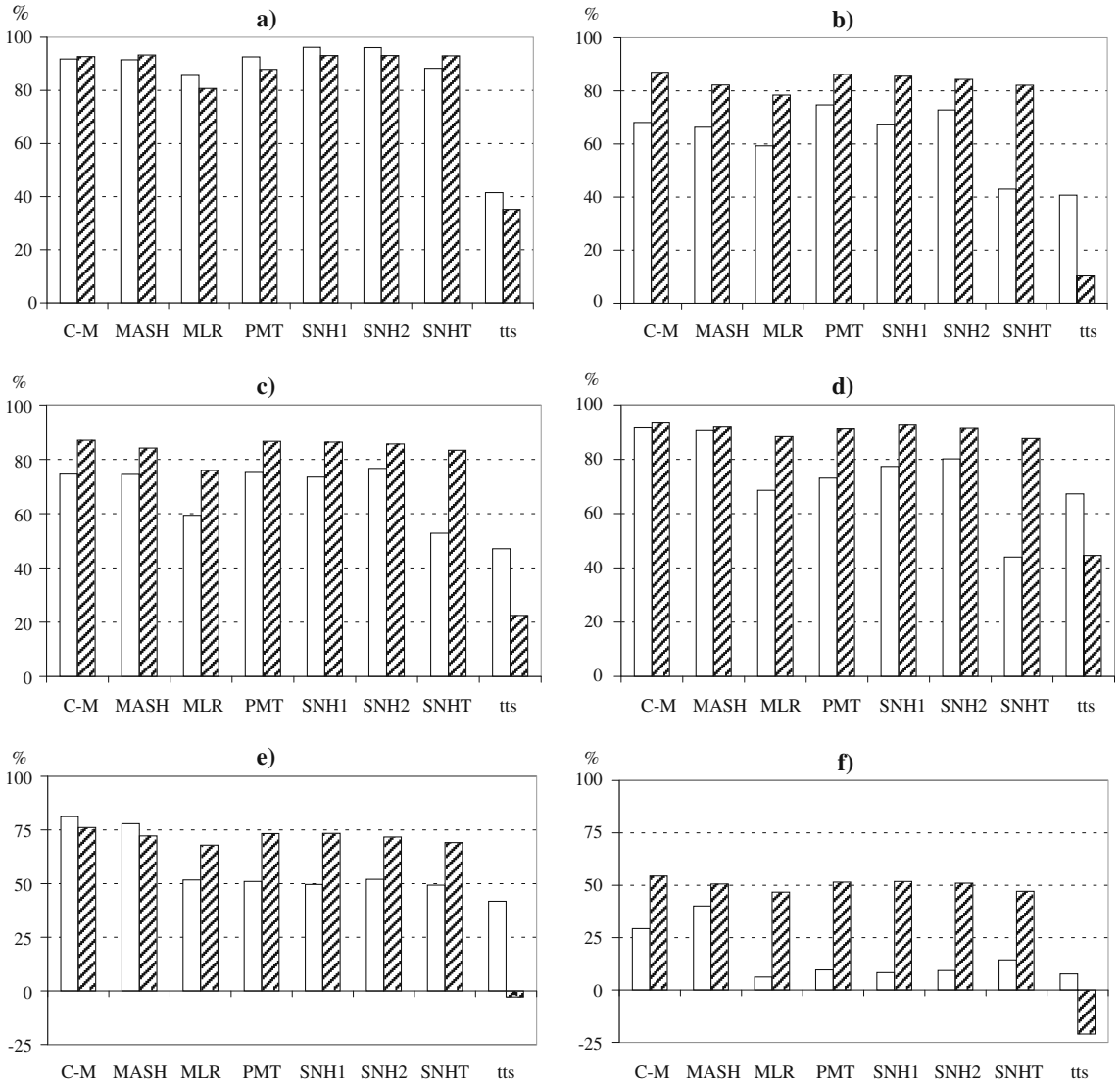

Fig. 3 **a**–**f** Detection skill (E_D, *empty columns*) and skill of linear trend estimation (E_T, *striped*) with different OHOMs in datasets (**a**)…(**f**), respectively

B and C, PMT, SNH1 and SNH2 perform similarly or even better than C-M and MASH; in experiments D, E and F, C-M and MASH have considerably higher detection skill than any other OHOMs do. No such difference is observed between the two groups of experiments when testing the skill of trend estimation. What is more, the distribution of E_T values has several common features in all but the "A" experiments: (a) Only small differences between the skills of C-M, MASH, PMT, SNH1 and SNH2 can be observed, (b) C-M always shows the highest skill, (c) PMT or SNH1 has the second highest E_T, (d) Skills of SNHT and MLR are always slightly lower than the five best OHOMs from the examined eight and (e) skill of tts is markedly smaller than that of the other OHOMs.

Results show also that (1) MASH has slightly lower skill than C-M does, with few exceptions only. (2) In spite of the fact that MLR and SNHT are capable of detecting both abrupt shifts and gradual linear changes, the skill of these OHOMs is not higher than that of the other OHOMs, even in experiments including linear changes with 25% rate of persistent IHs (datasets E and F). A possible reason of this result is that these OHOMs often fail to identify the kind of the IH. For example, they detect a linear change instead of two abrupt shifts of the same direction. However, it is hard to explain why the skills of trend estimation for SNHT and MLR are lower than that for SNH1 and SNH2. (3) tts has generally much lower skill than the other OHOMs examined do, especially in trend estimation. For standard and quasi-standard datasets, this efficiency measure shows negative skill. The explanation is that tts examines sections of time series and detects change-points without coherence between the examinations and time series properties for

individual sections. (4) Application of SAMA has no considerable positive effect on the skills relative to the common cutting algorithm. Although SNH2 always has slightly higher detection skill than SNH1 does, the relation for E_T values is just the opposite.

Powers and false alarm rates are also examined. In Fig. 4, the results are grouped according to the test-datasets used. It can be seen that differences between the efficiency characteristics are often higher according to the time series properties than among the OHOMs. In Fig. 4, the best results (high power and low false alarm rate) can be found in the upper left corner, while moving right and down in the figure, the skill is decreasing. It is not a surprise that the highest skills are usually achieved with dataset "A" in which one change-point per time series is included only. Skills are usually high also for dataset D in which large IHs occur at high frequency. Results for datasets B and C have an interesting feature, namely OHOMs with the lowest false alarm rate have rather low power, while OHOMs with the highest power have considerably high false alarm rate. For the standard and quasi-standard datasets, 2-2 results have distinguishably higher power than the other OHOMs do. In accordance with the results of Fig. 3, these OHOMs are the C-M and MASH. There is a big difference between the results for dataset F and the other results that false alarm rate is markedly higher for this dataset with all the OHOMs examined. The likely explanation is that the low rate of large, persistent IHs severely reduces the rate of correct detection, and supposing the unchanged amount of false detection, it results in a great increase in the false alarm rate.

Figure 5 presents again the *Pw*-*Fa* characteristics, but in this figure the results are separated according to the test-dataset used, and the characteristics with different OHOMs are for comparison. This figure confirms that, while PMT and SNH are the best OHOMs for datasets A, B and C, C-M and MASH have the highest skill for the other three datasets. Seeing the details, we can find that, using datasets A–C, Fa is considerably lower and *Pw* almost the same for PMT and SNH than for C-M and MASH, while, for D–F, the power is much higher with C-M and MASH than with any other method, and differences of *Fa* do not compensate for the differences of *Pw*.

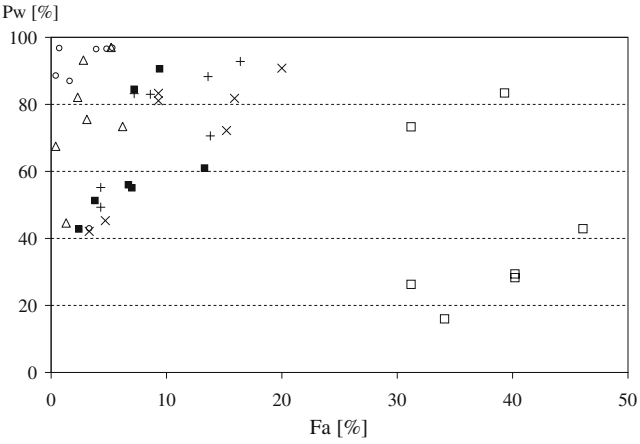

Fig. 4 Power of detection (Pw) and false alarm rate (Fa) value-pairs. Dataset A: *empty circle*, B: *multiplication sign*, C: *plus sign*. D: *empty triangle*, E: *filled square*, F: *empty square*

In spite of the big differences in the absolute values according to the test-dataset used, there are several common features in the relative position of *Pw*-*Fa* value-pairs: (1) The differences between the characteristics of SNH1 and SNH2 show the impact of the SAMA. The results with SNH2 are closer to the upper left corner of the figure indeed than that of SNH1, but the differences are disappointing small. (2) Efficiency characteristics of PMT hardly differ from those of SNH1 and SNH2. (3) MLR can be characterised by moderately high power and rather high false alarm rate. (4) Results for SNHT and tts show favourably low false alarm rate, but insufficiently low power. (5) Comparing the results of C-M and MASH, C-M consistently has higher power than MASH, but the false alarm rates are also always higher with C-M. As the differences between the *Fa* values are usually small, C-M has the higher detection skill. However, because of the mixture of favourable and unfavourable features and the relatively small absolute differences, the rank order between C-M and MASH may depend on the weighting of the individual efficiency characteristics.

The last figure of this section (Fig. 6) shows the rates of first type error in pure white noise process for the OHOMs examined. FTE values scattered in a rather wide range, and they often differ from the OHOM-constructors' intention, e. g. in case of MASH, the applied parameterisation is the one which was developed for approaching 5% FTE, but the results presented indicate a higher than 10% first type error. On the other hand, there is no substantial direct connection between the FTE in white noise and *Fa* in inhomogeneous series. For instance, MASH and PMT have higher FTE than C-M and SNH1, respectively. For dataset A (with only one IH), MASH and PMT display higher false alarm rates indeed than C-M and SNH1 do, but, for all the other test-datasets, C-M and SNH1 produce higher *Fa* values than MASH and PMT, respectively.

4 Discussion

Although the differences between efficiencies for individual OHOMs often seem to be small in Figs. 3, 4 and 5, the examination presented yielded results of theoretical importance. First of all, the sharp drop of detection skill changing the test-dataset from A to B. Although there are five change-points per time series in dataset B, the detection skill for the one with the highest magnitude was tested only,

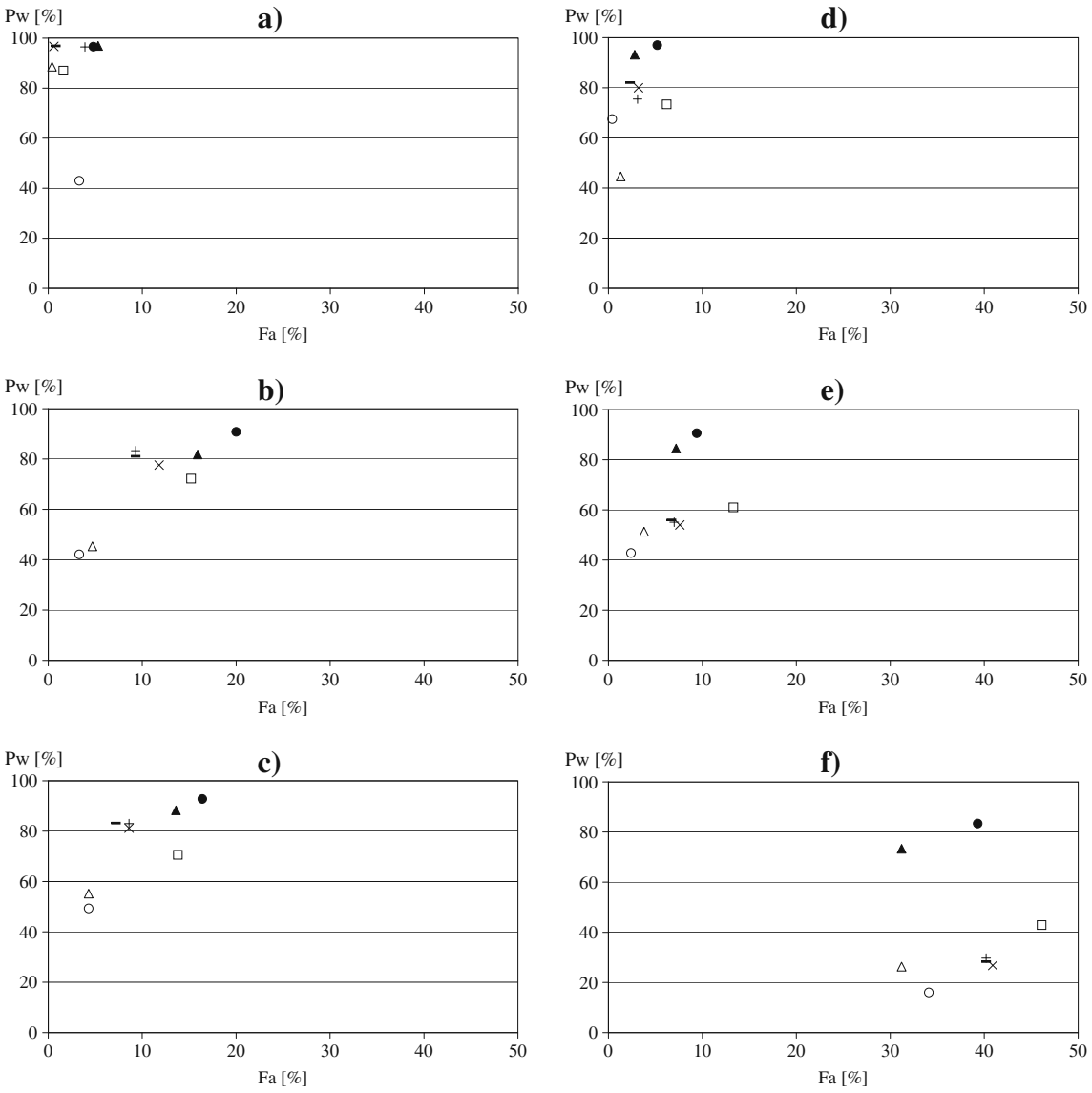

Fig. 5 **a**–**f**. Pw-Fa value-pairs for individual OHOMs in datasets (**a**) to (**f**), respectively. C-M: *filled circle*, MASH: *filled triangle*, MLR: *open square*, PMT: *plus sign*, SNH1: *multiplication sign*, SNH2: *minus sign*, SNHT: *open triangle*, tts: *open circle*

since IHs of $m^*<2$ are not considered in the evaluation. Thus, in both A and B datasets, the detection skill for one IH with $m=3$ was tested, and the four small IHs (with $m=1.5$) in dataset B are considered as part of the noise. (There is an exception from the latter rule, when a change-point was detected within ±2 years time shift relative to the factual timing of a small IH, the detection was neither considered to be right nor to be false; it follows from the definitions given in section 2). Consequently, the differences between the detection skills for A and B datasets indicate the effect of different kind noises on detection skill. The results show that small IHs may have a robust impact on the detection skill of large IHs. Considering this finding and the fact that observed time series generally contain large number of small IHs (see section 2.3), it can be

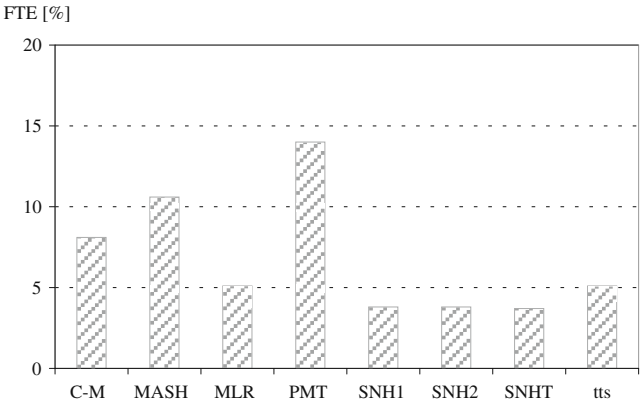

Fig. 6 Rates of first type error (FTE) in pure white noise

concluded that test-datasets simulating "real world" should contain multiple IHs with a wide range of magnitudes.

High false alarm rates for dataset F (Fig. 5f) require some explanation. The amount of change-points with considerable magnitude ($m^*>2$) in dataset F is approximately the half of that in dataset E (Fig. 2). Therefore, the large difference between the *Fa* values for datasets E and F may be surprising. Here, another figure is presented with the magnitude distribution of persistent IHs. In the construction of Fig. 7, only change-points with at least 10 years persistence in both sides of the shift were considered. It can be seen that the amount of persistent IHs with $m^*>2$ is much lower in dataset F than in dataset E. Interestingly, the change between datasets E and F involves only a few percent of all IHs in the datasets, but the efficiency characteristics belonging to them are very different, because of the dominant impact of large persistent IHs on the detection skill. On the other hand, positive skills for dataset F prove that the use of OHOMs generally results in quality improvement, even if the rate of persistent IHs of considerable size is dramatically reduced.

Relying on the theoretical bases, one may suppose four general relations among the efficiencies: (1) OHOMs that examine the whole series as one unit perform better than the ones that examine sections of series with arbitrary borders (*t* test in this study); (2) OHOMs capable of detecting both shifts and trends are better than OHOMs detecting shifts only; (3) OHOMs with semihierarchic algorithm perform better than OHOMs with the common cutting algorithm (Gérard-Marchant et al. 2008; Menne and Williams 2009); (4) Direct detection of multiple IHs is more effective than both SAMA and the common cutting algorithm. However, the real world is different, and the results justify only the first and fourth suppositions. Moreover, there are exceptions even for "rule 4", i.e. in certain cases the performance of direct algorithms is poorer than that of some other methods. Note that other semihierarchic algorithms than SAMA are also known (Lanzante 1996; Gérard-Marchant et al. 2008), but it seems to be unlikely that markedly different efficiencies could be achieved with them than by applying SAMA. The limited skill of OHOMs with cutting algorithm or SAMA for datasets D, E and F likely arises from the fact that the application of very simple models in individual steps is far not the most appropriate way of detection when large number of IHs are present in time series.

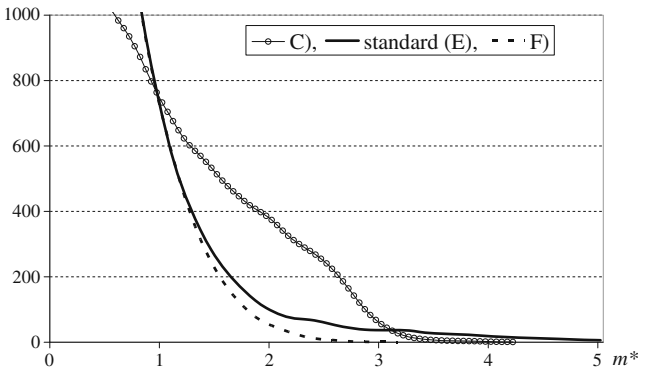

Fig. 7 Distribution of magnitudes (m^*) of persistent IHs for some selected test-datasets

I would like to stress that for time series with large number of IHs, which is typical for observed time series, the direct methods – C-M and MASH – provide the highest efficiency because of their markedly higher power relative to the other OHOMs. In the comparison of efficiencies between C-M and MASH, the results are mixed. Examining the characteristics of E_D, E_T and *Pw*, it can be found that C-M performs better than MASH with only few exceptions. In contrast, the false alarm rate is always higher with C-M than with MASH. *Pw*-*Fa* characteristics in Mestre et al. (2008) are partly similar to the results shown in Fig. 5d. That study also found a higher false alarm rate with C-M than with MASH, but the difference there is very little. The differences in *Fa* values for C-M and MASH are generally not large, that is why detection skill is higher with C-M than with MASH. The use of a mixed characteristic "general efficiency" in Domonkos (2006a) also resulted in a higher efficiency for C-M than for MASH. However, it is clear that the rank order between C-M and MASH depends on the weighting of the individual efficiency characteristics and perhaps on the details of the definitions of correct detection and false detection. Especially when false alarm rate is given an enhanced importance, MASH may easily be more efficient than C-M.

It is hard to find appropriate weightings for efficiency characteristics because the appropriateness depends on the purpose of the homogenisation and, on the other, unexamined parts of the OHOMs. We reject the theory that making an additional error in an observed time series by a false detection is generally more harmful than keeping an existing error unchanged. So-called raw time series have usually been checked several times by several investigators before homogenisation, and some suspicious values might had been altered by the time the "raw" time series is presented to the analysers for the first time. Another argument is that the origination of the errors has no impact on the climate variability investigations, while the frequency, magnitude and the possible spatial coherence of the errors have real importance. On the other hand, if iteration is applied during the homogenisation of a dataset, both of false detections and lack of correct detections may have negative feed-back on the results. Therefore, further examinations are needed to find realistic weightings for the efficiency characteristics.

The results show that there is no direct connection between the rate of first type error in white noise process and the efficiency characteristics examined. OHOMs with high FTE may have low false alarm rate, and vice versa.

Therefore, although a low FTE is a traditional requirement from homogenisation methods, it is not necessary at all, if one applies an auxiliary method to select inhomogeneous time series from datasets for investigation and starts to work with the chosen OHOM afterwards. A consequence of this freedom is that the optimal parameterisation of OHOMs may be far outside from the range which is bounded by the application of some low FTE target-values. Optimal parameterisations can be determined empirically only (Domonkos 2006a). A serious problem of the optimisation is that the optimum depends on several factors, i.e. depends on the purpose of the homogenisation, on the characteristics of the dataset for examination and on the operation of other components of the OHOMs than the detection part. The problem is the same when the Caussinus–Lyazrhi criterion is compared with other criterions in the C-M method to find the optimal number of change-points.

5 Conclusions

A challenging task is to evaluate the efficiency of OHOMs. Even the concept of "efficient" may depend on the goal of the user. A high skill in finding the timings and magnitudes of IHs with substantial sizes does not guarantee similarly high reliability in reproducing climatic trends and variability, as it was illustrated by several examples of the paper. Consequently, a profound evaluation of efficiency must be based on sub-evaluations relying on different aspects of the homogenisation task. Efficiency evaluations shown in this paper are examples only; yet, some general conclusions can be drawn relying on them.

- The use of OHOMs is generally beneficial for the quality of time series. When OHOMs are used for time series with at least one change-point of substantial size, the residual error is mostly lower than 50% of the error in time series without homogenisation.
- Small IHs have robust impact on the detection results, even if the skill of detecting large IHs is examined only.
- Often, larger differences appear between efficiencies for different kind datasets than for different OHOMs. The use of different efficiency evaluations may result in different rank-orders of the performances of OHOMs.
- Several relations among efficiency characteristics for different OHOMs seem to be stable; they depend little on the test data set and efficiency measure applied.
- Comparing the results of different detecting methods with various approaches towards detecting multiple IHs, it seems that direct algorithms are much more beneficial in identifying complex IH-structures than the cutting algorithm or the SAMA. Although the differences are small in the skill of trend estimation, the differences in the power are striking, and the impact of these differences seems to be robust on the detection skill and other efficiency measures. Using the test-dataset of fairly high similarity to relative time series derived from observational data in Central Europe, the Caussinus–Mestre method and MASH show the highest efficiency. Other tests can also be useful in special tasks, e.g. in checking the significance of a presumed change-point relying on metadata.

This study aimed only at the efficiency of the detection parts of OHOMs and does not address differences related to time series comparison methods, iterative improvements through repeated application of OHOMs, as well as from possible supplements with the use of metadata information or other subjective steps. The results show that there are still many tasks that the climatologists have to accomplish in order to obtain reliable knowledge about the efficiency of homogenisation methods.

Acknowledgements The research was partially funded by the COST ES0601 project. The author thanks Matthew Menne and an anonymous reviewer for their useful comments.

Appendix I

Simulation of the standard test-dataset

1. 196-year-long series are generated, and always, the slices of years 48–147 are the target series.
2. IHs and noises are introduced in each year (but their values can be 0, naturally).
3. Types of the terms for introduction to time series: (a) long-term IH (y), (b) short-term IH (z) and (c) white noise (w). A certain part of y- and z-type terms is handled as noise (cf. step 10).
4. Forms of the IHs: (a) sudden shift, (b) gradual change, (c) platform-like change, (d) bias for one specific year. Form (d) is a specific case of class (c).
5. Introduction of long-term IHs.

5.1: Size and direction of the IH

This term includes an IH whose magnitude can be large, with the probability given in K_1, as well as a small IH with the probability given in K_2:

$$\Delta y_i' = K_1(q_1) \cdot \mathrm{sign}\,(0.5 - q_2) \cdot (8 + 4p) \cdot q_3^{6+4p}) + K_2(q_4) \cdot G_1, \qquad \text{(A1)}$$

where $K_1(a)=1$, if $a<0.012$, and $K_1(a)=0$ otherwise; $K_2(a)=1$, if $a<0.07$, and $K_2(a)=0$ otherwise; q (with all indices): variable of the uniform distribution over the period [0,1) p has the same distribution as q does, but p is constant for a given time series. Δ denotes that (A1) is not for

substituting, but for modifying the earlier value of y_i. Apostrophe above y shows that values gained by (A1) are modified in certain cases (see below) before the introduction of Δy_i . If $\Delta y_i'=0$ the steps 5.2 and 5.3 are omitted.

5.2: Form of the IH

The form of Δy_i' is (A) sudden shift, (B) gradual change or (C) platform-like change, with 0.4, 0.25 and 0.35 probability, respectively.

For (A)- and (B)-form IHs a negative autocorrelation is present:

$$\Delta y_i = \sqrt{1-r^2}\cdot \Delta y_i' + r\cdot F, \tag{A2}$$

where $F=0$ for the first (A)- or (B)-form IH of the series, and $F=\Delta y_k$ otherwise, k indicates the year of the previous introduction of (A)- or (B)-form IH, and r=–0.5.

For (C)-form IHs:

$$\Delta y_i = \Delta y_i' \tag{A3}$$

5.3: Calculation of the y_i components of the series

(A)-form IHs:

$$y_j = y_{j,-1} + \Delta y_i \quad \text{for each } j\in[i,n], \tag{A4}$$

where $y_{j,-1}$ denotes the value of term y_j before the ongoing modification.

For (B)- and (C)-form IHs, duration-values must be paired at first. For B-form, IHs the duration D_1 is:

$$D_1 = 5 + 2\cdot \mathrm{Int}\left(48\cdot q_5^{1.5}\right) \tag{A5}$$

("Int" denotes integer part), and the appearance of the IH is:

$$y_j = y_{j,-1} + \frac{(j-i+0.5D_1)\Delta y_i}{D_1} \tag{A6}$$

$$\text{for each } j\in[i-0.5D_1, i+0.5D_1-1],$$

while for (C)-form IHs:

$$D_2 = \mathrm{Int}\left(30\cdot q_6^{1.5}\right), \tag{A7}$$

$$y_j = y_{j,-1} + \Delta y_i \ \text{ for each } j\in[i,i+D_2]. \tag{A8}$$

6. Introduction of short-term IHs

The size and the direction of this term is calculated by the same functions as those of long-term IHs (A1), but the frequencies (determined by the K-functions) are different:

$$\Delta z_i' = K_3(q_7)\cdot sign(0.5-q_8)\cdot(8+4p)\cdot q_9^{6+4p}) + K_4(q_{10})\cdot G_2, \tag{A9}$$

where $K_3(a)$=1, if $a<0.04-0.03p$, and $K_3(a)$=0 otherwise; $K_4(a)$=1, if $a<0.5-0.4p$, and $K_4(a)$=0 otherwise. The ongoing modification has a negative autocorrelation (r=–0.5) with the z value accumulated prior.

$$\Delta z_i = \sqrt{1-r^2}\cdot \Delta z_i' + r\cdot z_{i,-1} \tag{A10}$$

The form of this term is always platform-like change. Its duration is given by D_3.

$$D_3 = \mathrm{Int}\left(\frac{12\cdot q_{11}^3}{1+0.3|\Delta z_i|}\right), \tag{A11}$$

$$z_j = z_{j,-1} + \Delta z_i \ \text{ for each } j\in[i,i+D_3]. \tag{A12}$$

7. Introduction of white noise term:

$$w_i = G_3 \tag{A13}$$

8.

$$\mathbf{X} = \mathbf{Y} + \mathbf{Z} + \mathbf{W} \tag{A14}$$

9. Serial correlation of **X** is calculated, and the series is added to the test-dataset if the value is not lower than 0.4, while it is discarded otherwise.

10. A part of long-term IHs (**Y**) and short-term IHs (**Z**) is not considered to be errors of the candidate series, so it is handled as noise. The rate of this type noise increases with decreasing IH magnitudes, and it is higher for platform-like changes than for change-points and gradual changes. As a consequence of these noise terms, the model series of the standard dataset is

$$\mathbf{X} = \mathbf{H} + \mathbf{W} + \mathbf{W^*} \tag{A15}$$

where

$$\mathbf{W^*} = \mathbf{Y_w} + \mathbf{Z_w} \tag{A16}$$

$$\mathbf{H} = \mathbf{Y} - \mathbf{Y_w} + \mathbf{Z} - \mathbf{Z_w} \tag{A17}$$

The index w denotes noise part. The probability (P) of that a given term is considered to be noise, is determined according to the rules below:

For platform-like IHs, the probability P_1 is given by:

$$P_1 = \max(0.6 - 0.4\cdot|\Delta y_i|, 0), \tag{A18}$$

where Δy_i is determined by Formulae (A1) and (A3). (A18) is applied also for Δz-type IHs.

For change-points and gradual changes

$$P_2 = \max(0.3 - 0.4 \cdot |\Delta y_i|, 0). \quad \text{(A19)}$$

where Δy_i is determined by Formulae (A1) and (A2).

Appendix II

Simulation of the quasi-standard test-dataset

The procedure is the same as for the standard dataset, except that K_1 is always equal to 0 in formula (A1). As a result of this change, the frequency of persistent large IHs is much lower in this dataset than in the standard dataset.

References

Aguilar E, Auer I, Brunet M, Peterson TC, Wieringa J (2003) WMO Guidelines on climatemetadata and homogenization. WMO, Geneva, WCDMP-No. 53, WMO-TD No 1186

Alexandersson H (1986) A homogeneity test applied to precipitation data. J Climatol 6:661–675

Alexandersson H, Moberg A (1997) Homogenization of Swedish temperature data. Part I: homogeneity test for linear trends. Int J Climatol 17:25–34

Auer I et al (2005) A new instrumental precipitation dataset for the greater Alpine region for the period 1800-2002. Int J Climatol 25:139–166. doi:10.1002/joc.1135

Beaulieu C, Seidou O, Ouarda TBMJ, Zhang X, Boulet G, Yagouti A (2008) Intercomparison of homogenization techniques for precipitation data. Water Resour Res 44:W02425. doi:10.1029/2006WR005615

Brunet M, Saladié O, Jones P, Sigró J, Aguilar E, Moberg A, Lister D, Walther A, Lopez D, Almarza C (2006) The development of a new dataset of Spanish daily adjusted temperature series (SDATS) (1850-2003). Int J Climatol 26:1777–1802. doi:10.1002/joc.1338

Buishand TA (1982) Some methods for testing the homogeneity of rainfall records. J Hydrol 58:11–27

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

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

DeGaetano AT (2006) Attributes of several methods for detecting discontinuities in mean temperature series. J Climate 19:838–853. doi:10.1175/JCLI3662.1

Domonkos P (2006a) Testing of homogenisation methods: purposes, tools and problems of implementation. In: Szalai S (ed) Proceedings of the fifth seminar for homogenization and quality control in climatological databases. Hungarian Meteorological Service, Budapest, pp 126–145

Domonkos P (2006b) Application of objective homogenization methods: inhomogeneities in time series of temperature and precipitation. Időjárás 110:63–87

Domonkos P, Štěpánek P (2009) Statistical characteristics of detectable inhomogeneities in observed meteorological time series. Studia Geoph et Geod 53:239–260. doi:10.007/s11200-009-0015-9

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. doi:10.1007/s00704-004-0088-x

Ducré-Robitaille J-F, Vincent LA, Boulet G (2003) Comparison of techniques for detection of discontinuities in temperature series. Int J Climatol 23:1087–1101. doi:10.1002/joc.924

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

Gérard-Marchant PGF, Stooksbury DE, Seymour L (2008) Methods for starting the detection of undocumented multiple change-points. J Climate 21:4887–4899. doi:10.1175/2008JCLI1956.1

Hawkins DM (1972) On the choice of segments in piecewise approximation. J Inst Math Appl 9:250–256

Lanzante JR (1996) Resistant, robust and non-parametric techniques for the analysis of climate data: theory and examples, including applications to historical radiosonde station data. Int J Climatol 16:1197–1226

Menne MJ, Williams CN Jr (2005) Detection of undocumented changepoints using multiple test statistics and composite reference series. J Climate 18:4271–4286. doi:10.1175/JCLI3524.1

Menne MJ, Williams CN Jr (2009) Homogenization of temperature series via pairwise comparisons. J Climate 22:1700–1717. doi:10.1175/2008JCLI2263.1

Mestre O, Domonkos P, Lebarbier E, Picard F, Robin S (2008) Comparison of change-point detection methods in the mean of Gaussian processes. In: Sixth seminar for homogenization and quality control in climatological databases (in print)

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 17:35–54

Peterson TC et al (1998) Homogeneity adjustments of in situ atmospheric climate data: a review. Int J Climatol 18:1493–1517

Sneyers R (1997) Climate chaotic instability. Statistical determination – theoretical backgrounds. Environmetrics 8:517–532

Syrakova M (2003) Homogeneity analysis of climatological time series – experiments and problems. Időjárás 107:31–48

Szentimrey T (1999) Multiple Analysis of Series for Homogenization (MASH). In: Szalai S, Szentimrey T, Szinell CS (ed) Proceedings of the second seminar for homogenization of surface climatological data. World Meteorological Organization, WCDMP-41, WMO-TD 932: 27-46

Titchner HA, Thorne PW, McCarthy MP, Tett SFB, Haimberger L, Parker DE (2009) Critically reassessing tropospheric temperature trends from radiosondes using realistic validation experiments. J Climate 22:465–485. doi:10.1175/2008JCLI2419.1

Vincent LA (1998) A technique for the identification of inhomogeneities in Canadian temperature series. J Climate 11:1094–1104

Wang XL, Wen QH, Wu Y (2007) Penalized maximal t test for detecting undocumented mean change in climate data series. J Appl Meteor Climatol 46/6: 916-931. doi:10.1175/JAM2504.1