IDŐJÁRÁS

QUARTERLY JOURNAL
OF THE HUNGARIAN METEOROLOGICAL SERVICE

Special Issue of the COST-ES0601 (HOME) ACTION: Advances in homogenization methods of climate series: an integrated approach

Guest Editors: **Mónika Lakatos** and **Tamás Szentimrey**

CONTENTS

****

http://www.met.hu/Journal-Idojaras.php

VOL. 117* NO. 1 * JANUARY – MARCH 2013

IDŐJÁRÁS

Quarterly Journal of the Hungarian Meteorological Service

Editorial Office: Kitaibel P.u. 1, H-1024 Budapest, Hungary
P.O. Box 38, H-1525 Budapest, Hungary
E-mail: journal.idojaras@met.hu
Fax: (36-1) 346-4669

Indexed and abstracted in Science Citation Index Expanded™ and Journal Citation Reports/Science Edition
Covered in the abstract and citation database SCOPUS®

Subscription by

mail: IDŐJÁRÁS, P.O. Box 38, H-1525 Budapest, Hungary
E-mail: journal.idojaras@met.hu

Editorial

Special Issue of the COST-ES0601 (HOME) Action: Advances in homogenization methods of climate series: an integrated approach

Long term instrumental climate records are the basis of climate research. However, these series are usually affected by inhomogeneities (artificial shifts), due to changes in the measurement conditions (relocations, instrumentation). As the artificial shifts often have the same magnitude as the climate signal, such as long-term variations, trends, or cycles, a direct analysis of the raw data series can lead to wrong conclusions about climate change. In order to deal with this crucial problem, many statistical homogenization procedures have been developed for detection and correction of these inhomogeneities.

The large number of different homogenization methods and the need for a realistic comparative study was the reason to start a coordinated European initiative, the COST Action ES0601: Advances in Homogenization Methods of Climate Series: an integrated approach (HOME). Its main objective was to review and improve common homogenization methods, and to assess their impact on climate time series. As one of the high importance achievements of the Action a benchmark dataset was generated for comparing monthly homogenization algorithms. The main results of this examination were published in the journal Climate of the Past.

The COST HOME Action ended in October 2011. The final meeting of the Management Committee was organized in Budapest together with the 7th Seminar for Homogenization and Quality Control in Climatological Databases. The Homogenization Seminars are traditionally held in Budapest and hosted by the Hungarian Meteorological Service from 1996. The jointly organized Seminar and the final MC meeting was a good occasion for conversation between the participants of the HOME Action and other researchers of the homogenization community. During this meeting, publishing a special issue of the COST HOME Action was suggested. It is a pleasure for us that this publication has been realized at the Quarterly Journal of the Hungarian Meteorological Service as a Special Issue of the Action.

This Special Issue includes eight papers which are covering wide range of topics on homogenization. The first five articles are connected mainly with the homogenization on monthly scale, while the other three ones focus rather on the homogenization of daily series. In both cases, theoretical aspects and practical applications are discussed and presented alike.

We are very grateful to the Editor–in-Chief of IDŐJÁRÁS supporting the progress on the field of homogenization, thank to the authors of the articles for their high scientific level work, and also to the reviewers supporting the improvement of papers with their critical comments and recommendations keeping the high standards of the Journal. We have to underline the hard work of the Executive Editor of the Journal, therefore, we express our thanks together with the authors for that.

Tamás Szentimrey and *Mónika Lakatos*
Guest Editors

IDŐJÁRÁS

Quarterly Journal of the Hungarian Meteorological Service
Vol. 117, No. 1, January–March 2013, pp. 1-34

On the multiple breakpoint problem and the number of significant breaks in homogenization of climate records

Ralf Lindau* and **Victor Venema**

Meteorological Institute, University of Bonn
Auf dem Hügel 20, D-53121 Bonn, Germany

**Corresponding author E-mail: rlindau@uni-bonn.de*

(Manuscript received in final form November 8, 2012)

Abstract—Changes in instrumentation and relocations of climate stations may insert inhomogeneities into meteorological time series, dividing them into homogeneous subperiods interrupted by sudden breaks. Such inhomogeneities can be distinguished from true variability by considering the differences compared to neighboring stations. The most probable positions for a given number of break points are optimally determined by using a multiple-break point approach. In this study the maximum external variance between the segment averages is used as decision criterion and dynamic programming as optimization method. Even in time series without breaks, the external variance is growing with any additionally assumed break, so that a stop criterion is needed. This is studied by using the characteristics of a random time series. The external variance is shown to be beta-distributed, so that the maximum is found by solving the incomplete beta function. In this way, an analytical function for the maximum external variance is derived. In its differential form our solution shows much formal similarities to the penalty function used in *Caussinus* and *Mestre* (2004), but differs numerically and exhibits more details.

Key words: Climate records, homogenization, multiple break point detection, stop criterion for search algorithms, dynamic programming, penalty term.

1. Introduction

Multiple-century long instrumental datasets of meteorological variables exist for Europe (*Brunetti et al.*, 2006; *Bergström* and *Moberg*, 2002; *Slonosky et al.*, 2001). Such series provide invaluable information on the evolution of the climate. However, between the Dutch Golden Age, the French and the industrial

revolution, the rise and the fall of communism, and the start of the internet age, inevitably many changes have occurred in climate monitoring practices (*Aguilar et al.*, 2003; *Trewin*, 2010). The typical size of temperature jumps due to these changes is similar to the global warming in the 20th century, and the average length of the periods between breaks in the climate records is 15 to 20 years (*Auer et al.*, 2007; *Menne* and *Williams*, 2009). Clearly, such changes interfere with the study of natural variability and secular trends (*Rust et al.*, 2008; *Venema et al.*, 2012).

Technological progress and a better understanding of the measurement process have led to the introduction of new instruments, screens, and measurement procedures (*MeteoSchweiz*, 2000). In the early instrumental period, temperature measurements were often performed under open shelters or in metal window screens on a North facing wall (*Brunetti et al.*, 2006), which were replaced by Montsouri (*Brunet et al.*, 2011), Wild, and various Stevenson screens (*Nordli et al.*, 1997; *Knowles Middleton*, 1966), and nowadays more and more by labor-saving automatic weather stations (*Begert et al.*, 2005). Every screen differs in their protection against radiation, wetting, as well as their quality of ventilation (*Van der Meulen* and *Brandsma*, 2008). Initially many precipitation observations were performed on roofs. As it was realized that many hydrometeors do not enter the gauge due to wind and turbulence, especially in case of snow, the observations were taken nearer the ground, and various types of wind shields were tested leading to deliberate inhomogeneities (*Auer et al.*, 2005). Due to the same effect, any change in the geometry of a rain gauge can lead to unintended inhomogeneities.

Inhomogeneities are frequently caused by relocations, either because the voluntary observer changed, because the observer had to move or because the surrounding was no longer suited for meteorological observations. Changes in the surrounding can lead to gradual or abrupt changes, for example gradual increases in urbanization or growing vegetation or fast changes due to cutting of vegetation, buildings that disrupt the flow or land-use change.

Changes in the observations should be documented in the station history. It is recommended to perform several years of parallel measurements in case of changes (*Aguilar et al.*, 2003). However, it is not guaranteed that metadata is complete, thus statistical homogenization should always be performed additionally. The dominant approach to homogenize climate networks is the relative homogenization method. This principle states that nearby stations are exposed to almost the same climate signal, and thus, the differences between nearby stations can be utilized to detect inhomogeneities (*Conrad* and *Pollack*, 1950). By computing the difference time series, the interannual weather noise, decadal variability, and secular trends are strongly reduced. Consequently, a jump in single station becomes much more salient.

The two fundamental problems of homogenization are that the nearby stations are also inhomogeneous and that typically more than one break is

present. Recent intercomparison studies by *Domonkos* (2011a) and *Venema et al.* (2012) showed that the best performing algorithms are the ones that attack these two problems directly. This study will focus on the multiple-breakpoint problem.

Traditionally the multiple-breakpoint problem is solved by applying single-breakpoint algorithms multiple times. Either a cutting algorithm is applied: the dataset is cut at the most significant break and the subsections are investigated individually until no more breaks are found or the section become too short; see, e.g., *Easterling* and *Peterson* (1995). A variation on this theme is a semi-hierarchical algorithm, in which potential breakpoints are found using the cutting algorithm, but before correcting a potential break its significance is tested anew (*Alexanderson* and *Moberg*, 1997). According to *Domonkos* (2011a), this improvement has a neutral effect on the efficiency of homogenization.

The first algorithms solving the multiple-breakpoint problem directly are MASH (*Szentimrey*, 1996, 1999) and PRODIGE (*Caussinus* and *Mestre*, 1996, 2004). MASH solves the problem with a computationally expensive exhaustive search (*Szentimrey*, 2007). PRODIGE solves the problem in two steps. First, the optimal position of the breaks for a given number of breaks is found using a computationally fast optimization approach called dynamic programming (*Bellman*, 1954; *Hawkins*, 1972). Second, the number of breaks is determined by minimizing the internal variance within the subperiods between two consecutive breaks plus a penalty for every additional break (*Caussinus* and *Lyazrhi*, 1997). The penalty term aims to avoid adding insignificant breaks.

Recently, *Domonkos* (2011b) expanded ACMANT, which is based on the generic PRODIGE algorithm, by searching for common breaks in the annual mean and the size of the annual cycle. *Picard et al.* (2011) developed an alternative version, in which not only pairs, but all data in the network are jointly taken into account for optimization. ACMANT, PRODIGE, and the joint detection method of *Picard et al.* (2011) are implemented in the software package HOMER (*Mestre et al.*, 2012). *Nemec et al.* (2012) used PRODIGE with three different criteria for the assessment of the number of breaks. Beyond dynamic programming, genetic algorithms (e.g., *Li* and *Lund*, 2012; *Davis et al.*, 2012) and simulated annealing (*Lavielle*, 1998) are alternatively used for reducing the computational demand.

Not all inhomogeneities are abrupt changes, some changes are more gradual (*Lindau*, 2006). Such trends are explicitly considered by some homogenization algorithms (*Vincent*, 1998; *Menne* and *Williams*, 2009). Using the HOME benchmark dataset in which 10% of the stations contained a local trend inhomogeneity, a blind experiment with two versions of PRODIGE has been performed (*Venema et al.*, 2012). In the main version only breaks have been used for homogenization, and in the alternative version multiple breaks in one direction are combined into a slope correction. These two versions have a very similar performance. One of the reasons may be that not many local trends

had to be introduced. Still this suggests that trend inhomogeneities can be reasonably well modeled by multiple breaks. Consequently, this paper will only consider break inhomogeneities.

To characterize breaks within a time series, it is helpful to decompose the total variance of the time series into two terms: the internal and the external variance. Consider a time series with k breaks dividing it into $k + 1$ subperiods (*Fig. 1*). In this concept, the variance within the subperiods is referred to as the internal variance, whereas the variance between the means of different subperiods is the external variance. The decomposition with the maximum external variance defines the optimum positions of breaks for a given number of breaks.

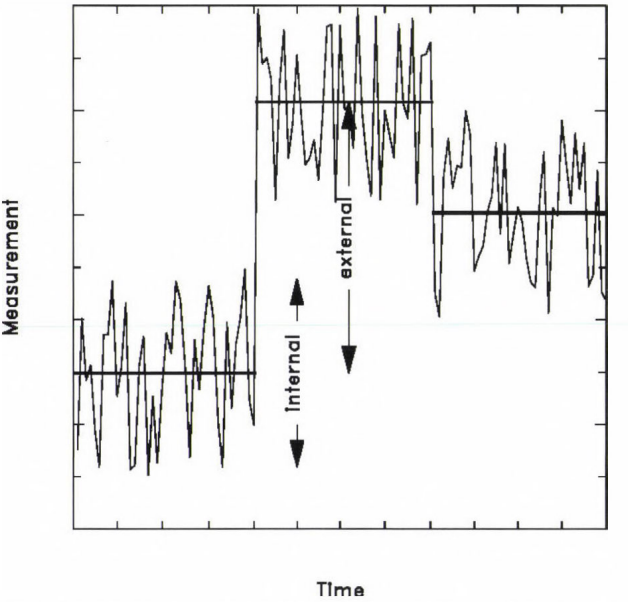

Fig. 1. Sketch to illustrate the occurrence of breaks in climate records and the related expressions, internal and external variance.

As we use internal and external variance as the basic concept to characterize breaks, an exact quantitative formulation is necessary. *Lindau* (2003) discussed the decomposition of variance and showed that the total variance of a time series can be divided into three parts:

$$\frac{1}{n-1}\sum_{i=1}^{N}\sum_{j=1}^{n_j}\left(x_{ij}-\bar{x}\right)^2=\frac{1}{n}\sum_{i=1}^{N}n_i(\bar{x}_i-\bar{x})^2+\frac{1}{n}\sum_{i=1}^{N}\sum_{j=1}^{n_j}\left(x_{ij}-\bar{x}_i\right)^2$$

$$+\frac{1}{n(n-1)}\sum_{i=1}^{N}\sum_{j=1}^{n_j}\left(x_{ij}-\bar{x}\right)^2 \quad . \tag{1}$$

In Eq. (1), the variance of a time series of length n is considered. It contains N subperiods, each comprising n_i members. Individual members are denoted by x_{ij}, where i specifies the subperiod and j the temporal order within the respective subperiod. The mean of the ith subperiod is denoted by $\bar{x}_i$ and the overall mean of the entire time series by $\bar{x}$, without any index. The total variance on the left hand side is decomposed into the three parts on the right hand side of Eq. (1). These are equal to the external and the internal variance plus, as third term, the error of the total mean. As the last term is constant for a given time series, the sum of internal and external variance is constant, too. Consequently, we can formulate an alternative criterion for the optimum decomposition of a time series into subsegments being a minimum internal variance.

However, two problems arise. The first is of practical nature. The number of possible decompositions is normally too large for a simple test of all permutations. The second is rather fundamental. For a fixed number of breaks, the maximum external variance is actually a reasonable criterion for the optimum decomposition. However, it is obvious that for zero breaks, the entire variance is internal, whereas it is fully external for n–1 breaks. During the transition from 0 to n–1 breaks, more and more variance is converted from internal to external, so that the internal variance is a monotonously falling function of the break number k. Consequently, we need a second criterion for the optimum *number* of breaks. As this is the critical problem for any multiple-breakpoint detection algorithm, the discussion and proposed solution of this problem built the major part of this paper. However, initially also the first minor problem and its solution are shortly described in the following.

There exists a large number of possibilities to decompose a time series of length n into a fixed number of N subsegments: it is equal to $\binom{n-1}{N-1}$. Even for a moderate length of $n=100$ and ten subsegments, there are already more than 10^{12} combinations, so that the testing of all permutations is mostly not feasible. This problem is already solved by the so-called dynamic programming method, firstly inspired by *Bellman* (1954). Originally designed for economic problems, this method is by now established in many different disciplines, in climate research (*Caussinus* and *Mestre*, 2004) as well as in biogenetics (*Picard et al.*, 2005). As we will also use dynamic programming later on, we describe shortly how we applied this technique.

2. *Dynamic programming*

We begin with the optimum solution for a single break point. In this case, simple testing of all possibilities is still feasible as only n–1 permutations exist. Afterwards, the best break position together with its respective internal variance is known. The basic idea is now to find an optimum decomposition not only for the entire time series, but also for all truncated variants of any length l. There exist n–1 variants, all beginning with the first time point. The first variant ends at the second time point, the second at the third time point, and the last variant is equal to the entire time series. For each of these variants an optimum position of a single breakpoint is searched and stored together with the criterion on which the decision is made, i.e., its internal variance. In the next step we consider what happens if the truncated variants are filled up to the original length n. In this case the internal variance consists of two contributions: that of the truncated variant, plus that of the added rest. For this step, it is, of course, necessary that the used criterion is additive, which is fulfilled for variances. Consequently, we can test a number of n–1 filled-up variants. That variant, where the combined internal variance is minimal, is then the optimum solution for two break points. The first break is situated within the truncated time series; the second is equal to the length l of the truncated series itself, because here is the break between the two combined time series.

To expand the method from two to three and more breaks, some more work is necessary already at the beginning. So far the truncated variants are always filled up to the entire length n. But the starting point for the proceeding from one to two breaks are, as described above, known previous solutions for all lengths. Consequently, to proceed from two to three breaks, we need not only the best two-break solution for the entire length n, but the solutions for every length. Thus, also all shorter fillings are performed so that we obtain the optimum two-break solution not only for the final time series length n, but also for every shorter length between 2 and n. This set of solutions is then used accordingly as basis to find the three-break solution. Filling up the time series to the full length would be sufficient if we want to stop at three breaks. However, if the method should be continued to higher break numbers, again a full set of three-break solutions is needed.

Thus, the solution for k breaks is found by testing only n–1 truncated and refilled optima, where the truncated part contains already the optimum distribution of k–1 breaks. To perpetuate the method for k+1 breaks, each truncated optimum has to be refilled to all possible length so that a number of cases in the order of n^2 has to be calculated. This reduces the number of cases from the order of $\binom{n}{k}$ to n^2, which facilitates a much faster processing.

3. *Outline of the paper*

In the above described way, the optimum positions for a given number of breaks can be calculated. Minimum internal variance is serving as criterion, and dynamic programming avoids a time consuming exhaustive search. However, as mentioned above, there is still a problem left. The absolute minimum of internal variance being equal to zero would be attained by inserting n–1 breaks into the time series, which is obviously not the optimum solution. Instead, we need to define which number of breaks is appropriate.

A state-of-the-art method for detecting breaks is PRODIGE (*Caussinus* and *Mestre*, 2004). Although using a log-likelihood method, it is based on the minimization of the internal variance and does not differ essentially from the procedure described here so far. PRODIGE uses a penalty term to ensure that the search stops at a reasonable number of breaks. This penalty term is adopted from *Caussinus* and *Lyazrhi* (1997). Similar to PRODIGE, *Picard et al.* (2005) applied a log-likelihood method to minimize the internal variance, but developed a specific penalty term. Before, they discussed different commonly used penalty terms, such as the Information Criteria AIC and BIC, based on *Akaike* (1973), and found that these penalty terms suffer from different weaknesses.

In the remaining part of this study, we derive an alternative stop criterion based on the idea that the external variance is the key parameter, which defines the optimum solutions. We will use the characteristics of a random standard normal distributed time series as reference. Only if the optimum solution for an additionally inserted break gains significantly more external variance than the expected amount for a random time series, an increased break number is justified. Thus, it is necessary to describe mathematically how the external variance of random data increases with increasing number of breaks, so that it can be used as reference for real data.

In a first step, we derive the statistical distribution that can be expected for the external variance v. In Section 4, we show theoretically that the χ^2 distribution would be a good candidate. In Section 5, we show by empirical tests that the related Beta distribution is even better suited to describe the external variance. To identify the optimum solution for the decomposition, we use, as mentioned, the maximum external variance. Consequently, we have to find the maximum value within a Beta distribution, identical to its exceeding probability, which is performed in Section 6. For that purpose, the definite integral of the Beta distribution, known as the incomplete Beta function, has to be solved. From this formulation, the rate of change of the external variance v for growing break numbers k is derived in Section 7. This derivative dv/dk is then integrated and a formulation for $v(k)$ is presented.

In its differential form our solution shows much formal similarities to the penalty function used in *Caussius* and *Mestre* (2004), but it differs numerically

and exhibits more details. In Section 8 we discuss these differences and propose finally a revision.

4. Theoretical characteristics of random data

Consider a random standard normal distributed time series $N(0,1)$ with k breaks inserted, so that the number of segments N is:

$$N = k + 1 \ . \tag{2}$$

According to Eq. (1) the external variance v is:

$$v = \frac{1}{n}\sum_{i=1}^{N} n_i(\bar{x}_i - \bar{x})^2 \quad . \tag{3}$$

As standard normal distributed data is considered, $\bar{x}=0$ and $\sigma_x=1$. Furthermore, we are interested here in the statistics of external variance for many realizations as produced by $\binom{n-1}{k}$ permutations of break positions for a fixed number of breaks. Averages over these permutations are denoted by brackets, whereas averages over individual data points within a time segment are overlined.

$$[v] = \left[\frac{1}{n}\sum_{i=1}^{N} n_i\,\bar{x}_i^{\,2}\right] \quad . \tag{4}$$

Consider now the segment averages $\bar{x}_i$, which are the critical constituents of $[v]$ in Eq. (4). Their expected mean is equal to zero, since random data with $\bar{x} = 0$ is assumed. Only the finite number of segment members causes the segment means to scatter randomly around zero. As the members of a segment are standard normal distributed, the standard deviation of any segment mean is equal to $1/\sqrt{n_i}$.

$$\bar{x}_i \sim N\left(0, \frac{1}{n_i}\right) \quad . \tag{5}$$

If the segment means are multiplied by the square root of the number of segment members (Eq. (6)), the distribution is broadened in such a way that a standard normal distribution is obtained. These modified means can be defined as to y_i.

$$y_i := \sqrt{n_i}\,\bar{x}_i \sim N(0,1) \quad . \tag{6}$$

Inserting this definition into Eq. (4) leads to:

$$[v] = \left[\frac{1}{n}\sum_{i=1}^{N} {y_i}^2\right] = \frac{N-1}{n} = \frac{k}{n} \quad . \tag{7}$$

The second equal sign in Eq. (7) follows, because the squared sum over standard normal distributed data is N–1, which is directly evident from the definition of standard deviation. Furthermore, the brackets can be omitted as both the total length of the time series n and the number of segments N are constants for all permutations subsumed under the brackets. The last equal sign follows from Eq. (2), which just states that there is always one segment more than breaks.

From Eq. (7), we can conclude the following. The average external variance increases linearly with the number of inserted breaks k. Such a linear increase of v could be expected if one of the $\binom{n-1}{k}$ segmentation possibilities for a given number of breaks is chosen randomly. However, actually we select always the optimum segmentation as given by the above described dynamic programming. Consequently, we are less interested in the expected mean, but in the best of several attempts. In order to conclude such an extreme, the distribution has to be known.

For this purpose, let us go back to Eq. (3) where we insert again Eq. (6). It follows the same relationship as given in Eq. (7), but without averaging brackets, according to:

$$v = \frac{1}{n}\sum_{i=1}^{N} {y_i}^2 \quad . \tag{8}$$

It is striking that Eq. (8) is nearly identical to the definition of a χ^2 distribution, which is as follows: N values are randomly taken out of a standard normal distribution, which are then squared and added up. By repeating this procedure several times, these square sums form a χ^2 distribution with N being the number of degrees of freedom. Remembering that y_i is standard normal distributed, it becomes obvious that Eq. (8) reproduces this definition. The difference is that we divide finally by n. But, hereby, no substantial change is performed, as n is a constant equal to the length of the considered time series.

Consequently, we can conclude that v (actually nv), must be χ^2 distributed with k being the degree of freedom, according to:

$$f(x) = \frac{x^{\frac{k-2}{2}} e^{-\frac{x}{2}}}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)} \quad . \tag{9}$$

However, there is an important restriction of this rule. As v is normalized, it is confined between 0 and 1, whereas normal distributed data have no upper limit. The number of breaks k is inversely proportional to the number of segment members n_i. Therefore, the standard deviation of segment averages (Eq. (5)), is small compared to 1 for low break numbers. In this case the normal distribution is a good approximation for $\overline{x_k}$. However, with increasing break number, n_i decreases so that the standard deviation is approaching 1. Assume, e.g., a time series of length 100 with 25 breaks. n_i is then in the order of 4, so that the standard deviation of the $\overline{x_k}$ becomes 0.5 (Eq. (5)). Assuming still a normal distribution is no longer appropriate, as the true frequency for $\overline{x_k}$=1 is zero by definition, whereas the normal distribution at 2 standard deviations is not exactly zero. For the distribution of v it means that we have to expect a kind of confined χ^2 distribution, which is defined exclusively between zero and one. In the next chapter, we will show empirically that this is a Beta distribution. For this purpose, we verify in the following our theoretical considerations by practical tests with random data.

5. Empirical tests with random data

Typical climate time series contain at least 100 data points, which is preventing in general the explicit calculation of the entire distribution as discussed above. However, for n=20, this is still possible and carried out in the following to check our theoretical conclusions. *Fig. 2* shows the development of the external variance v as a function of the number of inserted breaks k.

To obtain statistical quantities, 100 repetitions have been performed. The mean amount of v increases linearly with k, as stated in Eq. (7). Additionally, the minimum and maximum are given for each number of breaks. In realistic cases, i.e., for larger n, the maximum can only be determined by dynamic programming; here the entire distribution could be explicitly calculated. In the following, it is our aim to find a mathematical function determining how the maximum external variance is growing with increasing number of breaks. A first approximation of this solution is already visible in *Fig. 2*. Three estimates are given for the maximum external variance. The central one, where an exponent of 4 is assumed, is in good agreement with the data. Obviously, the external variance v is connected to the break number k by the approximate function:

$$1 - v = \left(1 - \frac{k}{n-1}\right)^4 \quad . \tag{10}$$

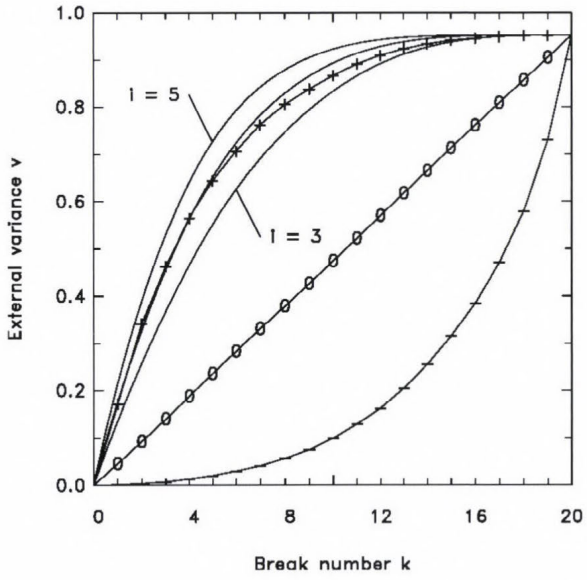

Fig. 2. Mean (0), maximum (+), and minimum (–) external variance as a function of inserted breaks for an $n = 21$ year random time series. For the maximum, three estimates are given: $1-v=(1-k^*)^i$, for $i=3, 4, 5$, where $k^*= k/(n-1)$ is the normalized break number. For 20 breaks, v reaches not 1, but 0.95, because a fraction of $1/(n-1)$ is covered by the error of the total mean, as given in Eq. (1).

For each break number, *Fig. 2* gives minimum and maximum of the external variance for 100 repetitions. Between these extremes we expect a kind of confined χ^2 distribution. As the shown result is based on numerical calculations, we are able to check our theory. *Fig. 3* shows exemplarily the distribution as obtained from a Monte Carlo experiment for 7 breaks. Differences to the corresponding χ^2 distribution are not large, but noticeable, especially at the tail of the distribution, where the maximum value, we are interested in, occurs. In contrast, the Beta distribution with 7 degrees of freedom is in good agreement with the data. Confirmed by tests with further break numbers, we assume in the following that the external variance is generally Beta distributed. The Beta distribution is formally given by:

$$p(v) = \frac{v^{\frac{k}{2}-1}\,(1-v)^{\frac{n-1-k}{2}-1}}{B\left(\frac{k}{2}\,,\,\frac{n-1-k}{2}\right)}\,, \tag{11}$$

with p denoting the probability density, v the external variance, and B the Beta function defined as:

$$B(a,b) = \frac{\Gamma(a)\,\Gamma(b)}{\Gamma(a+b)}\,, \tag{12}$$

with Γ denoting the Gamma function.

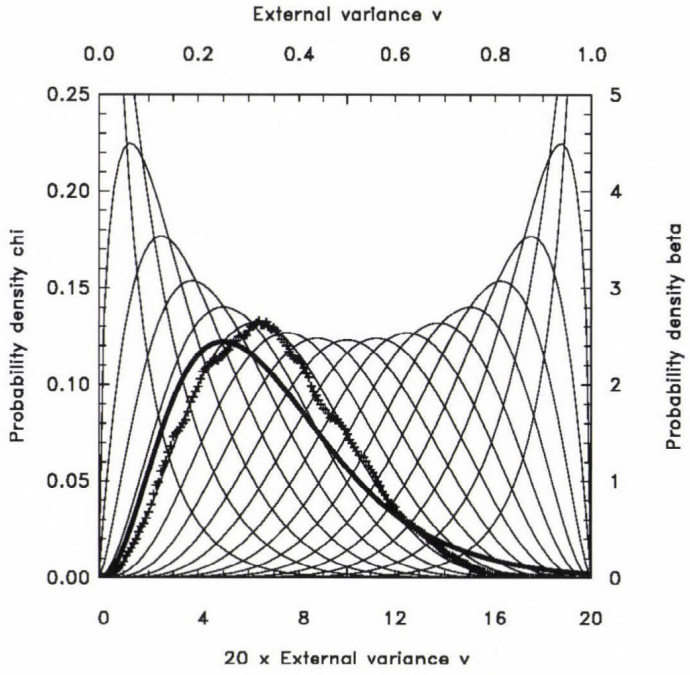

Fig. 3. Probability density for the χ^2 distribution, as given in Eq. (9) for k=7 (thick line). Furthermore, the 20 Beta distributions (thin), and the distribution of random data (crosses) are given. As expected, the data deviates slightly from the χ^2–7 and fits well to the Beta–7 curve. The lower abscissa and the left ordinate is valid for the χ^2 distribution. The upper v-abscissa and the right ordinate are valid for the Beta distribution and the normalized random data.

6. *The incomplete Beta function*

By Eq. (11) we are so far able to describe the distribution of the external variance v depending on length n and break number k. However, it is the maximum of v, which defines the optimum decomposition. Therefore, we need to find the maximum value of Eq. (11), or in other words, the exceeding probability of the Beta distribution, as given by:

$$P(v) = 1 - \int_0^v p dv \quad , \tag{13}$$

where the definite integral over a Beta distribution has to be solved, which is referred to as the incomplete Beta function $B(a,b,v)$. With this substitution Eq. (13) reads:

$$P(v) = 1 - \frac{B\left(\frac{k}{2}, \frac{n-1-k}{2}, v\right)}{B\left(\frac{k}{2}, \frac{n-1-k}{2}\right)} \quad . \tag{14}$$

For whole numbers the incomplete Beta function is obviously solvable by integration by parts, and the solution is:

$$\frac{B(i, m-i+1, v)}{B(i, m-i+1)} = \sum_{l=i}^{m} \binom{m}{l} v^l (1-v)^{m-l} \quad . \tag{15}$$

By comparing the arguments of the Beta function in Eq. (14) with those in Eq. (15), it follows:

$$i = \frac{k}{2} \quad , \tag{16}$$

and

$$\frac{n-1-k}{2} = m - i + 1 \quad . \tag{17}$$

Inserting Eq. (16) in Eq. (17) we have:

$$m = \frac{n-3}{2} \quad . \tag{18}$$

Since the variables i and m are defined as integers, Eq. (14) is solvable for even k and odd n. Replacing n and k in Eq. (14) by i and m, it follows:

$$P(v) = 1 - \frac{B(i,\ m-i+1,\ v)}{B(i,\ m-i+1)} \quad . \tag{19}$$

Using Eq. (15), the solution is:

$$P(v) = 1 - \sum_{l=i}^{m} \binom{m}{l} v^l (1-v)^{m-l} \quad . \tag{20}$$

Now we are aiming to replace the 1 in Eq. (20) by using the binomial definition, which is as follows:

$$\sum_{l=0}^{m} \binom{m}{l} a^l b^{m-l} = (a+b)^m \quad . \tag{21}$$

With a being v and b being 1- v it follows:

$$\sum_{l=0}^{m} \binom{m}{l} v^l (1-v)^{m-l} = \big(v + (1-v)\big)^m = 1 \quad , \tag{22}$$

so that it is actually possible to replace the 1 in Eq. (20) by a sum from zero to m:

$$P(v) = \sum_{l=0}^{m} \binom{m}{l} v^l (1-v)^{m-l} - \sum_{l=i}^{m} \binom{m}{l} v^l (1-v)^{m-l} \quad . \tag{23}$$

Calculating the sum from zero to m minus the sum from i to m, the sum from zero to i–1 is remaining:

$$P(v) = \sum_{l=0}^{i-1} \binom{m}{l} v^l (1-v)^{m-l} \quad . \tag{24}$$

Eq. (24) gives the exceeding probability as a function of external variance for any even break number k=2i. Let us again check the obtained equation

numerically by a Monte Carlo computation. For this purpose we create a random time series of the length n=21 and search for the combination of 4 breaks that produces the maximum external variance. Fig. 4 shows the result as obtained by 1000 repetitions. As each individual time series contains $\binom{n-1}{k}=\binom{20}{4}=4845$ possibilities of decomposition, we are dealing with a sample size of 4,845,000. Two conclusions can be drawn. First, the data is in good agreement with Eq. (24). Second, the effective number of combinations is much smaller than the nominal.

To the first conclusion: In *Fig. 4*, vertical lines from ln(0)=1 are drawn down to the exceeding probability that is found in the numerical test data. Thus, the edge of the shaded area gives the probability function for a certain maximum external variance. The according theoretical function as derived from Eq. (24) is given alternatively as a curve. The chosen numbers of n=21 and k=4 can be transformed by Eqs. (16) and (18) to m=9 and i=2. Inserted into Eq. (24) it follows for the depicted example:

$$P(v) = (1-v)^9 + 9v\,(1-v)^8 \quad . \tag{25}$$

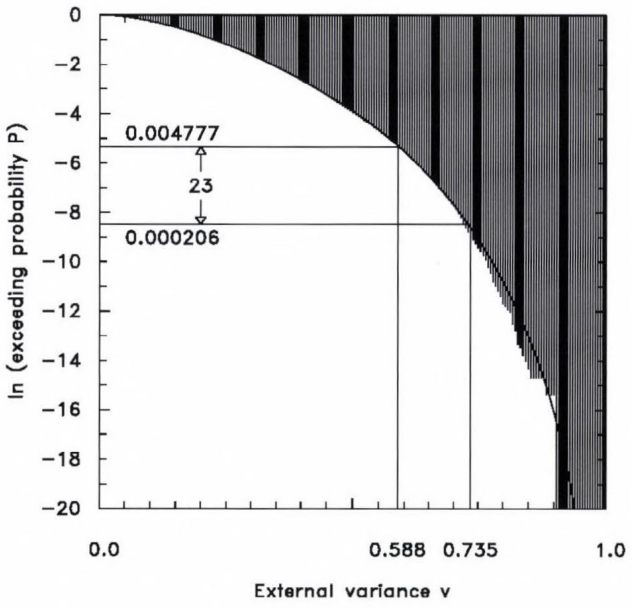

Fig. 4. Logarithmic exceeding probability as a function of external variance for 4 breaks within a 21-year time series. Vertical lines are drawn down from ln(0)=1 to the probability found for random data. The theoretical probability as generally given in Eq. (24) and specified in Eq. (25) is given by a curve. Two special data pairs are indicated, which are discussed in the text.

Fig. 4 shows that the data fits well to Eq. (25) if the probability is not too extreme. For such low probability it is not surprising that the limited Monte Carlo dataset shows more scatter and randomly deviates from the theory.

To the second conclusion: Two reading examples are given in *Fig. 4*. One starts from the exceeding probability of 2.064×10^{-4} (ln (0.0002) = –8.5). This value is equal to $\binom{20}{4}^{-1}$, the reciprocal of the nominal number of combinations for $n=21$ and $k=4$. If all combinations were independent, we could expect a maximum external variance of 0.7350. However, this is not the actually true value, which is already determined as 0.5876 (*Fig. 2*). But we can draw the reverse conclusion: What must be the effective number of combinations for the known external variance? We obtain a value of 4.777×10^{-3}, which is 23 times larger than the starting point. The conclusion is that the effective number of combinations for this special case ($n=21$, $k = 4$) is 23 times smaller than the nominal one, which is equal to $\binom{20}{4}$. The dependency of different solutions is reasonable. Shifting only one break position by one time step creates already a new break combination. However, its external variance will not deviate much from the original.

7. The relative change of variance as a function of increased break number

After confirming Eq. (24) by test data, we can assume its general validity and turn towards more realistic lengths. *Fig. 5* shows the graphs of Eq. (24) for $n=101$ and all even k from 2 to 20. As in *Fig. 4,* the number of independent combinations is estimated by a reversal conclusion from the known results of the maximum external variance. (In this case the results stem from a dynamic programming search as the length of $n = 101$ is too large for an explicit all-permutations-search of the maximum as it was possible for $n = 21$.)

The following question arises: What is the rate of change of the variance, if the number of breaks is increased? Obviously, there are two contributions. First, we skip from the graph in *Fig. 5* valid for k breaks to the next one valid for $k+2$. This causes a certain increase in the external variance, even if the number of combinations would remain constant. Second, there *is* certainly an increased number of permutations, although we showed that the effective number is always smaller than the nominal one.

Fig. 6 gives a sketch of the situation to illustrate how the mathematical formulations for the two components are derived in detail. The exceeding probability P for two arbitrary even break numbers is depicted. To determine the first contribution, we need to know the distance between two neighboring curves in v-direction for a fixed P ($v1 - v0$ in *Fig. 6*).

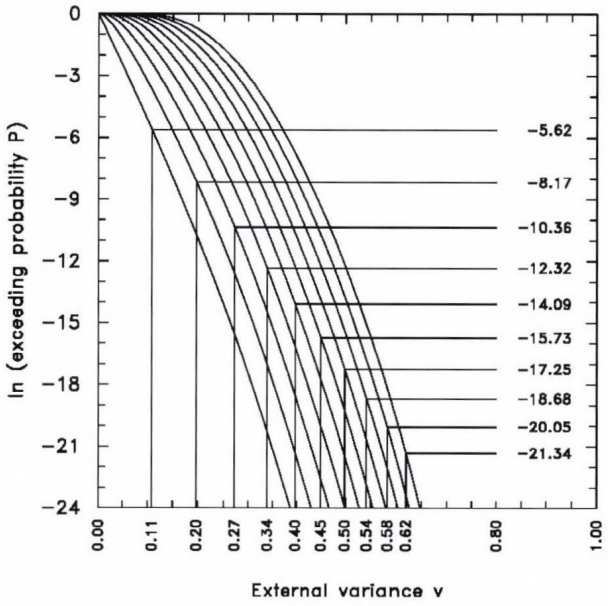

Fig. 5. As *Fig. 4*, but for a 101-year time series and for the ten different break numbers from 2, 4, 6, … , 20. The known external variances for each break number are retranslated into the observed effective exceeding probabilities given as the column at the right edge.

As Eq. (24) is difficult to solve for v, we estimate the v -distance by the P-distance, which is divided by the slope s:

$$\left(\frac{dv}{di}\right)_1 = v1 - v0 = \frac{ln(P1) - ln(P2)}{s} \quad . \tag{26}$$

Using the respective i-indices for $P1$ and $P0$ (see *Fig. 6*) we can rewrite:

$$\left(\frac{dv}{di}\right)_1 = \frac{\left(ln\left(P_i(v)\right) - ln\left(P_{i+1}(v)\right)\right)_{v=const}}{s} \quad . \tag{27}$$

This first part of dv/di arises because different functions of $P(v)$ has to be used. We introduce C_f and refer to it the following as the function contribution:

$$C_f = \left(ln\left(P_{i+1}(v)\right) - ln\left(P_i(v)\right)\right)_{v=const} \quad , \tag{28}$$

so that Eq. (27) can be rewritten:

$$\left(\frac{dv}{di}\right)_1 = -\frac{C_f}{s} \quad . \tag{29}$$

The second contribution is the increase of v due to the total decrease of P ($v2 - v1$ in *Fig. 6*). Geometrically, this can be perceived as a walk down the respective curve.

$$\left(\frac{dv}{di}\right)_2 = v2 - v1 = \frac{ln(P2) - ln(P1)}{s} \quad . \tag{30}$$

Using i-indices for $P2$ and $P1$, it follows:

$$\left(\frac{dv}{di}\right)_2 = \frac{ln\big(P_{i+1}(v)\big) - ln\big(P_i(v)\big)}{s} \quad . \tag{31}$$

This second part of dv/di depends on the increased number of decomposing permutations with growing i. Consequently, we refer to the numerator as number contribution C_n, according to:

$$C_n = ln\big(P_{i+1}(v)\big) - ln\big(P_i(v)\big) \quad , \tag{32}$$

and it follows:

$$\left(\frac{dv}{di}\right)_2 = \frac{C_n}{s} \quad . \tag{33}$$

In both cases, changes in P are translated into v by the slope of the curves. This is appropriate if the curvatures are small and the slopes remain nearly constant. For the relevant parts of the curves this is a good approximation (*Fig. 5*).
Finally, we can summarize Eq. (29) and Eq. (33) to:

$$\frac{dv}{di} = \left(\frac{dv}{di}\right)_2 + \left(\frac{dv}{di}\right)_1 = \frac{C_n - C_f}{s} \quad . \tag{34}$$

To determine dv/di, we obviously need three terms, the slope s, the function contribution C_f, and the number contribution C_n. These three terms are derived in the following subsections.

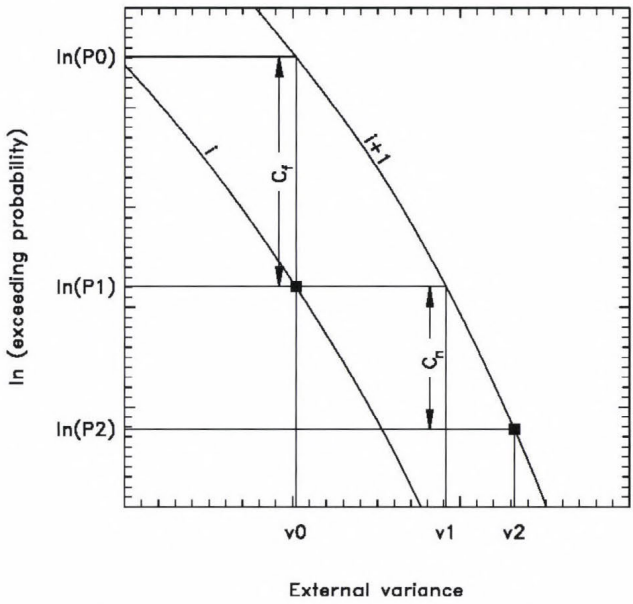

Fig. 6. Sketch to illustrate the total gain of external variance from v0 to v2, when the number of breaks k is increased by 2, i.e., from i to i+1. The first contribution ($v1 - v0$) depends on the horizontal distance of the two curves. This contribution is derived in the text by the vertical distance C_f and the slope of the curve. The second contribution ($v2 - v1$) occurs due to the increase of possible combinations when the break number is increased. As for the first contribution, it is translated from C_n by using the slope of the depicted curves.

7.1. The slope

The slope s of the logarithm of Eq. (24) as it is depicted in *Figs. 5* and *6* is equal to:

$$s = \frac{d}{dv}\Big(ln\big(P(v)\big)\Big) = \frac{1}{P(v)}\frac{dP(v)}{dv} \quad . \tag{35}$$

With Eq. (13) it follows:

$$s = -\frac{p(v)}{P(v)} \quad . \tag{36}$$

Replacing n and k by m and i and using the result of Appendix A we can rewrite Eq. (11) to:

$$p(v) = v^{i-1}\,(1-v)^{m-i}\,(m-i+1)\,\binom{m}{i-1} \quad . \tag{37}$$

Inserting Eq. (24) and Eq. (37) into Eq. (36), it follows:

$$s = -\frac{v^{i-1}\,(1-v)^{m-i}\,(m-i+1)\,\binom{m}{i-1}}{\sum_{l=0}^{i-1}\left(\binom{m}{l} v^l (1-v)^{m-l}\right)} \quad . \tag{38}$$

In Appendix B we show that the last summand is a good approximation for the sum occurring in the denominator and it follows:

$$s = -\frac{v^{i-1}\,(1-v)^{m-i}\,(m-i+1)\,\binom{m}{i-1}}{\binom{m}{i-1}\,v^{i-1}\,(1-v)^{m-i+1}} \quad , \tag{39}$$

which can be reduced to:

$$s = -\frac{m-i+1}{1-v} \quad . \tag{40}$$

After replacing again m and i by n and k it follows:

$$s = -\frac{n-1-k}{2\,(1-v)} \quad . \tag{41}$$

7.2. The function contribution

With Eq. (24) the vertical distance between two neighboring curves as given in *Fig. 6* is:

$$C_f = ln(P_{i+1}) - ln(P_i) = ln\left(\frac{\sum_{l=0}^{i}\binom{m}{l} v^l (1-v)^{m-l}}{\sum_{l=0}^{i-1}\binom{m}{l} v^l (1-v)^{m-l}}\right). \tag{42}$$

We use again Appendix B and approximate the sums by their last summand:

$$C_f = ln\left(\frac{\binom{m}{i}\,v^i\,(1-v)^{m-i}}{\binom{m}{i-1}\,v^{i-1}\,(1-v)^{m-i+1}}\right) \quad , \tag{43}$$

which can be reduced to:

$$C_f = ln\left(\frac{\binom{m}{i} v}{\binom{m}{i-1} (1-v)}\right) \quad . \tag{44}$$

The ratio of consecutive binomial coefficients is equal to $(m\text{-}i+1)/i$:

$$C_f = ln\left(\frac{(m-i+1)\, v}{i\,(1-v)}\right) \quad . \tag{45}$$

Replacing m and i again by n and k, it follows:

$$C_f = ln\left(\frac{(n-1-k)\, v}{k\,(1-v)}\right) \quad . \tag{46}$$

7.3. The number contribution

The nominal number of combinations grows with growing k from $\binom{n-1}{k}$ to $\binom{n-1}{k+1}$. This corresponds to a factor of $(n\text{–}1\text{–}k)/k$. However, in *Fig. 4* we show exemplarily for $k = 4$ that the effective number of combinations is lower. In *Fig. 5* the decrease of $\ln(P(v))$ due to the increase of the effective number of combinations is given in a column at right edge for the even k from 2 to 20. From these numbers we derived the actual decreasing factor $C_n = \Delta\ln(P(v))$ and compared it with the nominal (*Table 1*). The nominal decreasing factor for $\Delta k = 1$ is equal to the reciprocal of the growth of combinations $\frac{\binom{n-1}{k}}{\binom{n-1}{k+1}} = \frac{k}{n-1-k}$. Here we need its logarithm; and as the effective decreasing factor is only available for every second k, $nom = -2\ln((n\text{–}1\text{–}k)/k)$ is the proper reference.

From *Table 1* we can extract that the ratio between the effective and nominal factor is rather constant with $r \approx 0.4$, but slightly growing with increasing break number. The growth will be discussed in detail in Appendix C, for the time being we can summarize:

$$C_n = r\ nom = -2r\ ln\left(\frac{n-1-k}{k}\right) \quad . \tag{47}$$

Table 1. From *Fig. 5*, C_n, the effective decrease of $\ln(P(v))$ for the transition from k to $k+2$ is taken. It is compared to the nominal decrease equal to $-2\ \ln((n-1-k)/k)$. Finally, the ratio r between the effective and nominal factor is given

k_1	k_2	k	eff = Δ ln (P(v))	nom = –2 ln((n–1–k)/k)	r = eff/nom
2	4	3	–2.552	–6.952	0.367
4	6	5	–2.186	–5.889	0.371
6	8	7	–1.963	–5.173	0.379
8	10	9	–1.765	–4.627	0.381
10	12	11	–1.645	–4.181	0.393
12	14	13	–1.514	–3.802	0.398
14	16	15	–1.435	–3.469	0.414
16	18	17	–1.363	–3.171	0.430
18	20	19	–1.292	–2.900	0.446

7.4. The differential equation and its solution

The rate of change of v with regard to k is only half of that with regard to i (compare Eq. (16)):

$$\frac{dv}{dk} = \frac{dv}{di}\frac{di}{dk} = \frac{1}{2}\frac{dv}{di} \quad . \tag{48}$$

Using Eq. (34) it follows:

$$\frac{dv}{dk} = \frac{1}{2}\frac{C_n - C_f}{s} \quad . \tag{49}$$

Inserting our findings for the slope s (Eq. (41)) and for the two contributions C_f and C_n (Eqs. (46) and (47)), the growth of v with growing k is given by:

$$\frac{dv}{dk} = \frac{1-v}{n-1-k}\left(2r\ ln\left(\frac{n-1-k}{k}\right) + ln\left(\frac{(n-1-k)\,v}{k\,(1-v)}\right)\right) . \tag{50}$$

Reducing the fractions under the logarithms by n–1 leads to the normalized break number k^*, defined as:

$$k^* = \frac{k}{n-1} \quad . \tag{51}$$

At the same time, the differential dk has to be replaced by:

$$dk = (n-1)\, dk^* \quad , \tag{52}$$

so that Eq. (50) may be rewritten in normalized form:

$$\frac{dv}{dk^*} = \frac{1-v}{1-k^*}\left(2r\, ln\left(\frac{1-k^*}{k^*}\right) + ln\left(\frac{(1-k^*)\, v}{k^*\,(1-v)}\right)\right) \quad . \tag{53}$$

The final main question is now: What is the solution of Eq. (53)? Let us make a first approach to the solution by a very rough estimate for small k^*.

$$\frac{1-k^*}{1-v}\,\frac{dv}{dk^*} = 2r\, ln\left(\frac{1-k^*}{k^*}\right) + ln\left(\frac{(1-k^*)\, v}{k^*\,(1-v)}\right) = \alpha = -C_n + C_f \quad . \tag{54}$$

The first logarithm constituting α, i.e., $-C_n$, is for small k^* in the order of $\ln(n)$ and it decreases with increasing k^*. The second, C_f, is in the order of $\ln(v/k^*)$. Because we know already the approximate solution being $1-v \approx (1-k^*)^4$, we can estimate the second term to about $\ln(4)$ (compare Eq. (78) in Appendix B). In contrast to the first term, this term increases with increasing k^* (see Appendix C), because $1-v$ is decreasing faster than $1-k^*$. Assuming $n = 101$, an estimate for α is:

$$\alpha \approx 2r\, ln(n) + ln(4) \approx 2 \cdot 0.4\, \ln(100) + ln(4) = 5.07 \quad . \tag{55}$$

If α were actually constant, the integration of Eq. (54) would be easy:

$$\frac{1}{1-v}\, dv = \frac{\alpha}{1-k^*}\, dk^* \quad , \tag{56}$$

$$-\, ln(1-v) = -\,\alpha\, ln(1-k^*) \quad , \tag{57}$$

$$1-v = (1-k^*)^{\alpha} \quad , \tag{58}$$

which is rather similar to the already known approximate solution (Eq. (10)), except that the exponent found in Eq. (55) is higher. This already shows that the assumptions made to estimate s, C_f, and C_n were reasonable.

For a more accurate solution let us go back to the performance of the random data that we already used above to verify our theory. By these data we can check how well the rough estimate of a constant α is fulfilled in reality. *Fig. 7* shows that such an estimate is actually not too bad, which is the reason

for Eq. (58) being rather close to the true solution. For a more precise solution, we fit a function to $\alpha(k^*)$ and obtain:

$$\frac{1-k^*}{1-v}\,\frac{dv}{dk^*} \;=\; \frac{1-k^*}{2}\,ln\left(\frac{1-k^*}{k^*}\right) + 2\,ln(5) \quad . \tag{59}$$

Eq. (59) may be rewritten as:

$$\frac{1}{1-v}\,dv \;=\; \left(\frac{1}{2}\,ln\left(\frac{1-k^*}{k^*}\right) + \frac{2\,ln(5)}{1-k^*}\right)\,dk^* \quad , \tag{60}$$

which is easy to integrate. Its solution is:

$$1-v \;=\; (1-k^*)^a\left(\frac{1-k^*}{k^*}\right)^{bk^*} \quad , \tag{61}$$

with $a = 2\ \ln(5) + 1/2$ and $b = -1/2$.

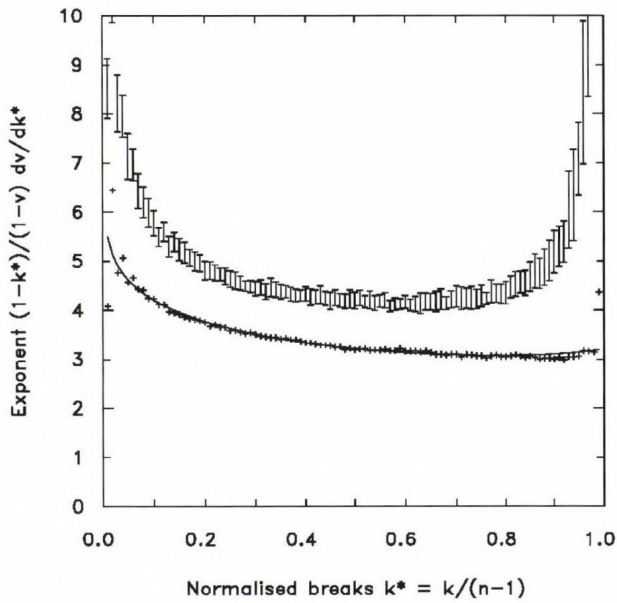

Fig. 7. Exponent α as given in Eq. (54) as a function of the normalized break number k^* for random data (crosses). These data consists of 1000 random 101-year time series. The vertical bars connect the 90 and 95 percentiles. The thin line is giving the function according to Eq. (59).

In *Fig. 7*, the function for the exponent α as given in Eq. (59) fits well to the results derived from random data. However, the relative gain of external variance is larger for even values compared to their uneven neighbors, especially for low values. This feature is reasonable, as it needs always a pair of breaks to isolate a subsegment. To produce the data, we performed 1000 repetitions, mainly to reduce the scatter. However, the repetitions can also be exploited to derive the variability of the solution. Consequently, not only the mean, but also the 90 and 95 percentiles are given. The average exponent starts for low normalized break numbers at about 5. This means that the external variance grows at the beginning 5 times faster than the normalized break number. This behavior is found for random data. When such a variance growth will occur in real data, we can be rather sure that no true break is present as it is normal for random data which has by definition no real break. The 95 percentile is for the first breaks as large as nearly 10. Thus, in only 5% of the cases, the external variance grows by a factor of more than 10 times faster than k^*. Hence, this value can be used as limit to distinguish true from spurious breaks. For the first break numbers it reaches nearly 10, decreasing rapidly to about 5 for $k^* = 0.1$.

8. Discussion of the penalty term

Within the homogenization algorithm PRODIGE (*Caussinus* and *Mestre*, 2004), the following expression is minimized to estimate the number of predicted breaks.

$$C_k(Y) = \ln\left(1 - \frac{\sum_{j=1}^{k+1} n_j\left(\overline{Y}_j - \overline{Y}\right)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}\right) + \frac{2\,(k+l)}{n-1}\,ln(n) = \min \quad . \quad (62)$$

The numeric value of $C_k(Y)$ depends on the data Y and the number of breaks k and consists of two opposite contributions. Firstly, the logarithm of the normalized internal variance, and secondly, a penalty term, originally proposed by *Caussinus* and *Lyazrhi* (1997). Whereas the first is decreasing with larger k, the second is increasing. Using our notations for the same terms, Eq. (62) can be rewritten as:

$$\ln(1-v) + \frac{2k}{n-1}\,ln(n) = \min \quad . \quad (63)$$

In Eq. (63), we combined k and l, the number of breaks and the number of outliers to a single number. Splitting off an outlier is identical to the separation of a subperiod of length 1. Consequently, it is not necessary to treat outliers

separately. If we further use the normalized break number k^* according to Eq. (51) instead of k, we can rewrite:

$$\ln(1 - v) + 2k^* ln(n) = \min \quad . \tag{64}$$

To find the break number k^* for which the expression is minimal, the first derivative with respect to k^* is set to zero:

$$-\frac{1}{1 - v}\frac{dv}{dk^*} + 2\, ln(n) = 0 \quad , \tag{65}$$

which can be rewritten to:

$$\frac{1}{1 - v}\frac{dv}{dk^*} = 2\, ln(n) \quad . \tag{66}$$

For a given time series, the length n is constant. Consequently, we can conclude from Eq. (66), that PRODIGE uses a fixed number, equal to 2 ln(n), as stop criterion. If the relative gain of the external variance falls below that constant, no further breaks are added and the final break number is reached. However, from Eq. (59) we know the function for the relative gain of external variance in detail; it just has to be divided by $1-k^*$.

$$\frac{1}{1 - v}\frac{dv}{dk^*} = \frac{1}{2} ln\left(\frac{1 - k^*}{k^*}\right) + \frac{2\, ln(5)}{1 - k^*} \quad . \tag{67}$$

Fig. 8 shows this function for a time series of length 101. Additionally, six exceeding values for probabilities from 1/4 to 1/128 are given, based on 5000 repetitions. These curves are approximately equidistant. For comparison, the constant as proposed by *Caussinus* and *Mestre* (2004) and rewritten in Eq. (66) is given, which is about 9 (exactly 2 ln(101)) for $n = 101$.

As the exceeding values are computed for random data, they can be interpreted as error probability. The 1% error line (exactly 1/128) at the upper end of the family of curves in *Fig. 8* starts at a variance gain of about 15, and reaches, for $k^* = 0.1$, a value of about 8.

In the climatologically interesting range of small k^*, the numeric value of the mean variance gain (lowest line in *Fig. 8*) is equal to about 5 and can be interpreted as following. For random data, which contains no break by definition, the relative external variance grows on average with each additionally inserted break 5 times faster than expected by a simple linear approach. Such a linear approach just supposes that each break adds the same amount of external variance. For $n = 101$ this would be one percent per break. In reality, the data contains larger jumps just by chance, comprising not only 1%,

but 5% of the remaining variance. In seldom cases, these highest jumps contain even 15% of the remaining variance, but the probability for that is only about 1% (uppermost line in *Fig. 8*). As the number of tentatively inserted breaks is growing, the highest jumps are already used before, so that the amount of the remaining decreases. Increasing the break number from 9 to 10 ($k^* = 0.1$) gains in average still 5%, as for the lowest break numbers, but the maximum value, exceeded in 1% of the cases, drops from 15% to 8%.

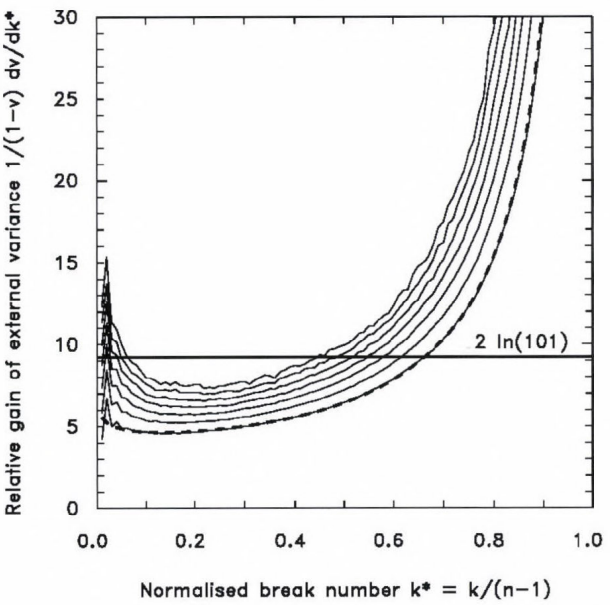

Fig. 8. Relative gain of external variance as a function of normalized breaks k^* for a time series length of $n = 101$. The dashed fat curve denotes the theoretical value as given by Eq. (67). The solid thin curves are showing the data results as obtained by 5000 repetitions. The lowest indicates the mean, which is largely congruent with the theory. The upper ones give the exceeding value for probabilities from 2^{-2}, 2^{-3}, ... , 2^{-7}. For comparison, the constant 2 ln(n) proposed as stop criterion by *Caussinus* and *Lyazrhi* (1997) is given by the horizontal line.

The Lyazrhi constant of 2 ln(n) as proposed by *Caussinus* and *Mestre* (2004) is equal to about 9 for $n = 101$. At the beginning, i.e. for one break, this value lies in the middle of the family of error curves in *Fig. 8*. Thus, it corresponds here to an error of about 5%. At $k^* = 0.08$, i.e. for 8 breaks, the horizontal line is leaving the area covered by error curves. Thus, the error level decreases below 1%. Assuming continued equidistance, the horizontal line will reach areas with errors of less than 0.1% at $k^* = 0.15$. Thus, for low break numbers, PRODIGE accepts breaks, even if the error is relatively high (about

5%). In contrast, higher break numbers are effectively suppressed. Only breaks are accepted that add an amount of variance, which would occur randomly with a probability of less than 1%.

The choice of the Lyazrhi constant appears to be rather artful. For the first breaks, it allows errors of about 5%, which is a widely accepted error margin. However, for more than 8 breaks (within a time series of 101 data points), the method is much more rigid. Obviously, the preexisting knowledge is used that such high numbers of breaks are per se unlikely, so that a suppression is reasonable.

In *Fig. 9*, the corresponding features for shorter time series (n = 21) are given. Compared to n = 101, the average variance gain remains unchanged, showing that Eq. (67) is universally valid. However, the exceeding values increase, and the distances between the error curves grow by a factor of 5. This indicates that the growing factor is inversely proportional to the time series length n. In contrast, the Lyazrhi constant even decrease, although only slightly due to its logarithmic form. The direction of change of the Lyazrhi constant for different time series length is contradicting our findings for random data and should be studied further. However, instrumental climate records comprise often about 100 data points, and for such lengths the constant is chosen rather well.

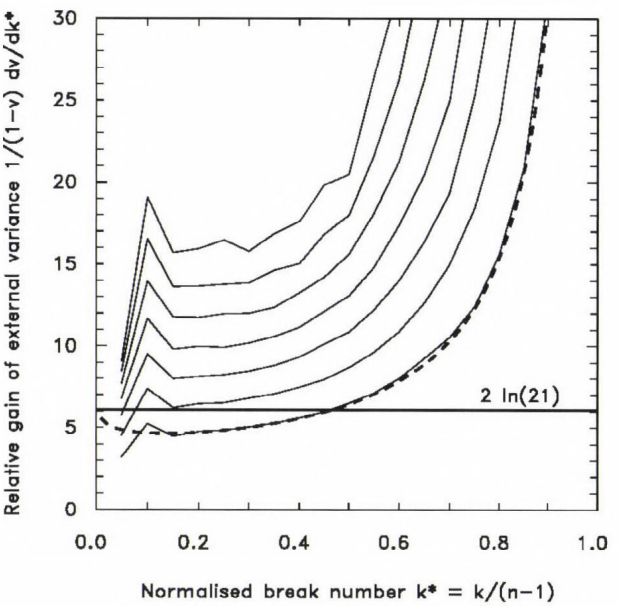

Fig. 9. As *Fig. 8*, but for n = 21. The average variance gain remains unchanged compared to *Fig. 8*, because Eq. (67) is universally valid. However, the exceeding values increase inversely proportional to n. In contrast, the constant of *Caussinus* and *Lyzrhi* (1997) decreases with decreasing n.

9. Conclusions

The external variance, defined as the variance of the subperiods' means, is shown to be the key parameter to detect breaks in climate records. Maximum external variance indicates the most probable combination of break positions. We analyzed the characteristics of the external variance occurring in random data and derived a mathematical formulation (Eq. (61)) for the growth of its maximum with increasing number of assumed breaks. As random data includes by definition no break, this knowledge can be used as null hypothesis to separate true breaks in real climate records more accurately from noise. In this way, it helps to enhance the valuable information from historical data.

Acknowledgement—We are grateful for advice from *Peter Domonkos* and *Tamas Szentimrey*. This work has been performed within the project Daily Stew supported by the Deutsche Forschungsgemeinschaft DFG (VE 366/5).

Appendix A

Consider the Beta function in Eq. (11):

$$B\left(\frac{k}{2}, \frac{n-1-k}{2}\right) = B(i, m-i+1)$$
$$= \frac{\Gamma(i)\,\Gamma(m-i+1)}{\Gamma(i+m-i+1)} = \frac{(i-1)!\;(m-i)!}{m!} \quad . \tag{68}$$

Multiplication of both the numerator and denominator with $m–i+1$ leads to:

$$B\left(\frac{k}{2}, \frac{n-1-k}{2}\right) = \frac{(i-1)!\;(m-i+1)!}{(m-i+1)\,m!} \quad . \tag{69}$$

Remembering the definition of binomial coefficients being $\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$, we can write:

$$B\left(\frac{k}{2}, \frac{n-1-k}{2}\right) = \left((m-i+1)\binom{m}{i-1}\right)^{-1} \quad . \tag{70}$$

Appendix B

Consider the individual summands of the sum as defined in Eq. (24). The factor of change f between a certain summand and its successor is:

$$f = \frac{\binom{m}{l_i}\, v}{\binom{m}{l_i - 1}(1 - v)} \quad , \tag{71}$$

where l_i runs from zero to i. The ratio of consecutive binomial coefficients can be replaced, and it follows:

$$f = \frac{(m - l_i + 1)\, v}{l_i(1 - v)} \quad . \tag{72}$$

m and i can be replaced by n and k:

$$f = \frac{(n - 1 - l_k)\, v}{l_k(1 - v)} \quad . \tag{73}$$

Inserting k instead of l_k is a lower limit for f because $(n–1–l_k)/l_k$, the rate of change of the binomial coefficients, is decreasing monotonously with k:

$$f > \frac{(n - 1 - k)\, v}{k(1 - v)} \quad . \tag{74}$$

Normalize k by $1/(n–1)$:

$$f > \frac{(1 - k^*)\, v}{k^*(1 - v)} \quad . \tag{75}$$

The approximate solution is known with $1–v = (1–k^*)^4$, see Eq. (10).

$$f > \frac{(1 - k^*)\ (1 - (1 - k^*)^4)}{k^*(1 - k^*)^4} \quad , \tag{76}$$

$$f > \frac{1 - (1 - k^*)^4}{k^*(1 - k^*)^3} \quad , \tag{77}$$

for $k \to 0$:

$$f > \frac{1 - (1 - 4k^*)}{k^*(1 - 3k^*)} = \frac{4k^*}{k^*(1 - 3k^*)} = \frac{4}{1 - 3k^*} = 4 \quad , \tag{78}$$

for $k \to 1$:

$$f > \frac{(1-k^*)^{-3} - (1-k^*)^4}{k^*} = \frac{\infty - 0}{1} = \infty \quad . \qquad (79)$$

We can conclude that each element of the sum given in Eq. (24) is by a factor f larger than the prior element. For small k^* the factor f is greater than about 4 and grows to infinity for large k^*. Consequently, we can approximate the sum by its last summand according to:

$$P(v) = \sum_{l=0}^{i-1} \binom{m}{l} v^l (1-v)^{m-l} \approx \binom{m}{i-1} v^{i-1} (1-v)^{m-i+1} \quad . \qquad (80)$$

Appendix C

Once the solution for $v(k^*)$ is available (Eq. (61)), a more accurate estimation of the function contribution C_f is possible. So far, we approximated the sum given in Eq. (24) by its last summand, as discussed in Appendix B. Now we are able to check the impact of this approximation. Using the known solution, we calculated two versions of C_f. First, by taking into account only the last summand as in Eq. (43) and alternatively the complete term, as given in Eq. (42). *Fig. 10* shows these two estimates of C_f as dashed lines. The upper one denotes the full solution, the lower the approximation. Their difference remains limited, which confirms our findings in Appendix B. As discussed in Eq. (55), C_f starts for low k^* at about ln(4) and rises to infinity for high k^*.

Concerning the number contribution C_n, we applied so far only a rough estimate as given in Eq. (47), assuming a constant ratio between effective and nominal combination growths. Actual values for C_n are listed in *Table 1* for low break numbers. However, they are numerically computable up to about $k^* = 0.75$. In *Fig. 10*, these values for C_n are given as crosses. They are multiplied by –1, as $-C_n$ contributes to the exponent α. We fitted a function of the form:

$$(ak^* + b)\, ln\left(\frac{1-k^*}{k^*}\right) + c \quad , \qquad (81)$$

to the data, which is depicted by the lower full curve in *Fig. 10*, and obtained for the coefficients:

$$a_1 = 0.5, \quad b_1 = 0.55, \quad c_1 = 0.4 \quad .$$

A similar fit to C_f is given by the upper full curve in *Fig. 10*. Here the coefficients are:

$$a_2 = -1.0, \quad b_2 = -0.15, \quad c_2 = 2.7 \quad .$$

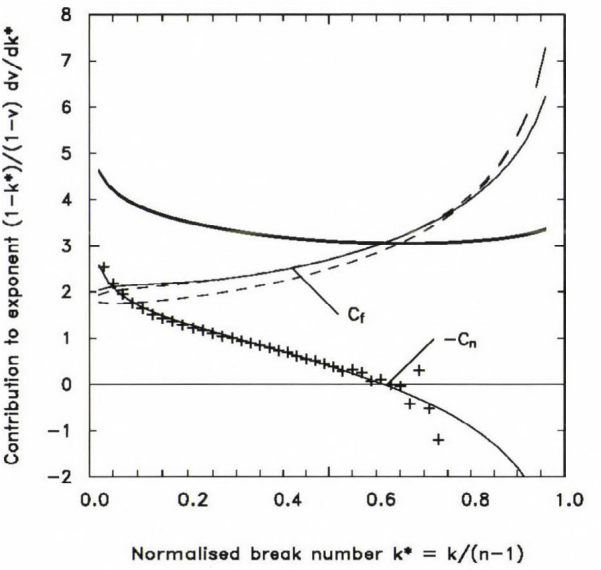

Fig. 10. Contributions of C_f and $-C_n$ to the exponent $= \frac{1-k^*}{1-v}\frac{dv}{dk^*}$. The two dashed lines are reconstructions of C_f from the known solution of $v(k^*)$, as given in Eq. (61). The solid line gives a fitted function for C_f. Crosses denote data for C_n connected likewise by a fitted curve. The sum of the two contributions is given by the fat line.

The sum of two curves yields then an alternative estimation for the exponent α. It is depicted as a fat line in *Fig. 10* and characterized by the sum of the coefficients:

$$a_3 = -0.5, \quad b_3 = 0.4, \quad c_3 = 3.1 \quad .$$

This alternative estimate is in good agreement (please compare *Fig. 7* lowest line with *Fig. 10* uppermost fat line) with the solution derived directly from the data as given in Eq. (59), where the coefficients are:

$$a_4 = -0.5, \quad b_4 = 0.5, \quad c_4 = 2\ln(5) = 3.2 \quad .$$

We see that Eq. (59), so far directly based on a fit to the data, is as well understandable from the theory as the sum of the two contributions C_f and $-C_n$.

References

Aguilar, E., Auer, I., Brunet, M., Peterson, T.C., and *Wieringa, J.,* 2003: Guidelines on climate metadata and homogenization. World Meteorological Organization, *WMO-TD No. 1186*, WCDMP No. 53, Geneva, Switzerland, 55 pp.

Akaike, H., 1973: Information theory and an extension of the maximum likelihood principle, In *2nd International Symposium on Information Theory*, Akademiai Kiado, Budapest, 267–281.

Alexandersson, H. and *Moberg, A.,* 1997: Homogenization of Swedish temperature data. 1. Homogeneity test for linear trends. *Int. J. Climatol. 17*, 25–34.

Auer, I., Böhm, R., Jurkovic, A., Orlik, A., Potzmann, R., Schöner, W., Ungersböck, M., Brunetti, M., Nanni, T., Maugeri, M., Briffa, K., Jones, P., Efthymiadis, D., Mestre, O., Moisselin, J.M., Begert, M., Brazdil, R., Bochnicek, O., Cegnar, T., Gajic-Capkaj, M., Zaninovic, K., Majstorovic, Z., Szalai, S., Szentimrey, T., and *Mercalli, L.*,2005: A new instrumental precipitation dataset for the Greater Alpine Region for the period 1800–2002, *Int. J. Climatol. 25*, 139–166.

Auer, I., Böhm, R., Jurkovic, A., Lipa, W., Orlik, A., Potzmann, R., Schöner, W., Ungersböck, M., Matulla, C., Briffa, K., Jones, P., Efthymiadis, D., Brunetti, M., Nanni, T., Maugeri, M., Mercalli, L., Mestre, O., Moisselin, J.M., Begert, M., Müller-Westermeier, G., Kveton, V., Bochnicek, O., Stastny, P., Lapin, M., Szalai, S., Szentimrey, T., Cegnar, T., Dolinar, M., Gajic-Capka, M., Zaninovic, K., Majstorovic, Z., and *Nieplova, E.,* 2007: HISTALP – historical instrumental climatological surface time series of the Greater Alpine Region, *Int. J. Climatol.* 27, 17–46.

Begert, M., Schlegel, T. and *Kirchhofer, W.,* 2005: Homogeneous temperature and precipitation series of Switzerland from 1864 to 2000. *Int. J. Climatol. 25*, 65–80.

Bellman, R., 1954: The Theory of Dynamic Programming, *Bull. Am. Math. Soc. 60*, 503–516. doi: 10.1090/S0002-9904-1954-09848-8, MR 0067457.

Bergström, H. and *Moberg, A.,* 2002: Daily air temperature and pressure series for Uppsala (1722–1998). *Climatic Change, 53*, 213–252.

Brunet, M., Asin, J., Sigro, J., Banon, M., Garcıa, F., Aguilar, E., Esteban Palenzuela, J., Peterson, T.C., and *Jones, P.,* 2011: The minimization of the screen bias from ancient Western Mediterranean air temperature records: an exploratory statistical analysis. *Int. J. Climatol. 31*, 1879–1895.

Brunetti, M., Maugeri, M., Monti, F., and *Nannia, T.*,2006: Temperature and precipitation variability in Italy in the last two centuries from homogenised instrumental time series. *Int. J. Climatol. 26*, 345–381.

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

Caussinus, H. and *Mestre, O.,* 1996: New mathematical tools and methodologies for relative homogeneity testing. *Proc. First Seminar for Homogenization of Surface Climatological Data*, Budapest, Hungary, Hungarian Meteorology Service, 63–72.

Caussinus, H. and *Mestre, O.,* 2004: Detection and correction of artificial shifts in climate series. *Appl. Statist. 53*, part 3, 405–425.

Conrad, V. and *Pollak. C.,* 1950: *Methods in climatology*, Harvard University Press, Cambridge, MA, 459 pp.

Davis R.A.,. Lee, T.C.M., and *Rodriguez-Yam, G.A.,* 2012: Structural break estimation for nonstationary time series models. *J. Am.. Stat. Assoc. 101*, 223–239.

Domonkos, P., 2011a: Efficiency evaluation for detecting inhomogeneities by objective homogenization methods. *Theor. Appl. Climatol. 105*, 455–467.

Domonkos, P., 2011b: Adapted Caussinus-Mestre Algorithm for Networks of Temperature Series (ACMANT). *Int. J. Geosci. 2*, 293–309.

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

Hawkins, D.M., 1972: On the choice of segments in piecewise approximation. *J. Inst. Maths. Applics., 9*, 250–256.

Knowles Middleton, W.E., 1966: A history of the thermometer and its use in meteorology. The John Hopkin Press, Baltimore, Maryland. 249 pp.

Lavielle, M., 1998: Optimal segmentation of random processes. *IEEE Trans. Signal Processing, 46*, 1365–1373.

Li, S. and *Lund, R.*, 2012: Multiple Changepoint Detection via Genetic Algorithms. *J. Climate 25*, 674–686.

Lindau, R., 2003: Errors of Atlantic Air-Sea Fluxes Derived from Ship Observations., *J. Climate, 16*, 783–788.

Lindau, R., 2006: The elimination of spurious trends in marine wind data using pressure observations. *Int. J. Climatol.* 26, 797–817.

Menne M.J. and *Williams Jr., C.N.*, 2009: Homogenization of temperature series via pairwise comparisons. *J. Climate, 22*, 1700–1717.

Mestre O., Domonkos, P., Picard, F., Auer, I., Robin, S., Lebarbier, E., Böhm, R., Aguilar, E., Guijarro, J., Vertachnik, G., Klancar, M., Dubuisson, B., and *Stepanek, P.*, 2012: HOMER: homogenisation software in R –methods and applications. *Időjárás 117*, 47–67.

MeteoSchweiz, 2000: Alte meteorologische Instrumente (Old meteorological instruments). Bundesamt für Meteorology und Klimatologie (MeteoSchweiz), Zürich, 190 p.

Nemec J., Gruber, G., Chimani, B., and *Auer, I.*, 2012: Trends in extreme temperature indices in Austria based on a new homogenized dataset. *Int. J. Climatol.*, DOI: 10.1002/joc.3532.

Nordli, P.O., Alexandersson, H., Frich, P., Förland, E.J., Heino, R., Jonsson, T., Tuomenvirta, H., and Tveito, O.E., 1997: The effect of radiation screens on Nordic time series of mean temperature. Int. J. Climatol. 17, 1667–1681.

Picard, F., Robin, S., Lavielle, M., Vaisse, C., and Daudin, J.-J., 2005: A statistical approach for array CGH data analysis, BMC Bioinformatics 6, 27.

Picard F., Lebarbier, E., Hoebeke, M., Rigaill, G., Thiam, B., and *Robin, S.*, 2011: Joint segmentation, calling, and normalization of multiple CGH profiles. *Biostatistics 12*, 413–428.

Rust, H.W., Mestre, O., and *Venema, V.K.C.*, 2008: Less jumps, less memory: homogenized temperature records and long memory. *J. Geophys. Res. Atmos. 113*, D19110.

Slonosky, V.C., Jones, P.D. and *Davies, T.D.*, 2001: Instrumental pressure observations and atmospheric circulation from the 17th and 18th centuries: London and Paris, *Int. J. Climatol. 21*, 285–298.

Szentimrey, T., 1996: Statistical procedure for joint homogenisation of climatic time series. *Proceedings of the First seminar of homogenisation of surface climatological data*, Budapest, Hungary, 6–12 October 1996, 47–62.

Szentimrey, T., 1999: Multiple Analysis of Series for Homogenization (MASH). *Proceedings of the second seminar for homogenization of surface climatological data*, Budapest, Hungary; WMO, WCDMP-No. 41, 27–46.

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

Trewin, B., 2010: Exposure, instrumentation, and observing practice effects on land temperature measurements. *WIREs Clim. Change, 1*, 490–506.

Van der Meulen, J.P. and *Brandsma, T.*, 2008: Thermometer screen intercomparison in De Bilt (The Netherlands), part I: Understanding the weather-dependent temperature differences. *Int. J. Climatol. 28*, 371–387.

Venema, V., Mestre, O., Aguilar, E., Auer, I., Guijarro, J.A., Domonkos, P., Vertacnik, G., Szentimrey, T., Stepanek, P., Zahradnicek, P., Viarre, J., Müller-Westermeier, G., Lakatos, M., Williams, C.N., Menne M.J., Lindau, R., Rasol, D., Rustemeier, E., Kolokythas, K., Marinova, T., Andresen, L., Acquaotta, F., Fratianni, S., Cheval, S., Klancar, M., Brunetti, M., Gruber, Ch., Prohom Duran, M., Likso, T., Esteban, P., and *Brandsma, Th.*, 2012: Benchmarking homogenization algorithms for monthly data. *Clim. Past 8*, 89–115.

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

IDŐJÁRÁS

Quarterly Journal of the Hungarian Meteorological Service
Vol. 117, No. 1, January–March 2013, pp. 35-45

Climatological series shift test comparison on running windows

José A. Guijarro

State Meteorological Agency, Moll de Ponent s/n, Portopi,
07015-Palma de Mallorca, Spain
jguijarrop@aemet.es

(Manuscript received in final form October 3, 2012)

ABSTRACT–The detection and correction of inhomogeneities in the climate series is of paramount importance for avoiding misleading conclusions in the study of climate variations. One simple way to address the problem of multiple shifts in the same series is to apply the tests on windows running along the series of anomalies. But it is not clear which of the available tests works better. 500 Monte Carlo simulations have been done for the ideal case of a 600 normally distributed terms (a 50 years series of monthly differences), with a single shift in the middle and magnitudes of 0 to 2 standard deviations (s) in steps of 0.2 s. The compared tests have been: 1) classical t-test; 2) standard normal homogeneity test; 3) two-phase regression; 4) Wilcoxon-Mann-Whitney test; 5) Durbin-Watson test (lag-1 serial correlation), and 6) squared relative mean difference (simpler than t-test and hence faster to compute). The criterion for qualifying the performance of each test was the ability to detect shifts without false alarms and to locate them at the correct point. Results indicate that, under these precise simulated conditions, the best test are the classical t-test, Alexandersson's SNHT and SRMD, with almost identical results, followed by the Wilcoxon-Mann-Whitney test, while two phase regression and Durbin-Watson performances are very poor.

Key-words: homogenization, shift tests comparison, climatological series.

1. Introduction

Climatological series are very important for studying climate variability at all scales, but the climate signal is too often merged with unwanted variations due to changes in the type or exposure of the instruments, methods of observation, relocations of the stations, or changes in their surroundings.

Many methodologies have been proposed so far to detect and correct these inhomogeneities, which commonly appear as either sudden shifts or smooth trends in relative series. These relative series are usually computed as difference or ratio series between the problem series and a reference, that can be an observed trusted homogeneous series or a synthetic one compiled from a selection of the nearest or more correlated stations. Reviews of the different methods can be seen in *Easterling* and *Peterson* (1992), *Peterson et al.* (1998), *Aguilar et al.* (2003), and *Beaulieu et al.* (2008).

Several comparisons of shift detection methods have been undertaken so far (*Easterling* and *Peterson*, 1995; *Bosshard* and *Baudenbacher*, 1997; *Ducre-Robitaille et al.*, 2003; *Beaulieu et al.*, 2008), their results being influenced by the type (shifts and/or local trends), number and position of the simulated inhomogeneities, differences in station variance and between-station correlation structure, series length, autocorrelation, and nonstationarity.

The frequent concurrence of several jumps in the same series makes their detection problematic. One simple way to address this problem is to apply the test on time moving windows. During the development of an automated homogenization function for the CLIMATOL R contributed package (*Guijarro*, 2011a), the chosen approach for the detection of multiple change points was the application of a two-sample t-test for equal means to windows running along the series of anomalies (differences between the tested series and a synthetic reference series computed from neighboring stations). At this point, the question whether there were better detection tests emerged, but the available reviews are not fully conclusive, since the performance of the tests depends on the particular settings of the simulations and the significance threshold values chosen in each case, as it happens in the differing results of *Ducre-Robitaille et al.* (2003) and *Beaulieu et al.* (2008).

Therefore, new Monte Carlo experiments were designed to test the sensitivity and correctness of several algorithms in detecting and locating a shift in repeated series of white noise that simulate the ideal case of series of differences between a tested series with a single abrupt change in the mean and a homogeneous well correlated reference series. In this way we avoid the problems of simulating networks of observation or pairs of tested and reference stations as in the aforementioned evaluation exercises. Moreover, no a priori level of significance will be imposed, and location errors of the break point will be studied with no established thresholds of good/bad location. Next sections will explain this methodology, and the results of the tested algorithms will be discussed.

2. Methodology

500 series of 600 normally distributed terms (equivalent to tested minus reference monthly series of 50 years) were generated with the help of the *R*

function *rnorm* (*R Development Core Team*, 2010). Single shifts were added to all of them just in the middle (from term 301) with magnitudes from 0 to 2 standard deviations (s) in steps of 0.2 s, yielding a total of 5500 testing series. Six shift detection algorithms were applied on them, but not over the whole series, but on fixed width windows running along them. Different sample sizes were tried, from n=1 to 5 years (12 to 60 terms in steps of 12), and since two samples were involved in the shift tests, window widths of 2, 4, 6, 8, and 10 years were used. In this way, for n years sample size, every algorithm was tested $600–24 \cdot n+1$ times in each of the 5500 series (from 577 times with 1 year samples to 481 for samples of 5 years). *Fig. 1* shows an example series with a 0.8 s shift.

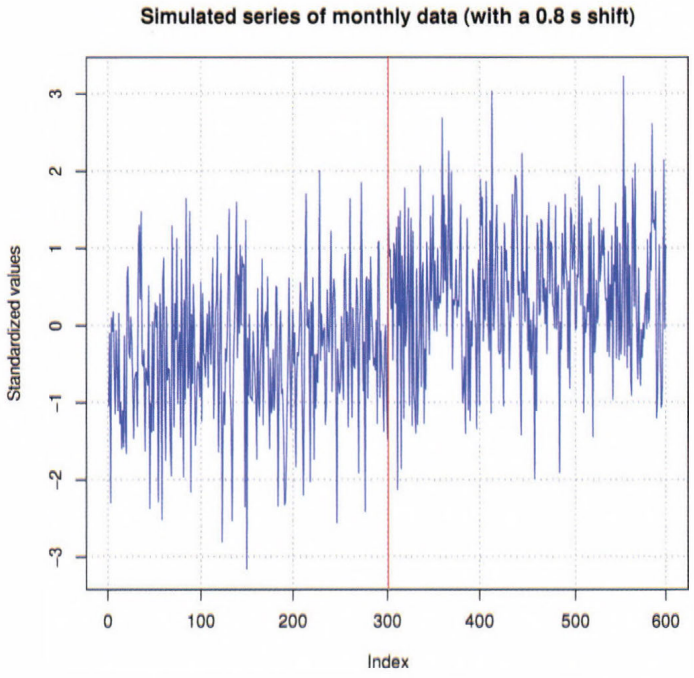

Fig. 1. Example of white noise difference series of 600 terms with a shift of 0.8 standard deviations in term 301.

The six algorithms tested were the following:

1. t-test: the classical test of mean differences of two samples.
2. SNHT: *Alexandersson*'s (1986) algorithm, but modified to test the middle point of the window only.

3. TPR: two-phase regression, as formulated by *Easterling* and *Peterson* (1995).

4. WMW: Wilcoxon-Mann-Whitney test, which is similar to the Wilcoxon rank sums applied by *Karl* and *Williams* (1987) but as formulated by *Gérard-Marchant* and *Stooksbury* (2008), and divided by the number of terms to make it less dependent on the sample size.

5. DW: lag-1 Durbin-Watson test for serial correlation.

6. SRMD (squared relative mean difference): $z = [\,(m_1\text{-}m_2)\cdot s^{-1}]^2$, where m_1 and m_2 are the sample means and s is the standard deviation of the whole window.

The reference values of DW and t-test were their returned p-values, but $\log_{10}$ transformed and sign reversed to allow more friendly figures (they are called *pV*, by analogy with the alkalinity index *pH* used in chemistry). *Fig. 2* displays the values returned by the six algorithms after being applied to a series similar to that in *Fig. 1* on running windows of 10 years (sample sizes of 5 years, i.e., 60 terms). Only the maximum value reached along the series, and its location (the middle point of the window giving that value) were retained for the statistical analysis of the results.

3. Results and discussion

The frequencies of the maximum values returned by the tests on each series and the errors of their corresponding locations (diagnosed break term minus 301) were analyzed statistically, and the results are shown graphically in form of boxplots, where each box summarizes 500 results. *Fig. 3* shows the influence of window size on the results yielded by the t-Test. It is clear that sample sizes of 12 terms are too small to allow the detection of shifts. If we take the value of the top whisker of the first box (homogeneous series) as a reasonable threshold to avoid false break detection, only roughly half of the 2 s shifts would be identified. With wider windows the power of detection improves: the half of the breaks detection reference is achieved with 0.8 and 0.6 s shifts for samples of 3 and 5 years respectively. (The intermediate 4 year sample graph can be seen in Figure 4). These results are in accordance with those of *Beaulieu et al.* (2008), who found that shifts under 1 s were difficult to identify, while all techniques tested by them worked well for breaks greater than 2 s.

The performance of the six tests with samples of 4 years can be seen in *Fig. 4*. As every test has its own metric, the units displayed in the vertical axis are all different, but it is easy to see that some tests reach higher values quicker than others as the shift magnitude increases, showing their greater power of detection.

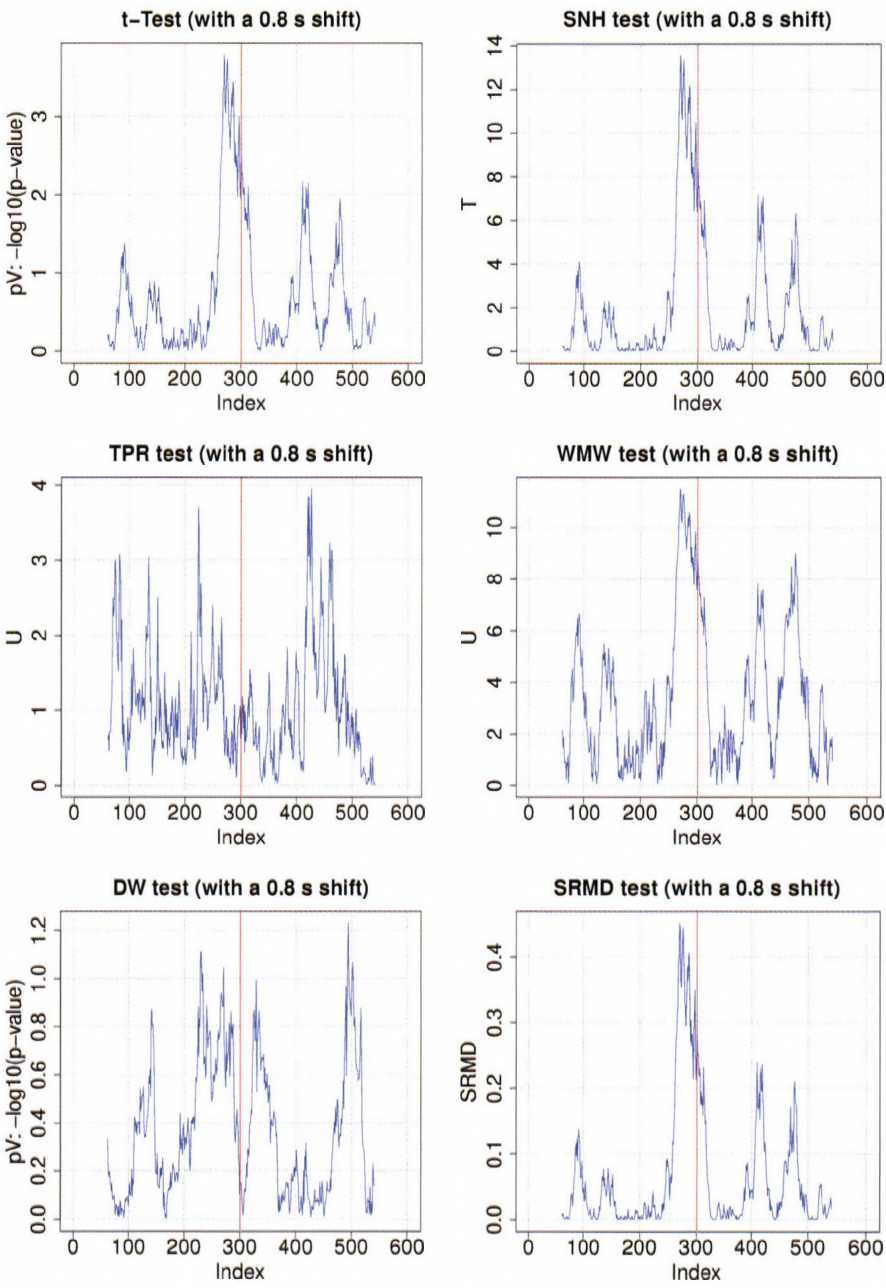

Fig. 2. Graphs of the values returned by the tests when applied to a series similar to that in *Fig 1* on running windows of 120 terms (two samples of 5 years). The vertical bar in the middle of the series indicates the true possition of the shift.

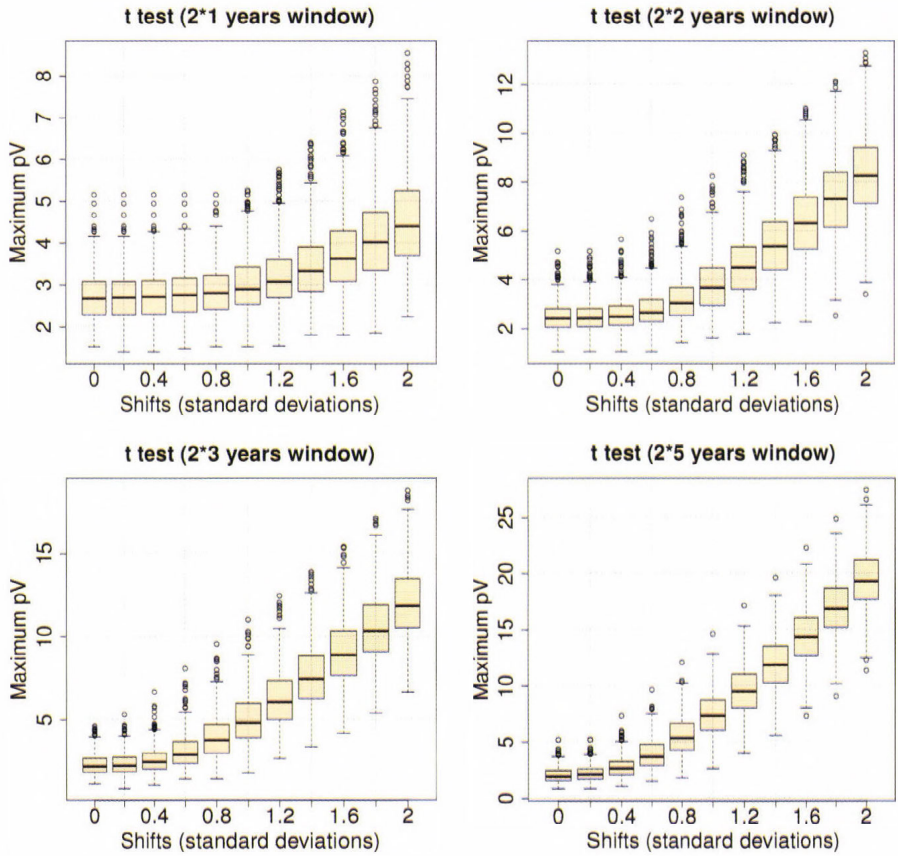

Fig. 3. Influence of window size on the results yielded by the t-test.

Table 1 presents the percentage of shift detection of every algorithm for each shift, for the 5 years samples, when the threshold detection is placed: a) at the maximum value obtained with the homogeneous series (no false detection is allowed); b) at the 99 percentile of the homogeneous values (permitting 1% of false detection). The best performances correspond to t-test, SNHT and SRMD, that give almost identical results, showing that they belong to the same family of tests. WMW follows, with good results form 1 *s* shift onwards, while DW and TPR both yield similar discouraging scores. Note that the thresholds of any test applied hundreds of times on every series through such a running window procedure, must be higher than their corresponding significant levels when applied only once on each series. E.g., the 14.23 of SNHT allowing 1% of false detection is higher than the 13.813 published by *Khaliq* and *Ouarda* (2007) for a 99% confidence level and sample size of 600 values (the whole simulated series).

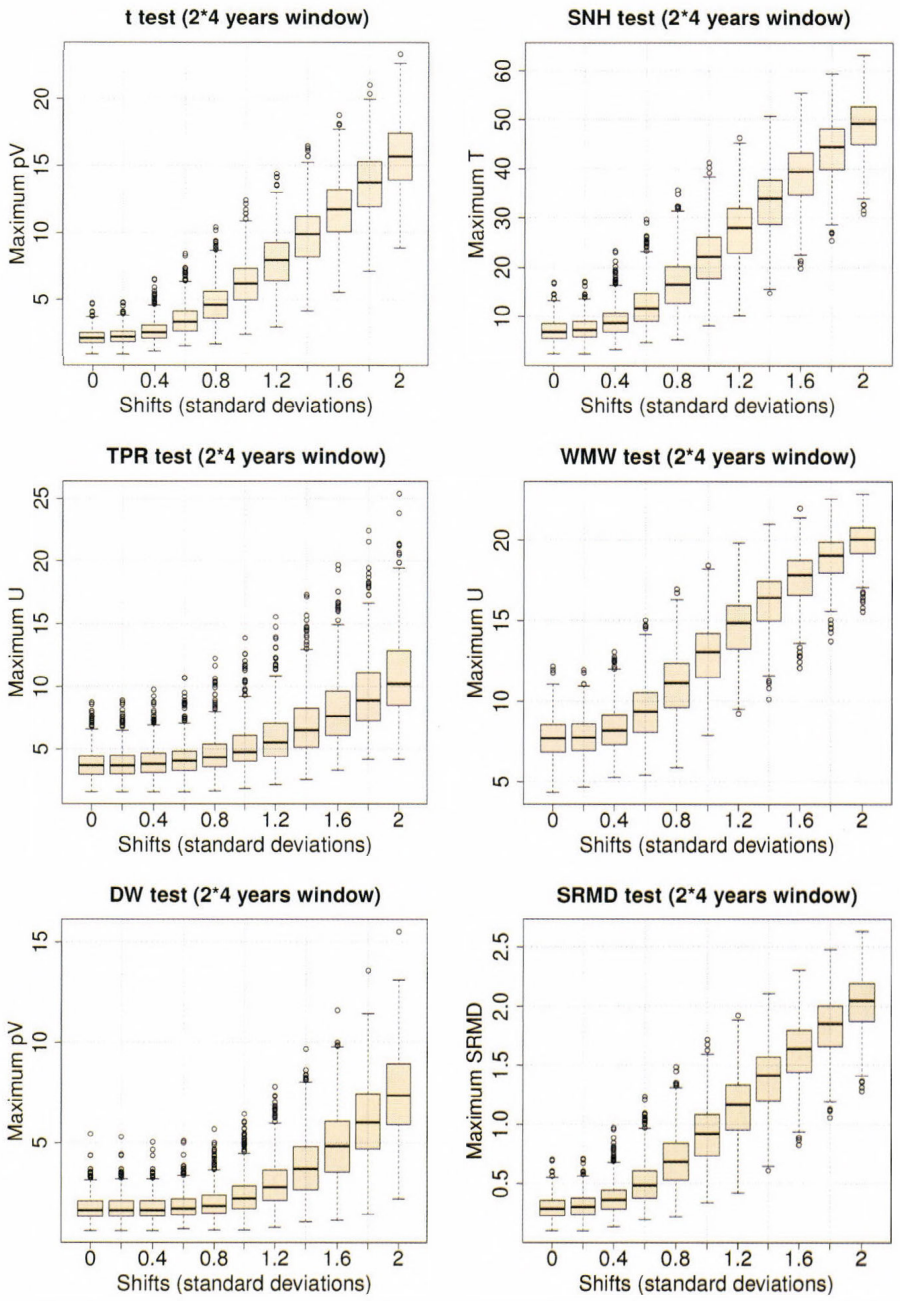

Fig. 4. Values of the six algorithms for shifts ranging from 0 to 2 standard deviations.

Table 1. Threshold values and percentage of shift detection in the cases of no false detection and allowing 1% of false detections, for a 5 years sample size (running windows of 10 years, i.e., 120 terms)

		Shift (standard deviations)									
Test	Thresh.	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0
					No false detection:						
t-test	5.22	0.0	2.4	16.6	54.4	88.8	98.8	100.0	100.0	100.0	100.0
SNHT	19.06	0.0	2.4	16.8	54.6	88.8	98.8	100.0	100.0	100.0	100.0
TPR	8.69	0.0	0.6	1.6	2.8	7.8	16.6	34.4	55.0	74.8	88.4
WMW	13.78	0.0	2.0	12.6	47.4	82.8	98.4	100.0	100.0	100.0	100.0
DW	4.98	0.0	0.0	0.0	0.6	3.6	14.8	37.8	65.2	86.0	95.8
SRMD	0.635	0.0	2.4	16.6	54.4	88.8	98.8	100.0	100.0	100.0	100.0
					1% false detection:						
t-test	3.96	2.2	13.2	44.6	81.4	97.8	100.0	100.0	100.0	100.0	100.0
SNHT	14.23	2.4	13.2	44.4	81.4	97.8	100.0	100.0	100.0	100.0	100.0
TPR	6.87	1.6	3.0	4.6	11.6	20.4	41.0	62.2	79.4	91.0	97.4
WMW	12.04	0.8	8.6	35.0	73.4	95.2	99.8	100.0	100.0	100.0	100.0
DW	3.59	1.0	1.2	1.8	5.4	17.4	41.2	66.4	87.0	96.4	99.6
SRMD	0.474	2.4	13.2	44.4	81.4	97.8	100.0	100.0	100.0	100.0	100.0

With respect to the location errors, *Fig. 5* shows the corresponding box plots for the 4 years sample size (running windows of $2 \cdot 4 \cdot 12 = 96$ terms). Again, the t-test family (including SNHT and SRMD) reaches the best results, with small location errors for shifts greater than 0.6 standard deviations. Location errors of WMW are only slightly higher, but those of DW and specifically TPR are very big.

As CLIMATOL must apply the chosen test many times in iterative runs during the homogenization of a climatological network, computing efficiency is also important, and therefore, the time used by each of the tests was accounted for. Those adjusting regression models (TPR and DW) were the most time consuming using the R *lm* function. The R implementation of the t-test is much faster, but at the same time much slower than SNHT, probably due to its higher complexity and the inherent computation of p-values and other statistical parameters. This is why SRMD was introduced, achieving identical results as SNHT (in this two sample version), but at 20% higher speed. If TPR or DW had given better results, rewriting the regression algorithm to shorten their computing time would have been explored.

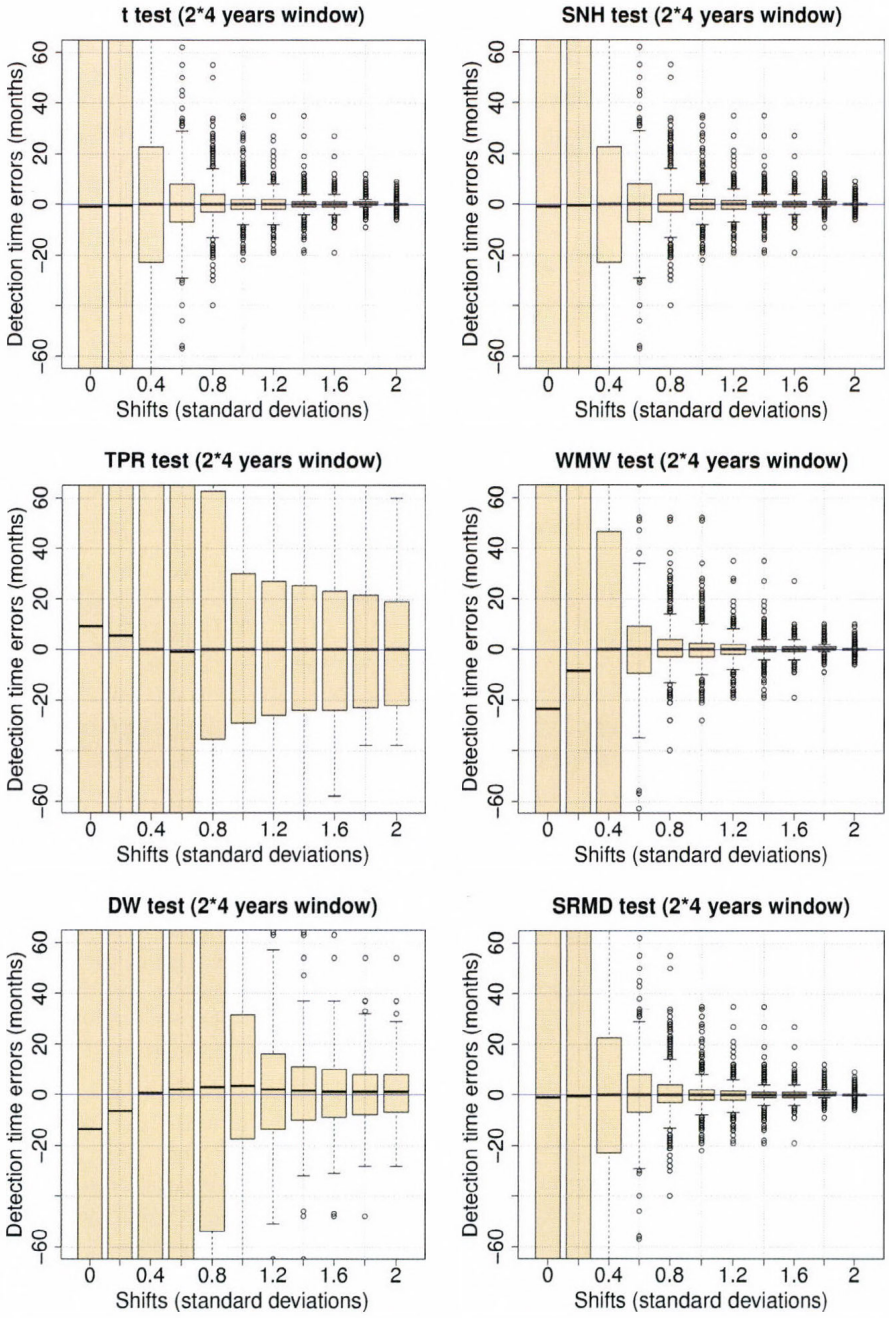

Fig. 5. Location errors of the six algorithms for shifts ranging from 0 to 2 standard deviations.

The combination of several of these tests was also tried, but when the best algorithm is used, there is no advantage in adding the results of any others. Therefore, CLIMATOL 2.0 implemented SRMD on running windows (4 years samples by default). Nevertheless, practical applications of that version showed that clear inhomogeneities spanning less than 3 years are common in real climatological series, and they were difficult to correct automatically due to the constraint of the minimum 3 years sample size required by the algorithm. Hence, the following 2.1 version dropped SRMD in favor of the popular and well tested SNTH which, freed from the window size restriction, is able to resolve close shifts. To avoid possible masking effects when multiple shifts are present in the same series, this test was implemented in two stages. In the first stages SNHT is applied on shifted windows of user defined width, and when significant shifts detected in this way have been corrected, SNHT is applied to the whole series in the second stage (*Guijarro*, 2011b).

4. Conclusions

The results of the simulations performed in this work indicate that, under these precise conditions of detection of a single shift in the middle of the series by means of fixed width windows running along the series, the best tests are the classical t-test and SNHT. SMRD is a simple derivative of the t-test with the same performance. The Wilcoxon-Mann-Whitney test yields acceptable results, but the two-phase regression and Durbin-Watson performances are very poor (although they can be better in other situations, e.g., in detecting local trends).

Nonetheless, windows need to have a minimum width of 6 years (two samples of 3 years), and that restrains the time resolution at which two close shifts can be identified. As a result, the t-test procedure of comparing the means of two samples was abandoned in favor of the standard formulation of SNHT, applied on stepped windows to avoid misleading results in the presence of multiple breaks in a first stage, then followed by an application on the whole series.

References

Aguilar, E., Auer, I., Brunet, M., Peterson, T.C., and *Wieringa, J.,* 2003: *Guidelines on climate metadata and homogenization*. WCDMP-No. 53, WMO-TD No. 1186. World Meteorological Organization, Geneve.

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

Beaulieu, C., Seidou, O., Ouarda, T.B.M.J., Zhang, X., Boulet, G., and *Yagouti, A.,* 2008: Intercomparison of homogenization techniques for precipitation data. *Water Resour. Res, 44*, 20.

Bosshard, W. and *Baudenbacher, M.,* 1997: Evaluation of various homogeneity tests by simulation of climatological time series. In: *Proceedings of the First Seminar for Homogenization of Surface Climatological Data*, Budapest, 6–12 October 1996, Hungarian Meteorological Service, 19–34.

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

Easterling, D.R. and *Peterson, T.C.,* 1992: Techniques for detecting and adjusting for artificial discontinuities in climatological time series: a review. *5th International Meeting on Stat. Climatology*, June 22–26, 1992, Toronto.

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

Gérard-marchant, P.G.F. and *Stooksbury, D.E.*, 2008: Methods for Starting the Detection of Undocumented Multiple Changepoints. *J. Climate 21*, 4887–4899.

Guijarro, J.A., 2011a: http://cran.r-project.org/web/packages/climatol/index.html.

Guijarro, J.A., 2011b: User's guide to Climatol. 40 pp. http://webs.ono.com/climatol/climatol-guide.pdf

Karl, T.R. and *Williams, C.N.,* 1987: An approach to adjusting climatological time series for discontinuous inhomogeneities. *J. Clim. Appl. Meteor. 26*, 1744–1763.

Khaliq, M.N. and *Ouarda, T.B.M.J.*, 2007: On the critical values of the standard normal homogeneity test (SNHT). *Int. J. Climatol. 27*, 681–687.

Peterson, T.C., Easterling, D.R., Karl, T.R., Groisman, P., Nicholls, N., Plummer, N., Torok, S., Auer, I., Boehm, R., Gullett, D., Vincent, L., Heino, R., Tuomenvirta, H., Mestre, O., Szentimrey, T., Salinger, J., Førland, E., Hanssen-bauer, I., Alexandersson, H., Jones, P., and *Parker, D.*, 1998: Homogeneity Adjustments of „In Situ" Atmospheric Climate Data: A Review. *Int. J. Climatol. 18*, 1493–1518.

R Development Core Team, 2010: *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

IDŐJÁRÁS

Quarterly Journal of the Hungarian Meteorological Service
Vol. 117, No. 1, January–March 2013, pp. 47-67

HOMER : a homogenization software – methods and applications

Olivier Mestre[1]*, Peter Domonkos[2], Franck Picard[3], Ingeborg Auer[4], Stéphane Robin[5,6], Emilie Lebarbier[5,6], Reinhard Böhm[4], Enric Aguilar[7], Jose Guijarro[8], Gregor Vertachnik[9], Matija Klancar[9], Brigitte Dubuisson[1], and Petr Stepanek[10]

[1]*Meteo-France, Direction de la Production, 42 avenue Coriolis, 31057 Toulouse cedex, France*

[2]*Center for Climate Change, Univ. Rovira i Virgili, Av. Remolins, 13-15, 43500-Tortosa, Spain*

[3]*UCB Lyon 1, UMR 5558, Villeurbanne, France*

[4]*Zentralanstalt für Meteorologie und Geodynamik, Wien, Austria*

[5]*AgroParisTech, UMR 518, Paris, France*

[6]*INRA, UMR 518, Paris, France*

[7]*Center for Climate Change, Univ. Rovira i Virgili, Tarragona, Spain*

[8]*Agencia Estatal de Meteorologia, Palma de Mallorca, Spain*

[9]*Environmental Agency of the Republic of Slovenia, Meteorology, Ljubljana, Slovenia;*

[10]*Czech Hydrometeorological Institute, Brno, Czech Republic*

**Corresponding author E-mail: Olivier.Mestre@meteo.fr*

(Manuscript received in final form October 24, 2012)

Abstract–Between 2007–2011, the European COST Action ES0601 called HOME project was devoted to evaluate the performance of homogenization methods used in climatology and produce a software that would be a synthesis of the best aspects of some of the most efficient methods. HOMER (HOMogenizaton softwarE in R) is a software for homogenizing essential climate variables at monthly and annual time scales. HOMER has been constructed exploiting the best characteristics of some other state-of-the-art

homogenization methods, i.e., PRODIGE, ACMANT, CLIMATOL, and the recently developed joint-segmentation method (*cghseg*). HOMER is based on the methodology of optimal segmentation with dynamic programing, the application of a network-wide two-factor model both for detection and correction, and some new techniques in the coordination of detection processes from multiannual to monthly scales. HOMER also includes a tool to assess trend biases in urban temperature series (UBRIS). HOMER's approach to the final homogenization results is iterative. HOMER is an interactive method, that takes advantage of metadata. A practical application of HOMER is presented on temperature series of Wien, Austria and its surroundings.

Key-words: Homogenization, optimal segmentation, joint segmentation, ANOVA, temperature, precipitation, urban trend bias

1. Introduction

The accuracy of climatic observations is often affected by inhomogeneities due to changes in the technical or environmental conditions of the measurements (station relocations, changes of the type, height or sheltering of the instruments, etc., *Aguilar et al.*, 2003, *Auer et al.*, 2005). Most of such changes cause sudden shifts (change-points) in the series of local climatic data, while some others (particularly urban development) result in gradually increasing biases from the real macroclimatic characteristics. Correction of inhomogeneities before any climate variability analyses is highly desirable, and for this purpose, a large number of homogenization methods have been developed in the recent decades (*Peterson et al.*, 1998; *Ducre-Robitaille et al.*, 2003; *Beaulieu et al.*, 2008; among others).

HOMER is a recently developed method for homogenizing monthly and annual temperature and precipitation data. It includes the best features of some other state-of-the-art methods, namely PRODIGE (*Caussinus* and *Mestre*, 2004), ACMANT (*Domonkos*, 2011), and *cghseg* a joint segmentation method that was developed originally by bio-statisticians in the context of DNA segmentation (*Picard et al.*, 2011). PRODIGE and ACMANT have the same theoretical base regarding the optimal segmentation with dynamic programming DP (*Hawkins*, 2001), an information theory based formula for determining the number of segments in time series (hereafter: C&L criterion, *Caussinus* and *Lyazrhi*, 1997), and a network-wide unified correction model (ANOVA, *Caussinus* and *Mestre*, 2004). The results of blind test experiments conducted during COST Action ES0601 (*Venema et al.*, 2012) validates these approaches, since PRODIGE and ACMANT rank among the best methods for homogenizing monthly and annual climate data (*cghseg* and HOMER were not tested during the HOME action). The joint segmentation is an extension of the optimal segmentation for finding network-wide optima by means of an iterative procedure, a modified BIC criterion being used for determining the number of changes (*Zhang* and *Siegmund*, 2007; *Picard et al.*, 2011).

HOMER is an interactive semi-automatic method. In applying HOMER, users may choose between the *cghseg* detection results whose generation is fully automatic on the one hand, and a partly subjective pairwise comparison technique that is adapted from PRODIGE on the other hand. This freedom allows users to add subjective decisions based on metadata or research experiences. HOMER includes also some innovations of ACMANT in the coordination of working on different time scales. Basic quality control and network analysis are adapted from CLIMATOL (*Guijarro*, 2011).

Our paper is organized as follows: first, Section 2 describes the main models and procedures of HOMER. The methodology of characterizing urban trends (UBRIS) and the main properties of ACMANT are also presented there, together with a discussion. An application of HOMER on Wien temperature series is then shown in Section 3.

2. *HOMER main procedures*

In this section, we will focus on functions used during the homogenization process: statistical tools for pairwise detection (2.1), two factor model for joint detection and correction (2.2), UBRIS model for urban trend bias assessment (2.3), ACMANT functions (2.4). Usefulness of each task is discussed in 2.5, and a workflow of tasks is provided.

2.1. Detection of changes in pairwise series (univariate detection)

2.1.1. Model

Let Y be the annual or seasonal difference between two series. We model $Y_i, i=1,\ldots,n$ as a series of Gaussian variables of constant variance σ^2, but with varying mean μ from sub-period to subperiod. The number and positions of change-points are unknown.

Let k the number of changes and $\tau_1, \tau_2, \ldots, \tau_k$ their positions. We denote $K=\{\tau_1,\ldots,\tau_k\}$ the set of changes in the series. At most cases old data are adjusted relative to the modern data, and for simplicity $\tau_o=0$ is fixed at $\tau_{k+1}=n$. Further notations are:

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i ,$$

$$\overline{Y}_j = \frac{1}{n_j} \sum_{\tau_{j-1}+1}^{\tau_j} Y_i \,,$$

where $n_j = \tau_j\text{-}\tau_{j-1} + 1 \;;\; j = 1, \ldots, k+1$,

$$\overline{Y}_{hm} = \frac{1}{m-h+1} \sum_{i=h+1}^{m} Y_i \,,$$

$$W_{hm} = \sum_{i=h+1}^{m} \left(Y_i - \overline{Y}_{hm}\right)^2 .$$

Changes in the mean are $\mathrm{IE}[Y_i] = \nu_j$ for $\tau_{j-1} + 1 < i < \tau_j$

Maximum likelihood estimates of the ν_j's are straightforwardly given by $\hat{\nu}_j = \overline{Y}_j$. For a given number k, we wish to maximize the likelihood, which is equivalent to minimize deviance D:

$$D_k = \frac{\sum_{j-1}^{k+1} \sum_{\tau_{j-1}+1}^{\tau_j} \left(Y_i - \overline{Y}_j\right)^2}{\sigma^2} + 2n \log\left(\sqrt{2\pi}\,\sigma\right) \quad . \qquad (1)$$

2.1.2. Dynamic programming

The naive way to minimize deviance D is to consider every combination of the position of the change-points. But the number of hypotheses rises very fast with n, the length of the series, and k, the number change-points. When detection is performed for change-points in a normal sample, a DP algorithm can be used (*Lavielle*, 1998; *Hawkins*, 1972, 2001; etc.). Computation time then becomes only linear in k and quadratic in n. It is based on a recursion between optimal k and k–1 solutions. DP allows us to find an optimal solution without computing all possibilities. For k changes, the problem is to minimize:

$$Q = \sum_{j=1}^{k+1} \sum_{i=\tau_{j-1}+1}^{\tau_j} \left(Y_i - \overline{Y}_j\right)^2 = \sum_{j=1}^{k+1} W_{\tau_{j-1}\,\tau_j} \,. \qquad (2)$$

The solution is given by the following recursion:

- $F_{1,m} = W_{0m}$ for $m = 1, n$,
- for each $r = 2, \ldots k+1$, let us compute $F_{r,m} = \mathrm{MIN}_{0<h<m}\left[F_{r-1,h} + W_{h,m}\right]$ for $m = 1, n$,
- for each $F_{r,m}$ value, let us keep in table $H_{r,m}$ the h value that corresponds to the minimum of $F_{r,m}$,
- the change-point estimates are given by: $\tau_{k+1} = n$, and for $r = k, k-1, \ldots, 1$ we get $\hat{\tau}_r = H_{r+1,\tau_{r+1}}$.

2.1.3. Selecting the number of changes.

The fit of the change-point model increases monotonously with k ($Q = 0$ for $k = n$). The model selection is guided finding the most parsimonious model that gives a "good" explanation of data vector Y. Several penalized likelihood criteria can be found in the literature. In the latest version of HOMER, we take the advantage of the uniseg procedure from the R package which uses the modified BIC criterion of *cghseg*. As in *Schwarz*'s BIC (1978), *Zhang* and *Siegmund* approach this problem by deriving an asymptotic approximation of the Bayes factor, using a uniform prior on change-points location (among other hypotheses).

The procedure is as follows: for each value of k, DP allows us to select the optimal position for the k change-points $\{0, \hat{\tau}_1, \hat{\tau}_2, \ldots, \hat{\tau}_k, n\}$. For each k value, $MBIC(Y; k)$ is computed:

$$MBI(Y;k) = \left(\frac{n-k+1}{2}\right)\log\left[1 - \frac{\sum_{j=1}^{k+1} n_j\left(\overline{Y}_j - \bar{Y}\right)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}\right] + \log\left[\frac{\Gamma\left(\frac{n-k+1}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)}\right]$$

$$+\frac{k}{2}\log\left[\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right] - \frac{1}{2}\log\left[\sum_{i=1}^{n}\left(\overline{Y}_i - \bar{Y}\right)^2\right] \qquad (3)$$

$$-\frac{1}{2}\sum_{j=1}^{k+1}\log(n_j) + \left(\frac{1}{2} - k\right)\log(n) \ ,$$

where Γ denotes the Gamma function. The model selection consists in selecting the number of change-points k that minimizes *MBIC*:

$$\text{select } k^* \text{ such that } k^* = \text{Argmin}_k(MBIC_k(Y;k)). \quad (4)$$

This criterion is more complex than the classical BIC or C&L criteria used in PRODIGE, but does not require any user-chosen shrinkage parameters like in *Tibshirani* (1996), *Birgé* and *Massart* (2001) or *Gu* and *Wang* (2003). The first term in Eq. (3) corresponds to a likelihood ratio term, the subsequent ones are the penalty. One has to note that the penalty depends on n, k, but also on the closeness of the changes via the sum of $\log(n_j)$ term: close change-points are more penalized. Simulations (not shown here) show that *MBIC* criterion is slightly less powerful than the C&L, but less sensitive to small autocorrelation that still might be present in the pairwise comparisons.

Standard deviation of the residuals is then estimated by:

$$\hat{\sigma}^* = \sqrt{\frac{1}{n-k^*}\sum_{j=1}^{k^*+1}\sum_{i=\hat{\tau}_{j-1}+1}^{\hat{\tau}}\left(Y_j-\bar{Y}_j\right)^2} = \sqrt{\frac{1}{n-k^*}\sum_{j=1}^{k^*+1}W_{\hat{\tau}_{j-1}\hat{\tau}_j}} \; . \quad (5)$$

We will see in practice that this estimation of noise is very useful, since detection power is directly related to the signal (i.e., amplitude of changes) to noise ratio. Smaller values of noise ensure more accurate detection.

2.2. ANOVA: a two-factor model for joint-detection and correction

2.2.1. Model

Let us consider p series belonging to the same climate area in such a way that all the series are affected by the same climatic conditions at the same time. This assumption is realistic when considering monthly or annual observations of the same geographical region. We assume that each series of observations is the sum of a climatic effect, a station effect, and random white noise. This is a simple two-factor analysis of variance model without interaction, and we will denote it by ANOVA in the following.

Let X be a matrix of n observations X_{ij} on p series where $i=1,...,n$ is the time index and $j=1,...,p$ is the station index. Let k_j be the number of change-points, let $\tau_{1,j}, \tau_{2,j}, \ldots, \tau_{k_j,j}$ be the positions of these k_j change-points. Let $K_j=(\tau_{1,j}, \ldots, \tau_{k_j,j})$ be the set of change-points for series j. To simplify the

notation, we set again $\tau_{o,j}$=0, and $\tau_{k_{j+1},j} = n$, so that K_j becomes $K_j = \left(0, \tau_{1,j}, \ldots, \tau_{k_j,j}, n\right)$.

The station effect is constant if the series is homogeneous. If not, the station effect is constant between two shifts. In the following, level denotes a homogeneous sub-period between two discontinuities of a given series. For a series j with k_j breaks, let L_{jh} be the hth level (h=1,...,k_j+1), thus L_{jh} is the interval: $\left[\tau_{h-1,j} + 1, \tau_{h_j,j}\right]$. Note that the level h for the observation X_{ij} depends both on time i and station j: when necessary it will be written $h(i,j)$.

Let μ_i be the climate effect at time i and ν_{jh} the station effect of station j for level L_{jh}. If there are no outliers, the data are described by the linear model:

$$\mathrm{IE}(X_{ij}) = \mu_1 + \nu_{jh(i,j)} \quad , \quad \mathrm{Var}(X) = \sigma^2 I_{np} \quad . \tag{6}$$

One parameter of the model can be freely chosen and it is done with introducing the condition $\sum_{i=1}^{n} \mu_i = 0$, so that μ_i are defined as climate anomalies.

The number of independent parameters of the model without discontinuities is $n+p-1$.

Examples:

- No break in series 1: $\mathrm{IE}(X_{i1}) = \mu_i + \nu_1$,
- One break at i_o for series 2: $\begin{cases} \mathrm{IE}(X_{i2}) = \mu_i + \nu_{21} \text{ , for } i \leq i_0 \\ \mathrm{IE}(X_{i2}) = \mu_i + \nu_{22} \text{ , for } i > i_0 \end{cases}$.

Some further characteristics of the model:

a) Estimation can be performed with missing data with the following conditions: there should be at least one non-missing value per year on the whole network (estimation of the μ's) and one non-missing value between two breaks for each subperiod on each series.

b) Climate signal is treated as a fixed parameter so that no assumption is made about the shape of this signal.

c) Conditionally to the climate signal, the disturbances are considered independent.

d) Local variabilities are very similar, which leads to the expression of Var(X).

Note that conditions c) and d) are approximately true within the same climatic region. Small spatial autocorrelation may be observed in the residuals.

So far, this model has been used in PRODIGE and ACMANT mainly for correction purposes – although *Caussinus* and *Mestre* (2004) propose some clue to use it for detection. It has been shown that the inclusion of ANOVA correction improves significantly the results of other methods participated in HOME blind test experiments (*Domonkos et al.*, 2012b). Using HOME benchmark and the set of break-points detected using for example standard normal homogeneity test (SNHT), correcting the inhomogeneities by ANOVA allowed a much better homogenization than the standard SNHT correction method. We will see below that this model can be used for detection as well, allowing for joint detection of a whole set of series.

2.2.2. Joint-detection

The change-point model Eq. (6) can theoretically be used for joint detection of the changes on the whole set of series. However, due to the introduction of factor μ, the classical DP algorithm cannot be applied (*Caussinus* and *Mestre*, 2004) and until recently, joint segmentation was considered computationally intractable. Adapted algorithms allow us to solve this problem in a reasonable computing time. *Picard et al.* (2011) rely on two "computational tricks". The first one solves the problems caused by segmentation of multiple series. Let us set all $\mu_i = 0$. Since DP complexity is quadratic with the size of the data, just considering segmentation of the ν factor may become problematic when considering multiple series. *Picard et al.* (2011) propose a "two-stage" DP algorithm that significantly reduces the computation time. Briefly, the first stage consists in finding all optimal solutions for each ν_j factor separately, from $k = 1$ to $kmax_j$. The second stage uses outputs from the first stage to optimally allocate the number of segments to each factor $\nu_1,\ldots,\nu_p$, in order to maximize the overall fit. The model selection is provided by a multivariate version of *Zhang* and *Siegmund* criterion derived in *Picard et al.* (2011).

The second strategy consists in iteratively estimating μ_i and the segmentation of factor ν: at step $(s+1)$, μ_i is estimated by:

$$\hat{\mu}_i^{(s+1)} = \frac{1}{p}\sum_{j=1}^{p} Y_{ij} - \hat{\nu}_{jh(i,j)}^{(s)} \quad , \tag{7}$$

where the segmentation of factor ν is updated using two-stage DP on $X_{ij} - \hat{\mu}_i^{(s+1)}$.

2.2.3. Correction and reconstitution of missing data

Once segmentation has been achieved, correction can be computed. Estimates $\hat{v}_{jh(i,j)}$ are used in the following way: let L_{jk_j} be the last level of series j, and $\hat{v}_{jk_j}$ the corresponding estimation of the station effect. Then, for every $X_{ij} \in L_{jh}$ $\left(1 \leq h \leq k_j + 1\right)$, corrected X_{ij} (denoted by X_{ij}^*) is given by:

$$X_{ij}^* = X_{ij} - \hat{v}_{jh(i,j)} + \hat{v}_{j,k_j+1} \,. \tag{8}$$

Note that the model allows the imputation of missing data and the correction of outliers. For any missing data or outlier (i,j), the imputation is naturally given by $\hat{X}_{ij} = \hat{\mu}_i + \hat{v}_{jh(i,j)}$. Since the two-factor model takes into account the change-points in the series, this allows an unbiased reconstitution of missing values, contrary to classical regression or interpolation methods.

2.3. Characterization of urban trends: UBRIS

UBRIS (urban bias remaining in series) procedure allows us to characterize artificial trends – in most cases related to urbanization, which are sometimes present in the climate series. UBRIS works jointly analyzing time series with potential artificial trends ("urban") and without potential artificial trends ("rural"). This is an improvement compared to traditional urban trend characterization, where rural and urban series are homogenized separately, before being compared (*Peterson*, 2003 for example). This requires a large set of both rural and urban series, which may be problematic on earlier periods for example.

UBRIS relies on an extension of model Eq. (6). Let us assume that the $j < m < p$ series are free of urban trends, and that for $m \leq j \leq p$, an additional trend may affect the series.

$$\begin{aligned} \mathrm{IE}(X_{ij}) &= \mu_i + v_{jh(i,j)} \,, \qquad \text{for} \quad 1 \leq j < m < p \,, \\ \mathrm{IE}(X_{ij}) &= \mu_i + v_{jh(i,j)} + \beta_j i \,, \quad for \quad m \leq j \leq p \,, \\ \mathrm{Var}(X) &= \sigma^2 I_{np} \,. \end{aligned} \tag{9}$$

Practically, UBRIS model is slightly more complicated than Eq. (9), since trend may not affect the whole period of the series. For computation, at least one series has to be free of trend, otherwise there is no unique solution when estimating climate factor μ and trend term β. Estimation is performed via ordinary least squares. Standard student t-test allows us to test significance of the trends (β_j). UBRIS ensures a posterior estimation of those additional trends.

Prior to UBRIS analysis, HOMER has to be run in order to detect abrupt changes.

UBRIS relies on the knowledge of climatologists who decide *a priori* which series may or may not be affected by urban trends. This human expertise is important. If series corrupted by artificial trends enter the "rural" group, they will bias the estimates of climate factor μ and trend term β.

2.4. ACMANT

ACMANT (adapted caussinus mestre algorithm for homogenizing networks of monthly temperature data, *Domonkos*, 2011) was developed from PRODIGE during the HOME period. However, in contrast with PRODIGE and HOMER, ACMANT is fully automatic and it applies reference series built from composites for time series comparisons. The other main novelties of ACMANT are i) it applies pre-homogenization in a way that the double use of the same spatial connection is excluded, ii) it coordinates the operations on different time scales (from multiannual to monthly) in a unique way.

2.4.1. ACMANT bivariate detection

Observed temperature data often have inhomogeneities with significant seasonal cycles in the resulted bias (*Drogue et al.*, 2005; *Brunet et al.*, 2011; etc.). Therefore, change-points are searched by fitting step-functions to two annual characteristics, i.e., to annual means (Y) and to the range of the seasonal cycle (R) in relative time series, that is, candidate series minus reference series. In HOMER, the reference series are the climate signals (μ coefficients in ANOVA model) or, with other words, the reference series for ACMANT detection are always pre-homogenized. Adapting notations of Section 2.1. to R series, ACMANT detection procedure aims at minimizing:

$$Q_{YR} = \sum_{j=1}^{k+1} \sum_{i=\tau_{j-1}+1}^{\tau_j} \left(Y_i - \bar{Y}_j\right)^2 + \frac{1}{2}\left(R_j - \bar{R}_j\right)^2 . \qquad (10)$$

The ½ factor in Eq. (10) was chosen empirically. Solutions with common timings of change-points on Y and R are considered only, so that the standard DP algorithm applies the cost function Q_{YR}. In order to set the number of changes, the C&L criterion is used both in original ACMANT and in its adaptation to HOMER:

$$C_0(Y,R) = 0 \quad \text{and}$$

$$C_k(Y,R) = \log\left[1 - \frac{\sum_{j=1}^{k+1} n_j\left[\left(\bar{Y}_j - \bar{Y}\right)^2 + \frac{1}{2}\left(\bar{R}_j - \bar{R}\right)^2\right]}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2 + \frac{1}{2}(R_i - \bar{R})^2}\right] + \frac{2k}{n-1}\ln(n) \quad . \quad (11)$$

The selection rule is: select k^* such that k^*=Argmin$_k$(C_k(Y)). In many cases, this procedure will allow us to detect changes hardly noticeable in annual means.

2.4.2. Month of change specification

Another feature of ACMANT that has been included in HOMER is its procedure for finding the most likely month of a change-point. If the precise month of the change is not known, since detection is mainly performed on annual indices, the default is to validate the break at the end of the year. At the end of the homogenization procedure, a more precise detection is made, using the monthly series serially (that is, the sequence of January, February, March, etc, for each year). Both candidate monthly series and reference series (computed from monthly μ factors) are deseasonalized; when analyzing change τ_j, standard DP algorithm is run on series of differences on interval $[\tau_{j-1}, \tau_{j+1}]$. Algorithm allows us to change the position of the change in a range of +/–2 years (in the original ACMANT the range is +/–12 months). Alternatively, the monthly precision can be determined by metadata. In HOMER, a flag marks whether a detected break is validated by metadata or not.

2.5. Discussion

The different methods contributing to the operation of HOMER have their own strengths and weaknesses. PRODIGE relies on a pairwise strategy for detection of the changes. A candidate series is compared to its neighbors in the same climatic area by computing series of differences. These difference series are then tested for discontinuities. On such a difference series without metadata, the detected changes may have been caused by the candidate or the neighbor. But, if a detected change-point remains constant throughout the set of comparisons of a candidate station with its neighbors, it can be attributed to this candidate station: this is called "attribution phase". There are two advantages in this approach. First, we avoid creating composite reference series averaging non-homogeneous series. Second, detection relies on an efficient univariate detection procedure whose level and power are well controlled. But, because of the randomness of the difference series, the change-points of weak amplitude will lead to less

accurate detection and sometimes no detection at all for some comparisons (in particular in the case of simultaneous breaks). At most cases, however, the induced ambiguity can be removed by considering the whole set of comparisons and using the metadata archives of the climate stations when available, as well as the knowledge of climatologists. This break-points detection phase has been considered the main drawback of PRODIGE, since it has to be performed manually, a process which may be tedious and time consuming, thus very difficult to apply to a large dataset and requiring a high level of regional climate knowledge and homogenization expertise.

To overcome the detection problem, an alternative approach is obtained by using the overall two-factor model, that allows the analysis and correction of a whole set of series (Section 2.2.). The *multiseg* (*cghseg* package) function determines the proper number of change-points using the MBIC criterion. This detection process with DP is quick and automatic. However model selection in a multivariate framework is a complex task, and the power of this procedure is sometimes lower than expected. In HOMER, function *multiseg* allows the automate attribution of the changes to a large extent, and in some cases the pairwise detection allows us to put into evidence changes that were not detected by *multiseg*.

ACMANT helps finding changes with a strong seasonal behavior in temperature series. In many cases, changes in observation conditions (location, sheltering, etc.) may have effects of opposing signs regarding the seasons, for example a positive effect in summer and a negative effect in winter. Such inhomogeneities are often hardly detectable on annual means, but clearly detectable with the ACMANT bivariate detection. A useful additional feature of ACMANT is the detection with monthly preciseness. The structure of HOMER has built in a way that it intends to exploit optimally the positive characteristics of the contributing methods. The tasks flow chart of HOMER is given in *Fig. 1*.

Detection is an iterative process. The initial detection phase usually reveals the most obvious changes which are corrected. Analyzing the result of this correction allows us to create an updated set of detected changes on a network. The joint detection is accompanied by the pairwise detection for allowing the use of metadata and for checking the results. The ACMANT detection follows the first cycle of detection and correction, since ACMANT detection needs pre-homogenized reference series. Note that correction is always performed on the initial data, simply by updating the set of the validated change-points before running ANOVA.

The process ends, whenever pairwise, joint-detection, and ACMANT bivariate detection find no additional changes on corrected series. In practice, the user may tolerate some pairwise comparisons still exhibiting unattributed isolated breaks, probably due to 1st kind errors.

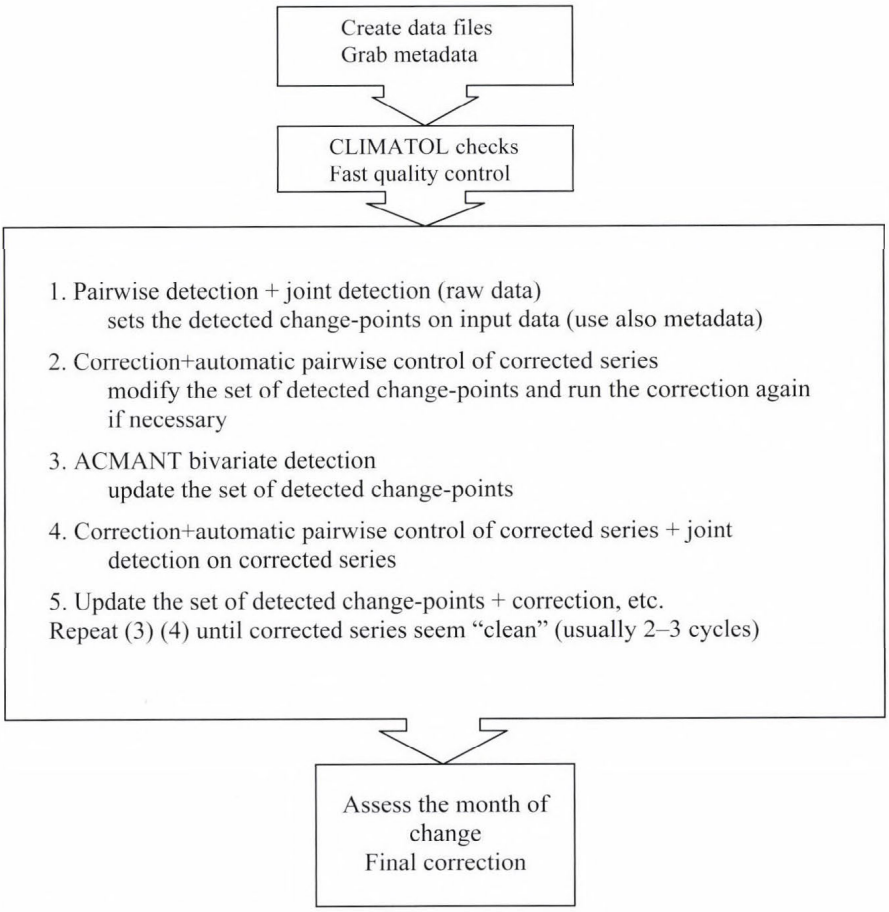

Fig. 1. Tasks flow chart of HOMER.

3. *Case study*

3.1. Homogenization using HOMER

A set of 13 series from Wien, Austria and its surroundings is provided by Zentralanstalt für Meteorologie und Geodynamik (ZAMG). Stations marked with (r) are considered rural: Fuchsenbigl$^{(r)}$, Gross-Enzersdorf$^{(r)}$, Klosterneuburg, Langenlebern$^{(r)}$, Schwechat, Wien-Innere-Stadt, Wien-Laaerberg, Wien-Mariabrunn$^{(r)}$, Rosenhügel, Rathauspark, Stadlau, Wien-Unterlaa, Wien-Hohe-Warte (*Fig. 2*).

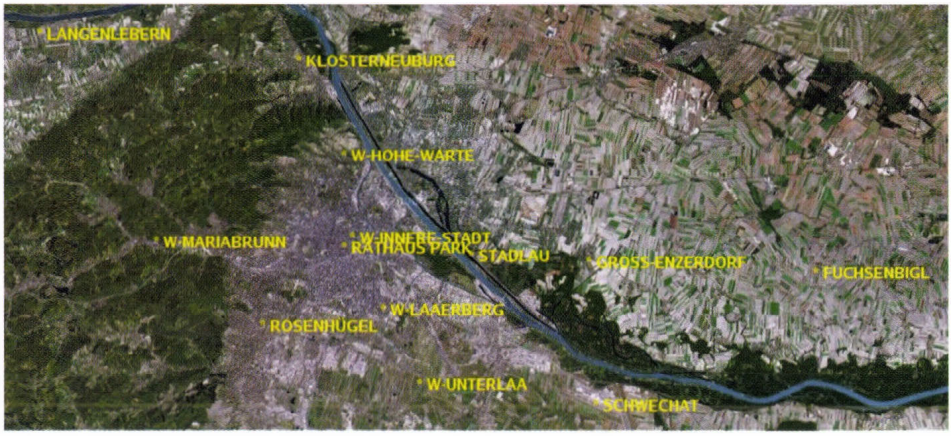

Fig. 2. Map of Wien series

Let us take Stadlau as an example: results of pairwise detection are given in *Fig. 3*. A quick examination of pairwise comparisons puts into evidence changes in 1969 or 1970, 1984, 2001 or 2002, and potential additional changes in 1953, and 1979.

The second step consists in running *cghseg* joint-detection (*multiseg* function). Combining pairwise and joint detection allows a quick attribution of the changes: 1954, 1969, 1984, and 2002 (*Fig. 4*). Note also the good agreement in the amplitudes of the changes detected in pairwise comparisons (triangles are black for breaks detected on pairwise annual series, blue for winter, and red for summer) and joint detection (green ⊕). However, the automatic joint-detection is not perfect. On Wien series, *multiseg* tends to detect a change around 1985-1987, which is not supported at all by pairwise comparisons, and thus, it is rejected manually by the user (large red cross in the same year). During estimation of μ and segmentation ν, *multiseg* iterative algorithm has wrongly attributed a climatic feature to the ν factor. Furthermore, the rather obvious change in 1979 (when considering pairwise comparisons) was not detected by *multiseg*. User has to validate it manually using the graphical user interface. When clicking on the window, the user adds red crosses to remove or validate breaks. The y axis is not important, only the date (x axis) is taken into account. Clicking on a date selected by *multiseg* (symbol ⊕ is present) removes the corresponding date, while clicking elsewhere validates a new change-point. Metadata allow us to validate changes in 1980 (relocation of the weather station) and 2002 (changes in instrumentation). There are also sufficient statistical clues to validate the other changes, even if metadata are lacking.

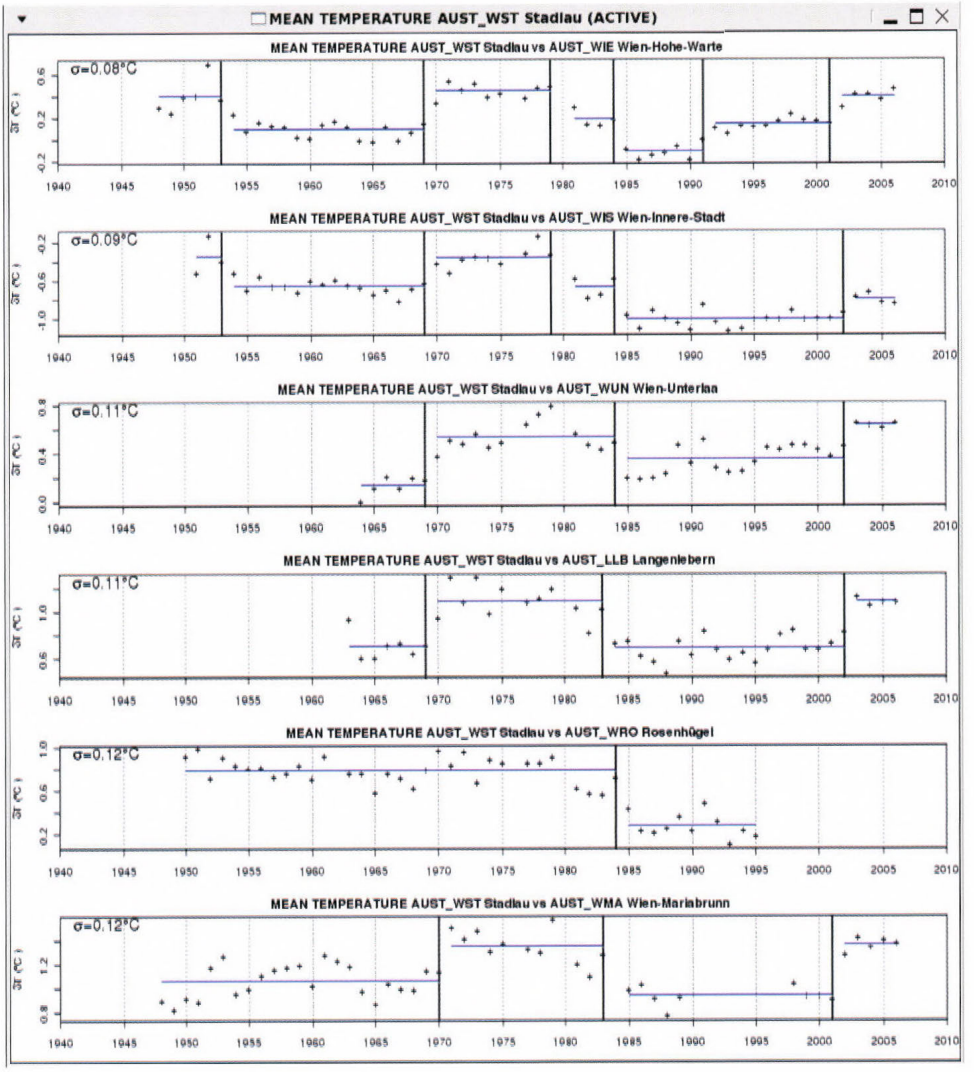

Fig. 3. Screen capture of HOMER outputs: Stadlau series compared to its neighbours. Pairwise comparison are sorted according to the increasing values of the noise standard deviation (upper left corner of each plot), computed using Eq. (5). For clarity reasons, only 6 comparisons with the smallest noise are shown.

After a correction step, ACMANT bivariate detection confirms the selected changes on Stadlau series (not shown). The raw and corrected Stadlau series after the final correction are shown in *Fig. 5* (upper panel for the raw, lower panel for the corrected series).

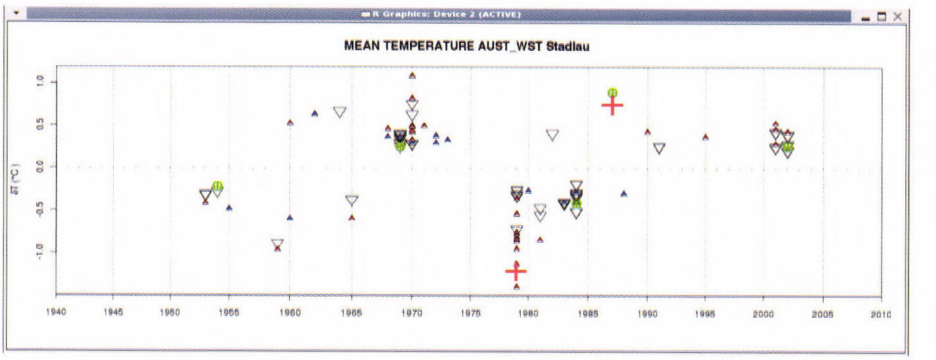

Fig. 4. Screen capture of HOMER outputs: date (x axis) and amplitude (y axis) of change-points detected on the whole set of pairwise comparisons: annual comparisons (black), winter (blue) and summer (red) triangles. Joint detection results are pointed as green ⊕ symbols. Red crosses mark user's interventions.

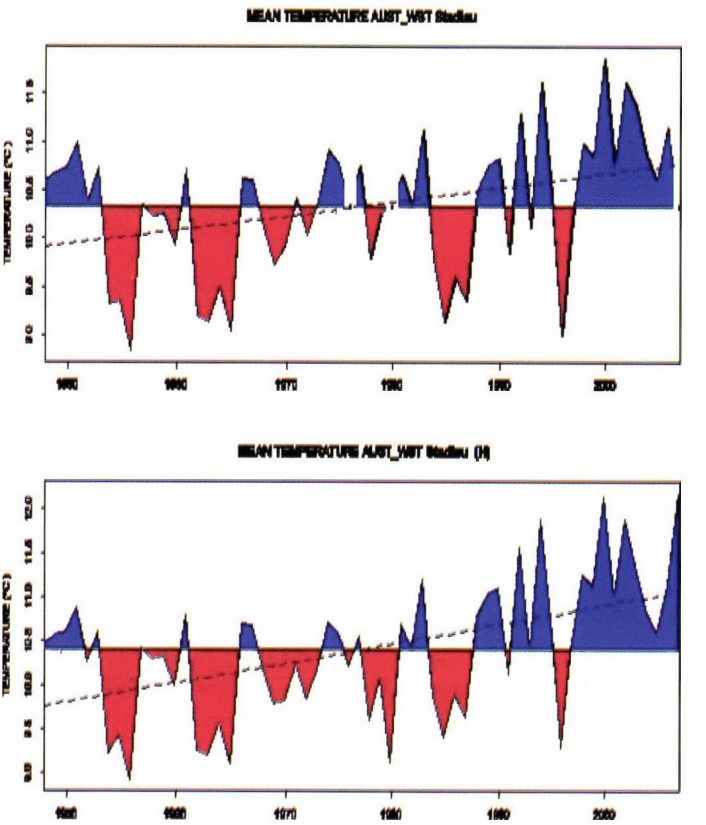

Fig. 5. Raw (up) and corrected (down) series of Stadlau.

Pairwise comparison of corrected series is characteristic of a good homogenization (*Fig. 6*).

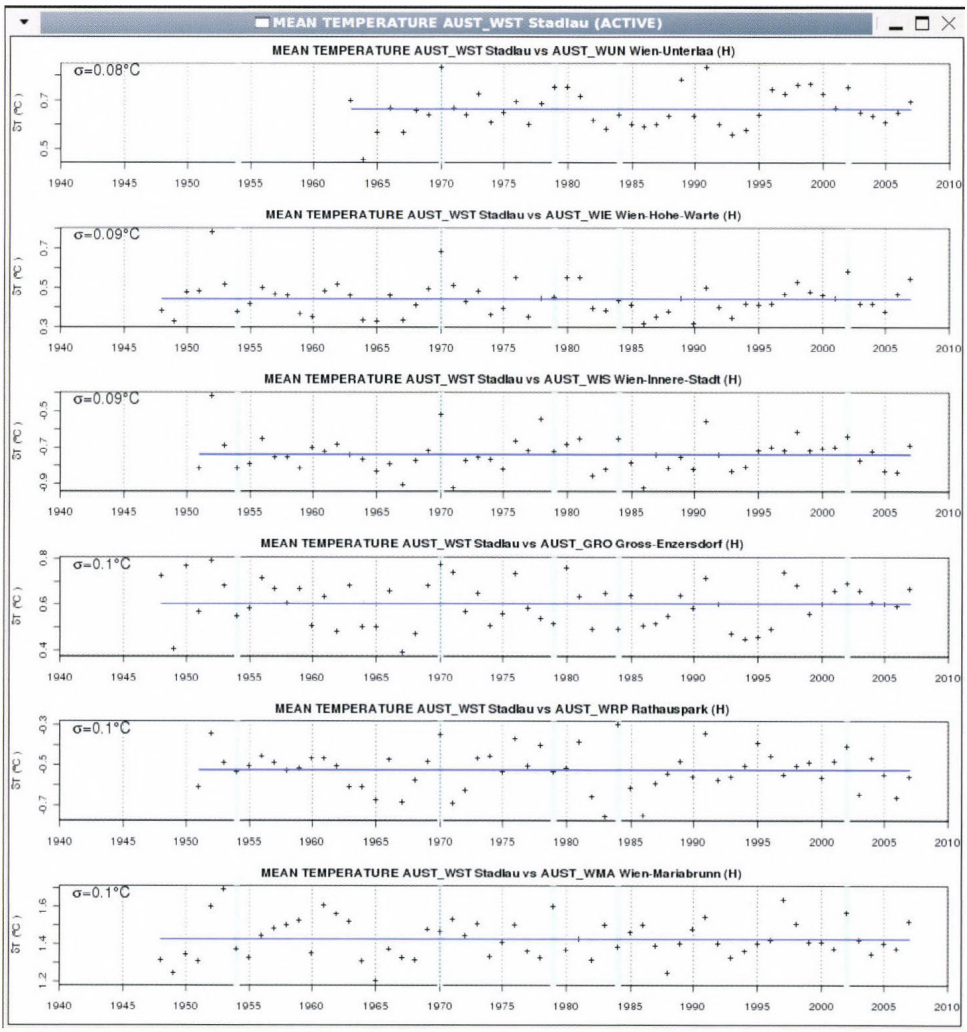

Fig. 6. The same as *Fig. 3*, but for the corrected Stadlau series compared to its corrected neighbors. The list of pairwise comparisons changed a little bit, since estimates of noise standard deviation slightly varied.

Another example of the effect of correction is shown for Rathauspark series (*Fig. 7* upper panel for the raw, lower panel for the corrected series).

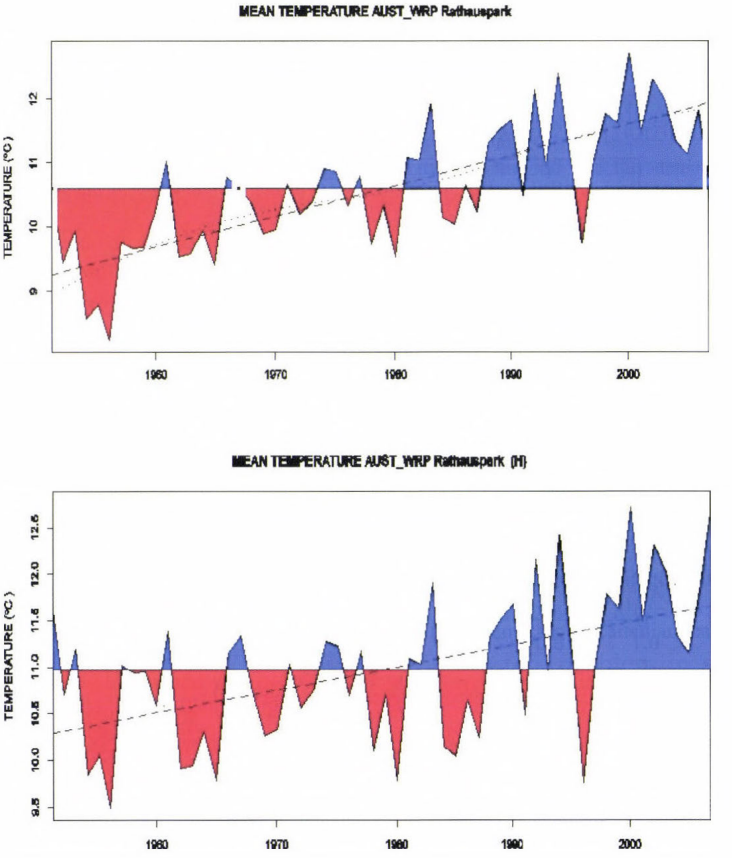

Fig. 7. Raw (up) and corrected (down) series of Rathauspark.

3.2. UBRIS characterization of urban trends

Running UBRIS allowed us to estimate jointly the effect of abrupt changes and the potentially significant urban trends on Wien series. UBRIS procedure is run in the following way: a first estimation allows us to put into evidence some urban series having no additional trend (large p values of the Student t-test for corresponding β). Those series are included into the rural set, and trends are re-estimated. At the end of the process, central temperature series exhibit no significant urban trends at level 0.05. Only suburban series (Wien Laaerberg, +0.10°C/decade, Rosenhügel +0.08°C/decade) exhibit significant positive trends (with student t-test p values lower than 10e–4). Corrected series of Laaerberg, with and without urban trend, is shown in *Fig. 8*.

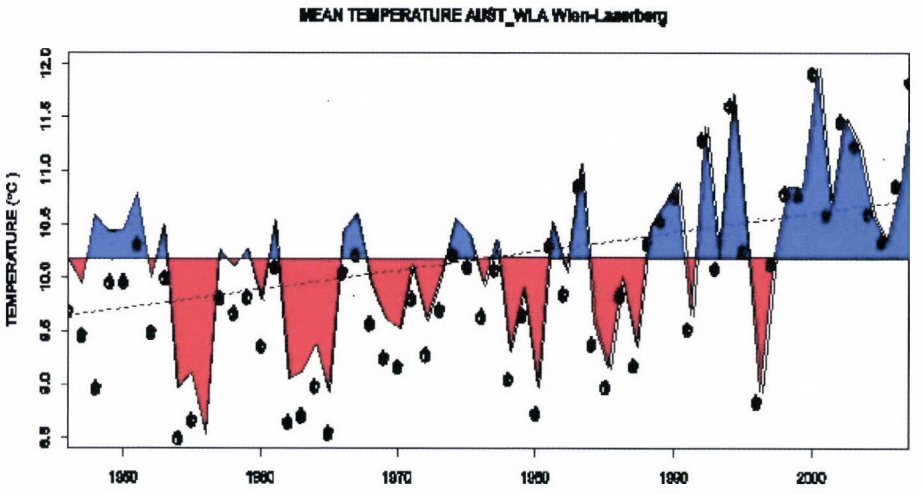

Fig. 8. Homogenized series of annual mean temperature of Laaerberg, with urban trend (series of ⊕ symbols) and removed urban trend(solid line).

These results are consistent with those obtained by *Böhm* (1998), who used a more traditional homogenization technique, and analyzed the trends of the series of differences of central series minus mean of the rural series. Note that those conclusions may not apply to other cities, since Wien population is remarkably stable since 1950 for example. UBRIS model should be run on each case study.

Additionally, Klosterneuburg series (not shown here) exhibits a remarkable feature, a highly significant decreasing trend for summer months (−0.02°C/decade). This site should be investigated for a potential shadowing effect.

4. Conclusion and perspectives

This paper presents a set of homogenization procedures integrated in the new software package HOMER (available at www.homogenization.org). This package was built relying on the results of the 4-year long COST-HOME project, so it implements the most significant findings achieved by its different working groups. The evolution of PRODIGE, combined with ACMANT and CLIMATOL procedures and supported by the R-package *cghseg* into HOMER provides a state-of-the-art homogenization tool for monthly to annual data

applicable to most essential climate variables. However, HOMER shall not be considered as an automatic method, since manual input is still required in order to control the homogenization process.

HOMER is recommended by the COST Action ES0601, together with *Craddock* (1979), MASH (*Szentimrey*, 2007), USHCN (*Menne* and *Williams*, 2005), ACMANT (2011) software that got valuable results during COST benchmark experiments (*Venema et al.*, 2012).

The addition of UBRIS procedures adds value to the package since artificial trends have remained a problematic issue in homogenization.

Further development planned in this work is using a generalized least squares estimation for the correction model, in order to take into account the spatial dependency of the residuals. Although this technique is expected to have a weak effect on the correction estimates themselves, it may provide more accurate confidence intervals. A Bayesian criterion for automatic attribution of changes detected in pairwise comparison is also in development.

Acknowledgements–HOMER has been developed with support of the European Union, through the COST Action ES0601 – Advances in Homogenization Methods of Climate Series: an Integrated Approach (HOME).

References

Aguilar, E., Auer, I., Brunet, M., Peterson, T.C. and *Wieringa, J.*, 2003: WMO Guidelines on climate metadata and homogenization. WCDMP-No. 53, WMO-TD No 1186, WMO, Geneva.

Auer, I., Böhm, R., Jurkovic', A., Orlik, A., Potzmann, R., Schöner, W., Ungersböck, M., Brunetti, M., Nanni, T., Maugeri, M., Briffa, K., Jones, P., Efthymiadis, d., Mestre, O., Moisselin, J.M., Begert, M., Bradzil, R., Bochnicek, O., Cegnar, T., Gagić-Ĉapka, M., Zaninović, K., Majstorović, Z., Szalai, S.,Szentimrey, T. and *Mercalli M.*, 2005: A new instrumental precipitation dataset for the greater Alpine region for the period 1800–2002. *Int. J. Climatol. 25*, 139–166.

Beaulieu, C., Seidou, O., Ouarda, T.B.M.J., Zhang, X., Boulet, G. and *Yagouti, A.*, 2008: Intercomparison of homogenization techniques for precipitation data. *Water Resour. Res. 44*, W02425, doi:10.1029/2006WR005615.

Birgé, L. and *Massart, P.*, 2001: Gaussian model selection. *J. Eur. Math. Soc. 3*, 203–268.

Böhm, R. 1998: Urban bias in temperature time series – a case study for the city of Vienna, Austria. *Climatic Change, 38*, 113–128.

Brunet, M., Asin, J., Sigró, J., Bañon, M., García, F., Aguilar, E., Palenzuela, J.E., Peterson, T.C., and *Jones, P.*, 2011: The minimization of the screen bias from ancientWestern Mediterranean air temperature records: an exploratory statistical analysis, *Int. J.Climatol. 31*, 1879–1895.

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

Caussinus, H. and *Mestre, O.*, 2004: Detection and correction of artificial shifts in climate series. *J. Roy. Stat. Soc. Series C53*, 405–425.

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

Domonkos, P., 2011: Adapted Caussinus-Mestre Algorithm for Networks of Temperature series (ACMANT). *Int. J. Geosci. 2*, 293–309.

Domonkos, P., Venema, V., Auer, I., Mestre, O., and *Brunetti, M.*, 2012a: The historical pathway towards more accurate homogenization. *Adv. Sci. Res. 8*, 45–52.

Domonkos, P., Venema, V., and *Mestre, O.*, 2012b: Efficiencies of homogenization methods: our present knowledge and its limitation. *Proceedings of the 7th Seminar for Homogenization and Quality Control in Climatological Databases* in press.

Drogue, G., Mestre, O., Hoffmann, L., Iffly, J-F., and *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.

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

Gu, C. and *Wang, J.*, 2003: Penalized likelihood density estimation: Direct cross-validation and scalable approximation. *Statistica Sinica 13*, 811–826.

Guijarro, J.A., 2011: User's guide to CLIMATOL. http://www.meteobal.com/climatol/climatol-guide.pdf

Hawkins, D.M., 1972: On the choice of segments in piecewise approximation. *J. Inst. Math. Appl. 9*, 250–256.

Hawkins, D.M., 2001: Fitting multiple change-points to data. *Comput. Statist. Data Anal. 37*, 323–341.

Lavielle, M., 1998: Optimal segmentation of random processes. *IEEE Trans. on Signal Proc. 46*, 1365–1373.

Menne, M. J. and *Williams, C.N.Jr.*, 2005: Detection of undocumented changepoints using multiple test statistics and composite reference series. *J. Climate 18*, 4271–4286.

Peterson, T.C., Easterling, D.R., Karl, T.R., Groisman, P., Nicholls, N., Plummer, N., Torok, S., Auer, I., Böhm, R., Gullett, D., Vincent, L., Heino, R., Tuomenvirta, H., Mestre, O., Szentimrey, T., Salinger, J., Førland, E. J., Hanssen-Bauer, I., Alexandersson, H., Jones, P., and *Parker, D.*, 1998: Homogeneity adjustments of in situ atmospheric climate data: a review. *Int. J. Climatol. 18*, 1493–1517.

Peterson, T.C., 2003: Assessment of Urban Versus Rural In Situ Surface Temperatures in the Contiguous United States: No Difference Found. *J. Climate 16*, 2941–2959.

Picard, F., Lebarbier, E., Hoebeke, M., Rigaill, G., Thiam, B. and *Robin, S.*, 2011: Joint segmentation, calling and normalization of multiple CGH profiles. *Biostatistics 12*, 413–428.

Schwartz, G., 1978: Estimating the dimension of a model. *Ann. Statist. 6*, 2745–2756.

Szentimrey, T., 2007: Manual of homogenization software MASHv3.02, Hungarian Meteorological Service.

Tibshirani, R., 1996: Regression shrinkage and selection via the lasso. *JRSS Series B 58*, 267–288.

Venema, V., Mestre, O., Aguilar, E., Auer, I., Guijarro, J.A., Domonkos, P., Vertacnik, G., Szentimrey, T., Štěpánek, P., Zahradnicek, P., Viarre, J., Müller-Westermeier, G., Lakatos, M., Williams, C.N., Menne, M., Lindau, R., Rasol, D., Rustemeier, E., Kolokythas, K., Marinova, T., Andresen, L., Acquaotta, F., Fratianni, S., Cheval, S., Klancar, M., Brunetti, M., Gruber, C., Duran, M.P., Likso, T., Esteban, P. and *Brandsma,T.*, 2012: Benchmarking monthly homogenization algorithms, *Climate of the Past 8*, 89–115.

Zhang, N.R. and *Siegmund, D.O.*, 2007: A Modified Bayes Information Criterion with Applications to the Analysis of Comparative Genomic Hybridization Data. *Biometrics 63*, 22–32.

IDŐJÁRÁS

Quarterly Journal of the Hungarian Meteorological Service
Vol. 117, No. 1, January–March 2013, pp. 69–90

Homogeneity of monthly air temperature in Portugal with HOMER and MASH

Luís Freitas[1]*, **Mário Gonzalez Pereira**[1,2], **Liliana Caramelo**[1], **Manuel Mendes**[3], and **Luís Filipe Nunes**[4]

[1] *Centre for Research and Technology of Agro-Environment and Biological Sciences (CITAB), University of Trás-os-Montes and Alto Douro, Apartado 1013, 5001-801 Vila Real, Portugal*

[2] *Instituto Dom Luiz – Universidade de Lisboa, Faculdade de Ciências da Universidade de Lisboa, Campo Grande, Edifício C8, Piso 3, 1749-016 Lisboa, Portugal*

[3] *Instituto de Meteorologia, IM, I.P., Rua C, Aeroporto de Lisboa, 1749-077 Lisboa, Portugal*

[4] *Laboratory for Systems, Instrumentation and Modeling in Science and Technology for Space and the Environment, Faculty of Sciences of University of Lisbon, Campo Grande, Edifício C1, Gabinetes 1.4.21 e 1.4.39, P-1749-016 Lisboa, Portugal*

**Corresponding author E-mail: pedro-fafe@hotmail.com*

(Manuscript received in final form November 25, 2012)

Abstract–In this paper we focus on the homogeneity of Portuguese monthly mean air temperature with two purposes: i) to detect and correct eventual inhomogeneities in the dataset; and, ii) to compare the homogenized time series with different methods. The dataset used in this study comprises time series of minimum (TN) and maximum (TX) monthly mean air temperature recorded in weather stations located in the northern region of the continental part of Portugal, from 1941 to 2010. MASH and HOMER were the methods used in this study to homogenize the Portuguese air temperature database. The former was selected for being one of the most widely used by the homogenization community, while the latter was selected because it is one of the most recent homogenization methods, and the combination of detection methods resulted in that, along with MASH, HOMER exhibited the best results in the comparative analysis performed within the COST Action ES0601 (HOME). A high number of break points were identified in both minimum and maximum air temperature time series, but differences in the number, size and temporal location of the breaks detected by both methods must be underlined. The homogenization process was assessed by comparing results obtained with correlation, trend, and principal component analysis using non-homogenized (NH) and homogenized datasets with both methods. Correlation analysis

reveals a higher increase in the similarity in homogenized TX than in TN in relation with NH time series. Decrease in the amplitude of the tendencies and in the number of statistically significant trends is higher in homogenized TX than in TN, independently of the homogenization method. On the other hand, the number of statistically significant principal components tend to decrease with the application of homogenization procedures, while the explained variance by the first principal components of homogenized datasets is tendentiously higher than for non-homogenized datasets.

Key-words: Homogenization, temperature, MASH, HOMER, Portugal.

1. Introduction

The existence of long and reliable instrumental climate records registered in a sufficiently dense network is fundamental to assess climate variability and climate change and to validate climate models. Climate research results are also dependent on the quality of the datasets, in particular on its homogeneity (*Venema et al.*, 2012). A homogeneous climate time series can be defined as the one whose variability is only caused by changes in weather and climate (*Aguilar et al.*, 2003). However, long instrumental records are rarely homogeneous because they include non-climatic signals which must be removed. Results from the homogenization of Western Europe climate records points to the existence of inhomogeneities in mean temperature series every 15 to 20 years (*Venema et al.*, 2012). In fact, any weather observation network, that operates for a long period of time, undergoes changes in its functioning due, for example, to instrumentation failure or damage, changes on its surrounding (e.g., urbanization), relocation and substitution of weather stations. For these reasons, it is expected that the Portuguese maximum and minimum air temperature datasets present heterogeneities that need to be detected and corrected.

In the last decades, inhomogeneity detection techniques have been developed based on classical statistical tests (*Alexandersson*, 1986; *Gullett et al.*, 1990), regression models (*Vincent*, 1998), or Bayesian approaches (*Perreault et al.*, 2000). More recently, new procedures were particularly developed to detect and correct multiple change-points using reference series (*Szentimrey*, 1999; *Mestre*, 1999; *Caussinus* and *Mestre*, 2004; *Menne* and *Williams*, 2005) and changes in the mean and variance (*Toreti et al.*, 2012). Review papers and comparison studies of homogenization methods have been published regularly (*Peterson et al.*, 1998; *Ducré-Robitaille*, 2003, *Reeves et al.*, 2007, *Venema et al.*, 2012). Some authors have been focusing their interest in specific aspects of the homogenization procedure such as the cause of inhomegeneities (*Trewin*, 2010), use of reference series (*Menne* and *Willians*, 2005, *Domonkos*, *et al.*, 2012), ability of homogenization methods (*Menne* and *Willians*, 2005), or to test automatic homogenization methods by the introduction of perturbed parameter experiments (*Williams et al.*, 2012).

The inventory and evaluation of existing detection and correction methods and the need of an objective comparative analysis to assess their performance was included in the scientific programme of the COST Action HOME ES0601: Advances in Homogenization Methods of Climate Series: an integrated approach (HOME). HOME results include the publication of a comparison study, based on 25 blind contributions and 22 contributions made after knowing the location and size of the heterogeneities, performed with a large number of different versions of 9 main methods (*Venema et al.*, 2012). This study was based on a benchmark dataset of monthly air temperature and precipitation and on different error metrics to assess the performance of the methods. Results of this comparison suggests that: (i) the assessment of the methods is dependent on the error metric considered; (ii) in general, all relative methods contribute to homogenized temperature data; but, (iii) only the methods with best performance are able to improve the quality of precipitation datasets; and, (iv) the list of methods with better performance includes Craddock (*Craddock*, 1979), PRODIGE (*Caussinus* and *Mestre*, 2004), MASH (*Szentimerey*, 2007), ACMANT (*Domonkos*, 2011), and USHCN methods (*Menne* and *Williams*, 2009).

HOME main objective was to develop a general homogenization method for homogenizing climate and environmental datasets which was accomplished in 2011 with the release of a free software package (HOMER), implemented in R language (*HOME*, 2011). It should be noted that ACMANT is a modified and automated version of PRODIGE, and that HOMER integrates PRODIGE, ACMANT, and USHCN.

Consequently, the purpose of this study is twofold: (i) to analyze the homogeneity of minimum and maximum air temperatures in northern Portugal; and, (ii) to compare the homogenized maximum and minimum air temperatures Portuguese datasets with HOMER and MASH. A review of the main characteristics of the procedures used to control the quality of the data and methods of homogenization will be undertaken in order to justify the options taken in this study and to highlight the methodological differences between MASH and HOMER.

2. Dataset description

The dataset that we analyze here is representative of the monthly mean maximum and minimum air temperature fields (hereafter TX and TN, respectively) in the northern region of the continental part of Portugal for the 1941–2010 period. Monthly time series were calculated from daily values, following the WMO directives in what concerns to the existence of missing values in daily time series. Specifically, a monthly value should only be computed if no more than five consecutive daily values or less than ten daily values throughout the month are missing (*WMO*, 2011).

Daily values of TX and TN were recorded at weather stations managed by the Portuguese Meteorological Institute (IM). Location and characteristics of these weather stations are presented in *Fig. 1* and *Table 1*, respectively. This network comprises both classical weather stations (CWS), collecting data since the mid-1800s, and automatic weather stations (AWS), installed in the end of the 20th century. In cases where AWS were installed in approximately the same location of the CWS, the time series from both weather stations were merged, the type of station in *Table 1* was set to CWS/AWS, and the date of the fusion was stored as metadata. Maximum distance between an AWS and CWS used to produce the merged time series was 4.7 km (in Vila Real), which is a much lower distance than those used in previous studies (*Stepanek* and *Mikulova*, 2008; *Vicente-Serrano et al.*, 2010).

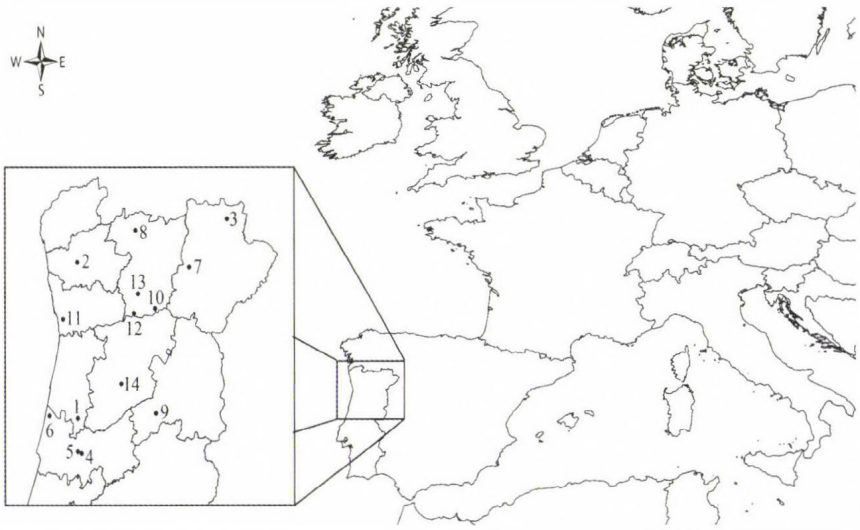

Fig. 1. Location of the weather stations of the Portuguese Institute of Meteorology (IM) network, in northern Portugal. Addition characteristics of these stations are provided in *Table 1*.

In this network, weather stations are well distributed and located both in low and high altitude (ranging from 14 m to 1380 m), in densely populous coastal areas and sparsely populated inner regions within the country territory (*Fig. 1*). The northern Portugal is characterized for being the region with the highest density of mountains and river basins in the country as well as by a diverse land use/occupation (*Freitas et al.*, 2012). Independently of the

proximity to the Atlantic Ocean or the altitude, all weather stations considered in this study are located in a region of C_s type of climate, which is a temperate climate with dry period in summer (*AEMET-IM*, 2011). In more detail, the climate of this northern region is essentially of C_{sb} type, which corresponds to a temperate climate with dry or temperate summer, except a small part, in the northeast, which is of type C_{sa}, also temperate but with dry or hot summer. The recently published Iberian Climate Atlas (*AEMET-IM*, 2011) provides a brief history of the complete IM network and additional description and characteristics of the temperature dataset. Results of the exploratory preliminary statistical analysis of minimum and maximum air temperature datasets for the 1941–2010 period are presented and discussed in *Freitas et al.* (2012).

Table 1. Characteristics of the weather stations of the Portuguese Institute of Meteorology (IM) network, located in northern Portugal including: identification code (ID); stations name; station type; altitude (m); start and ending dates; and, amount of missing values (in %), accounted for the 1941–2010 period. When the entire time series results from measurements from a CWS (or AWS), the type is simply CWS (or AWS); in the cases where a CWS was replaced by a AWS, the type is CWS/AWS

ID	Station Name	Type	Altitude (m)	Start year	End year
1	Anadia (AN)	AWS	45	1941	2010
2	Braga (BR)	CWS/AWS	65	1931	2010
3	Bragança (BG)	CWS	690	1932	2010
4	Coimbra B. (CB)	CWS	35	1941	2010
5	Coimbra G. (CG)	CWS	141	1864	1996
6	Dunas Mira (DM)	CWS	14	1935	2005
7	Mirandela (MI)	CWS/AWS	250	1926	2010
8	Montalegre (MO)	CWS/AWS	1050	1880	2010
9	Penhas D. (PD)	CWS/AWS	1380	1932	2010
10	Pinhão (PI)	CWS/AWS	130	1941	2010
11	Porto S.P. (PS)	CWS	93	1863	2005
12	Régua (RE)	CWS	56	1933	2010
13	Vila Real (VR)	CWS/AWS	561	1928	2010
14	Viseu (VI)	CWS/AWS	443	1925	2010

3. *Methodological procedures*

This section is devoted to the description of the methods used to perform the quality control of the data, the homogeneity analysis and to compare the homogenized datasets with those methods. The quality control of the datasets comprises a preliminary exploratory statistical analysis to characterize the potential and limitations of the datasets as well as to identify and correct missing values and outliers. Main technical features of the procedures used in this study will be briefly discussed to validate the followed methodology and to underline the major differences between the approaches of the two selected methods to homogenize the Portuguese air temperature dataset.

3.1. Quality control with homogenization methods packages

In this study two homogenization methods were used: (i) the most recent version of MASH, (Version MASHv3.03), initially developed in the Hungarian Meteorological Service by *Szentimrey* (1994, 1999); and, (ii) HOMER, developed in the framework of COST Action ES0601 (*HOME*, 2011). We start with presenting the homogenization methods because, in addition to being able to detect and correct inhomogenieties, these softwares comprise additional functions to perform fast quality control. On this subject, with MASH it is obligatory to use available functionalities to fill the missing values and perform automatic correction of outliers. On the other hand, HOMER provides a fast quality control of the data, which includes functions of the CLIMATOL R package (*Guijarro*, 2011), which allow the user to perform/estimate station density, correlogram, histograms, boxplots, and cluster analysis. With respect to the detection of heterogeneities, MASH relies on multiple references series while HOMER combines three detection algorithms: pairwise – univariate detection (*Caussinus* and *Lyazrhi*, 1997), joint detection (*Picard et al.*, 2011), and ACMANT – bivariate detection (*Domonkos et al.*, 2012). To correct the datasets, MASH uses multiple comparison techniques whereas HOMER uses ANOVA. MASH is provided with a user guide, while a brief description of HOMER can be found in *Mestre* and *Aguilar* (2011) or in *Freitas et al.* (2012).

3.2. Outlier detection

It is recommended to use different methods for outlier detection because, in general, one single method/criteria is not sufficient to identify real outliers nor to exclude false detections (*Stepanek et al.*, 2009). Consequently, in this study, abnormal high and low values were only classified as outliers if two criteria were simultaneously verified: (i) values above/below the upper/lower thresholds defined as the upper/lower quartiles plus/minus the interquartile range times a coefficient (usually equal to 1.5 to detect outliers and equal to 3.0 to detect extreme values); and, (ii) pairwise comparison which is based on the difference

time series between candidate and best neighbor time series, which can be defined as the closer stations and/or those presenting higher correlation (*Stepanek et al.*, 2009; *Syrakova* and *Stefanova*, 2009). This latter procedure can be performed in HOMER by visual inspection of the plots of the difference between candidate and best neighbor time series. As mentioned in the previous section, in addition to this analysis, MASH has an independent and automatic procedure to detect and correct outliers that is executed before detection procedures.

3.3. Missing values correction

The existence of missing data in climate time series can be solved with temporal interpolation, using data of the same time series before and after the data gap, or with spatial interpolation, using data from nearby weather stations (*WMO*, 2011). Complex estimation methods, such as weighted averages, spline functions, linear regression, and kriging, which take into account the correlations with other elements, can also be used to complete the time series. *Brunetti et al.*, (2006) adopted a procedure to fill the gaps on monthly precipitation and temperature Italian time series, with estimates based on the highest correlated reference series. For temperature, this method is based on the differences between incomplete and reference temperature series. *Staudt et al.* (2007), replace the missing values on monthly time series of Spanish minimum and maximum temperatures by weighted means of the best-correlated synchronous data. The method used by *Syrakova* and *Stefanova* (2009) to fill the gaps in Bulgarian monthly temperature is based on the stability of the differences between the time series at neighboring highly correlated stations. More recently, *Vicente-Serrano et al.* (2010) tested three different procedures to fill missing data in daily precipitation time series: (i) the nearest neighbor, (ii) inverse distance weighted interpolation; and, (iii) linear regression methods, concluding that the nearest-neighbor method provided the best results. Both homogenization methods used in this study (MASH and HOMER) have corrected databases as final result with respect to inhomogeneities and missing values using multiple comparison and ANOVA, respectively.

3.4. Reference time series

Reference series or reference sections are used in detection procedures in many homogenization methods, such as ACMANT, AnClim/ProClimDB, Climatol, RHTestV3, and MASH (*WMO*, 2011). Reference series are also used to assess the quality of the homogenization (*Kuglitsch et al.*, 2009). These reference series do not need to be homogeneous (*Szentimrey*, 1999; *Zhang et al.*, 2001; *Causinus* and *Mestre*, 2004), but must encompass the same climatic signal as the candidate series (*Della-Marta* and *Wanner*, 2006) and, in this sense, are usually produced as weighted averages of the time series from surrounding stations

(*Peterson* and *Easterling*, 1994; *Sahin* and *Cigizoglu*, 2010). *Stepanek* and *Mikulova* (2008) discuss the advantages and disadvantages producing weighted reference series based on the distance between stations or on the correlation between candidate and potential time series, while *Della-Marta* and *Wanner* (2006) argue about the benefits of using weighted reference series in comparison with a single reference station. The selection procedure of the surrounding stations to produce the reference series can be based on the distance between stations or on the correlation between candidate and potential time series. Both criteria present advantages and disadvantages that must be underlined. Distance-based methods preserve the geographical vicinity, but time series from near stations with different climatic signals (e.g., due to altitude) can be selected. Using high correlated neighbor time series, both the candidate and reference series present similar variability (which reduces differences/ratios time series variability), but stations affected with similar/coincident inhomogeneities with the candidate can be selected (*Stepanek* and *Mikulova*, 2008). Weighted reference series are considered more representative of the climatic region and, for being less prone to potential inhomogeneities in the neighbor series than single reference station, are more characteristic of the climate variability at smaller scale (*Della-Marta* and *Wanner*, 2006).

In this study, reference time series are used in the detection procedure, because this is the methodology adopted in MASH and ACMANT, and to assess the quality of the homogenized time series. For the reasons presented before, weighted reference series were produced with AnClim software (*Stepanek*, 2008) using difference series to evaluate the correlation coefficients as suggested in *Alexandersson* and *Molberg* (1997), *Peterson et al.* (1998), *Stepanek* and *Mikulova* (2008), and *Domonkos et al.* (2012). Since our database is affected by only a few number of missing values and the objective is to assess the quality of the homogenization process not of the data completion process, reference series were produced to present the same data gaps than the uncorrected time series. This is achieved by using uncorrected time series (with the data gaps) and neighbor time series without missing values (in order to exclude neighbor time series missing value in the reference series).

3.5. Homogenization methods performance assessment

In contrast to comparative studies performed with synthetic databases, when type, size, and location of inhomogeneities are known a priori (as in *Venema et al.*, 2012), the homogenization methods performance assessment must be executed with real data, by comparing the results obtained with different techniques using non-homogenized (hereafter NH) and homogenized data with MASH (hereafter HM) and HOMER (hereafter HH). This section is devoted to present the methodology used to assess the quality of the corrected dataset and, consequently, methods used in the homogenization process.

i. Correlation analysis

The main objective of correlation analysis is to evaluate the strength of the temporal linear relationship through the computation of the Spearman correlation coefficient, SCC (*Pereira et al.*, 2011). In this sense, to assess potential improvement in the similarity between time series before and after the homogenization process, correlation analysis was applied to annual time series to compute: (i) the correlation matrix between time series of non-homogenized and homogenized time series with MASH and with HOMER datasets; and, (ii) the SCC between each candidate and corresponding reference series. Since our objective is to assess the quality of the homogenization process, and not of the interpolation procedures used in MASH and HOMER to fill the data gaps, SCC was computed between time series with the same missing values than in NH datasets.

ii. Trend analysis

The existence of trends is in the basis of climate change studies (*Raj* and *Azeez*, 2012). In this study, the Mann-Kendal non-parametric test is used to estimate the existence, magnitude and statistical significance of potential trends in the NH, HM, and HH time series, in order to assess the impacts of homogenization methods. This test is suggested for trend analysis by the WMO (*Sneyers*, 1990) and has been used in many published works on climate change and climate variability (e.g., *Moberg* and *Jones*, 2004; *Brunetti et al.*, 2006; *Rodrigo* and *Trigo*, 2007).

iii. Principal component analysis (PCA)

When PCA is applied on a dataset, a new set of time series is produced as linear combination of the original ones. The new time series are the so-called principal components (PC), while the coefficients used to compute them are the elements of the empirical orthogonal functions (EOF). From the mathematical point of view, EOFs are the eigenvectors of the variance-covariance or the correlation matrix of the original dataset, the PCs are obtained by projecting the original time series into the EOF, and the eigenvalues are a measure of the explained variance, i.e., the proportion of the total variance explained by each PC. Obtained PCs are uncorrelated and sorted by decreasing order of variance, while EOFs are orthogonal to each other and constitute a vector base. There are different versions of this multivariate statistical technique, but it is easy to find their description/characteristics (*Jolliffe*, 2005; *Wilks*, 2011). PCA has multidisciplinary applications and is used in data analysis as an exploratory tool (for outlier detection, cluster identification, data visual examination, and interpretation), data preprocessing (dimensionality and noise reduction), modeling, and to identify spatial and temporal patterns and modes of variability such as NAO and ENSO (*Wold et al.*, 1987; *Jolliffe*, 2005; *Pozo-Vazquez et al.*,

2005). PCA results are dependent on the scaling of the original matrix (*Wold et al.*, 1987; *Jolliffe*, 2005), but statistical significance can be assessed, e.g., with cross-validation, bootstrap, or jackknifing techniques (*Romanazzi*, 1993; *Jolliffe*, 2005). PCA outputs, in particular the amount of explained variance by each PC, are dependent on the similarity of the time series (*Jolliffe*, 2005). This characteristic of PCA will be used in this study to assess homogenization results.

4. Obtained results

Preliminary exploratory statistical analysis reveals the existence of a very small number of missing values. Time series most affected by this problem present multiple consecutive missing values or their last record (end date) is before 2010. Results for maximum temperature are very similar to that for minimum temperature. The great majority of the low number of outliers detected above and below the defined thresholds based on the quartiles of their own time series was not confirmed with pairwise comparison with neighboring time series. The final number of outliers considered in HOMER for minimum and maximum temperatures were 10 and 11, respectively, which corresponds to 0.1% of total number of monthly values in each dataset or to less than 1 missing values per time series in each dataset. As mentioned in Section 3.2, MASH has an automatic procedure to detect and correct outliers which is not controlled by the user.

Temporal location and size of the breaks detected in minimum and maximum air temperature time series with MASH and HOMER are shown in *Table 2*. It should be pointed out that breaks marked with a star (*), noticeable only in the detection list of MASH, correspond to shifts of equal value but opposite sign in two consecutive years, that will most likely be an annual outlier than a break point and, from this point forward, will not be considered as breaks. Consequently, the number of breaks detected with HOMER (39 in TN and 32 in TX) is higher than with MASH (32 in TN and 24 in TX). Since the original data only have one significant decimal digit, the physical meaning of a great number of these breaks can be questioned. The number of shifts smaller than 0.1ºC detected with MASH is much higher (12 breaks in TN and 19 in TX) than with HOMER (5 breaks in TN and 1 in TX). On the other hand, the number of coincident breaks detected in TN with both methods is 18 (which corresponds to 56% and 46% of total number of breaks detected with MASH and HOMER, respectively) and 9 in TX (37% of MASH and 28% of HOMER total breaks, respectively). If the analysis is restricted to breaks with shifts greater or equal to 0.1ºC, the number of coincident breaks in TN is 14 (which corresponds to 70% and 41% of total number of breaks detected with MASH and HOMER, respectively) and 5 in TX (100% of MASH and 16% of HOMER total number of detected breaks, respectively). These results suggest that MASH could be able to detect smaller shifts but an overall small number of break points.

Table 2. Location (and magnitude) of break points detected on minimum and maximum air temperature during 1941-2010 period, with MASH and HOMER. Coincident detections with both methods, defined with utmost 18 months apart are presented in bold. Detections in two consecutive years with symmetrical shifts are marked with a star (*)

ID	Minimum air temperature (TN)		Maximum air temperature (TX)	
	MASH	HOMER	MASH	HOMER
1	**1963**(0.11), 1996(0.12)	1944(-0.08), 1950(0.04), **1964**(-0.45), 1970(0.49), 1984(0.29)	1944*(0.12)	1977(-0.22)
2	-	1963(-0.37), 1987(0.24) 1992 (0.53)	**1949**(0.14)	**1950**(-0.76), 1959(-0.49) 1971(-0.27), 1981(0.23)
3	1947*(0.17), 1972(0.08), **1980**(0.14)	1962(0.27), **1980**(-0.49)	1962(-0.05), 1969(-0.14), 1977(-0.03)	1965 (0.27) 1972(0.39) 1993 (0.45)
4	**1979**(-0.31)	1966(-0,18), **1979**(0,78)	1943*(0.46), 1949(-0.03), 1961(0.06), 1963(0.01), 1988*(-0.03), 2000*(0.14)	1953(-0.26), 1992(0.42)
5	**1982**(-0.04)	1950(-0.41), 1967(0.19) **1982**(0.14)	1969(-0.06), **1971**(0.15)	1949(-0,35), **1971**(-0,51)
6	1965(0.15), **1971**(0.15), 1976 (0.09), **1979**(0.26), 1985*(0.39), **1993**(0.05), 1996 (-0.07)	**1969**(-0.75), **1980**(-1.21), 1987(1.26), **1994** (-0.20)	1949*(0.09), 1986*(-0.16)	-
7	1951(-0.28), **1966**(0.21)	**1967**(-0.60), 1989(0.31), 1998(-1.54)	-	-
8	-	1950(-0.67)	**1953**(-0.14), 1976(-0.02), 1979(-0.04), 1994*(-0.18), 1996*(-0.16), 1998*(0.08)	**1951**(0.41), 1974(0.25) 19992(0.35)
9	1963*(-0.128)	-	-	1972(0.12), 1988(0.40)
10	**2003**(-0.54), 2007(0.16)	1958(-0.33) **2004**(1.11)	**1953**(0.04), 1965*(0.06), 1977*(0.12), 1995(0.06), **1998**(-0.27), 2000(-0.12), 2003(0.09)	**1951**(-0.01), 1974(-0.59), 1991(-0.39), **1996**(1.14)
11	-	1986(0.12), 1990(0.32)	-	1951(0.40), 1955(0.19), 1973(-0.42), 1990(0.41)
12	1968(0.11), **1977**(-0.05), 1980(-0.17), **1984**(-0.12), **1986***(-0.12), 1996(-0.03)	**1978**(0.56), **1984**(-0.03), **1987**(0.80), 2000 (0.28)	-	1995 (0,52)
13	**1943***(-0.28), 1948(-0.06), 1966 (0.07), **1974**(0.12), **1986**(-0.09)	**1944**(0.81), 1946(0.45), **1973**(-0.49), **1986** (0.04), 1993 (0.04)	**1954**(0.09), 2000(0.15)	**1953**(-1.22), 1959(0.25), 1991(-0.41)
14	1955(-0.21), **1958**(-0.42), 1969 (0.11), **1982**(-0.58), **1994**(0.80),1996(-0.08), 1999(-0.08)	**1957**(0.88), **1982**(0.27), **1994**(0.90)	1950(0.05), **1978**(-0.08), **1981**(-0.06), 1992*(0.26), **1994** (0.45)	**1977**(0.29), **1982**(0.35), **1994**(-1.70)

Correlation matrices between non-homogenized time series of maximum (minimum) air temperature, TXNH (TNNH), as well as between homogenized time series with MASH, TXHM (TNHM) and with HOMER, TXHH (TNHH) were computed. Boxplots of the Spearman correlation coefficient (SCC) values obtained for homogenized time series with HOMER are higher having lower dispersion in relation to non-homogenized and homogenized with MASH (*Fig. 2*). In general, SCC values between homogenized TX and TN time series is higher than those obtained between non-homogenized times series. Median value of the difference between TXHH and TXNH correlation matrix is higher (0.09, which corresponds to an increase of 9%) than the difference between TXHM and TXNH correlation matrix (0.02, which corresponds to a general increase of 2%). For minimum temperature, the median of the difference between TNHH and TNNH is equal to 0.13, while between TNHM and TNNH it is equal to 0.05.

Spearman correlation coefficient values obtained between reference series and non-homogenized and homogenized with MASH and HOMER corresponding time series (*Fig. 3*) reveals: (i) higher SCC values between reference and homogenized time series with MASH in every stations and for both TX and TN than between reference and non-homogenized time series; (ii) higher SCC values between reference and homogenized time series with HOMER for TX than between reference and non-homogenized time series but lower values for TN in 6 weather stations. Median of the SCC values obtained for maximum and minimum air temperature homogenized time series with HOMER are similar (94.3% and 89.3%) to those obtained with MASH (91.8% and 88.8%) but higher than for non-homogenized time series (88.2% and 86.4%), in particular for maximum air temperature. At this respect, the increase in the SCC can be underlined computed between the reference and one of the corresponding series: (i) TXHM and TXHH time series in Vila Real and Viseu (of 14.7% and 13.0%, respectively); and, (ii) TNHH and TNHM time series in Vila Real (of 13.7% and 8.4%, respectively).

Trend analysis for TN performed with Mann-Kendal test assuming a statistical significance level of 99% (Table 3) reveals that: (i) only a small number of non-homogenized times series presents statistically significant trends (5 in TNNH and TNHM datasets and only 1 in TNHH dataset); (ii) almost all time series present positive trends except Mirandela and Dunas de Mira; (iii) a reduction in the number of statistical significant trends is only verified for TNHH dataset; and, (iv) with the homogenization procedures, the trend of two time series, after being homogenized, became statistically significant (time series of Bragança, with MASH and of Montalegre with HOMER). Results obtained for TX shows that: (i) there is a lower number of statistically significant trends (2 in TXNH and only 1 in TXHM); (ii) the number of non-homogenized and homogenized time series with negative and positive trends are similar; but, (iii) all statistically significant trends are positive; and, (iv)

homogenization procedures lead the loss of statistical significance of the trends in one homogenized time series with MASH and in two homogenized time series with HOMER.

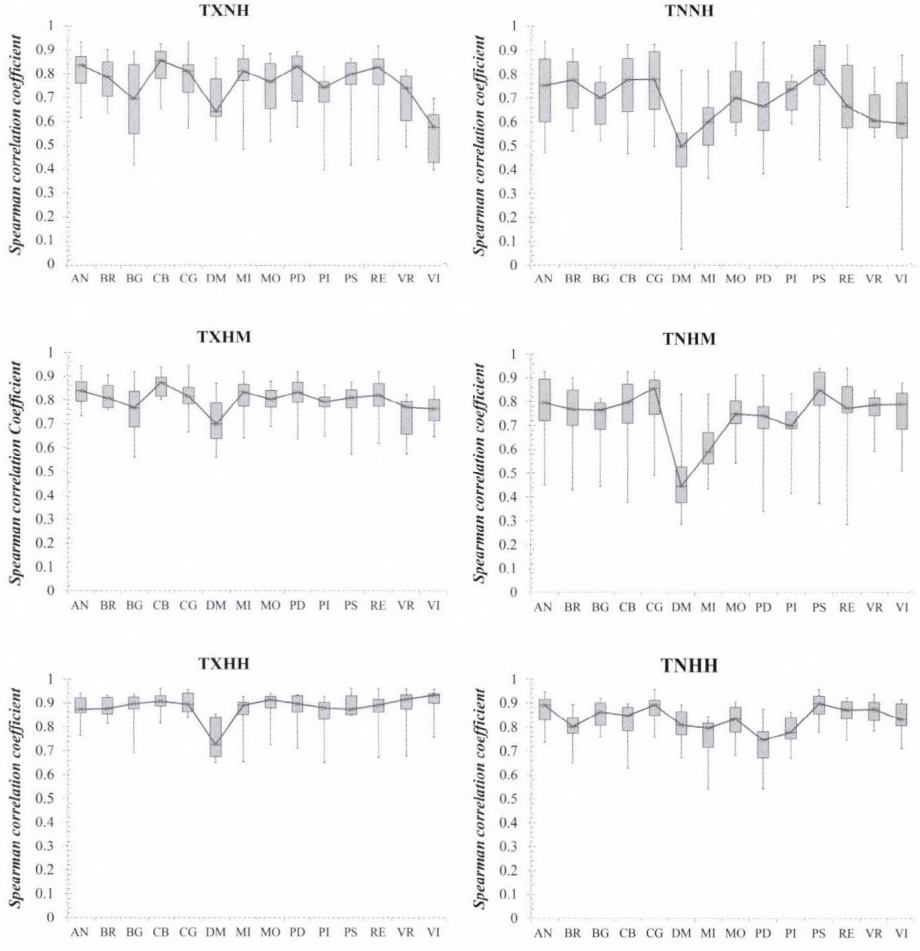

Fig. 2. Boxplot of Spearman correlation coefficient (SCC) between annual time series of non-homogenized (NH), homogenized with MASH (HM) and with HOMER (HH) maximum (top panel) and minimum air temperatures (bottom panel), from weather stations located in northern part of the continental Portugal (*Table 1* and *Fig. 1*), for 1941–2010 period. SCC was evaluated taking into account missing values of NH time series. The bottom/top indicates the lower/upper quartiles, and the band near the middle of the box is the median. The lower/upper end of the whiskers represents the minimum/maximum values.

Results obtained with PCA performed on non-homogenized and homogenized datasets (Table 3) can be summarized as follows: (i) only a small number of PCs are statistical significances (1 PC for TXHM, TXHH, and TNHH and 2 PCs for TNNH, TXNH, and TNHM); (ii) the explained variance by the first PC of homogenized datasets is greater than the explained variance by the first PC of non-homogenized ones; (iii) explained variance of first PC are higher for homogenized datasets with HOMER than with MASH.

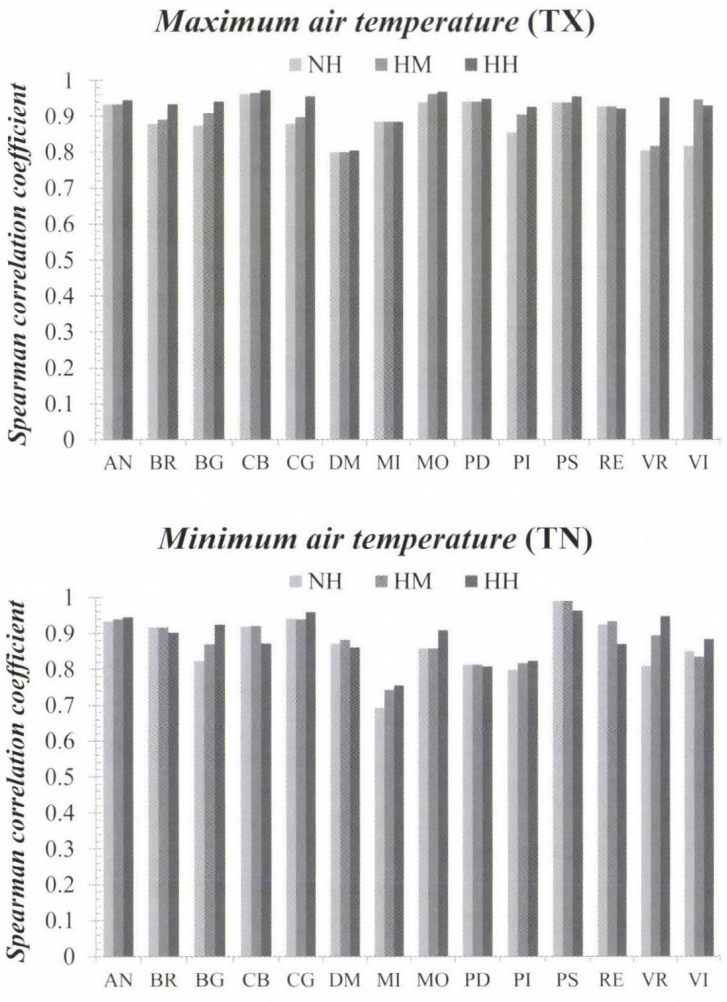

Fig. 3. Spearman correlation coefficient (SCC) between annual reference series and time series of non-homogenized (NH), homogenized with MASH (HM) and HOMER (HH) of maximum air temperature (top panel) and minimum air temperature (bottom panel), for the 1941–2010 period.

5. *Discussion and conclusions*

The IM network analyzed here includes stations located near the coast and a few meters above sea level and inland stations at higher altitude. Moreover, all weather stations are located in the same climatic region (temperate with dry and hot summer), which is a necessary condition to perform homogenization analysis. Results of the preliminary exploratory data analysis reveals time series with no extremes and only a small amount of outliers and missing values except in cases where times series does not cover the entire analysis period of 1941–2010. This is an important characteristic of the dataset, because missing values can have profound impact on reference series and, consequently, in the detection procedures (*Menne* and *Williams*, 2005; *Syrakova* and *Stefanova*, 2009). In addition, since missing values are treated differently in MASH and HOMER, a small number of data gaps cannot be associated with potential significant differences between homogenized datasets with both methods. On the other hand, heterogeneities are to be expected in TX and TN datasets, since this network is in operation for a long time, and during this period experienced adjustments were carried out on its structure (e.g., replacement of instruments), on its type (changes from classical to automatic sensors), and spatial distribution (e.g., relocation, cessation, and installation of new stations). For these reasons, we may conclude that maximum and minimum air temperature datasets in northern Portugal are examples of databases in good position to be analyzed for homogeneity.

MASH and HOMER were the methods used to perform the homogeneity analysis of TX and TN datasets. The selection criterion was, primarily, the high performance shown by these two methods during the comparison study performed in the framework of the COST Action HOME, using monthly temperature benchmark databases but also the large methodological differences between these two methods, discussed in previous sections. In fact, HOMER was not compared with other methods in *Venema et al.* (2012), because it became available later, but its results from the combination of the methods had the best performance. Craddock method was also included in the list of algorithms with best performance, but because it is a subjective method (uses visual detection of breaks), was not used in this study.

Time series were corrected with both methods from the most recent observations to the oldest. This procedure is consistent with the general believe that current sensors and data acquisition systems are more reliable than previous ones. Both methods uses interpolation to produce homogenized time series without missing values, but MASH also uses extrapolation to fill the data gaps in the extremes of the time series. MASH identifies the location of the break with the year of the shift, while HOMER is able to estimate the month of the change also (not shown in *Table 2*).

The total number of breaks detected in TN with both methods is higher than in TX, and the number of breaks detected with HOMER is higher than with MASH,

in both climatic elements. The same conclusion is supported by considering the number of breaks with amplitudes above increasing thresholds. In addition, the amplitude of the breaks detected with HOMER is, in general, higher than the amplitude of the breaks detected with MASH (*Table 2*). The weather stations of Vila Real and Coimbra B were selected as examples of inland and coastal weather stations, located at higher and lower altitudes, respectively (*Fig. 4* and *Fig. 5*), to illustrate the differences between the non-homogenized and homogenized time series with MASH and HOMER. It should also be mentioned that maximum air temperature time series in Mirandela is the only one without inhomogeneities.

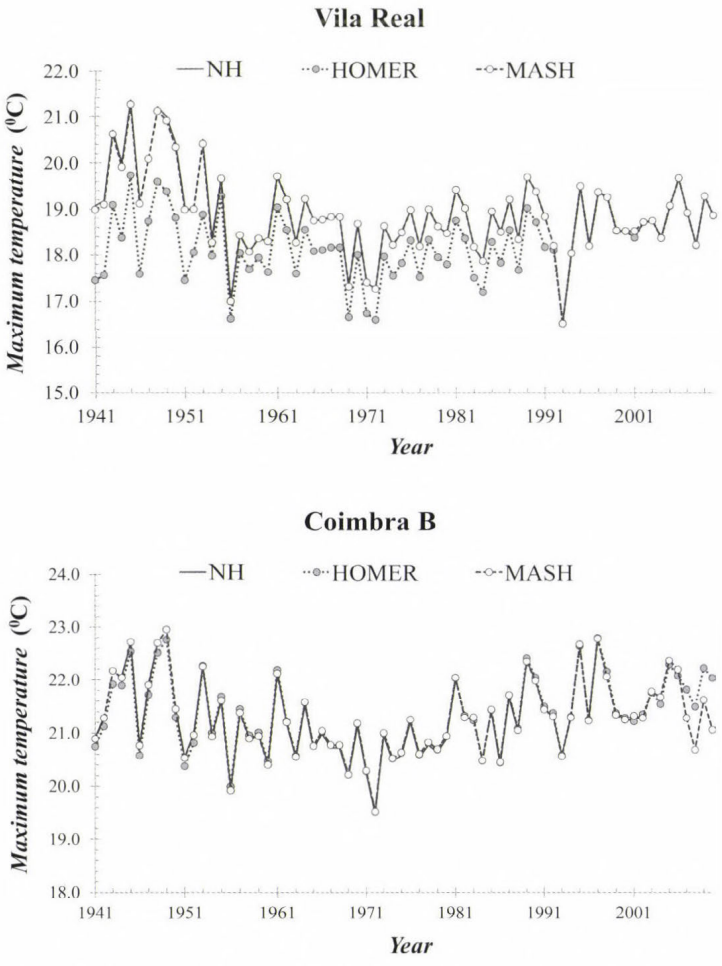

Fig. 4. Non-homogenized (NH) and homogenized time series (with HOMER and MASH) of maximum air temperature (TX) recorded in Vila Real and Coimbra B weather stations. Coimbra B is an example of weather station located in near the coast at low altitude, while Vila Real is an example of weather station located at mountainous region of the interior.

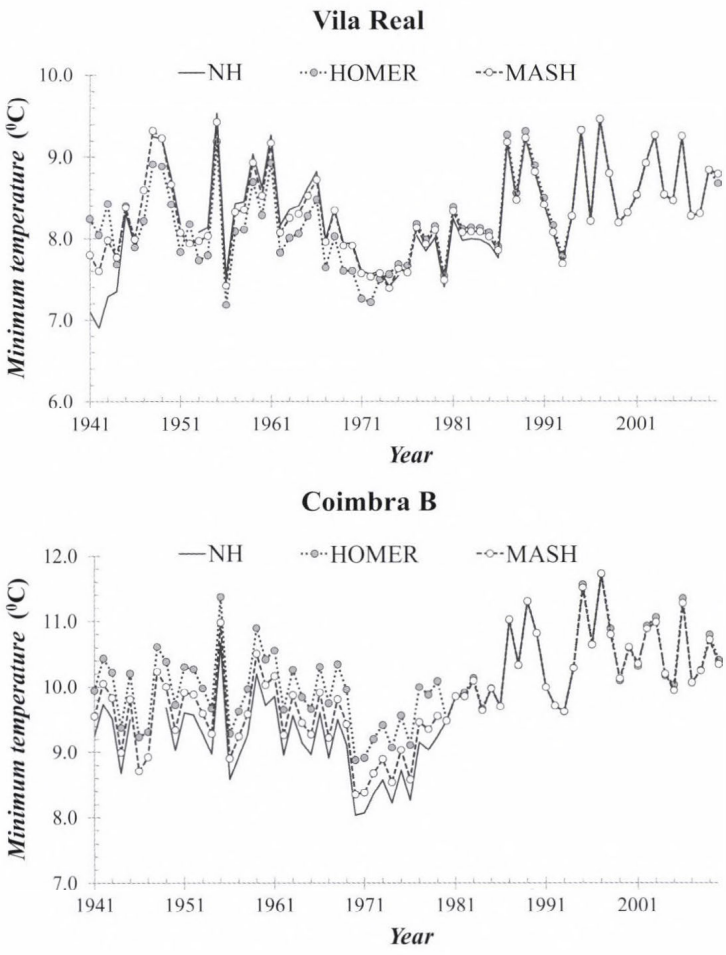

Fig. 5. As in *Fig. 4*, but for minimum air temperature (TN).

Correlation, trend, and principal component analysis were used to assess the homogenization process performance by comparing the results obtained with using non-homogenized and homogenized datasets. Boxplots of the Spearman correlation coefficient (SCC) statistical values obtained between homogenized time series with HOMER are higher and have much lower dispersion than those obtained between homogenized with MASH and non-homogenized time series (*Fig. 2*). For maximum air temperature, the homogenized time series of Dunas de Mira weather station (with both methods) presents the lowest values of the statistics, and it is responsible for the high dimension of the lower whisker. This result is more perceptible in homogenized time series with HOMER than with

MASH as boxplots for other time series are much more alike. For TN, the dispersion is much higher than for TX, and Dunas de Mira time series is also among those presenting lower statistics values. Results obtained with correlation analysis between reference, non-homogenized, and homogenized time series are also consistent with the increase of the similarity of the datasets with the application of both homogenization procedures. Values of SCC increases for all homogenized time series with MASH, but for a few time series (both in TX and TN), SCC values obtained for homogenized time series with HOMER are smaller than for non-homogenized. Notwithstanding this fact, an overall small increase in the median SCC values is conspicuous.

Trend analysis performed on TX and TN time series reveals a small reduction in the number of statistically significant tendencies after homogenization, but a general decrease in the slope, more significant for homogenized time series with HOMER than with MASH, must be underlined. Results obtained using different statistical significance levels (97.5% and 95%) are similar except for the expected higher number of statistically significant trends.

Results from PCA are consistent with those previously obtained with other methodologies and also suggests that homogenization leads to an increase of the resemblance in the spatial and temporal variability of both TN and TX. This behavior is more evident for TX than for TN. In general, the first EOF presents elements with equal sign, which reflects similar behavior in the entire region. Then, the following EOF represents small scale features of variability (e.g., contrast between north and south or between east and west). The magnitude of each feature can be measured by the explained variance of the corresponding mode of variability. In this study, the explained variance by the first PC is higher for homogenized than for non-homogenized datasets, independently of the climatic element (*Table 3*). This difference is higher for homogenized time series with HOMER than with MASH.

Table 3. Explained variance of the statistically significant principal components of non-homogenized (NH), homogenized with MASH (HM) and with HOMER (HH) minimum and maximum air temperature datasets, for the 1941–2010 period

	N	NH	HM	HH
Maximum temperature	1	78.6%	81.0%	89.4%
	2	8.6%	–	–
Minimum temperature	1	75.2%	76.3%	84.8%
	2	9.7%	8.3%	–

In resume, the most important conclusion from this study is that both methods contribute to correcting the inhomogeneities detected in both TN and TM datasets, and that there is no clear evidence of the better performance of one method relative to one another. Results obtained from the correlation analysis, trend analysis and principal component analysis point to a general increase on the spatial and temporal similarity of the time series as should be expected in datasets of the same climatic region. Apparently, these results are independent of the location and altitude of the weather stations. However, these conclusions should be taken with caution, because earlier studies reveal that the evaluation of methods performance is dependent on the metrics used for this purpose (*Venema et al.*, 2012) and on the quality and characteristics of databases (*Freitas et al.*, 2012).

Finally, it should be noted that, to the best of our knowledge, this study is the first effort to compare HOMER with other homogenization methods using observed datasets. The other known study assessing HOMER performance was recently presented in the 7th seminar for homogenization and quality control on climatological databases, but using the HOME benchmark datasets (*Domonkos*, 2012). Furthermore, besides the study of *Freitas et al.* (2012), to assess HOMER potential and limitations, this is the first consistent attempt to homogenize maximum and minimum air temperature Portuguese datasets, using more than one method, MASH and HOMER in particular. The other known homogenization study performed with Portuguese data, was performed by *Soares* and *Costa* (2009), which used precipitation data from stations located in the southern part of the country, as a case study, to compare the potential advantages of geostatistical techniques. As a final point, it should also be emphasized the number of different methods/measures used to compare the homogenized Portuguese air temperature datasets.

References

AEMET, IM, 2011: Iberian Climate Atlas. Closas-Orcoyen S.L, Madrid.

Aguilar E, Auer I, Brunet M, Peterson T.C, Wieringa J., 2003. *Guidelines on Climate Metadata and Homogenization*. World Meteorological Organization: *WMO–TD No. 1186, WCDMP No. 53*, Switzerland, Geneva.

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

Alexandersson, H. and *Moberg, A.*, 1997: Homogenization of Swedish temperature data. 1. Homogeneity test for linear trends. *Int. J. Climatol. 17*, 25–34.

Brunetti M., Maugeri M., Monti F., Nanni T., 2006: Temperature and precipitation variability in Italy in the last two centuries from homogenized time series. *Int. J. Climatol. 26, 345–381* .

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

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

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

Della-Marta, P.M. and *Wanner, H.*, 2006: A Method of Homogenizing the Extremes and Mean of Daily Temperature Measurements. *J. Climate 19*, 4179–4197.

Domonkos, P., 2011: Adapted Caussinus-Mestre Algorithm for Networks of Temperature series (ACMANT). *Int. J. Geosci. 2*, 293–309.

Domonkos, P., Venema, V., Auer, I., Mestre, O., and *Brunetti, M.*, 2012: The historical pathway towards more accurate homogenisation. *Adv. Sci. Res. 8, 45–52*

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

Freitas,L., Pereira M.G., Amorim L., Caramelo L., Mendes M., and *Nunes F.L*, 2012: Portuguese temperature dataset homogeneity with HOMER *(submitted to WMO serial World Climate Data and Monitoring Program)*

Guijarro, J.A., 2011: Climatol version 2.0, an R contributed package for homogenization of climatological series. State Meteorological Agency, Balearic Islands Office, Spain, http://webs.ono.com/climatol/climatol.html

Gullet, D.W., Vincent, L., and *Sajecki, P.J.F.*, 1990: Testing for Homogeneity in Temperature Time Series at Canadian Climate Stations. CCC Report No. 90-4, *Atmospheric Environment Service, Downsview, Ontario.*

HOME, 2011: Homepage of the COST Action ES0601 - *Advances in Homogenisation Methods of Climate Series: an Integrated Approach* (HOME), http://www.homogenisation.org.

Jolliffe, I.T , 2005:Principal component analysis. In (B.S. Everitt and D.C. Howell) *Encyclopedia of Statistics in Behavioral Science, vol. 3,* Wiley, New York, 1580–1584.

Kuglitsch, F.G., Toreti, A., Xoplaki, E., Della-Marta, P.M., Luterbacher, J., and *Wanner, H.*, 2009: Homogenization of daily maximum temperature series in the Mediterranean, *JGR 114*, 1–6.

Menne, M.J. and *Williams Jr., C.N.*, 2009: Homogenization of temperature series via pairwise comparisons. *J. Climate 22*, 1700–1717.

Menne, M.J. and *Williams jr. C.N.*, 2005: Detection of undocumented changepoints using multiple test statistics and composite reference series. *J. Climate 18*, 4271–4286.

Mestre O., 1999. Step-by-step procedures for choosing a model with change-points. In *Proceedings of the second seminar for homogenisation of surface climatological data*, Budapest, Hungary, *WCDMP-No.41, WMO-TD No.962*, 15–26.

Mestre, O. and *Aguilar E.*, 2011: Homogenization (personal communication). COST ES0601 Training School, 5–7th of October, Tarragona.

Moberg, A. and *Jones, P.D.*, 2004: Regional Climate Models simulations of daily maximum and minimum near-surface temperatures across Europe compared with observed station data for 1961-1990. *Clim. Dynamics 23*, 695–715.

Pereira, M.G., Caramelo, L., Gouveia, C., Gomes-Laranjo, J., and *Magalhaes, M.* 2011: Assessment of weather-related risk on chestnut productivity. *NHESS*, 2729–2739.

Perreault L., Bernier J., Bobée B., and *Parent E.*, 2000. Bayesian change-point analysis in hydrometeorological time series. Part. 1. The Normal model revisited. *J. Hydrol. 235*, 221–241.

Peterson, T.C. and *Easterling, D.R*, 1994: Creation of homogeneous composite climatological reference series, *Int. J. Climatol. 14*, 671–679.

Peterson T.C., Easterling, D.R., Karl, T.R., Groisman, P., Nicholls, N., Plummer, N., Torok, S., Auer, I., Boehm, R., Gullett, D., Vincent, L., Heino, R., Tuomenvirta, H., Mestre, O., Szentimrey, T., Salinger, J., Førland, E.J., Hanssen-Bauer, I., Alexandersson, H., Jones, P. and *Parker, D.*,1998: Homogeneity adjustments of in situ atmospheric climate data: A review. *Int. J. Climatol. 18*, 1493–1517.

Peterson, T.C., Easterling, D.R., Karl, T.R., Groismanm P., Nicholls, N., Plummer, N., Torok, S., Auer, I., Boehm, R., Gullett, D., Vincent, L., Heino, R., Tuomenvirta, H., Mestre, O., Szentimrey, T., Salinger, J., Forland, E.J., Hanssen-Bauer, I., Alexandersson, H., Jones, P., and *Parker, D.*, 1998: Homogeneity adjustments of in situ atmospheric climate data: A review. *Int. J. Climatol. 18*, 1493–1517.

Picard, F., Lebarbier, E., Hoebeke, M., Rigaill, G., Thiam B., and *Robin, S.*, 2011: Joint segmentation, calling, and normalization of multiple CGH profiles. *Biostatistics.* doi:10.1093/biostatistics/kxq076

Pozo-Vazquez, D., Gamiz-Fortis, S.R., Tovar-Pescador, J., Esteban- Parra, M.J., Castro-Diez, Y., 2005: North Atlantic winter SLP anomalies based on the autumn ENSO state. *J. Climatol. 18*, 97–103.

Raj, P.P.N. and *Azeez, P.A.*, 2012: Trend analysis of rainfall in Bharathapuzha River basin, Kerala, India. *Int. J. Climatol. 32,* 533–539.

Reeves, J., Chen, J., Wang, X.L., Lund, R., and *Lu, Q.*, 2007: A review and comparison of changepoint detection techniques for climate data. *J. Appl. Meteorol. Climatol. 46,* 900–915.

Rodrigo, F.S. and *Trigo, R.M.*, 2007: Trends in daily rainfall in the Iberian Peninsula from 1951 to 2002. *Int. J. Climatol. 27*, 513–529.

Romanazzi, M., 1993: Jackknife estimation of the eigenvalues of the covariance matrix. *Comput. Statist. Data Anal. 15*, 179–198.

Sahin, S. and *Cigizoglu, H.K.*, 2010: Homogeneity analysis of Turkish meteorological data set. *Hydrol. Process. 24*, 981–992.

Sneyers, R., 1990: On the statistical analysis of series of observations., WMO *Tech. Note 143, 145*, Geneva, Switzerland.

Staudt, M., Esteban-Parra, M.J., and *Castrı-Diez, Y.*, 2007: Homogenization of long-term monthly Spanish temperature data. *Int. J. Climatol. 27*, 1809–1823.

Štěpánek, P., 2008: AnClim – software for time series analysis (for windows). *Dept. of Geography, Fac. of Natural Sciences*, MU, Brno. http://www.climahom.eu/AnClim.html.

Štěpánek, P. and *Mikulová, K.*, 2008: Homogenization of air temperature and relative humidity monthly means of individual observation hours in the area of the Czech and Slovak Republic. In: *5th Seminar for Homogenization and Quality Control in Climatological Databases.* Hungarian Met. Service, Budapest, Hungary, 147–163.

Štěpánek, P., Zahradniéek, P., and *Skalák, P.*, 2009: Data quality control and homogenization of the air temperature and precipitation series in the Czech Republic in the period 1961–2007, *Adv. Sci. Res. 3*, 23–26.

Syrakova, M. and *Stefanova, M.*, 2009: Homogenization of Bulgarian temperature series. *Int. J. Climatol. 29*, 1835–1849.

Szentimrey, T., 1994: Statistical problems connected with the homogenization of climatic time series. In *Proceedings of the European Workshop on Climate Variations*, Kirkkonummi, Finland, *Publications of the Academy of Finland, 3/94*, 330–339.

Szentimrey, T., 1999: Multiple Analysis of Series for Homogenization (MASH). *Proceedings of the Second Seminar for Homogenization of Surface Climatological Data*, Budapest, Hungary, WMO, *WCDMP-No. 41*, 27–46.

Szentimrey, T., 2007: Manual of homogenization software MASHv3.02", Hungarian Meteorological Service.

Toreti, F.G., Kuglitsch, A., Xoplaki, E., and. Luterbacher, J., 2012: A Novel Approach for the Detection of Inhomogeneities Affecting Climate Time Series. *J. Appl. Meteorol. Climatol.,vol. 51*, 317–326.

Trewin, B., 2010: Exposure, instrumentation, and observing practice effects on land temperature measurements. *WIREs Clim. Change, 1*, 490–506.

Venema, V., Mestre, O., Aguilar, E., Auer, I., Guijarro, J.A., Domonkos, P., Vertacnik, G., Szentimrey, T., Stepanek, P., Zahradnicek, P., Viarre, J., Müller-Westermeier, G. Lakatos, M., Williams, C.N., Menne, M., Lindau, R., Rasol, D., Rustemeier, E., Kolokythas, K., Marinova, T., Andresen, L., Acquaotta, F., Fratianni, S., Cheval, S., Klancar, M., Brunetti, M., Gruber, C., Duran, M.P., Likso, T., Esteban, P., and Brandsma, T., 2012: Benchmarking monthly homogenization algorithms. *Climate of the Past 8*, 89–115.

Vicente-Serrano, S., Begueria, S., Lopez-Moreno, J.I., Garcia-Verac, M.A., and *Stepanek, P.*, 2010: A complete daily precipitation database for northeast Spain: reconstruction, quality control, and homogeneity. *Int. J. Climatol. 30*, 1146–1163.

Vincent, L.A., 1998: A technique for the indentification of inhomogeneities in Canadian temperature series. *J. Clim. 11*, 1094–1104.

Wilks, D.S., 2011. Statistical Methods in the Atmospheric Sciences, vol 100, 3rd Edition,*International Geophysics 100.* Academic Press, Oxford, UK

Williams, C.N., Menne, M.J., and *Thorne, P.W.*, 2012:. Benchmarking the performance of pairwise homogenization of surface temperatures in the United States. *J. Geophys. Res.-Atmos. 117, D05116.*

Wold, S., Geladi, P., Esbesen, K. and *Ohman, J.*, 1987: Multi-way principal components- and PLS-analysis. *J. Chemometrics 1*, 41–56.

World Meteorological Organization, 2011: Guide to Climatological Practices, WMO/No. 100., Geneva.

Zhang, X.B., Vincent, L.A., Hogg, W.D., and *Niitsoo, A.*, 2001: Temperature and precipitation trends in Canada during the 20th century, *Atmos. Ocean 38*, 395–429.

IDŐJÁRÁS
Quarterly Journal of the Hungarian Meteorological Service
Vol. 117, No. 1, January–March 2013, pp. 91-112

Measuring performances of homogenization methods

Peter Domonkos

University Rovira i Virgili, Centre for Climate Change, Campus Terres de l'Ebre, Av. Remolins, 13-15, 43500-Tortosa, Spain,
peter.domonkos@urv.cat

(Manuscript received in final form October 8, 2012)

Abstract–Climatologists apply various homogenization methods to eliminate the non-climatic biases (the so-called inhomogeneities) from the observed climatic time series. The appropriateness of the homogenization methods is varied, therefore, their performance must be examined. This study reviews the methodology of measuring the efficiency of homogenization methods. The principles of reliable efficiency evaluations are: (i) Efficiency tests need the use of simulated test datasets with similar properties to real observational datasets; (ii) The use of root mean squared error (RMSE) and the accuracy of trend-estimations must be preferred instead of the skill in detecting change-points; (iii) The evaluation of the detection of inhomogeneities must be clearly distinguished from the evaluation of whole homogenization procedures; (iv) Evaluation of homogenization methods including subjective steps needs blind tests. The study discusses many other details of the efficiency evaluation, recalls the results of the blind test experiment of the COST action ES0601 (HOME), summarizes our present knowledge about the efficiencies of homogenization methods, and describes the main tasks ahead the climatologist society in the examinations of the efficiency of homogenization methods.

Key words: time series homogenization, efficiency, surface climatic observations, upper air climatic observations

1. Introduction

Homogenization is a procedure to improve the quality of data. During homogenization, the temporal constancy of some characteristics is tested and the degree of constancy is a quality indicator. In contrast with the common data quality

control, homogenization examines the characteristics of segments of data instead of those of individual pieces of data. Homogenization is applied in several branches of science, e.g., economics, information systems, neurology, etc. (see some references in *Toreti et al.* 2012), but homogenization tasks often have peculiarities according to research fields and the variables examined.

In climatology, homogenization examines and adjusts temporal biases of climatic variables, caused by non-climatic factors. Various technical changes may cause non-climatic biases in observed surface climate (*Aguilar et al.*, 2003; *Auer et al.*, 2005; *Menne et al.*, 2009, etc.) and in radiosonde data (*Lanzante et al.*, 2003, *Gruber* and *Haimberger*, 2008; *Dai et al.*, 2011, etc.), and a large number of methods are applied for their correction. The purpose of homogenization is to obtain observed climatic datasets of the best possible quality for climate variability investigations. Relative homogenization, named also innovation of time series (*Haimberger*, 2007), examines the series of the differences or ratios of the observed data (relative time series hereafter) instead of examining directly the raw data (absolute homogenization). Relative homogenization is preferred when the spatial density and coherence of the observed data allows it, because in relative time series the climatic fluctuation that is common for the examined region does not appear. Note, however, that absolute homogenization is also applicable under certain conditions. Different homogenization methods often have markedly different efficiencies in finding and correcting the non-climatic biases. The objective interpretation of climate change and climate variability, assessment of risks of extreme climatic events, modeling of spatial and temporal evolution of weather and climate events all need accurate input data fields; therefore, the climatological community is interested in finding the best homogenization methods. The selection of the most appropriate methods requires the application of objective efficiency tests.

The COST action "Advances in homogenization methods of climate series: an integrated approach" (2007–2011) accelerated the progress of the methodological development of homogenization and its reliable testing in several ways. We refer to the COST action with its acronym "HOME", to its benchmark dataset for the surrogate European surface temperature dataset with "Benchmark", and often to its closing study written by the HOME group (*Venema et al.*, 2012). Under HOME, 25 versions of 9 statistical homogenization algorithms were subjected to blind tests, and their results were evaluated with 13 efficiency measures. Nevertheless, the scope of this paper is much wider than the analysis of HOME products. We review the contemporary methodology of tests applied in the efficiency evaluations for homogenization procedures in a wide range of homogenization tasks. The problems related to the choice of efficiency measure and the construction or selection of test

datasets are widely discussed. Reliable efficiency tests must be based on test datasets whose statistical properties mimic well the properties of observational datasets. In this respect the study has limitations, i.e., we do not deal with the peculiarities of individual homogenization tasks apart from some examples. We do not deal with the particularities of daily data homogenization either.

The organization of the paper is as follows. In the next section, the problems related to setting up efficiency evaluation methods with general reliability are listed and discussed. In Section 3, the efficiency measures and their properties are described. In Section 4, various kinds of efficiency tests with their different objectives are presented, while Section 5 deals with the problem of constructing realistic test datasets. Finally, the tasks for the future are discussed in Section 6.

2. Difficulties in producing reliable efficiency evaluations

As the most frequent type of inhomogeneities is the sudden shift in the means (referred also as change-points), e.g., for station relocations, change in the instrumentation, etc., the evaluation of efficiency might not seem to be a complicated task: the simplest assessment is to calculate the ratio of correctly identified change-points relative to all change-points (the so-called hit rate), since higher hit rate generally indicates better performance of homogenization method. However, this simple approach often fails and its causes are discussed in this section, grouping the problematic aspects into four subsections.

2.1. Complexity of homogenization methods

Homogenization is a complex procedure. It generally includes at least 3 segments (*Gruber* and *Haimberger*, 2008), they are:

(i) time series comparison,
(ii) detection of inhomogeneities,
(iii) adjustments of the detected biases.

Each segment can be objective (i.e., based on pure statistics) or subjective. Detection and adjustments may be partly or fully based on metadata. Note that in absolute homogenization segment (i) is missing. On the other hand, any segments or even all segments can be included multiple times in one homogenization procedure, because several procedures are iterative with the cyclically repeated application of their segments. From the point of view of efficiency evaluation, the problem is that a certain detection method can be applied with various time series

comparisons, adjustments, and iteration techniques. When, for instance, the hit rate is calculated, the result depends on all the segments and even on each of the parameters included in the procedure. For example, *Moberg* and *Alexandersson* (1997) presented the application of the standard normal homogeneity test (SNHT) through the homogenization of Swedish temperatures. They built reference series from 8 series of the neighbourhood with the highest spatial correlations for the increment (first difference) series, applying the cutting algorithm by *Easterling* and *Peterson* (1995) until the subsections had at least 10 years length, etc. A problem of testing the performance of homogenization methods is that the details of the methods, as for instance the ones cited from *Easterling* and *Peterson* (1995) and particularly the parameters are only recommendations, and some of the proposed details cannot even be applied for all homogenization tasks. However, the performance depends on all the particularities of the procedure.

A further problem is that most homogenization procedures include subjective steps that make the objective evaluation of the performance difficult (Section 4.3).

2.2. Indication of good performance

The simplest evaluation of performance is the calculation of the hit rate. The problem with hit rate (and with more advanced related metrics, e.g., detection skill, see Section 3) is that the accuracy of homogenized time series only partly depends on it. When the Hit rate is high, but some large shifts fail to be detected, the variability of the homogenised time series may substantially differ from the true climatic variability. By contrast, if the largest shifts are detected well, the final result of homogenization might be fair in spite of a relatively low hit rate. Note that the accumulated effect of inhomogeneities on the bias of variability characteristics is even more important than the shift-magnitudes at individual change-points (an example will be shown in Section 3). As the aim of homogenization is to have the climatic time series in the appropriate state for deriving climate variability characteristics with high accuracy, the best way is to use efficiency measures which directly evaluate the quality of homogenized time series from this point of view. Yet, there are two more problems. One is that a homogenization result may be excellent for the examinations of some climatic characteristics (e.g., linear trend, low frequency variability of the mean values, etc.), but might be poor for some other examinations (e.g., standard deviation, extreme events, etc.). For this reason, the use of one efficiency measure cannot be sufficient to evaluate the general performances of homogenization procedures. Another problem is that often the objective parts of homogenization methods are evaluated only (we consider homogenization procedures or their certain segments objective when subjective decisions by homogenisers are not allowed in them). Evaluations are often

restricted to the examination of the detection of change-points with statistical tests (*DeGaetano*, 2006; *Gerard-Marchant et al.*, 2008; *Bealieu et al.*, 2008; etc.). However, if one would like to know the connection between the skill in detecting change-points and the final quality of homogenization product, the inclusion of other segments of the homogenization procedure is necessary for the evaluation. A suggested solution will be described in Section 4.1.

2.3. Station effects: true or false?

Even if all time series are ideally homogeneous in a network of the same climatic region, some statistical properties of time series are still distinct for each individual time series due to the peculiarities of the observing station (e.g., exposure, land use, natural vegetation, etc.). Therefore, when the aim of homogenization is transformed to an exact mathematical task, it should include the elimination of change-points, but should exclude the cancellation of true station effects. In relative homogenization, temporally constant station effects can be preserved only, since relative homogenization is based on the equalization of differences or ratios of time series. The accuracy of mean station effects can hardly be controlled by efficiency tests, because a) there is no objective method for the estimation of mean station effects, b) it seems to be a challenge to construct test datasets with pre-defined realistic mean station effects.

The most usual way of determining station effects in homogenization procedures is to keep the last homogeneous section (the section between the change-point detected with the latest date and the end of the series) unchanged and adjust all the other parts of the series to that section. This assumption is correct when all the technical, personal, and environmental conditions were good to provide high quality observations in the last period of the series, but false in the opposite case. Especially, when only the late part of a time series is influenced by urbanization, that urban effect will be included for the whole homogenized time series if the adjustments are made relative to the last homogeneous section of the series. Note however, that from the point of view of macroclimatic examinations, the temporal changes of urban effect are undesired inhomogeneities, thus their elimination by homogenization is correct.

In the homogenization of surface climatic data, the assessment of mean station effect could be considered a task that is out of the scope of homogenization, since a series can be perfectly homogeneous in term of mathematical homogeneity in spite of the average station effect is false. An example for the latter case is when each value of the series of ideally accurate observational values is shifted with a constant error-term mimicking an erroneous mean station effect: the trends remain in line with the macroclimate, but the distribution function and its statistical characteristics

would be false. We incorporate this problem into the homogenization task, because with the expression "homogenized time series" climatologists mean high quality data that are applicable well for climate variability analyses. Note that large errors in the assessment of mean station effects in the principal surface climate variables are rare, therefore, the problem of their correct treatment is basically theoretical with relatively little practical importance.

In upper air measurements, the origin of true station effect is restricted to the geographical coordinates, as there is no particular effect from exposure, surface type, natural vegetation, etc. However, systematic local errors can be larger due to the more serious problems of instrumentation than in surface observations. In accordance with these facts, the homogenization of upper air time series includes the optimization of the spatial differences of data (*Haimberger*, 2007).

2.4. Dependence on the properties of test dataset

True observational time series cannot be used for evaluating efficiency, because we never know the exact characteristics of non-climatic biases. Even with the best homogenization methods, only a part of the inhomogeneities can be identified, and even false detections sometimes occur (*Venema et al.*, 2012). Therefore, artificial test datasets are needed with known positions and magnitudes of inhomogeneities for measuring the efficiency of homogenization methods. These test datasets should resemble the true climatic time series as much as it is possible, because otherwise the observed efficiencies in test experiments may not be valid in the real world. We illustrate the seriousness of this problem with the description of two experiments. Detection skill (see its definition in Section 3) was examined for 6 widely used change-point detection methods (*Fig. 1*). In the first experiment, one change-point was inserted into 100 years long stationary white noise processes. The shift-magnitude was 3 times larger than the standard deviation of the white noise. In the second experiment the only difference was that further four change-points were inserted with half-size shifts relative to the one large shift inserted earlier. The positions and directions of the small shifts were random. In both experiments, the detection skill for the one large shift was evaluated only. The difference between the obtained efficiencies is striking: while in the first experiment the efficiencies are between 88–96% for the five best methods (out of the examined six in *Fig. 1*), the values drop to 59–75% when 4 more change-points are present in the time series. We underline that the detection skill of small change-points did not contribute directly to the results shown, but small shifts generally act as a kind of noise, which substantially worsens the detection skill of large shifts.

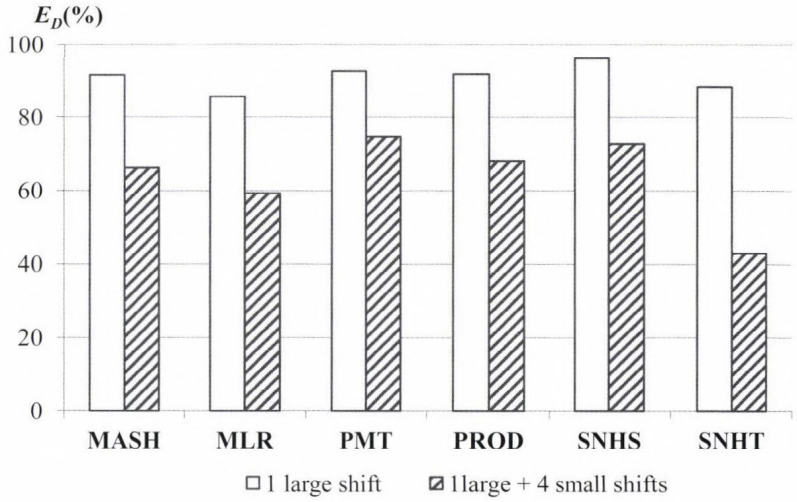

Fig. 1. Detection skill (E_D) of large shifts with the background of white noise and with that of white noise plus four small shifts. PMT – Penalized maximum t-test (*Wang et. al.*, 2007); PROD – PRODIGE, SNHS – SNHT for shifts only (*Alexandersson*, 1986); the explanation of other denotations is in the text.

For constructing realistic test datasets, we should know the statistical properties of inhomogeneities in the target climatic time series. There are two problems related to this point: a) We cannot learn the exact properties of inhomogeneities, because small inhomogeneities often cannot be detected with any kind of method; b) Even if we could determine the exact characteristics of some real climatic time series, that characteristics would not be projected without control to new homogenization tasks, since the properties for different time series, networks, and climatic variables are obviously diverse.

3. Efficiency measures

For characterizing the appropriateness of homogenization methods to make the climatic time series more suitable for accurate climate variability examinations, efficiency measures must evaluate the mean progress in the accuracy of the variability of homogenized time series. The root mean squared error (RMSE) is a known tool for characterizing skills and remaining errors:

$$\mathrm{RMSE}(\mathbf{X}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - t_i)^2}\,. \tag{1}$$

It can be formed to an efficiency measure at which 1 means perfect skill and 0 means neutrality (no improvement, neither destruction):

$$E = \frac{\mathrm{RMSE}(\mathbf{Z}) - \mathrm{RMSE}(\mathbf{X})}{\mathrm{RMSE}(\mathbf{Z})}\,, \tag{2}$$

where **Z**, **X**, **T** stand for raw, homogenized, and true (fully homogeneous) time series, respectively, n is the sample size, and E is the efficiency.

The RMSE can be calculated for various time units of the observed series. For instance, *Venema et al.* (2012) applied month, year, and decade time units. With RMSE of long time units, the evaluation is focused on the accuracy in long-term variability, while the meaning of RMSE of short time units is more general. Especially, the detection of seasonality of non-climatic biases can be evaluated with the comparison of monthly and annual RMSE results.

Venema et al. (2012) introduced a modified version of RMSE (centered RMSE, CRMSE), it calculates the RMSE of the anomalies relative to the mean bias:

$$\mathrm{CRMSE}(\mathbf{X}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - t_i - \overline{\mathbf{X}-\mathbf{T}}\right)^2}\,, \tag{3}$$

where the upper stroke means arithmetical average. The motivation of using CRMSE instead of RMSE in HOME was to eliminate the effect of unknown mean station effects, because the HOME tests did not incorporate this specific problem in any form.

In the evaluation of the accuracy of linear trends in homogenized time series, RMSE is also applicable to the comparison of trend slopes in **X** and **T** (*Venema et al.*, 2012; *Domonkos*, 2011a). Linear trend estimations and their accuracy have enhanced importance in climate studies, since linear trends indicate the sign and degree of mean systematic change of the climate variable under study over the period examined. Note that the RMSE of trend biases is not impacted by the possible errors in the estimation of mean station effects.

All RMSE and CRMSE characteristics described can be applied in the evaluations of entire time series or sections of time series. The accuracy of network-mean values is particularly important in the assessment of past climate

changes, and a novelty of HOME was that RMSE was calculated also for the series of network-means (*Venema et al.*, 2012). We mention that apart from RMSE or CRMSE, mean absolute error or rank order are also applicable (the latter is only for the comparison of performances, see, e.g., *Titchner et al.*, 2009).

The most frequently used efficiency measures for the homogenization of climatic time series are hit rate (S_R, referred also as detection power), false alarm rate (S_F) and their various combinations (*Buishand*, 1982; *Easterling* and *Peterson*, 1995, *Reeves et al.* 2007, etc.). Hit rate (false alarm rate) shows the ratio of correctly (falsely) detected change-points relative to the total number of change-points (S) that are present in time series. Large S_R and small S_F indicate good skill in detecting change-points, while the opposite case indicates poor skill. With detection skill (E_D, Eq. (4)), S_R and S_F can be examined jointly (*Menne* and *Williams*, 2005; *Domonkos*, 2011a).

$$E_D = \frac{S_R - S_F}{S} \quad . \tag{4}$$

Although hit rate and detection skill are the most traditional efficiency measures, there are several problems with their applications. Their main deficiency is that S_R and E_D do not indicate confidently the accuracy of homogenized time series and the appropriateness of time series for climate variability analysis. *Fig. 2* presents the imaginary results of a time series homogenization. The series consists of 100 years, and it contains three change-points in years 30, 50 and 70. The shift-magnitudes in years 30 and 70 are slightly larger than in year 50. The sign of the shift in year 30 is the opposite of the sign of the other two shifts. In the homogenization labelled with Y1, only the shift of year 70 was detected and corrected, while in Y2 two shifts, i.e.,the ones in years 30 and 70 (*Fig. 2*). It is clear that the hit rate and detection skill are better for Y2 than for Y1 (namely 1/3 for Y1 and 2/3 for Y2). However, the RMSE and remaining trend-bias are better for Y1 than for Y2. The RMSE (remaining trend bias for the entire series) are 1.45 (0.54/100yr) for Y1 and 2.12 (4.50/100yr) for Y2. We have these seemingly contradictory results in spite of the two largest shifts were detected in Y2, the time-lapses of the detected change-points are zero, no false detection occurred, and the assessment of shift-magnitudes is perfect for the detected shifts. The only thing that favored for Y1 is that the accumulated effect of inhomogeneities in the first 30 years section of the series is smaller if the bias of year 70 is corrected.

Another problem is that during the practical use of hit rate and detection skill, subjectively-set parameters are often applied. First, because a certain time lapse in the detection is usually accepted as correct detection, otherwise the evaluation

could be too strict and unrealistic (*Ducré-Robitaille et al.*, 2003; *Bealieu et al.*, 2008; etc.). Second, because the detection of small size biases is often not evaluated, and third, if test series include other kinds of inhomogeneities than sudden shifts (e.g., gradually increasing biases), even the calculation of *S* may need the incorporation of parameterized definition (*Domonkos*, 2011a). The thresholds for the allowed time lapse, minimum size of shift, and criterions for change-point in temporally irregular station effects all need subjective decisions which reduce the power and comparability of tests with hit rate and detection skill. On the other hand, these statistics must be considered as indications about the operation of the homogenization procedure, which indications may be important both for the users and constructors of the methods. Note that from hit rate and false alarm rate not only detection skill can be derived, but also several other scores that characterize the success of change-point detection in various ways (*Menne* and *Williams*, 2005; *Venema et al.*; 2012).

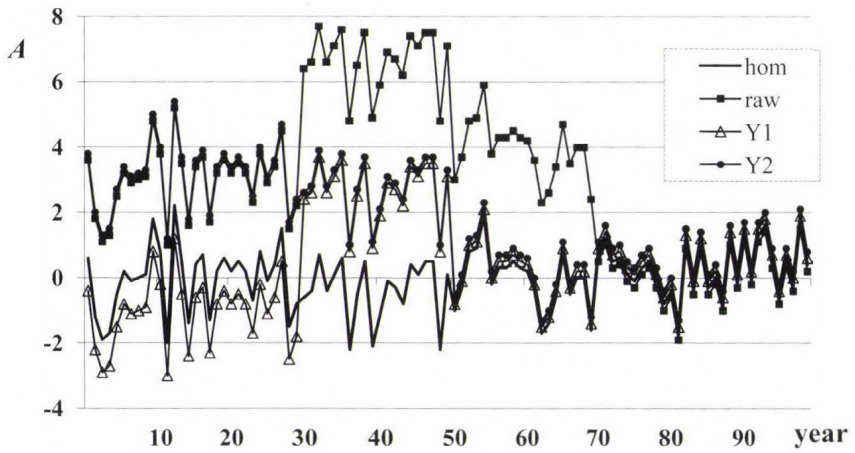

Fig. 2. Example of homogenization results. The raw time series is a 100 years long white noise with shifts in years 30, 50, and 70, whose magnitudes are +4, –3, –4, respectively. The unit of the values (*A*) is the standard deviation of white noise. hom = homogenous data, raw = "raw" (simulated) data, Y1 and Y2 are the results of partly successful homogenization. In Y1 only the shift of year 70, while in Y2 two shifts, i.e., the shifts in years 30 and 70 are detected and corrected. Hit rate and detection skill are better for Y2, but the RMSE and mean trend bias are better for Y1.

The skill of detection can also be characterized in other ways than with versions or combinations of hit rate and false alarm rate. Such indicators are the

ratio of experiments in which the exact number of change-points are detected (*Caussinus* and *Mestre*, 2004; *Bealieu et al.* 2009; etc.), the mean squared temporal distance between true change-points and detected change-points (*Bealieu et al.*, 2008, 2009), and the ratio of correctly chosen models in the detection process (*Reeves et al.*, 2007). Their connection with the method performance is similar to that of hit rate, i.e. they serve useful information about the operation of homogenization methods, but they cannot be applied directly for characterizing efficiency. In connection with the model selection during detection process, it has to be noted that there is no evidence that the use of more complex or more flexible models in the detection process would result in higher efficiency than the use of step function. In reality, *Domonkos* (2011a) found just the opposite relation when he compared the performances of Multiple Linear Regression (MLR, *Vincent*, 1998) and the second version of SNHT (*Alexandersson* and *Moberg*, 1997) with detection methods applying always step function model, namely with Multiple Analysis of Series for Homogenization (MASH, *Szentimrey*, 1999) and PRODIGE (*Caussinus* and *Mestre*, 2004). The likely explanation is that the selection of model type and its parameters is problematic from noisy, finite, and inhomogeneous samples, like true observed climatic time series.

4. Kinds of efficiency tests

Efficiency tests can be sorted at least into two groups according to their goals. One type is for measuring the performance of complete homogenization procedures and another type is when a particular segment of homogenization procedures is tested only. Both types of tests are important: while the tests of complete procedures inform us about the practical appropriateness of a method, the separated investigation of segments helps to reveal the positive features and deficiencies of the methods, and thus it may give suggestions for further, methodological developments. In this section we define more than two kinds of tests, but we admit that the classification is partly subjective.

4.1. Tests for detection methods

We have mentioned in Section 3 that mostly the detection parts of homogenization methods are tested only, and it is often the case even when studies promise tests for entire homogenization procedures. This inexactness in the use of terms might arise from the fact that a particular detection method is often paired with indefinite characteristics in the other segments of the homogenization procedure (some examples were mentioned in Section 2.1), and thus, often only the detection part is common in the different versions of the method. Another possible explanation is

that the detection segment might be expected to be the most influential part of homogenization procedure to the final efficiency. Note that the latter expectation is often not true, the comparison of efficiency results by *Venema et al.* (2012) and *Domonkos* (2011a) proves that the major error source is often in other segments of homogenization procedures than the detection part.

A seeming difficulty of finding the real effectiveness for detection methods is that hit rate, false alarm rate, and the characteristics that are derived from these two do not show the true efficiency accurately. On the other hand, the calculation of RMSE error needs the incorporation of the other segments of homogenization procedures. This problem can be solved with the application of standard procedures in all segments except for the detection part (*Domonkos*, 2011a). The idea is not new, since for calculating relative time series and shift magnitudes, standard procedures had been applied in earlier studies (*Ducré-Robitaille et al.*, 2003; *DeGaetano*, 2006). Recent examinations show that the most effective detection methods apply a relatively simple model, namely the step-function, and they select the most probable parameters of this model by the examination of all possible combination of change-point positions. Such detection segments are included in MASH, PRODIGE, Applied Caussinus-Mestre Algorithm for homogenizing Networks of Temperature series (ACMANT, *Domonkos*, 2011b), and HOMER (*Mestre et al.*, 2012). Note, however, that when the signal to noise ratio is small or when the frequency of change-points is very low, the advantage of the highlighted detection methods ceases.

4.2. Tests for specific segments others than the detection of inhomogeneities

There are three main kinds of time series comparisons: a) for each candidate series, building one reference series from composite series, b) using multiple reference comparisons for each candidate series, c) using multiple comparisons without defining which are the candidate and the reference. Their testing is problematic, because this segment contains subjective steps in many procedures. For fully objective procedures the testing would be straightforward with the inclusion of standardised detection and correction segments, but according to our knowledge such tests never have been done.

Objective correction methods can generally be tested applying the same logic as described for the testing of detection methods and time series comparisons. We know about one example of testing correction segment, i.e., the test of ANOVA (*Domonkos et al.*, 2012a). The testing of ANOVA is much easier than making any other segment-specific tests, because the input field of ANOVA is the list of change-point positions detected. Once such lists are available from different test experiments, there is no need of constructing test dataset, applying standard procedures for other segments than the

target segment, etc. HOME provided the required lists from different homogenization procedures and homogenizers, and these data are freely available for the climatologist community. The tests showed that the application of ANOVA always results in improvement in the final results of homogenization. It means that the performance of any homogenization procedures (at least, which were participating in this test) could be improved with the inclusion of ANOVA. This finding is in accordance with the fact that ANOVA provides the optimal estimation of correction terms when the climate is uniform in the network and when the detected change-point positions are correct (*Caussinus* and *Mestre*, 2004). Considering the contemporary homogenization methods, PRODIGE, ACMANT and HOMER include ANOVA. Note that MASH was one of the most successful methods of HOME and although MASH does not include ANOVA, there was no experiment of adding ANOVA to MASH, because MASH did not produce a usable list of change-point positions.

4.3. Tests for complete homogenization procedures

Testing whole procedures might not seem to be more challenging than testing selected segments only: it needs the use of a reliable test dataset and the calculation of some efficiency measures. However, most procedures contain subjective steps, which make it difficult to produce objective comparative tests for wide range of homogenization methods.

The testing of fully automatic procedures is relatively easy: Running an automatic program is simple, and nowadays, the computational time is usually fairly short. The results are objective, impersonal, and they can be reconstructed at any time. Although the application of appropriate test dataset is a critical point of the methodology, the doubts and difficulties can be fairly treated by the use of some variety of test datasets (*McCarthy et al.*, 2008; *Titchner et al.*, 2009). Note that the same works also when detection segments are tested only (*Ducré-Robitaille et al.*, 2003; *Domonkos*, 2011a; etc.); moreover, tests with moving parameters of the test dataset may clarify the roles of selected dataset characteristics in the performance of the examined methods (*DeGaetano*, 2006). The easy application of tests for automatic methods favors their development, since large number of variants of the same homogenization procedure can be executed with relatively little effort. Tests with moving parameters of the examined method show the sensitivity of the performance to changes in its parameters (*Gruber* and *Haimberger*, 2008; *Domonkos*, 2008, 2012), while ensemble experiments with random selection of parameter sets indicate the general stability of method performance (*McCarthy et al.*, 2008; *Titchner et al.*, 2009; *Williams et al.*, 2012).

The main problem with testing subjective or partly subjective methods is that the evaluation might be affected from the known truth, both in the construction of

test datasets and in the execution of tests. This influence can be unintentional, and it questions the objectivity of the test results. Further problems of subjective methods are that the test results are homogenizer dependent and usually cannot be reconstructed. Finally, the subjective homogenization of large datasets is sometimes very tiring, practically unmanageable.

One conclusion could be that the use of automatic homogenization procedures should be encouraged, because their performance is more easily controllable. However, even when automatic methods will be much better developed than at present, the best statistical homogenization will still need expert decisions at least in two cases: a) when the number of comparable time series or their spatial correlations are relatively low, b) in the use of certain kinds of metadata.

4.4. Blind tests

The most correct tool for the evaluation of homogenization procedures including subjective steps is the blind test, i.e., when homogenizers do not know the properties of the test series. Naturally, automatic methods may also be incorporated in such tests, and thus, the performances of various homogenization methods are objectively comparable. An appropriate test dataset is not only blind for homogenizers, but also realistic, which means that its properties are similar to the general properties (or at least to the properties of certain kinds) of true data in observational networks and time series. The development of such comparative tests needs wide cooperation of dataset developers, method developers, and homogenizers. In the blind tests of HOME in homogenizing the benchmark, large number of researchers worked together, and thus, HOME substantially improved our knowledge about the performances of homogenization methods. The results are particularly valuable in the homogenization of monthly and annual surface temperature data and in the homogenization of precipitation totals. The HOME tests proved that a) among objective and semi-objective methods the most sophisticated ones based on simple model structure, provide the best performance, namely MASH, PRODIGE, and ACMANT; b) the predominantly subjective homogenization with Craddock-test (*Craddock*, 1979) can compete with any objective method considering the mean performance, but not in the amount of accomplished tasks. Another important finding was that the United States Historical Climate Network homogenization (USHCN, *Menne* and *Williams*, 2009) produced the lowest rate of unnecessary adjustments, while its general performance was only slightly lower than the other best methods. The other methods participated in the HOME tests had significantly poorer performance than MASH, PRODIGE, ACMANT, Craddock and USHCN, therefore, in the final conclusions of *Venema et al.* (2012), these five methods are recommended for practical use. Note that the recently developed

HOMER likely has at least as good performance as the highlighted five, because HOMER adapts the best segments of PRODIGE and ACMANT and applies them in a sophisticated way. Note also that in specific tasks, the enhanced methods do not always show the best performance, e.g. in the example of *Fig. 1* (detection skill for one large shift) the early version of SNHT and PMT perform best.

Naturally, one set of blind tests as it was done under HOME could not answer all questions related to effective homogenization, because the kinds of homogenization tasks are diverse and not restricted to the homogenization of monthly surface temperature and precipitation data. The tasks ahead for the method developers will be discussed in Section 6.

5. Datasets for efficiency examinations

The numerical results of efficiency tests are most meaningful when they are based on the full understanding of the homogenization problem and the nature of inhomogeneities in the climate data, therefore, the use of test datasets of realistic properties is essential. In this section we deal with the construction, selection, and application of appropriate datasets for testing efficiencies. The appropriateness largely depends on the type of the homogenization task, but here some general aspects will be discussed only. In the first part of this section, some general problems of creating realistic test datasets and the properties of benchmark are discussed. In the second part, some examples are shown in which the test datasets do not contain simulated data.

5.1. Datasets of simulated time series

The simulation of time series for surface climatic variables is based on the constructors' knowledge of climatic and non-climatic properties of observed time series. By contrast, in the simulation of upper air data, general circulation models (GCM) are used, since GCM products provide more reliable data for the upper air conditions than for the surface climate. Both ways of dataset construction have advantages and weak points.

It is obvious that the more similar the test dataset to the real observational data, the more reliable conclusions can be drawn from its use. The problem is that we do not know exactly the properties of observed datasets. The last statement might sound strange, because thousands of studies have been devoted to examine and quantify the climatic and non-climatic characteristics (trends, low- and high-frequency variability, change-points, etc.) of observed data. However, the problem is not with the possible lack of scrutiny, but with the nature of data. In nature, magnitudes of inhomogeneities

can be either small or relatively large, and it seems to be a realistic approach that their distribution is normal with 0 mean (*Menne* and *Williams*, 2005; *Venema et al.*, 2012). However, small inhomogeneities cannot be detected (*Fig. 3*). The ratio of detected biases is particularly low for small and medium-size platform-shaped biases, i.e. when the duration of biases is limited (*Fig. 3b*). *Fig. 3* proves that the detection results of homogenization procedures do not provide realistic information neither about the rate of very small biases, nor about the rate of platform-shaped biases.

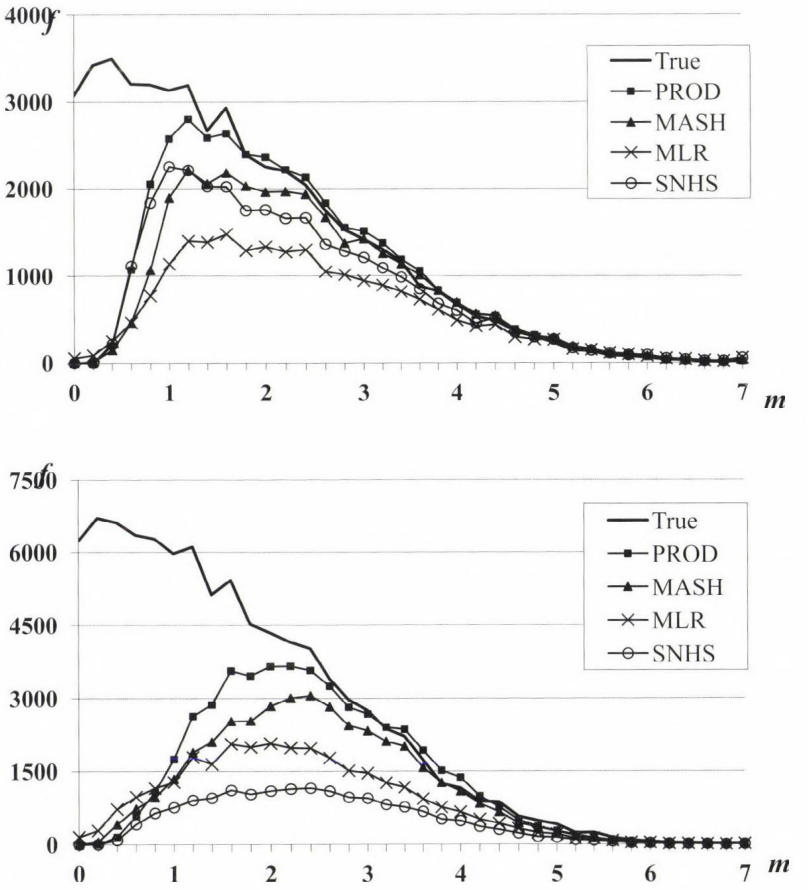

Fig. 3. Frequency (*f*) of detected change-points as a function of shift-magnitudes (*m*), when 5 shifts with random positions (top panel) and 5 platforms (pair of shifts with the same *m* and opposite directions, bottom panel) with random positions are inserted into 100 years long white noise process. The duration of platforms is evenly distributed between 1 and 10 year. *m* is shown in the ratio to the standard deviation of the background noise, while the unit of *f* is arbitrary.

Domonkos (2011a) presented an experiment in which the detection results for true and simulated observational datasets were empirically approached for large number of detection methods and surprisingly high rate of platforms, especially platforms of short duration was reported for the best approach achieved. However, the direct application of that structure of inhomogeneities for construction of test datasets is not recommended, because i) the results are valid for a specific temperature dataset (of Hungary), ii) small, persistent anomalies of short duration in the spatial gradients of a climatic variable may be components of the true climate, even when data of the same climatic region is examined, so that platform-like biases of relative time series may have climatic origin, iii) the mode of generating reference series applied by *Domonkos* (2011a) might have contributed to the amount of apparent small biases for the candidate series. In spite of the uncertainties related to the lately described experiment, it is very likely that the amount of short-term platform-shaped biases in observational time series is much larger than that exists in a simulated test dataset with randomly positioned shifts. This thesis also has non-statistical reasoning: a non-climatic shift and/or its technical cause is often realized after some periods have passed and thereafter, the bias does not appear in the time series, due to the elimination of the technical problem (see also *Rienzner* and *Gandolfi*, 2011; *Domonkos*, 2011a). However, with resetting the technical conditions, observed data are usually not corrected backwards, and even if they are corrected, they might still have systematic bias. We think that the described phenomenon and its consequences on time series properties are general for all observed climatic variables, although the frequency and intensity of platform-shaped biases as well as their impact on the quality of observed time series may substantially differ. Note that the test datasets generated by *Domonkos* (2008, 2011a; etc.) directly mimic relative time series instead of generating raw time series and their differences. This simplification is allowed only when detection segments are tested.

The properties of test datasets may have crucial impact on the observed efficiencies in test experiments (*Caussinus* and *Mestre*, 2004; *Titchner et al.*, 2009; *Domonkos*, 2011a; etc.). Unfortunately, the test dataset properties are often far from the real world in climatological studies, even sometimes the natural spread of shift-magnitudes is missing. In HOME, the benchmark was constructed in a way that it includes realistic climatic signal, the statistical momentums, spatial correlations, and low frequency fluctuations mimic the natural variability of surface temperature and precipitation data in Europe (*Venema et al.*, 2012). The statistical characteristics of inhomogeneities were established with expert decisions of some HOME participants, thus, the frequency and magnitude distributions of biases are likely realistic. However, the frequency of platform-shaped biases in the benchmark is lower than what would follow from the arguments of *Rienzner* and

Gandolfi (2011) and *Domonkos* (2011a). We emphasise that the necessity of inclusion of realistic amount of small biases and platform-shaped biases in test datasets is not because we should be able to detect such inhomogeneities, but because they influence the detection results for the larger and more persistent biases, as it is illustrated in *Fig. 1.*

5.2. Test datasets composed of real data

It was mentioned that the true positions and magnitudes of non-climatic shifts are not exactly known in real observed time series, therefore, efficiency tests usually need the use of simulated datasets. However, under specific conditions, there are some other options for testing efficiencies. The performance of an automatic homogenization method can be tested against a good quality real dataset that has been homogenized with a dense network and/or metadata (*Begert et al.*, 2008). The use of satellite data in the validation of radiosonde data homogenization method has been reported by *Sherwood et al.* (2008), although it must be noted that the homogeneity of satellite data is doubtful due to small temporal biases and calibration problems (*Mears et al.*, 2003). Metadata can be valuable either in the accomplishment or in the validation of homogenization procedures (*Auer et al.*, 2005; *Brunet et al.*, 2006; *Sherwood et al.*, 2008; etc.). Note, however, that sizes of non-climatic biases cannot be quantified from metadata, with few exceptions. This fact reduces the usability of metadata in making quantitative evaluations. Finally, we mention that in testing ANOVA, lists of the timings of detected change-points have been used as test datasets (*Domonkos et al.*, 2012a and Section 4.2. of this study).

6. Tasks for the future

The HOME blind test experiments showed that the differences between the efficiencies of homogenization methods are larger than that was thought earlier when detection parts were examined only. Although most efficiencies obtained in the HOME experiments are positive, some results show the opposite. Consequently, the impact of statistical homogenization on the final quality of observed climatic datasets is often significantly positive, but sometimes nearly neutral or negative. The success depends on the signal to noise ratio (*Ducré-Robitaille et al.*, 2003; *Caussinus* and *Mestre*, 2004; *DeGaetano*, 2006; etc.) and the homogenization method applied (*Venema et al.*, 2012). The blind tests of HOME have brought a large number of valuable new results. Supplying the test results with the details of the historical methodological development of

homogenization methods (*Domonkos et al.*, 2012b), our knowledge has become more complete about some fundamental rules of homogenization. Yet, there are still a large number of open questions that indicate the tasks ahead the developers of homogenization methods.

- We have limited knowledge about the method performances when the signal to noise ratio is not high. HOME results showed the best resistance for USHCN against applying spurious adjustments, but, on the other hand, certain segments of USHCN are suboptimal. These two facts together show that we have not found yet the most appropriate method for treating the cases of moderate signal to noise ratio.
- In HOME, only surface temperature data and precipitation total data were homogenized, and even for these two variables, daily scale homogenization was not included apart from some sporadic examinations.
- Several widely used methods were not tested by HOME, e.g., MLR, the method of *Easterling* and *Peterson* (1995), the family of Bayes methods (*Perreault et al.*, 2000a,b), etc. Most of them have similar statistical structures to the tested methods, therefore, the appearance of substantially new, highly efficient homogenization methods is not envisaged at present. However, some of the methods which were found to be the best in the HOME tests are still under development. ACMANT and Climatol (www.climatol.eu) have newer versions than that were tested in HOME, and the availability of a fully automated MASH version has been reported (www.homogenization.org). HOMER has been developed after the HOME experiments, thus it had not been subjected to the blind tests of HOME. There are promising experiments with developing the detection segment of PRODIGE and HOMER to a network-wide joint segmentation algorithm (*Picard et al.*, 2011). The strategy of USHCN against applying unnecessary adjustments should likely be combined with segments of other homogenization methods of better general performance.

New blind test experiments could produce the largest amount of new and objective information about the performance of homogenization methods. However, blind test experiments such that accomplished under HOME are not economic in costing time, money, and human effort. Perhaps an alternative could be producing an automatic version for each promising homogenization method with subjective steps in a way that default options would be included in them at steps that may incorporate subjective decisions. Its advantage would be that with tests for the automated versions one could easily filter the possible false expectations and common software errors. The weak point of this idea is that it is a

challenge to find relatively simple but intelligent defaults (otherwise, there would no need to subjective steps). Note that at present, the International Surface Temperature Initiative works on developing a benchmark dataset for surface temperature data of all over the world (www.surfacetemperatures.org).

Testing automatic methods is much simpler and more productive than organizing and performing blind tests. On the other hand, the development of homogenization methods is worth some investment. The observed climatic datasets is of huge value to the human society. This value has been accumulated during decades and centuries. The costs of gaining as-optimal-as-possible climatic information from the data via their homogenization are much lower than the costs of many other steps in producing and archiving reliable climatic data.

Acknowledgements–The research was supported by the projects COST ES0601 and EURO4M FP7-SPACE-2009-1/242093. The author thank Constanta Boroneant and Dimitrios Efthymiadis for their contribution to finding the final form of the paper.

References

Aguilar, E., Auer, I., Brunet, M., Peterson, T.C., and *Wieringa, J.*, 2003: WMO Guidelines on climate metadata and homogenization. WCDMP-No. 53, WMO-TD.No:1186, WMO, Geneva.

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

Auer, I., Böhm, R., Jurković, A., Orlik, A., Potzmann, R., Schöner, W., Ungersböck, M., Brunetti, M., Nanni, T., Maugeri, M., Briffa, K., Jones, P., Efthymiadis, D., Mestre, O., Moisselin, J.- M., Begert, M., Brazdil, R., Bochnicek, O., Cegnar, T., Gajić-Čapka, M., Zaninović, K., Majstorović, Ž., Szalai, S., Szentimrey, T., and Mercalli, L., 2005: A new instrumental precipitation dataset for the greater Alpine region for the period 1800–2002. *Int. J. Climatol. 25*, 139–166.

Beaulieu, C., Seidou, O., Ouarda, T.B.M.J., Zhang, X., Boulet, G., and *Yagouti, A.*, 2008: Intercomparison of homogenization techniques for precipitation data. *Water Resour. Res. 44*, W02425.

Beaulieu, C., Seidou, O., Ouarda, T.B.M.J. and *Zhang, X.*, 2009: Intercomparison of homogenization techniques for precipitation data continued: Comparison of two recent Bayesian change point models. *Water Resour. Res. 45*, W08410, pp15.

Begert, M., Zenklusen, E., Häberli, C., Appenzeller, C., and *Klok, L.*, 2008: An automated procedure to detect discontinuities; performance assessment and application to a large European climate data set. *Meteorol. Z. 17*, 663–672.

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

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

Caussinus, H. and *Mestre, O.*, 2004: Detection and correction of artificial shifts in climate series, *J. Roy. Stat. Soc. Series C53*, 405–425.

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

Dai, A., Wang, J., Thorne, P.W., Parker, D.E., Haimberger, L., and *Wang, X.L.*, 2011: A new approach to homogenize daily radiosonde humidity data. *J. Climate 24*, 965–991.

DeGaetano, A.T., 2006: Attributes of several methods for detecting discontinuities in mean temperature series. *J. Climate 19*, 838–853.

Domonkos, P., 2008: Testing of homogenization 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, WMO, Geneva, 126–145.

Domonkos, P., 2011a: Efficiency evaluation for detecting inhomogeneities by objective homogenization methods, *Theor. Appl. Climatol. 105*, 455–467.

Domonkos, P., 2011b: Adapted Caussinus-Mestre Algorithm for Networks of Temperature series (ACMANT). *Int. J. Geosci. 2*, 293–309.

Domonkos, P., 2012: ACMANT: Why is it efficient? In *Proceedings of the 7th Seminar and Quality Control in Climatological Databases*. WMO-HMS, www.c3.urv.cat/publicacions/publicacions2012.html

Domonkos, P., Venema, V. and *Mestre, O.*, 2012a: Efficiencies of homogenization methods: our present knowledge and its limitation. In *Proceedings of the 7th Seminar for Homogenization and Quality Control in Climatological Databases* in press, www.c3.urv.cat/publicacions/publicacions2012.html

Domonkos, P., Venema, V., Auer, I., Mestre, O. and *Brunetti, M.*, 2012b: The historical pathway towards more accurate homogenization. *Adv. Sci. Res. 8*, 45–52.

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

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

Gérard-Marchant, P.G.F., Stooksbury, D.E. and *Seymour, L.*, 2008: Methods for starting the detection of undocumented multiple changepoints. *J. Climate 21*, 4887–4899.

Gruber, C. and *Haimberger, L.*, 2008: On the homogeneity of radiosonde wind time series. *Meteorol. Z. 17*, 631–643.

Haimberger, L., 2007: Homogenization of radiosonde temperature time series using innovation statistics. *J. Climate 20*, 1377–1403.

Lanzante, J.R., Klein, S.A. and *Seidel, D.J.*, 2003: Temporal homogenization of monthly radiosonde temperature data. Part I: Methodology. *J. Climate 16*, 224–240.

McCarthy, M.P., Titchner, H.A., Thorne, P.W., Tett, S.F.B., Haimberger, L., and *Parker, D.E.*, 2008: Assessing bias and uncertainty in the HadAT adjusted radiosonde climate record. *J. Climate 21*, 817–832.

Mears, C.A., Schabel, M.C. and *Wentz, F.J.*, 2003: A reanalysis of the MSU channel 2 tropospheric temperature record. *J. Climate, 16*, 3650-3664.

Menne, M.J. and *Williams Jr., C.N.*, 2005: Detection of undocumented changepoints using multiple test statistics and composite reference series. *J. Climate 18*, 4271–4286.

Menne, M.J. and *Williams Jr., C.N.*, 2009: Homogenization of temperature series via pairwise comparisons. *J. Climate, 22*, 1700–1717.

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

Moberg, A. and *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.

Perreault, L., Bernier, J., Bobée, B., and *Parent, E.* 2000a: Bayesian change-point analysis in hydrometeorological time series. Part 1. The normal model revisited. *J. Hydrology 235*, 221–241.

Perreault, L., Bernier, J., Bobée, B., and *Parent, E.* 2000b: Bayesian change-point analysis in hydrometeorological time series. Part 2. Comparison of chage-point models and forecasting. *J. Hydrology 235*, 242–263.

Picard, F., Lebarbier, E., Hoebeke, M., Rigaill, G., Thiam, B., and *Robin, S.*, 2011: Joint segmentation, calling and normalization of multiple CGH profiles. *Biostatistics 12*, 413–428.

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

Rienzner, M. and *Gandolfi, C.*, 2011: A composite statistical method for the detection of multiple undocumented abrupt changes in the mean value within a time series. *Int. J. Climatol. 31*, 742–755.

Sherwood, S.C., Meyer, C.L., Allen, R.J., and *Titchner, H.A.*, 2008: Robust tropospheric warming revealed by iteratively homogenized radiosonde data. *J. Climate 21*, 5336–5352.

Szentimrey, T., 1999: Multiple Analysis of Series for Homogenization (MASH). In *Second Seminar for Homogenization of Surface Climatological Data (Eds.: Szalai, S., Szentimrey, T. and Szinell, Cs.) WCDMP 41, WMO-TD 962*, WMO, Geneva, 27–46.

Titchner, H.A., Thorne, P.W., McCarthy, M.P., Tett, S.F.B., Haimberger, L., and *Parker, D.E.*, 2009: Critically reassessing tropospheric temperature trends from radiosondes using realistic validation experiments. *J. Climate 22*, 465–485.

Toreti, A., Kuglitsch, F.G., Xoplaki, E., and *Luterbacher, J.*, 2012: A novel approach for the detection of inhomogeneities affecting climate time series. *J. Appl. Meteor. Climatol. 51*, 317–326.

Venema, V., Mestre, O., Aguilar, E., Auer, I., Guijarro, J.A., Domonkos, P., Vertacnik, G., Szentimrey, T., Stepanek, P., Zahradnicek, P., Viarre, J., Müller-Westermeier, G. Lakatos, M., Williams, C.N., Menne, M., Lindau, R., Rasol, D., Rustemeier, E., Kolokythas, K., Marinova, T., Andresen, L., Acquaotta, F., Fratianni, S., Cheval, S., Klancar, M., Brunetti, M., Gruber, C., Duran, M.P., Likso, T., Esteban, P., and *Brandsma, T.*, 2012: Benchmarking monthly homogenization algorithms. *Climate of the Past 8*, 89–115.

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

Williams, C.N., Menne, M.J. and *Thorne, P.W.*, 2012: Benchmarking the performance of pairwise homogenization of surface temperatures in the United States. *J. Geoph. Res. Atmos. 117, D5.*

IDŐJÁRÁS

Quarterly Journal of the Hungarian Meteorological Service
Vol. 117, No. 1, January–March 2013, pp. 113-

Theoretical questions of daily data homogenization

Tamás Szentimrey

Hungarian Meteorological Service,
P.O. Box 38, H-1525 Budapest, Hungary
E-mail: szentimrey.t@met.hu

(Manuscript received in final form December 10, 2012)

Abstract—The so-called variable correction methods form a special type of methods developed for daily data homogenization. Their common assumption is that in case of daily data series, the corrections for inhomogeneity have to vary according to the meteorological situation of each day in order to represent the extremes. In this paper we express our objections to these variable correction methods, especially to their underlying principles. Since the exact theoretical mathematical formulation of the question of daily data homogenization is generally neglected, we also try to formulate and analyze this problem in accordance with mathematical conventions.

Key-words: daily data series, homogenization, climate extremes, higher-order moments, distribution, mathematical formulation

1. Introduction

During the last years, the interest to the daily data series homogenization has increased dramatically. The main reason of this tendency is that daily data are essential for studying extremes of weather and climate, for example, computing extreme climate indices requires reliable daily data series. However, according to numerous climatologists homogenization of daily data is still in its infancy and is much more difficult problem than homogenization at monthly or annual scales. The essence of this argumentation is that the correction in mean is sufficient for monthly and annual series, but in case of daily data series, the corrections should vary according to the meteorological situation of each day in order to represent the extremes. This idea was published in the paper by *Trewin*

and *Trevitt* (1996), where parallel measurements were examined and compared to each other. Since then on the basis of the ideas formulated in the paper, a number of variable correction methods have been developed with the declared aim of being capable of correcting the daily data not only in mean (first moment) but also in the higher order moments. For example, we mention the following methods: higher order moments (HOM) method by *Della-Marta* and *Wanner* (2006) and spline daily homogenization (SPLIDHOM) method by *Mestre et al.* (2011), and there are numerous other similar methods applied in practice. But unfortunately, in this paper we have to make a criticism about these variable correction methods, especially about their underlying principles. In our humble opinion, during the examinations only some physical experiences were considered without any exact theoretical, mathematical formulation of the problem. The empiric interpretation and formulation seem to be a misunderstanding. Moreover, there are some mathematical statements at the description of the methods – e. g., capability to correct the higher order moments – but without any proof, and this practice is of course contrary to the mathematical conventions.

2. Examination of parallel measurements

2.1. Examinations by Trewin and *Trevitt (1996)*

First here is a quotation from the paper of *Della-Marta* and *Wanner* (2006): "One of the most robust methods capable of adjusting the higher-order moments of daily temperature data is that of *Trewin* and *Trevitt* (1996)." *Trewin* and *Trevitt* (1996) intended to homogenize daily data series in order to create composite temperature records. The following sentences are from their paper: "It is therefore necessary to make use of climatological records with inhomogeneities, and to develop a means of removing or minimizing the impact of inhomogeneities on these records. One way of doing this is by adjusting all parts of a record to be comparable with some 'reference period'. Standard procedures for such adjustments in mean temperatures have relied on the implicit assumption that, if two neighbouring stations both have homogeneous records over some period of time, the difference in daily maximum (or minimum) temperature between them will be a constant for any day in a given month of the year. This implies that the difference in monthly means will be a constant for that month from year to year." In general it is not true of course, but after some examination of real station data series they obtained the following result: "This is observed at Armidale (P. Burr, pers. comm.), ..,where the difference in minimum temperature between the town centre site used in this study and a second site approximately 2 km to the east, in the outer part of the town, has a mean value of 1.5 to 2 °C , but can increase to 4 °C on cold, clear

nights. The assumption that the temperature difference between any two nearby sites is always constant must therefore be questioned."
The above conclusion was all right, but the next conclusion is a little bit surprising for us: "The relationship between the temperature characteristics of the two sites in each pair was examined, with the aim of determining an appropriate method for use in extrapolating records at one site to records at the other."
Probably here is the origin of the methods that apply varying corrections per days, and at this step a regression or interpolation problem was obtained for homogenization instead of the adequate distribution problem. Three interpolation techniques were considered by *Trewin* and *Trevitt* (1996) namely: the 'traditional' constant-difference approach, the 'regression' method, and the frequency distribution matching. The methods will be detailed in Section 4.1.

2.2. Mathematical examinations of parallel measurements

What was the reason of the development of the variable correction methods? Essentially, an observed phenomenon at the extremes, namely the differences of parallel measurements are larger in case of extremes. In our opinion, this observed phenomenon has a simple and logical reason, and it is superfluous to look for some complicated physical explanation for the inhomogeneity. The simple reason is that the extremes may be expected at different moments in case of parallel measurements, or in other words, there may be systematic biases in rank order! It is a natural phenomenon, and for illustration a trivial example is presented according to the probability theory.

Example 2.2

Let $Y_1(t) \in N(0,1)$, $Y_2(t) \in N(0,1)$ $(t = 1,2,..,n)$ be standard normally distributed series with expected values $\mathrm{E}(Y_1(t)) = \mathrm{E}(Y_2(t)) = 0$, with standard deviations $\mathrm{D}(Y_1(t)) = \mathrm{D}(Y_2(t)) = 1$, and with correlation between the series $\mathrm{corr}(Y_1(t), Y_2(t)) = \rho$ $(t = 1,2,..,n)$.

Then the mean difference $\mathrm{E}(Y_1(t) - Y_2(t)) = 0$ of course, however, the difference $Y_1(t) - Y_2(t)$ is not independent from the elements $Y_1(t)$, $Y_2(t)$ if $\rho \neq 1$, and, e.g., the conditional expectation of difference $Y_1(t) - Y_2(t)$ given $Y_1(t)$, or equivalently the regression of difference $Y_1(t) - Y_2(t)$ on $Y_1(t)$ is $\mathrm{E}(Y_1(t) - Y_2(t) | Y_1(t)) = (1 - \rho) \cdot Y_1(t)$.

Consequently, the difference $Y_1(t) - Y_2(t)$ is an expectedly monotonous increasing function of $Y_1(t)$ if $\rho \neq 1$. This is the theory, but it can be demonstrated in practice too. We generated such standard normal series by the Monte Carlo method with parameters $\rho = 0.9$, $n = 1000$. In this case,

$E(Y_1(t)-Y_2(t)|Y_1(t))=0.1\cdot Y_1(t)$ and the difference series $Y_1(t)-Y_2(t)$ as a function of series $Y_1(t)$ is plotted in *Fig. 1.*

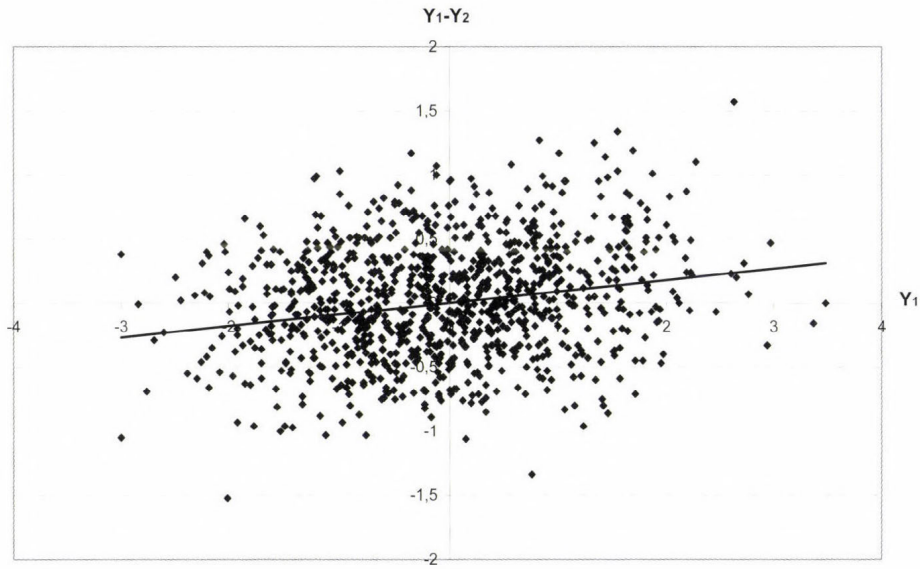

Fig. 1. Difference series $Y_1(t)-Y_2(t)$ as a function of series $Y_1(t)$

It is evident that the conditional expectation of difference $Y_1(t)-Y_2(t)$ is monotonous increasing function of $Y_1(t)$, consequently the difference may be larger mainly in the case of extreme values. It is a general phenomenon not only observed for meteorological measurements. Presumably this experience is the reason for the idea that the correction of daily data should vary according to the meteorological situation of each day, in particular on the basis of some regression models. But it is a misunderstanding of the homogenization problem.

3. Mathematical formulation of the daily data homogenization

Unfortunately, the exact theoretical, mathematical formulation of the problem of homogenization is generally neglected in meteorological studies. Therefore, we try to formulate this problem in accordance with mathematical conventions. First of all it is necessary to emphasize that homogenization is a distribution problem and not a regression one.

Notation

Let us assume we have daily data series:
$Y_1(t)\,(t=1,2,..,n)$: candidate time series of the new observing system.
$Y_2(t)\,(t=1,2,..,n)$: candidate time series of the old observing system.
$1 \le T < n$: change-point, series $Y_2(t)\,(t=1,2,..,T)$ can be used before and series $Y_1(t)\,(t=T+1,..,n)$ can be used after the change-point.

Definition

The aim of homogenization is the adjustment or correction of values $Y_2(t)\,(t=1,2,..,T)$ in order to have the corrected values $Y_{1,2h}(t)\,(t=1,2,..,T)$ with the same distribution as the elements of series $Y_1(t)\,(t=1,2,..,T)$, i.e.:

$$\mathrm{P}(Y_{1,2h}(t)<y)=\mathrm{P}(Y_1(t)<y), \quad y\in(-\infty,\infty)\,,\ t=1,2,..,T\,. \tag{1}$$

Eq. (1) means the equality in distribution: $Y_{1,2h}(t)\overset{d}{=}Y_1(t)\ \ (t=1,2,..,T)$.

Consequence

Within the same climate area, if the variables $Y_1(t), Y_2(t)\,(t=1,2,..,T)$ have identical distribution, i.e., $Y_2(t)\overset{d}{=}Y_1(t)\ \ (t=1,2,..,T)$, then the merged series $Y_2(t)\,(t=1,2,..,T)$, $Y_1(t)\,(t=T+1,..,n)$ is homogeneous.

Example

Let us assume we have parallel measurements $Y_1(t)$, $Y_2(t)\,(t=1,2,..,n)$ within the same climate area with distance 50 m between the locations. Then, as a consequence of micrometeorological processes, the series are probably different, $Y_2(t)\neq Y_1(t)\,(t=1,2,..,n)$, but they may be equal in distribution, $Y_2(t)\overset{d}{=}Y_1(t)$ $(t=1,2,..,n)$. In this case, the mixed series $Y_2(t)\,(t=1,2,..,T)$, $Y_1(t)\,(t=T+1,..,n)$ can be taken as a homogeneous series. This mixed series is equivalent with the homogeneous series $Y_1(t)\,(t=1,2,..,n)$ also in respect of the distribution of extremes.

Returning to the general question, we have to see clearly that the aim of homogenization is to correct the distribution of $Y_2(t)$ according to $Y_1(t)$, instead of the estimation or regression of $Y_1(t)$ on $Y_2(t)$! Moreover, the correction of distribution is equivalent in essence with the correction or adjustment of the moments. The aim of the homogenization expressed in k^{th} moments:

$$m_k = \mathrm{E}\left(\left(Y_{1,2h}(t)\right)^k\right) = \mathrm{E}\left(\left(Y_1(t)\right)^k\right) \quad k = 1,2,\ldots\ ;\ t = 1,2,..,T\,, \tag{2}$$

where E is the usual notation of the expected value or mean equivalently. Some remarkable formulas for the moments:

$$E = m_1\,,\ \ D^2 = m_2 - m_1^2 \tag{3}$$

where E denotes the expected value or mean, and D denotes the standard deviation.

In practice, numerous methods indicate the capability to correct the higher order moments but without any exact proof.

4. The variable correction methods

We return to the methods suggested by *Trewin* and *Trevitt* (1996) which was mentioned in Section 2.1. Essentially, the underlying principles of the variable correction procedures developed later were formulated based on these methods. We do not agree with these principles as explained by our argument in Sections 2.1 and 3, but let us see some details and properties of the mathematical consequences.

4.1. The Trewin and Trevitt (1996) methods for parallel measurements

The short description is cited word for word again from the paper of *Della-Marta* and *Wanner* (2006):

"Trewin and Trevitt (1996) present three different methods to build a composite daily temperature series. Essential to the methods is the existence of simultaneous (in time) observations from the *new* and *old* observing system. These parallel measurements had been taken based on the recommendations of Karl et al. (1995), who suggest that a minimum of a 2-yr overlap between the new and old observing systems be made. In Australia, for example, this practice has only become routine since around 1993 and so many inhomogeneities needed to be adjusted using the traditional constant difference techniques with neighboring reference stations. In this way, Trewin (2001) created a homogenized daily temperature dataset that has subsequently been used by Collins et al. (2000) to assess trends in the frequency of extreme temperature events in Australia.

The three methods they intercompared were constant difference, linear regression, and frequency distribution matching.

The constant difference approach simply adjusted the older data with the newer data using the mean of the daily differences in the simultaneous (parallel) measurements.

The linear regression method fitted a linear model to the difference in daily simultaneous measurements between the two observing systems and the temperature at the older station. This model could then be used to adjust daily temperatures at the older station differentially depending on the temperature, thereby adjusting the higher-order moments.

Their third method determines the frequency distribution of each site during the simultaneous measurement period. The adjustment for each desired percentile is calculated simply as the difference between the two percentiles. This method assumes that there is no systematic bias in the rank order of the temperatures at the two sites.

They show that both the regression method and the frequency distribution matching technique have certain advantages; however, if the homogenization of extreme events is most needed, then their frequency distribution matching technique is more accurate."

Our mathematical comments to the methods are as follows.

4.1.1. Constant difference approach

Yes, this approach is correct if the inhomogeneity is in mean or expected value or first moment, which are the same with different names.

4.1.2. Linear regression method

This procedure is absolutely wrong for homogenization. To demonstrate the problem, a trivial counter-example is presented.

Theorem

Let us assume that the different series $Y_1(t)$, $Y_2(t)$ $(t=1,2,..,n)$ have identical distribution, with expected values $\mathrm{E}(Y_1(t))=\mathrm{E}(Y_2(t))=0$, standard deviations $\mathrm{D}(Y_1(t))=\mathrm{D}(Y_2(t))=1$, and correlation between the series $\mathrm{corr}(Y_1(t),Y_2(t))=\rho$ $(t=1,2,..,n)$.

(i) Then the linear regression of difference $Y_1(t)-Y_2(t)$ on $Y_2(t)$ is $(\rho-1)\cdot Y_2(t)$, consequently, the homogenized series after the suggested adjustment, $Y_{1,2h}(t)=Y_2(t)+(\rho-1)\cdot Y_2(t)=\rho\cdot Y_2(t)$ and $\rho\cdot Y_2(t)$ is just the linear regression of $Y_1(t)$ on $Y_2(t)$.

(ii) Moreover, since the expected values $\mathrm{E}(Y_{1,2h}(t))=\mathrm{E}(Y_1(t))=\mathrm{E}(Y_2(t))=0$, therefore – using Eq. (3) – , the second moment of $Y_{1,2h}(t)$ is equal to the variance $\mathrm{D}^2(Y_{1,2h}(t))=\delta^2<1$, while the common second moment of $Y_1(t)$,

$Y_2(t)$ is equal to the variances $D^2(Y_1(t)) = D^2(Y_2(t)) = 1$. Therefore, the second moment was decreased from 1 to $\delta^2 < 1$ during the regression.
Summing up, according to (i) this procedure is equivalent with the simple linear regression of $Y_1(t)$ on $Y_2(t)$. Furthermore, according to (ii) the following statement about the method is absolutely false: "This model could then be used to adjust daily temperatures at the older station differentially depending on the temperature, thereby adjusting the higher-order moments." The truth is just the opposite, since the correct second moment was damaged at our counter-example.

4.1.3. Frequency distribution matching technique

The main problem is the following assumption which is the fundament of the method: "This method assumes that there is no systematic bias in the rank order of the temperatures at the two sites."
Unfortunately, the reality and the mathematics are much more complicated, and the above assumption cannot be accepted as it is demonstrated in *Fig. 1.* The bias in rank order depends on the stochastic connection, and there may be systematic bias, since $Y_1(t)$, $Y_2(t)$ are not monotonous increasing functions of each others. At this method, the adjusted $Y_{1,2h}(t)$ is obtained essentially by a simple exchange $Y_2(t)$ for $Y_1(t)$ according to the rank orders. Why? For example, if $Y_1(t)$, $Y_2(t)$ were equal in distribution then such an exchange would not be necessary.

4.2. The general type of variable correction methods applied in the practice

On the basis of the former principles described in Sections 4.1.2 and 4.1.3 (regression and frequency distribution matching), a number of variable correction methods have been developed during the last years. The new improvement of these methods is that they do not need overlap observations, instead of this they use information from nearby reference stations, for example higher order moments (HOM) method by *Della-Marta* and *Wanner* (2006) and spline daily homogenization (SPLIDHOM) method by *Mestre et al.* (2011). We do not want to criticize the details of these methods however, we express again our skepticism on their common fundamental principles which were based on a pseudo problem demonstrated in *Example 2.2.* Moreover, we repeat the following sources of errors for consideration.

- The assumption of the frequency distribution matching technique, i.e., there is no systematic bias in the rank, cannot be accepted.

- The regression methods are not adequate to correct the higher order moments.

Our last remark is connected also with the higher order moments. In general, the papers about these methods indicate the capability to correct the higher order moments, but this statement is always without any exact mathematical proof. We are skeptic, however if somebody could send us a nice proof, we would be grateful for it.

5. Some remarks about the homogenization in the higher-order moments

We suggest to consider the following remarks when developing homogenization methods with the capability to correct also the higher order moments.

Remark 1

There is a common assumption that the correction in mean is sufficient for monthly and annual series, and that the correction of higher order moments is necessary only in the case of daily data series. In general, it is tacitly assumed that the averaging is capable to filter out the inhomogeneities in the higher order moments. However, this assumption is false, for example, if there is an inhomogeneity in the standard deviation of daily data, we may have the same inhomogeneity in monthly data.

Proof

Daily data are $X(t)\ (t=1,2,..,30)$, monthly average is $\overline{X}=\frac{1}{30}\sum_{t=1}^{30}X(t)$.

Let us introduce an inhomogeneity in the standard deviation for the daily data:
$X_{ih}(t)=\alpha\cdot(X(t)-\mathrm{E}(X(t)))+\mathrm{E}(X(t)),\ (t=1,2,..,30)$.
The expected value is unchanged: $\mathrm{E}(X_{ih}(t))=\mathrm{E}(X(t))$, but the standard deviation has changed: $\mathrm{D}(X_{ih}(t))=\alpha\cdot\mathrm{D}(X(t))$.

Let us see the new monthly average: $\overline{X}_{ih}=\frac{1}{30}\sum_{t=1}^{30}X_{ih}(t)$.

The expected value is unchanged: $\mathrm{E}(\overline{X}_{ih})=\mathrm{E}(\overline{X})$, but the standard deviation changed with the same measure:

$$\mathrm{D}(\overline{X}_{ih})=\mathrm{D}\left(\frac{1}{30}\sum_{t=1}^{30}X_{ih}(t)\right)=\mathrm{D}\left(\frac{1}{30}\sum_{t=1}^{30}\alpha\cdot X(t)\right)=\alpha\cdot\mathrm{D}\left(\frac{1}{30}\sum_{t=1}^{30}X(t)\right)=\alpha\cdot\mathrm{D}(\overline{X}).$$

Remark 2

The correction in the first two moments or, equivalently, in mean and standard deviation can be formulated by using the notations defined in Section 3 as follows:

$$Y_{1,2h}(t) = E_1(t) + \frac{D_1(t)}{D_2(t)}\left(Y_2(t) - E_2(t)\right) \quad (t = 1,2,..,T), \tag{4}$$

where $E_1(t) = \mathrm{E}(Y_1(t))$, $E_2(t) = \mathrm{E}(Y_2(t))$ are the means, and $D_1(t) = \mathrm{D}(Y_1(t))$, $D_2(t) = \mathrm{D}(Y_2(t))$ are the standard deviations. Then $\mathrm{E}(Y_{1,2h}(t)) = E_1(t)$, $\mathrm{D}(Y_{1,2h}(t)) = D_1(t)$.

In general, the detection of the change points and the estimation of correction factors are suggested to be based on the examination of monthly data series because of the larger signal to noise ratio.

Remark 3

If the joint distribution of the series is normal, $Y_1(t) \in N\left(E_1(t), D_1(t)\right)$, $Y_2(t) \in N\left(E_2(t), D_2(t)\right)$ $(t = 1,2,..,n)$ and $Y_{1,2h}(t)$ $(t = 1,2,..,T)$, calculated according to Eq. (4), then $Y_{1,2h}(t), Y_1(t)$ $(t = 1,2,..,T)$ have identical distribution: $Y_{1,2h}(t) \overset{d}{=} Y_1(t)$ $(t = 1,2,..,T)$. Consequently, the mixed series $Y_{1,2h}(t)$ $(t = 1,2,..,T)$, $Y_1(t)$ $(t = T+1,..,n)$ is homogeneous, that means it is sufficient to correct only the first two moments in case of joint normal distribution.

Proof

Owing to Remark 2 and the joint normal distribution, $Y_{1,2h}(t) \in N\left(E_1(t), D_1(t)\right)$ $(t = 1,2,..,T)$.

6. Conclusion

It is necessary to define the exact mathematical theory for homogenization of climate data series. Homogenization is a probability distribution problem, and the methods applied in practice should be theoretically evaluated in this respect.

References

Della-Marta, P.M. and *Wanner, H.*, 2006: A Method for homogenizing the extremes and mean of daily temperature measurements. *J. Climate 19*, 4179–4197.

Mestre, O., Gruber, C., Prieur, C., Caussinus, H., and *Jourdain, S.*, 2011: Splidhom: a method for homogenization of daily temperature observations. *J. Appl. Meteor. Climatol. 50*, 2343–2358.

Szentimrey, T., 2008: Development of MASH homogenization procedure for daily data, *Proceedings of the Fifth Seminar for Homogenization and Quality Control in Climatological Databases,* Budapest, Hungary, 2006; *WCDMP-No. 68, WMO-TD NO. 1434*, 116–125.

Trewin, B.C. and *Trevitt, A.C.F.*, 1996: The development of composite temperature records. *Int. J. Climatol. 16*, 1227–1242.

IDŐJÁRÁS

Quarterly Journal of the Hungarian Meteorological Service
Vol. 117, No. 1, January–March 2013, pp. 123–141

Experiences with data quality control and homogenization of daily records of various meteorological elements in the Czech Republic in the period 1961–2010

Petr Štěpánek[1,2] *, **Pavel Zahradníček**[1,2] and **Aleš Farda**[2]

[1]*Czech Hydrometeorological Institute, regional office, Kroftova 43, 616 67 Brno, Czech Republic*

[2]*Global Change Research Centre AS CR, v.v.i, Bělidla 986/4a, 603 00 Brno, Czech Republic*

**Corresponding author E-mail: petr.stepanek@chmi.cz*

(Manuscript received in final form December 14, 2012)

Abstract–Quality control and homogenization has to be undertaken prior to any data analyses in order to eliminate any erroneous values and non-climatic biases in time series. In recent years, considerable attention was paid to daily data since it can serve, among other conventional climatological analyses, as non-biased input into extreme value analysis, correction of RCM outputs, etc. In this work, we describe and then apply our own approach to data quality control of station measurements, combining several methods: (i) by analyzing difference series between candidate and neighboring stations, (ii) by applying limits derived from interquartile ranges, and (iii) by comparing the series values tested with "expected" values – technical series created by means of statistical methods for spatial data (e.g., IDW, kriging). Because of the presence of noise in series, statistical homogeneity tests render results with some degree of uncertainty. In this work, the use of various statistical tests and reference series made it possible to increase considerably the number of homogeneity test results for each series and, thus, to assess homogeneity more reliably. Inhomogeneities were corrected on a daily scale. In the end, missing values were filled applying geostatistical methods; thus, the so-called technical series for stations were constructed, which can finally be used as quality input into further time series analysis. These methodological approaches are applied to daily data, for various meteorological elements within the area of the Czech Republic in the period 1961–2010, which allows demonstrate their usefulness. Series were processed by means of the developed ProClimDB and AnClim softwares (http://www.climahom.eu).

Key-words: data quality control, homogenization, statistical correction of inhomogeneities, daily data processing, climatological time series

1. Introduction

For any meaningful climate analysis, investigated time series should be homogeneous, which means that their variations are caused solely by variations in weather and climate (*Conrad* and *Pollak*, 1950). Thus, prior to any analyses, the need to homogenize data and check their quality arises. Unfortunately, most of the climatological series that span from decades to centuries, contain inhomogeneities caused by station relocations, change of observers, changes in the vicinity of the stations (e.g., urbanization), changes in instruments, observing practices (e.g., different formulas for calculating daily means, different observation times), etc. (*Aguilar et al.*, 2003). Another important requirement for climatological analyses is the quality of the individual values, where series should be free of errors and have a low number of missing values (*Vicente-Serrano et al.*, 2010).

In the Czech Republic, this topic has been a focus of interest for several years. The first studies devoted to the homogenization of long series of air temperature, precipitation, and relative humidity for individual stations (e.g., *Macková*, 1997; *Brázdil* and *Štěpánek*, 1998; *Brázdil et al.*, 1996, 2000, 2001), which makes their use difficult (their availability, purpose of the given study, etc.). Later, studies devoting to the whole county have emerged (*Štěpánek*, 2003; *Štěpánek* and *Mikulová*, 2009; *Štěpánek et al.*, 2009), and this interest has continued up to the time of this study dealing with all the basic climatological characteristics throughout the whole territory of the Czech Republic. In recent years, considerable attention has also been devoted to the analysis of daily data (e.g., *Klein Tank et al.*, 2002; *Vincent et al.*, 2002; *Wijngaard et al.*, 2003; *Brunet et al.*, 2006; *Brandsma* and *Können*, 2006; *Della-Marta* and *Wanner*, 2006; *Vicente-Serano et al.*, 2010), which then may be used for various analyses, including those that were not possible to apply when homogenizing only monthly data, like the analysis of extreme value (*Sacré et al.*, 2007; *Kyselý* and *Picek*, 2007; *Costa* and *Soares*, 2009).

The organization of meteorological observations (i.e., maintaining the station network) and administration of collected data belong among the main duties of the Czech Hydrometeorological Institute (CHMI). The climatological database CLIDATA (*Tolasz*, 2008) serves very well for the usual quality control of the collected data (using GIS), but for the historical records, we face a lack of human resources, since the system requires user input. The software tools ProClimDB and AnClim (*Štěpánek*, 2010a, 2010b) allow automation of the process for quality control, homogenization, and filling of missing data but, at the same time, give the user a variety of outputs from which he can easily read what happened during the data processing and can track back all the important changes made to data, and also he can change the parameters and re-run the calculation. Thus, the ProClimDB and AnClim complement the aforementioned CHMI database system very well. However, it can also be used as a stand-alone application.

The general scheme of data processing that we advise being performed before any analyses includes the detection, verification, and possible correction of outliers (at the sub-daily scale – using the measured values), creation of reference series, homogeneity testing applying various statistical tests to better account for uncertainties in the results, determination of inhomogeneities in the light of test results and metadata (in monthly, seasonal and annual scale), adjustment of inhomogeneities (in daily scale) and, finally, the filling of missing values (see *Fig. 1*). We applied this approach to various meteorological elements available in our area in the period 1961–2010: mean, maximum and minimum air temperature, precipitation totals, water vapor pressure, wind speed and sunshine duration.

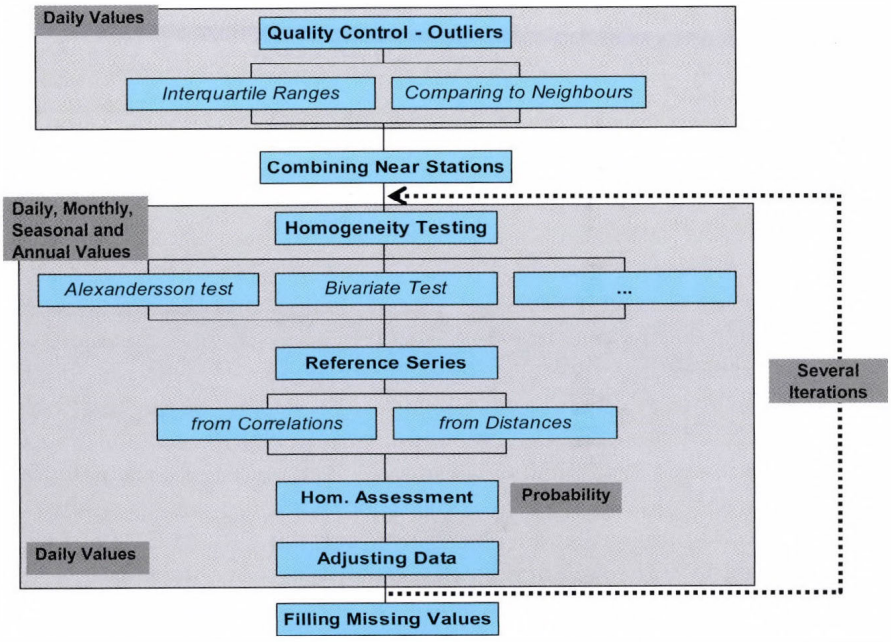

Fig. 1. Scheme of data processing – data quality control and homogenization.

2. Data Quality control

There is a lack of a generally accepted methodology for data quality control (contrary to homogenization). But, without outliers being properly treated, homogenization and analysis may render misleading results. Therefore, we devoted considerable attention to the methodology of outlier detection, to something that could, moreover, be automated to process large datasets of daily/sub-daily values (whole country dataset). This quality control was then

applied to historical records which had not yet been processed by the methods used nowadays (using GIS and user interaction).

In our approach, data quality control is carried out by combining several methods:

(i) analyzing series of differences between candidate and neighboring stations (i.e. pairwise comparisons);

(ii) applying limits derived from interquartile ranges (either to individual series, i.e. absolutely, or better, to series of the difference between candidate and reference series, i.e. relatively); and

(iii) comparing the series of tested values with "expected" (theoretical) values – "technical" series created by means of statistical methods for spatial data (e.g. IDW, kriging).

Neighboring stations (method, (i)) or reference series (method, (ii)) may be selected either by correlations or distances (in the case of temperature, the results are different, while for precipitation, the selection coincides). Correlation coefficients can be applied either to raw series or to series of first differences (see, e.g., *Peterson*, 1998). In our case, for comparison with neighbor stations, up to six of the nearest stations were selected, with significant correlation coefficients, a distance limit of 400 km and an altitude difference restricted to 500 m. The distance limit was set with the help of a preceding analysis about how correlation coefficients drop with distance and change in altitude.

A method for outlier detection that could be automated to the greatest extent was a priority, since millions of values had to be processed for each meteorological element. Such a method was finally found and successfully applied. It utilizes a combination of several characteristics and their limits are based on the methods mentioned above (details on the quality control process may also be found in the documentation for the ProClimDB software, see *Štěpánek*, 2010b). No method on its own was found adequate; only their combination leads to satisfying results, i.e. the discovery of real outliers and suppression of false alarms. Parameters (settings appropriate to methods) had to be individually found for each meteorological element. The setting of parameters for outlier detection was validated using stations selected within different parts of the Czech Republic and also representing different altitudes.

As for the number of found suspicious values, the wind speed seems to be the most problematic variable, while air temperature has a relatively low number of problematic values (see *Fig. 2*). The number of outliers has a clear annual cycle. For most of the elements (e.g., air temperature), a higher number of outliers was detected in the summer months than in the winter months (larger spatial differences in summer are related to the increased influence of radiation factors compared to winter patterns, prevailingly influenced by circulation factors). More outliers were detected in the morning (7:00 local mean time –

LMT) and evening (19:00 LMT) observation terms compared to 14:00 LMT (associated with steeper gradients in the former case). For precipitation, there are two maxima per year: in the summer months and in December–January (this pertains to problems with solid precipitation measurements in winter), while during spring and autumn, a lower number of outliers were detected. The number of detected outliers also changes with time. The higher number of temperature outliers since the late 1990s coincides well with the transition to automatic measurements. Our explanation is that all values coming from automatic measurements (including errors) are stored directly into the database, while in the case of manual measurements, the observer revises the measured values before sending them to CHMI. In the last years, the number of outliers is again lower, owing to improved data quality control in the database. Conversely, in the case of precipitation, no increase of errors after automation was encountered (*Fig. 3*).

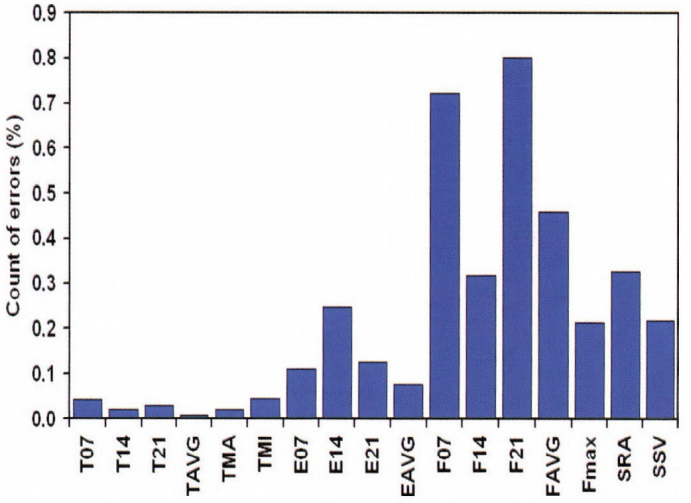

Fig. 2. Percentage portions of the number of errors detected in the total number of tested values for meteorological stations in the territory of the Czech Republic in the 1961–2009 period. Explanations: T – air temperature (T07, T14, T21 – observation terms, TAVG – daily mean), TMA – daily maximum air temperature, TMI – daily minimum air temperature, E – water vapor pressure (E07, E14, E21 – observation terms, EAVG – daily mean), F – wind speed (F07, F14, F21 – observation terms, FAVG – daily mean), Fmax – maximum daily wind gust, SRA – daily precipitation total, SSV – daily sunshine duration.

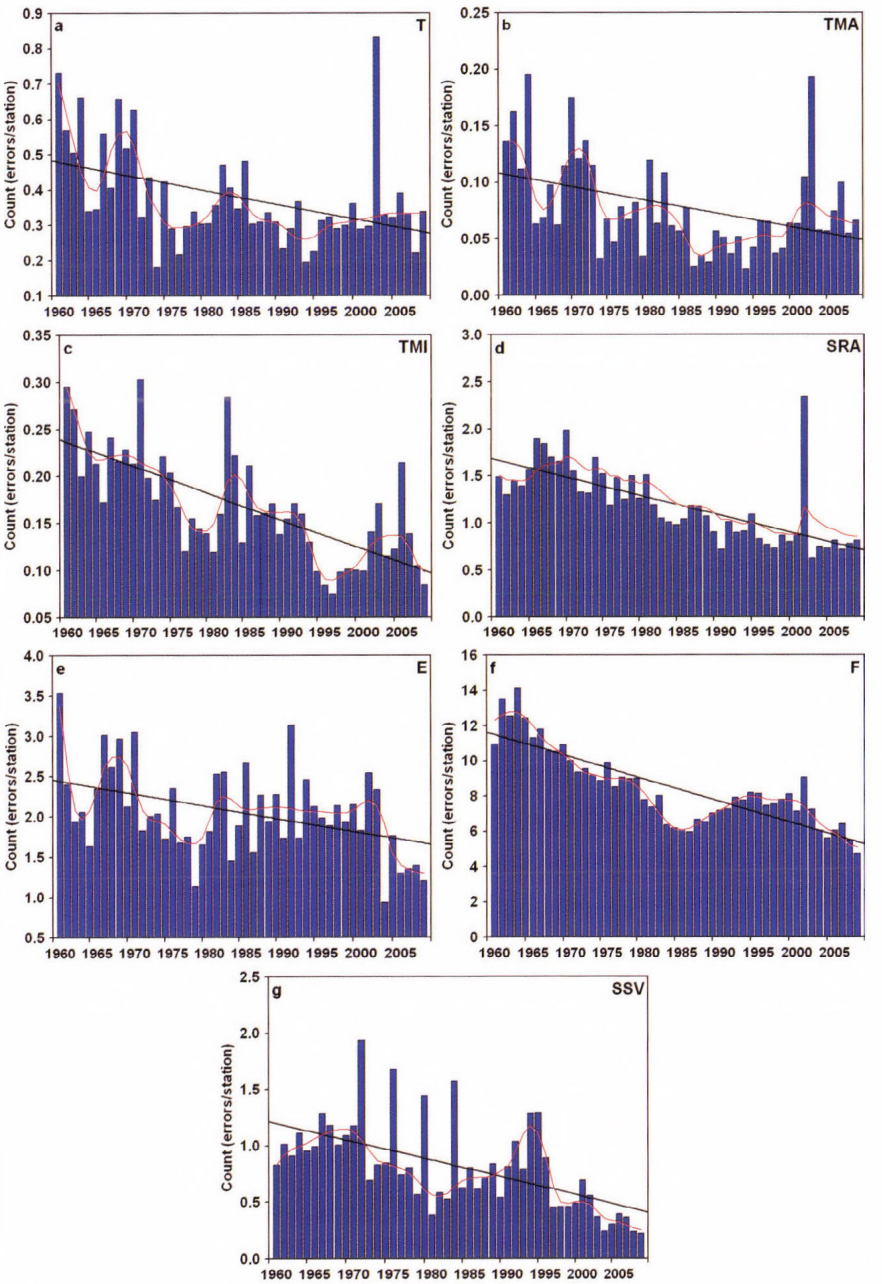

Fig. 3. Number of detected problematic values re-calculated per one meteorological station in the territory of the Czech Republic in the individual years of the 1961–2009 period: a) air temperature (observation terms and daily mean), b) maximum air temperature, c) minimum air temperature, d) precipitation total, e) water vapor pressure (observation terms and daily mean), f) wind speed (observation terms and daily mean), g) sunshine duration. The values are smoothed with a low-pass Gaussian filter for 10 years (red line) and complemented by the linear trend.

3. *Methodology of homogenization*

The general steps to be taken during homogenization consist of reference series creation (serving for comparison with tested series; this is a principal point of relative homogenization, see, e.g., *Conrad* and *Pollak*, 1950), applying statistical tests for testing the homogeneity of candidate series, homogenization (correction of inhomogeneities detected) and filling missing values (some prefer to fill missing values before homogenization). The individual steps are discussed, e.g., in *Štěpánek et al.* (2012), including a comparison of the results for various parameter settings (methods of weighting, number of stations used, individual statistical tests applied, method of correlation calculation for selection of neighbors, etc.)

Because of noise in time series, statistical homogeneity tests render results with some degree of uncertainty (see *Fig. 4*). In this work, the use of various statistical tests, types of reference series and time frames (monthly, seasonal, and annual series) allowed a considerable increase in the number of homogeneity test results for each series tested and thus to assess the homogeneity more reliably.

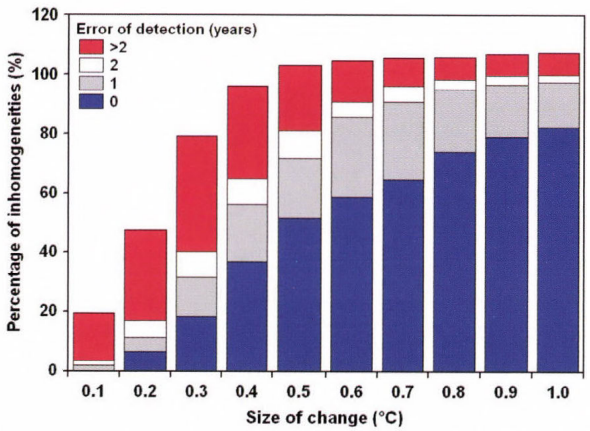

Fig. 4. The relative proportion (%) of the number of detected inhomogeneities of various sizes in the theoretically possible number of all inhomogeneities detected by the Alexandersson's SNHT test for the significance level of $\alpha = 0.05$. Generated series shorter than 50 years with annual standard deviation were used. Zero false detection (blue) corresponds to the exact inhomogeneity estimation in the given year; further false detections are given for 1, 2, or more years apart (grey, white, and red). A total of 180 series were used for each category of the inhomogeneity size in mean (shift). The proportion of inhomogeneities exceeding 100% is due to dividing series into more parts during the testing.

The relative homogeneity tests applied were as follows: the standard normal homogeneity test [SNHT] (*Alexandersson*, 1986, 1995), the Maronna and Yohai

bivariate test (*Potter*, 1981), and finally, the Easterling and Peterson test (*Easterling* and *Peterson*, 1995). Reference series were calculated as weighted means from the five nearest stations (measuring within the same period as the candidate series, they were also newly applied individually), with statistically significant correlations, a distance limit of 300 km, and an altitude difference limit of 500m. The weight (inverse distance) for temperature was taken as one and for precipitation as three. Neighbouring station values were standardized to the mean and standard deviation of the candidate station. The detection of inhomogeneities was performed for series divided into a maximum duration of 40 years, with an overlap for two consecutive periods of 10 years (due to requirements of SNHT to test only one shift in a series). The tests were applied for series of monthly, as well as seasonal and annual means (totals in the case of precipitation and sunshine duration).

The main criterion for determining a year of inhomogeneity was the probability of detection of a given year, i.e., the ratio between the count of detections for a given year from all test results for a given station (using type of reference series, range of tests applied, monthly, seasonal, and annual series) and the count of all theoretically possible detections (for more details of reference series creation and testing, see *Štěpánek et al.*, 2012).

After the evaluation of detected breaks and a comparison with metadata, a final decision on the correction of inhomogeneities was made. Data were corrected on a daily scale. The adjustment of such inhomogeneities was addressed by means of a reference series calculated in a similar way as described above.

We created our own correction method, an adaptation of a method for the correction of regional climate model outputs by *Déqué* (2007), itself based on assumptions similar to those implicit in methods described by *Trewin* and *Trevitt* (1996) and *Della-Marta* and *Wanner* (2006), which apply variable correction according to individual percentiles (or deciles). Our process is based on a comparison of percentiles (empirical distribution) of differences (or ratios) between candidate and reference series before and after a break. Percentiles are estimated from candidate and references series separately (not for the same date). Each month is processed individually, but the values of adjacent months before and after are also taken into account to ensure smoother passage from one month to another. Differences of candidate and reference series for individual percentiles are treated before and after a break and smoothed by low-pass filter to obtain a final adjustment based on a given percentile (see *Fig. 5* for illustration). Values before a break are then adjusted in such a way that we find a value for the candidate series before a break (interpolating between two percentile values if needed) and the corresponding correction factor, which is then applied to the values to be adjusted. Special treatment is needed for extremes at the ends of distribution. A comparison of the DAP approach and the "classic" one using monthly values is shown in *Fig. 6*.

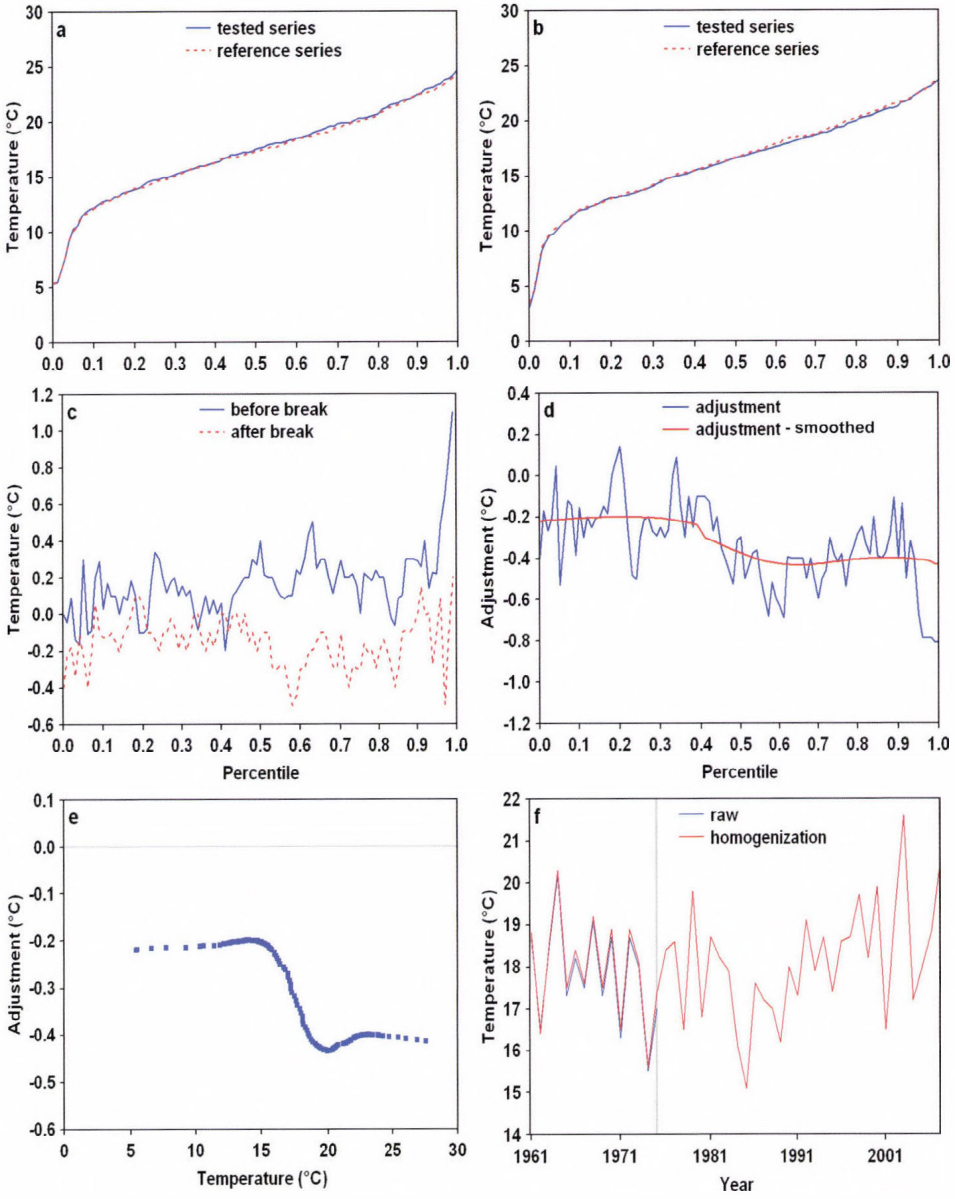

Fig. 5. Adjustment of series of daily mean air temperatures from the Velké Pavlovice station with an inhomogeneity detected in April 1975: a) quantiles (empirical distribution) of tested and reference series before the break, b) quantiles (empirical distribution) of tested and reference series after the break, c) the difference between tested and reference series for quantiles before and after the break, d) values of adjustments for quantiles (difference of tested and reference series differences before and after the break) and their smoothing by low-pass filter (final value of adjustments), e) adjustment values for a specific air temperature, f) original and homogenized series (monthly means).

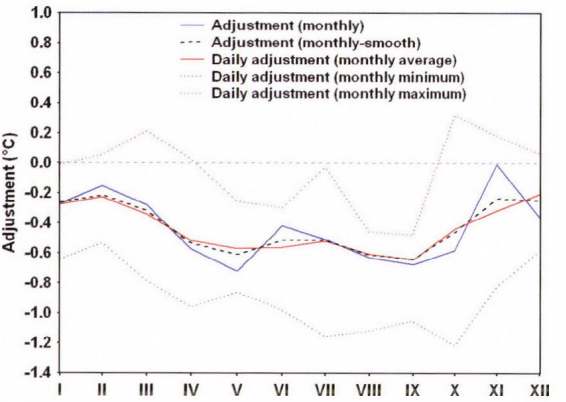

Fig. 6. Example of the series adjustment for inhomogeneity applying the classic approach with use of monthly data (solid blue line; smoothed corrections: dashed black line) and monthly means of daily corrections using DAP – distribution adjusting by percentiles method (solid red line; dotted red (blue) lines show the monthly maxima (minima) from daily corrections). The series is mean daily air temperature measured at station Bystřice pod Hostýnem with inhomogeneity on Januar 1, 1985.

Various characteristics were analyzed before applying the adjustments: the increment of correlation coefficients between candidate and reference series after adjustments; any change of standard deviation in differences before and after the change; the presence of linear trends, etc. In the event of any doubt, the adjustments were not applied.

Homogeneity testing, evaluation and correction of inhomogeneities detected were performed by several iterations, in which more precise results are gradually obtained. Missing values were filled after the homogenization and adjustment of inhomogeneities in the series. This means that the new values filled are estimated from data which are not influenced by possible shifts in the series. Filling missing data before homogenization may negatively influence inhomogeneity detection.

A preference for testing individual observation term series, if available, belongs among the recommendations for further homogenization improvement, since inhomogeneities are manifested in a different way in them (see *Fig. 7*). Further improvement can potentially be achieved by grouping values into categories, e.g., using weather types and testing individual categories alone.

Within the COST Action ES0601 ("Advances in homogenization methods of climate series: an integrated approach – HOME", 2006–2011), parameter settings used in this work were verified (at least for creation of reference series and detection of inhomogeneities that were run in the monthly mode) for air temperature and precipitation. The best parameter settings for air temperature were achieved applying a probability of detection equal to 20%, using correlations for neighbors selection calculated from the first differenced series,

pairwise comparison with neighbors, and running several iterations of homogeneity testing and correction. The second and third iterations improved the series the best, while further iterations meant only negligible improvement. As for precipitation parameter settings, the best results were gained applying probability of detection equal to 10%, using correlations for neighbors selection calculated from the first differenced series, and pairwise comparison with neighbors; however, improvement of statistical characteristics of the series after homogenization was not as profound as in the case of air temperature. These settings will be applied in the follow-up work dealing with historical records (before 1961).

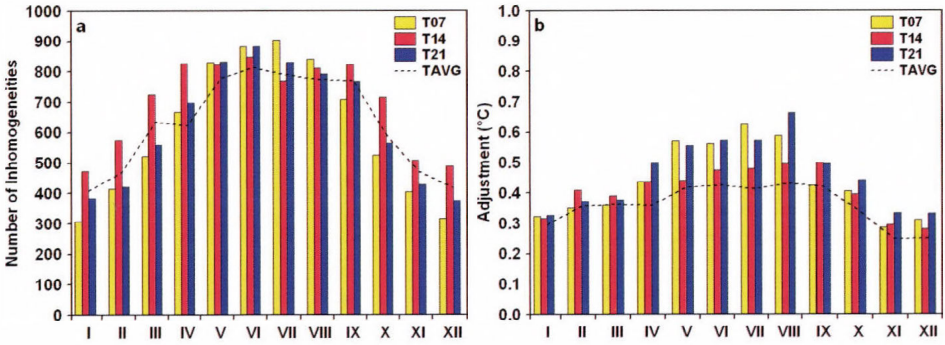

Fig. 7. a) Annual variation of the number of statistically significant inhomogeneities detected in air temperature series of observation terms (T07, T14, T21) and daily mean (TAVG) (Alexandersson's test – SNHT, bivariate test, reference series calculated using distances and correlations); b) annual variation of the size of corrections for individual observation terms and daily mean. Data refer to 230 stations analyzed in the Czech Republic and Slovak Republic in the 1961–2005 period.

4. Homogenization results for the Czech Republic

In the 1961–2007 period, 1750 series of seven climatological characteristics were tested and some inhomogeneities were found in 42% of them (*Table 1*). This value is underestimated, due to the low number of detections in precipitation series, in which breaks were detected in only 15% of series. For all other characteristics, this number is above 50%. The number of detected inhomogeneities varies according to the meteorological element (*Fig. 8*). For homogenization, just as for data quality control, the most problematic meteorological element is wind speed, where 75% of series were detected as inhomogeneous. Wind speed is a very specific meteorological element because, before automation, which took place since about 2000, it was estimated subjectively by observers using the Beaufort wind force scale.

Table 1. Number of breaks detected at meteorological stations in the Czech Republic in the 1961–2007 period for selected characteristics of meteorological variables: T – mean air temperature, TMA – maximum air temperature, TMI – minimum air temperature, SRA – precipitation total, E – mean water vapor pressure, F – mean wind speed, SSV – sunshine duration.

Meteorological element	Number of series	Number of series with break	Ratio (%)	Number of breaks in series			
				0	1	2	3
T	181	100	55,2	81	77	21	2
TMA	178	122	68,5	56	88	32	2
TMI	179	92	51,4	87	68	23	1
SRA	761	117	15,4	644	110	7	0
E	173	123	71,1	50	83	34	6
F	176	132	75,0	44	85	39	8
SSV	102	55	53,9	47	49	5	1
Total	1750	741	42,3	1009	560	161	20

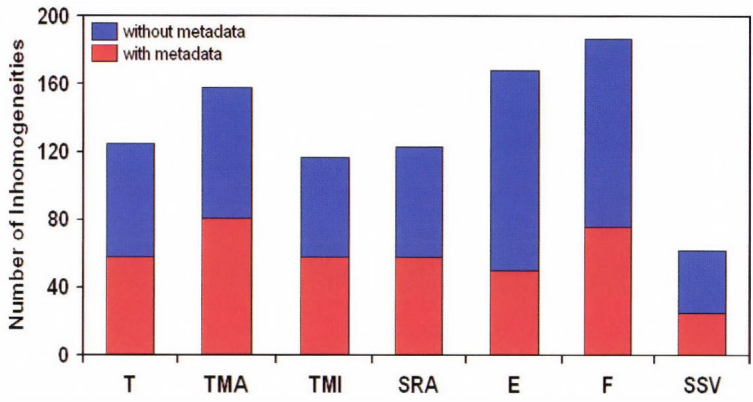

Fig. 8. Number of corrected inhomogeneities of selected characteristics of meteorological variables at stations in the territory of the Czech Republic in the 1961–2007 period: T – mean air temperature, TMA – maximum air temperature, TMI – minimum air temperature, SRA – precipitation total, E – mean water vapor pressure, F – mean wind speed, SSV – sunshine duration (the number of series tested for the individual characteristics is given in *Table 1*). Explanations for inhomogeneities: red – clarified by metadata, blue – no metadata.

For monthly values of air temperature and precipitation over the Czech Republic, the correlation coefficients between candidate and reference series are very high (median above 0.95 or 0.90, respectively; note that the rain-gauge station network is much denser than the climatological one). Along with mean wind speed, correlations are also very high in the case of the other characteristics.

As for inhomogeneity detection itself, more breaks occur in the summer months for air temperature and sunshine duration (the influence of relocation and other artificial changes is greater, resulting from the influences of the active surface, such as prevailing radiation factors and increased volume of vegetation), while for precipitation, it appears in the winter months (mainly due to problems associated with the measurement of solid precipitation). Water vapor pressure and wind speed do not show such a clear annual cycle (*Fig. 9*).

An annual variation is also clearly manifested in the correction of inhomogeneities. Considering the absolute values of corrections, the number of adjustments was higher during the summer months for temperature characteristics and water vapor pressure. After corrections, air temperature correlation coefficients increased mainly in the summer months and those for precipitation in the winter months. The largest increase in correlation coefficient after homogenization was observed in the case of wind speed.

The knowledge of metadata is an import factor for the proper correction of detected inhomogeneities. Out of all corrected breaks, 44% can be explained by metadata (*Fig 8*). There are some differences in the size of corrections according to the causes of the inhomogeneity: the size of correction was higher for inhomogeneities explained by metadata for all characteristics except minimum temperature and sunshine duration, where the mean size of corrections was similar to the case of missing metadata, and for precipitation, where it is even lower (however, for precipitation, only a small percentage of breaks can be detected). As it is evident from the results, the automation of measurements had a very strong influence on the homogeneity of series, as well as on the occurrence of outliers: for mean and maximum temperature and water vapor pressure, the size of the corrections was higher in the case of automatic measurements than the mean of over-all corrections. For example, the inhomogeneities in the series of maximum air temperature caused by automation are higher on average by 0.1°C than breaks not confirmed by metadata. Because the automation of measurements was introduced in the CHMI station network successively from the mid-1990s (see *Fig. 10*), it was possible to detect and make corrections without major problems.

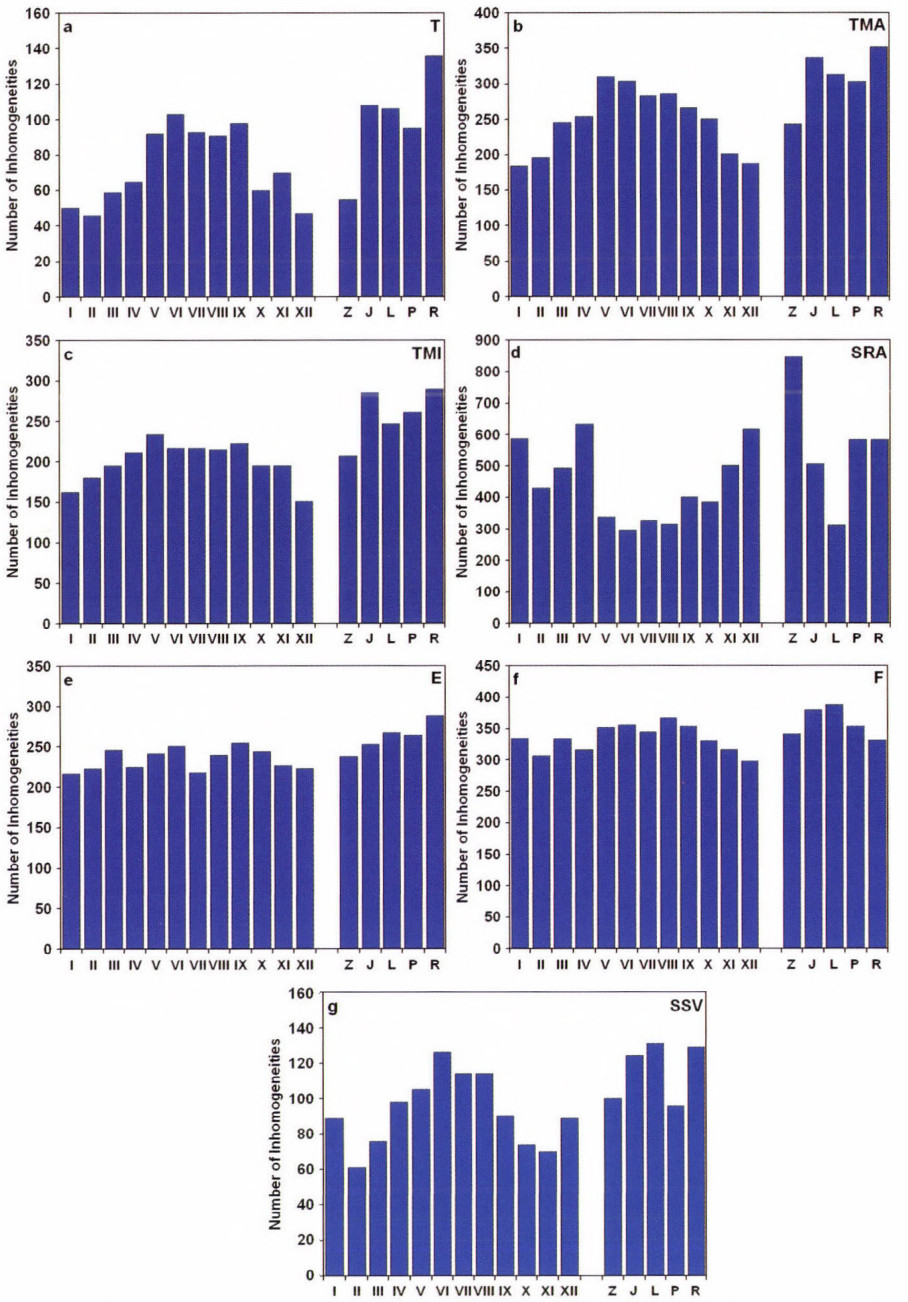

Fig. 9. Annual variation in the number of detected statistically significant inhomogeneities (α = 0.05) for selected climatological and rain-gauge stations in the territory of the Czech Republic in the 1961–2007 period: a) mean air temperature, b) maximum air temperature, c) minimum air temperature, d) precipitation total, e) mean water vapor pressure, f) mean wind speed, g) sunshine duration (Z – winter, J – spring, L – summer, P – autumn; R – year).

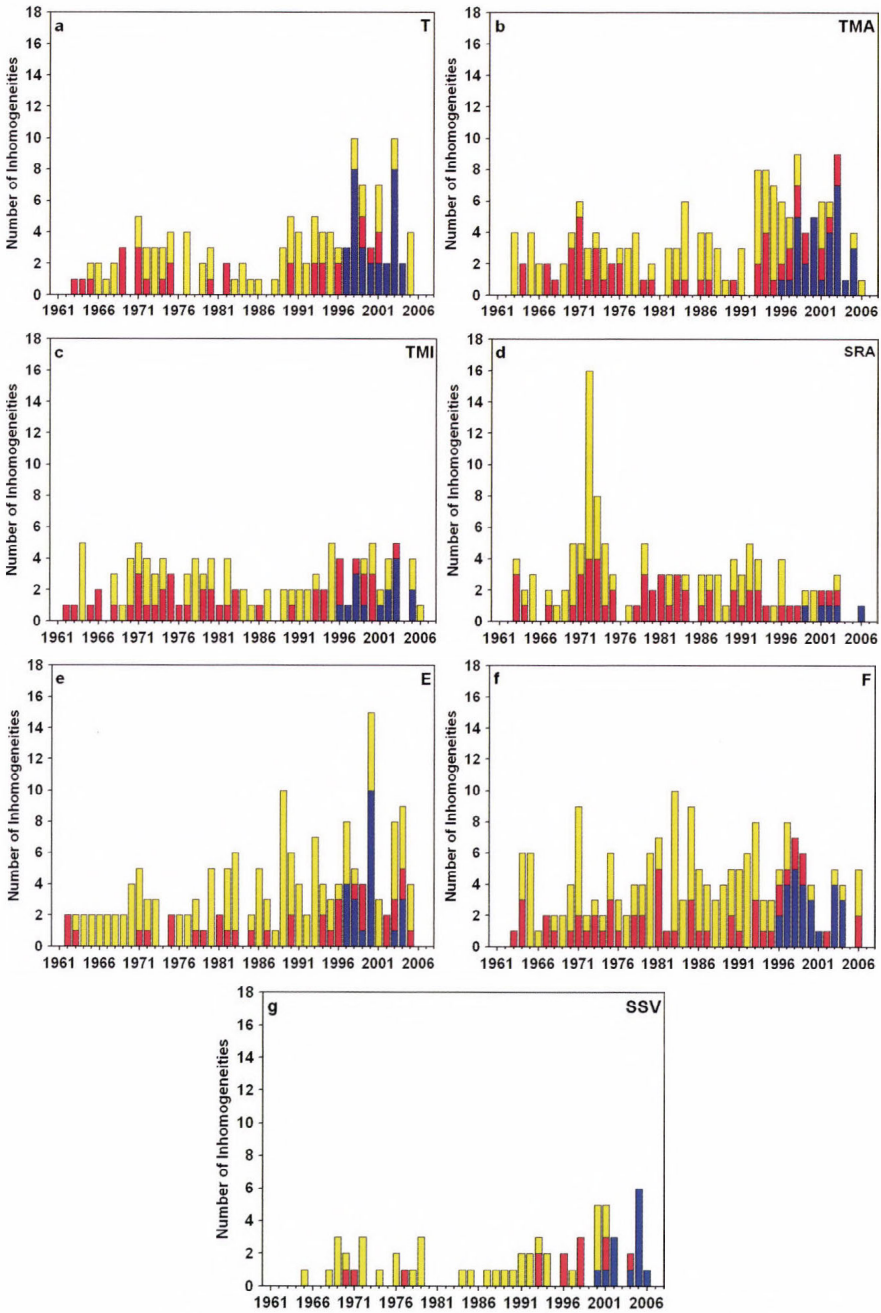

Fig. 10. Number of inhomogeneities detected in the series of climatological and rain-gauge stations in the territory of the Czech Republic in the 1961–2007 period: a) mean air temperature, b) maximum air temperature, c) minimum air temperature, d) precipitation total, e) mean water vapor pressure, f) mean wind speed, g) sunshine duration (yellow – break without metadata, red – break with metadata, blue – break with automation of measurements); AMS express a change to automatic measurements.

5. *Technical series*

Data quality control, homogenization and filling missing values lead to the creation of the so-called "technical" series for mean, maximum and minimum temperatures, precipitation totals, sums of sunshine duration, mean water vapor pressure, and wind speed. Such series may be used for further data analysis, because their values are consistent and complete over a given period. They were calculated for 268 climatological and 787 rain-gauge stations of the CHMI network in the 1961–2010 period, and actual values are continually added. Despite the fact that a smaller number of stations were available for some of the studied climatological characteristics (e.g. 196 stations for sunshine duration or 257 stations for water vapor pressure), "technical" series were completely calculated (for arbitrary station location or regular gridded network). In this way, we have a complex set of meteorological variables for each position of climatological stations, which can easily be further used (e.g., for evapo-transpiration calculation).

The possibility of calculating "technical" series for new positions, either in irregular or regular network, e.g. for grid points of regional climate model (RCM) outputs, allow their use for validation and correction of RCM outputs in each grid point. In the case of the RCM ALADIN-Climate/CZ (*Farda et al.*, 2010), series were calculated with 10x10 km resolution, specifically for 789 grid points over the Czech Republic (*Fig. 11*). The method for the "technical" series calculation is similar to the calculation of theoretical values during the data quality control (for more details, see, e.g., *Štěpánek et al.*, 2011).

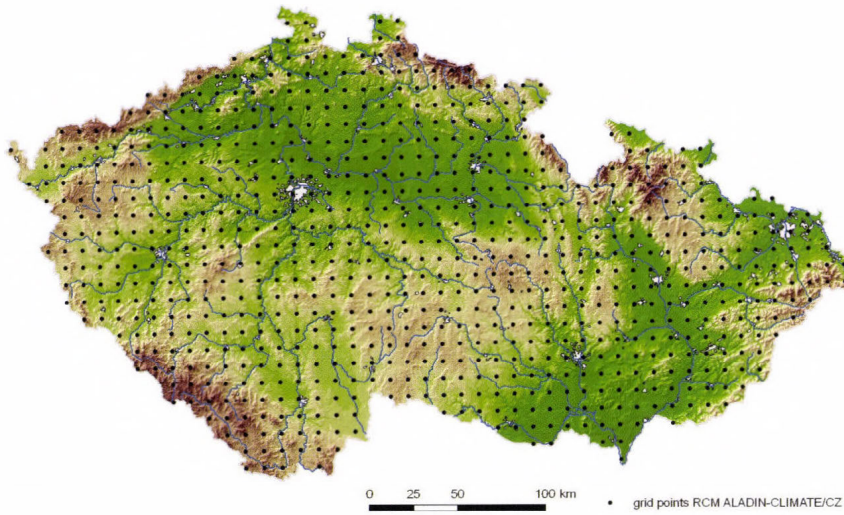

Fig. 11. Grid points of the outputs of RCM ALADIN-Climate/CZ for which "technical" series for the 1961–2009 period were calculated.

6. *Conclusions*

In the Czech Republic, experience with data quality control and homogenization has existed for several years. For data processing, software packages AnClim (*Štěpánek*, 2010a), LoadData and ProClimDB (*Štěpánek*, 2010b) were created. They offer complex solution, from tools for handling databases, through data quality control to homogenization of time series, as well as time series analyses, extreme value evaluation, and model output verification.

In this work, we summarize the effort and methodology behind outlier detection, series homogenization, and interpolation techniques for various climatological characteristics in the territory of the Czech Republic in the 1961–2010 period. In total, over 62 million values were data quality checked, for which the automation of the process was crucial. The final results are acceptable only because of the combination of several methods. The approach became part of the ProClimDB software (*Štěpánek*, 2010b). For correct outlier detection, it is necessary to work directly with measured values in the standard observing terms (e.g. 7:00, 14:00, 21:00 LMT), since possible errors can be masked in the "aggregated" values (daily means, monthly means, or sums).

Similar to the quality control, the aim of the created software for homogenization was to provide the user with support information for making quick, efficient, and correct decisions. Thanks to the COST Action ES0601 benchmark dataset, various parameter settings were checked in the software and recommended: finding neighbor stations using correlations of series of the first differences, performing homogeneity testing individually with each of the neighbors (pairwise comparison), weighting of the number of detected inhomogeneities for the homogeneity evaluation (weights of five for annual values, two for seasonal values, and one for monthly values). The correction of inhomogeneities was performed on a daily basis using our own approach (DAP –distribution adjusting by percentiles).

Quality control and correction of inhomogeneities have been performed on a daily (sub-daily) basis for all key meteorological variables over the territory of the Czech Republic in the 1961–2010 period. The homogenization of data before 1961 has only been carried out on the monthly basis so far (*Štěpánek*, 2003).

Due to the "technical" series calculated in both station and various regular grid point locations, we have gained a sufficiently large number of climatological series for subsequent analysis. These series are free of detectable outliers and inhomogeneities, have had their gaps filled, and are being applied to research in various projects in climatology and hydrology. From the "technical" series, we have also created maps for various meteorological variables for each month and day in the period of 1961–2010 (i.e., more than 130,000 maps).

Further steps will lead to the processing of series of individual observation terms and daily historical records before 1961, as well.

Acknowledgement–This paper was prepared with financial support from the Grant Agency of the Czech Republic for project No. P209/10/0605. The present work was also supported by the CzechGlobe project – the Centre for Global Climate Change Impacts Studies, Reg.No. Cz. 1.05/1.1.00/02.0073.

References

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

Alexandersson, A., 1995: Homogenity testing, multiple breaks and trends. In: *Proceedings of the 6th International Meeting on Statistical Climatology*, Galway, Ireland, 439–441.

Aguilar, E., Auer, I., Brunet, M., Peterson, T.C., and *Wieringa, J.*, 2003: Guidelines on Climate Metadata and Homogenization. WCDMP-No. 53, WMO-TD No. 1186. World Meteorological Organisation, Geneva.

Brandsma, T. and *Können, G.P.*, 2006: Application of nearest-neighbour resampling for homogenizing temperature records on a daily to sub-daily level. *Int. J. Climatol. 26*, 75–89.

Brázdil, R. and *Štěpánek, P.*, 1998: Kolísání teploty vzduchu v Brně v období 1891–1995. *Geografie – Sborník České geografické společnosti 103*, 13–30, (in Czech).

Brázdil, R., Štěpánek, P., and *Budíková, M.*, 1996: Homogenized air temperature series in Brno, 1891–1994. Zeszyty Naukowe Uniwersytetu Jagiellońskiego, MCLXXXXVI, *Prace Geograficzne 102*, 85–91.

Brázdil, R., Štěpánek, P., and *Květoň, V.*, 2000: Air temperature fluctuation in the Czech Republic in the period 1961–1999. Instytut Geografii UJ – *Prace Geograficzne 107*, 173–178.

Brázdil, R., Štěpánek, P., and *Květoň, V.*, 2001: Temperature series of the Czech Republic and its relation to Northern Hemisphere temperatures in the period 1961–1999. In: *Detecting and Modelling Regional Climate Change*(eds.: Brunet India, M., López Bonillo, D.), Springer, Berlin, 69–80.

Brunet, M., Saladie, O., Jones, P., Sigro, J., Aguilar, E., Moberg, A., Lister, D., Walther, A., Lopez, D., and *Almarza, C.*, 2006: The development of a new dataset of Spanish Daily Adjusted Temperature Series (SDATS) (1850–2003). *Int. J. Climatol. 26*, 1777–1802.

Conrad, V., Pollak, L.W.,1950: Methods in Climatology. Harvard University Press, Cambridge.

Costa, A.C. and *Soares, A.*, 2009: Trends in extreme precipitation indices derived from a daily rainfall database for the south of Portugal. *Int. J. Climatol. 29*, 1956–1975.

Della-Marta, P.M. and *Wanner, H.*, 2006: A method of homogenizing the extremes and mean of daily temperature measurements. *J. Climate 19*, 4179–4197.

Déqué, M., 2007: Frequency of precipitation and temperature extremes over France in an anthropogenic scenario: model results and statistical correction according to observed values. *Global Planet. Change 57*, 16–26.

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

Farda, A., Déqué, M., Somot, S., Horányi, A., Spiridonov, V., and *Tóth, H.* ,2010: Model ALADIN as Regional Climate Model for Central and Eastern Europe. *Studia Geophysica et Geodetica 54*, 313–332.

Klein Tank, A. M. G., Wijngaard, J. B., Können, G. P., Böhm, R., Demarée, G., Gocheva, A., Mileta, M., Pashiardis, S., Hejkrlik, L., Kern-Hansen, C., Heino, R., Bessemoulin, P., Müller-Westermeier, G., Tzanakou, M., Szalai, S., Pálsdóttir, T., Fitzgerald, D., Rubin, S., Capaldo, M., Maugeri, M., Leitass, A., Bukantis, A., Aberfeld, R., van Engelen, A. F. V., Forland, E., Mietus, M., Coelho, F., Mares, C., Razuvaev, V., Nieplova, E., Cegnar, T., López, J. A., Dahlström, B., Moberg, A., Kirchhofer, W., Ceylan, A., Pachaliuk, O., Alexander, L. V., and *Petrovic, P.*, 2002: Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. *Int. J. Climatol 22*, 1441–1453.

Kyselý, J. and *Picek, J.*, 2007: Regional growth curves and improved design value estimates of extreme precipitation events in the Czech Republic. *Climate Res. 33*, 243–255.

Macková, J., 1997: Homogenizace dlouhých teplotních řad v České republice. Geografický projekt. Katedra geografie PřF MU, Brno. (in Czech)

Peterson, T.C., Easterling, D.R., Karl, T.R., Groisman, P., Nicholls, N., Plummer, N., Torok, S., Auer, I., Boehm, R., Gullett, D., Vincent, L., Heino, R., Tuomenvirta, H., Mestre, O., Szentimrey, T., Salinger, J., Førland, E. J., Hanssen-Bauer, I., Alexandersson, H., Jones, P., and *Parker, D.*, 1998: Homogeneity adjustments of in situ atmospheric climate data: A review. *International J. Climatol. 18*, 1493–1517.

Potter, K.W., 1981: Illustration of a new test for detecting a shift in mean in precipitation series. *Mon. Weather Rev. 109*, 2040–2045.

Sacré, C., Moisselin, J.M., Sabre, M., Floria, J.P., and *Dubuisson, B.*, 2007: A new statistical approach to extreme wind speeds in France. *J. Wind Engineer. Indust. Aerodyn. 95*, 1415–1423.

Štěpánek, P., 2003: Homogeneización de las series de temperatura del aire en la República Checa durante el período instrumental. *Geographicalia 43*, 5–24.

Štěpánek P., 2010a: AnClim – software for time series analysis. Department of Geography, Faculty of Natural Sciences, MU, Brno, 1.47 MB (http://www.climahom.eu/AnClim.html).

Štěpánek, P., 2010b: ProClimDB – software for processing climatological datasets. CHMI, regional office Brno (http://www.climahom.eu/ProcData.html).

Štěpánek, P. and *Mikulová, K.*, 2009: Homogenization of air temperature and relative humidity monthly means of individual observation hours in the area of the Czech and Slovak Republic. In: (Eds: Lakatos, M., Szentimrey, T., Bihari, Z., Szalai, S.) *Proceedings of the Fifth Seminar for Homogenization and Quality Control in Climatological Databases* (Budapest, Hungary, 29 May – 2 June 2006). WCDMP-No. 71. World Meteorological Organization, Geneva, 149–164.

Štěpánek, P., Řezníčková, L., and *Brázdil, R.*, 2009: Homogenization of daily air pressure and temperature series for Brno (Czech Republic) in the period 1848–2005. In: (Eds: Lakatos, M., Szentimrey, T., Bihari, Z., Szalai, S.) *Proceedings of the Fifth Seminar for Homogenization and Quality Control in Climatological Databases* (Budapest, Hungary, 29 May – 2 June 2006). WCDMP-No. 71. World Meteorological Organization, Geneva, 107–122.

Štěpánek, P., Zahradníček, P., and *Huth, R.*, 2011: Interpolation techniques used for data quality control and calculation of technical series: an example of Central European daily time series. *Időjárás 115*, 87–98.

Štěpánek, P., Zahradníček, P., Brázdil, R., and *Tolasz, R.*, 2012: *Metodologie kontroly a homogenizace časových řad v klimatologii* (Methodology of quality control and homogenisation of time series in climatology - in Czech). ČHMÚ (in print) (in Czech)

Tolasz, R., 2008: *Databázové zpracování klimatologických dat*. Sborník prací ČHMÚ 52. Český hydrometeorologický ústav, Praha. (in Czech)

Trewin, B.C. and *Trevitt, A.C.F.*, 1996: The development of composite temperature records. *International Journal of Climatology, 16*, 1227–1242.

Vicente-Serrano, S., Beguería, S., Lopez-Moreno, J. I., García-Vera, M. A., and *Stepanek, P.*, 2010: A complete daily precipitation database for northeast Spain: reconstruction, quality control, and homogeneity. *Int. J. Climatol. 30*, 1146–1163.

Vincent, L.A., Zhang, X., Bonsal, B.R., and *Hogg, W.D.*, 2002: Homogenization of daily temperatures over Canada. *J. Climate 15*, 1322–1334.

Wijngaard, J.B., Klein Tank, A.M.G., and *Können, G.P.*, 2003: Homogeneity of 20th century European daily temperature and precipitation series. *Int. J. Climatol. 23*, 679–692.

AIDŐJÁRÁS

Quarterly Journal of the Hungarian Meteorological Service
Vol. 117, No. 1, January–March 2013, pp. 143–158

Creation of a homogenized climate database for the Carpathian region by applying the MASH procedure and the preliminary analysis of the data

Mónika Lakatos[1]*, Tamás Szentimrey[1], Zita Bihari[1], and Sándor Szalai[2]

[1]*Hungarian Meteorological Service,*
P.O. Box 38, H-1525 Budapest, Hungary

[2]*Department on Soil Sciences and Agrochemistry,*
Szent István University,
Páter Károly utca 1, H-2100 Gödöllő, Hungary

**Corresponding author E-mail:lakatos.m@met.hu*

(Manuscript received in final form December 12, 2012)

Abstract–Homogenization of the long term observation series is essential in climate change studies. The most important achievements of the COST Action ES0601 (HOME) are survey and the comparison of the available homogenization methods. A benchmark test was performed in the Action to choose the best recent methods. The MASH (Multiple Analysis of Series for Homogenization; Szentimrey) procedure which was developed at the Hungarian Meteorological Service (OMSZ) produced good results. The Short Term Scientific Missions (STSMs) supported by the COST established the wide usage of MASH in the neighboring countries. This is the main reason why MASH became the common homogenization method used to fulfil the Climate of the Carpathian Region tender service. The aim of the project is to improve the climate data source and data access in the Carpathian Region by creating a daily harmonized gridded dataset during the period between 1961 and 2010. The homogenization process executed and the verification of the homogenization along with the quality control results are introduced in this paper. Preliminary results of trend analysis carried out on the harmonized database are also presented.

Key-words: COST Action ES0601, homogenization, Climate of the Carpathian Region Project, climate indices

1. Introduction

Climate change is expected to result in significant changes in the Carpathian region to affect ecosystems and human activities *(UNEP, 2007)*. Investigation of the recent tendencies in the regional climate conditions is essential for coping with the consequences. It is essential that studying the spatio-temporal changes can be implemented through the analysis of the observations which are representative both in time and space. Climate change studies require long term, quality controlled, homogenized, high quality climate data series.

The COST (European Cooperation in Science and Technology) Action ES0601 titled "Advances in homogenization methods of climate series: an integrated approach (HOME)" focused on investigation of the homogenization methods and testing the recent used applications. Hungary contributed to the success of the COST HOME action with the experiences of the MASH (Multiple Analysis of Series for Homogenization; *Szentimrey*, 2011) homogenization procedure, which was developed at the Hungarian Meteorological Service. The main features of MASH are illustrated in this paper by its application in the framework of the "Climate of the Carpathian Region Project" (CarpatClim).

As result of a Hungarian initiative on creation a high quality dataset covering the Carpathian basin, the JRC (European Commission Joint Research Centre) Institute for Environment and Sustainability launched a tender call in 2010 for supplying the data demand of its Desert Action activity (*JRC,* 2010). The consortium lead by the Hungarian Meteorological Service together with 10 partner organizations from 9 countries in the Carpathians region is supported by the Joint Research Centre.

The main aim of the CarpatClim project is to improve a joint climate data source and data access in the Carpathian region for application such kind of regional climate studies like drought monitoring. The CarpatClim project investigates fine temporal and spatial structures of the climate in the Carpathian Mountains and the Carpathian basin with unified methodology. The results are 0.1°(~10 × 10 km) resolution gridded daily time series of various meteorological parameters from 1961 to 2010. The target area is partly includes the territory of Czech Republic, Slovakia, Poland, Ukraine, Romania, Serbia, Croatia, Austria, and Hungary (*Fig. 1*). Uniform process of data homogenization is crucial due to the fact that significant differences might be occurred between the measurements and data handling of participant countries during the examined fifty-year-long period.

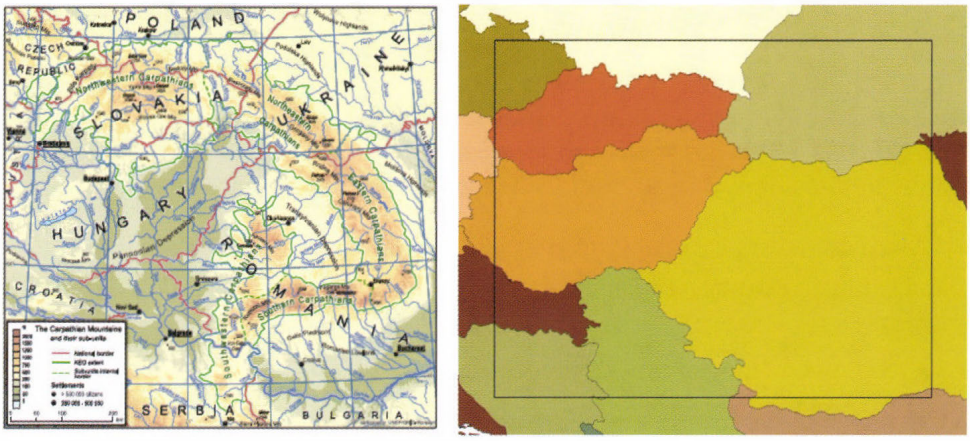

Fig. 1. The target area of the CarpatClim between latitudes 50°N and 44°N, and longitudes 17°E and 27°E approximately (left), and the political boundaries (right).

The project plan consists of three modules. Module 1 focuses on improving the availability and accessibility of homogeneous and spatially representative time series of climate data for the Carpathian Region through data rescue, quality control, and data homogenization. The activities in Module 2 ensure data harmonization with special emphasis on cross-border harmonization and production of gridded values for each country. A digital Climate Atlas as a basis for climate assessment and further applied climatological studies are developed in Module 3. The final outcome of the CarpatClim are the quality controlled, homogenized, in-situ daily time series and gridded data per country and the whole region as well, including a metadata catalogue with the documentation of the existing homogenized datasets. The daily grids with the metadata will be freely accessible for scientific purposes.

The consortium members agreed that the commonly used method for data homogenization and quality control in the project will be the MASH procedure. Using MASH is advantageous, because the COST HOME Action monthly benchmark results denoted that the MASH is one of the best monthly homogenization methods (*Venema et al,* 2012). Furthermore, several COST HOME delegates from the Carpathian region became familiar with the MASH software during STSMs supported by the COST.

The CarpatClim project is a well-accomplished cooperation for applying a single homogenization method in a region fragmented by boundaries and a pioneer work for countervailing against differences in measuring practice and strict data policies. The main features of MASH and the steps of the homogenization process along with the evaluation of the homogenization performed in Module 1 are presented in this study.

2. *Methodology*

The original MASH (*Szentimrey*, 1999) procedure was developed for homogenization of monthly series. The present version: MASHv3.03 (*Szentimrey*, 2011) has been expanded for daily series as well. The main features of the applied procedure to fulfil the tender service are summarized here.

The MASHv3.03 (*Szentimrey*, 2011) software consists of two parts.
Part 1: Quality control, missing data completion, and homogenization of monthly series:

- Relative homogeneity test procedure.
- Step by step procedure: the role of series (candidate or reference series) changes step by step in the course of the procedure.
- Additive (e.g., temperature) or multiplicative (e.g., precipitation) model can be used depending on the climate elements.
- Providing the homogeneity of the seasonal and annual series as well.
- Metadata (probable dates of break points) can be used automatically.
- Homogenization and quality control (QC) results can be evaluated on the basis of verification tables generated automatically during the procedure.

Part 2: Homogenization of daily series:

- Based on the detected monthly inhomogeneities.
- Including quality control (QC) and missing data completion for daily data. The quality control results can be evaluated by test tables generated automatically during the procedure.

These attributes are favorable to achieve the project goals in CarpatClim. The time resolution of variables is daily, the upgraded version of MASH is able to homogenize and control these daily data as well. Certain recently used daily homogenization methods take the monthly results for daily homogenization similarly to MASH (*Vincent*, 2002; *Szentimrey*, 2008). The excellent COST HOME monthly benchmark results and promising outcomes of the daily tests guarantee the high quality of times series got through the MASH procedure.

It has to underline that MASH is an automatically working software. Application of manual homogenization methods would be exceptionally labor intensive in handling huge data series. Moreover, the MASH is able to use the metadata (the date of moving of stations for example) automatically during the break point detection. This facility allows the effective usage of the existing metadata. We note that metadata were not used in CarpatClim.

Furthermore, the test results of the homogenization and quality control (e.g., detected errors, degree of inhomogeneity of the series system, number of break points, estimated corrections, and certain verification results) are documented in automatically generated tables during the homogenization process. Summary results of quality control and the homogenization performed in the project can be followed up and reported based on these tables. Verification statistics can be added to the homogenized series as the newly created metadata.

3. Homogenization process in the CarpatClim

The tasks in Module 1 are the data rescue, the digitization of the analogue datasets of climate observations, quality checking, including the data gap elimination of the existing climate time series, and homogenization of the data series. Completing the digitization of the measurements using MASHv3.03 is a proper way to perform the homogenization and the data quality control.

According to the tender specification, the elements listed in *Table 1* have to be homogenized in the period of 1961–2010. The chosen homogenization model is depending on the distribution of given element. Additive model is used except in case of precipitation and wind speed, where the appropriate model is multiplicative.

Table 1. Set of meteorological variables in daily temporal resolution to be homogenized (JRC, 2010)

Variable	Description	Units
T_a	2 m mean daily air temperature	°C
T_{min}	Minimum air temperature from 18:00 to 06:00	°C
T_{max}	Maximum air temperature from 06:00 to 18:00	°C
p	Accumulated total precipitation from 06:00 to 06:00	mm
DD	10 m wind direction	0°-360°
VV	10 m horizontal wind speed	m/s
Sunshine	Sunshine duration	hours
cc	Cloud cover	tenths
R_{global}	Global radiation	MJ/m^2
RH	Relative humidity	%
p_{vapour}	Surface vapour pressure	hPa
p_{air}	Surface air pressure	hPa

To ensure the most possible station usage, each contributor executed the necessary work phases individually. The cross border harmonization is guaranteed by bilateral data exchange. As the MASH is a relative homogenization method, the candidate series have to be compared to reference series which are in the nearby, within a given distance.

3.1. Steps of creation of the homogenized station data series in the CarpatClim

I. Compilation of the raw station data series of each country.
1. Selection of the stations (with the help of spherical coordinates: φ, λ).
2. Collecting the daily station data series (missing data are allowed) and the metadata per countries. Exchange of the near border raw data series and the existing metadata between the neighboring countries.

II. Homogenization, quality control, data completion of the station data series by MASH v3.03 on national level, using near border data.

1. Derivation of monthly station data series from the daily station data series collected in step I.2. Homogenization, quality control, data completion of the monthly station data series. Metadata (probable dates of break points) can be used automatically.
2. Daily station data series (step I.2): homogenization, quality control, data completion. This procedure is based on the results of step II.1.
3. Exchange of the near border homogenized data for cross-border harmonization and for gridding (Module 2 of the project: modeling, interpolation).
4. Evaluation of the verification results of the homogenization and quality control. Controlling of the cross-border harmonization of the data series. Note that further cross-border harmonization is achieved after the modeling part of the gridding procedure in Module 2.

Summary of the main steps of homogenization of daily data series with quality control and missing data completion in CarpatClim are as follows:

1. Monthly series derivation from daily series.
2. MASH homogenization procedure for monthly series, estimation of monthly inhomogeneities. (Metadata can be used automatically.)
3. Smooth estimation of daily inhomogeneities on the basis of estimated monthly inhomogeneities.
4. Automatic correction of daily series.
5. Automatic quality control (QC) of homogenized daily data.

6. Automatic missing daily data completion.
7. Monthly series derivation from the homogenized, quality controlled, and completed daily data.
8. Test of homogeneity for the new monthly series with using the automatic verification results.

The original time series of the variables listed in *Table 1* were homogenized, completed, and quality controlled by the participants individually. The automatically generated verification results were gathered and reported to the supporter. The following chapter is an overview of the evaluation of the implemented homogenization process.

4. Verification of the homogenization

Validation is an essential part of the process, to make sure that the data quality increased as a result of homogenization. Hence a verification part is integrated into the MASH system for interpretation of the outcomes, it makes the evaluation of the different phases of the homogenization possible from the initial to the final stage. The basic conception of the verification test is that the confidence in the homogenization may be increased by the joint comparative mathematical examination of the original and the homogenized data series.

Two types of outcomes of the MASH software can be separated. The first type of output is the files containing the homogenized, controlled, and completed series, inhomogeneity series, detected breaks, and detected errors. The second type of output is the files containing the test results and verification tables in order to evaluate the homogenization. The verification tables contain the test statistics values before and after homogenization, measures to characterize the modification of series, the spatial representativity of the station network, and the evaluation of metadata. The quality control results for the daily data are also included.

The verification procedure based on hypothesis test results. The null hypothesis is that examined series are homogeneous. The test statistics can be compared to the critical value before and after homogenization. The critical values belong to different significance levels are built in the MASH software (it is 20.86 on the 0.05 significance level in our case). The homogenization is successful if the test statistics after homogenization is low. The theoretical background and more details of the derivation of the verification statistics can be found in MASH manual (*Szentimrey*, 1999).

The test statistics before (TSb) and after homogenization (TSa) and characteristics of the modified series are presented in this paper. Annual

statistics are examined here; though all of them are produced automatically on the monthly and seasonal scales altogether. *Tables 2* to *4* contain the average measures for maximum and minimum temperatures and precipitation for each of the station systems and the QC results alike. Number of the partners in the header lines is as follows: Hungary and Croatia with their jointly handled dataset (1), Serbia (2), Romania (3), Ukraine (4), Slovakia (5), Poland (6), Czech Republic (7). The representativity is about 50 km for climate stations and 25 km for precipitation stations, respectively. Participants have contributed to the project with data of 415 climate stations and 904 precipitation stations in all.

The TSa has to be near to the critical value or much less than the TSb if the homogenization is acceptable. Moreover, the measures of the relative modification are expected to be in accordance with the relative change of the test statistics: (TSb-TSa)/TSb. The applied statistics for the measure of the relative modification is in fact the ratio of the RMSE (root mean square error) and the standard deviation. If the significant modification of series induces weak decreasing in the degree of inhomogeneity, overdrawing the series is unnecessary and erroneous. *Tables 2–4* containing the summary statistics and the complementary diagrams in *Figs. 2–4* support the evaluation of homogenization.

The degree of inhomogeneity of the raw minimum temperatures *(Table 3.)* is substantially higher for Serbia (2) and much higher for the Hungarian and Croatian (1) dataset than in case of the maximum temperatures. The relative modification (42%) for the Hungarian and Croatian (1) series is achieved the most, although the largest improvement (*Fig. 3.*). The Serbian (2) system has been upgraded in the same rate by less relative modification. The Slovakian (5) system is near to homogeneous after processing. Relative changes of the test statistics are small in the Romanian (3) and Ukrainian (4) series, in accordance with the low value of relative modification. At the Czech Republic (7), the degree of homogeneity increased with relatively high modification. It can be found that MASH reduced the inhomogeneity of all systems, but less than in the case of maximum temperatures. The QC results relating the minimum temperatures show that the number of erroneous data per station is the largest in the Ukrainian (4) system. The Romanian (3) and Ukrainian (4) series contained more than 400 (°C) negative error and almost 100 (°C) positive errors in the data. The smallest correction has to be performed in the Czech (7) system, although it is a minor system with 18 stations.

Table 2. Average test statistics and quality control (QC) results for maximum temperature

Maximum temperature							
No. of station system	1	2	3	4	5	6	7
Number of stations	68	39	140	53	59	38	18
Verification results of homogenization							
TS after homog. (TSa)	23.6	55.7	39.0	23.7	26.4	24.8	26.7
TS before homog. (TSb)	190.7	186.2	72.9	154.0	175.6	150.6	184.3
Relative modification (%)	21	14	9	13	23	21	29
Quality control results							
Total number of errors	6307	3811	10241	5444	4542	3288	1400
Maximal positive error (°C)	10.9	13.5	996.6	107.7	11.3	22.7	10.4
Minimal negative error (°C)	–2.3	–7.5	–21.0	–22.0	–14.5	–26.3	–6.2

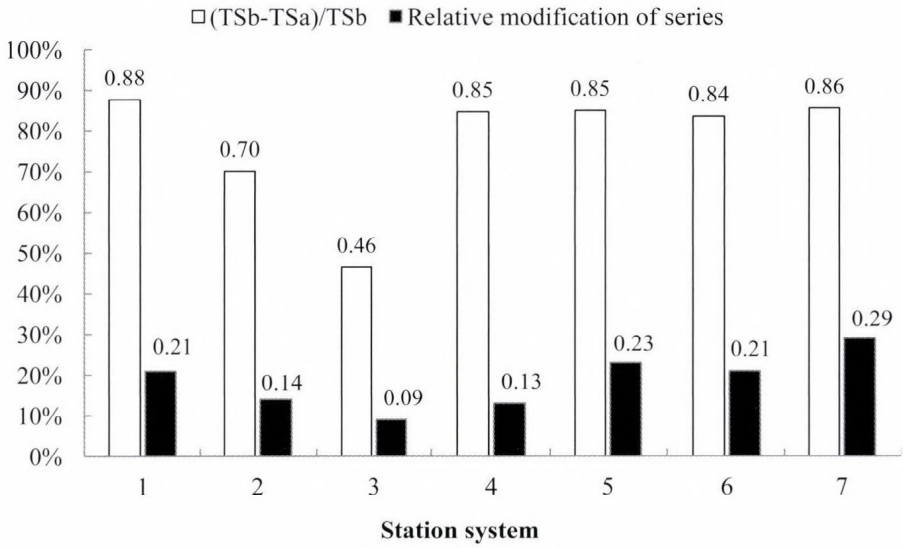

Fig. 2. Verification results for maximum temperature.

Table 3. Average test statistics and quality control (QC) results for minimum temperature

Station Sytem	1	2	3	4	5	6	7
Number of stations	68	39	140	53	59	38	18
TS after homog. (TSa)	24.3	52.5	52.5	51.9	28.5	43.5	37.8
TS before homog. (TSb)	227.5	484.7	128.3	120.3	179.7	141.3	93.9
Relative modification (%)	42	28	14	13	22	23	21
Total number of errors	4110	2161	6689	4111	3197	2592	375
Maximal positive error (°C)	23.7	11.8	95.1	79.3	14.9	15.9	0.7
Minimal negative error (°C)	–9.7	–8.0	–416.6	–417.6	–9.9	–10.0	–1.1

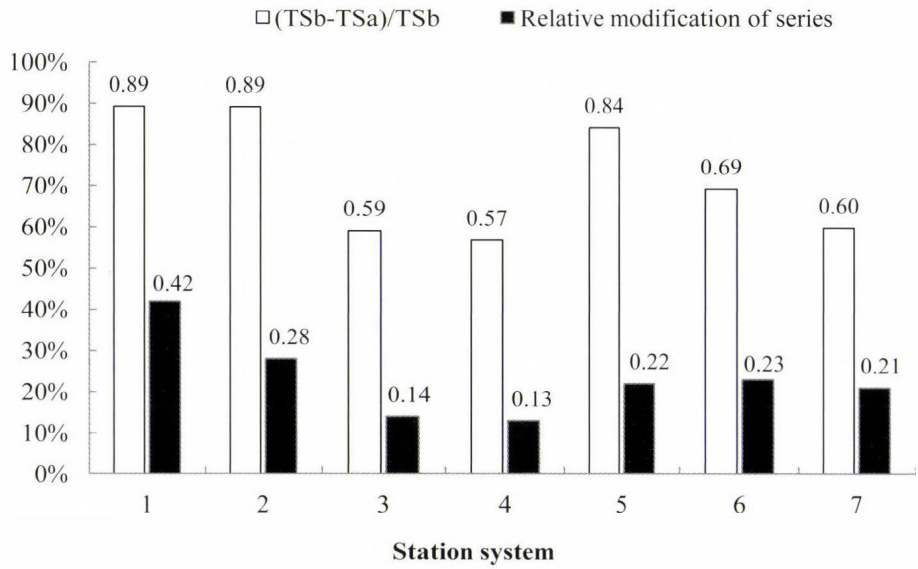

Fig. 3. Verification results for minimum temperature.

Analyzing the precipitation results, we have to take into consideration that the MASH procedure carefully detects the break points. Lower inhomogeneity arose for the precipitation series than for temperatures *(Table 4)*. During the homogenization, all of the networks became more homogeneous; nevertheless,

the modification was precautious. The test statistics indicates that the Polish (6) system was the most inhomogeneous, and the improvement is also little afterward, although the similar relative modification caused higher improvement than in the Romanian (3) system *(Fig. 4.)*. The Slovakian (5) dataset passed through the most advance, at the expense of remarkable modifications of the series comparing to the others. Resulting from the QC numerous errors were detected, about in the rate of the amount of contributed stations. The amplitude of the errors in several systems is higher towards extremely heavy precipitations.

Table 4. Average test statistics and quality control (QC) results for precipitation

Station sytem	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Number of stations	233	114	182	57	165	102	51
TS after homog. (TSa)	21.6	31.27	28.09	25.61	21.89	38.97	35.53
TS before homog. (TSb)	27.93	34.73	31.88	28.98	38.17	46.29	39.77
Relative modification (%)	4	5	6	3	10	5	4
Total number of errors	1531	672	975	313	803	408	223
Maximal positive error (mm)	71.94	230.27	10.27	179.46	94.29	93.36	60.38
Minimal negative error (mm)	23.24	–36.87	–1.52	–5.68	–59.46	–25.47	–11.41

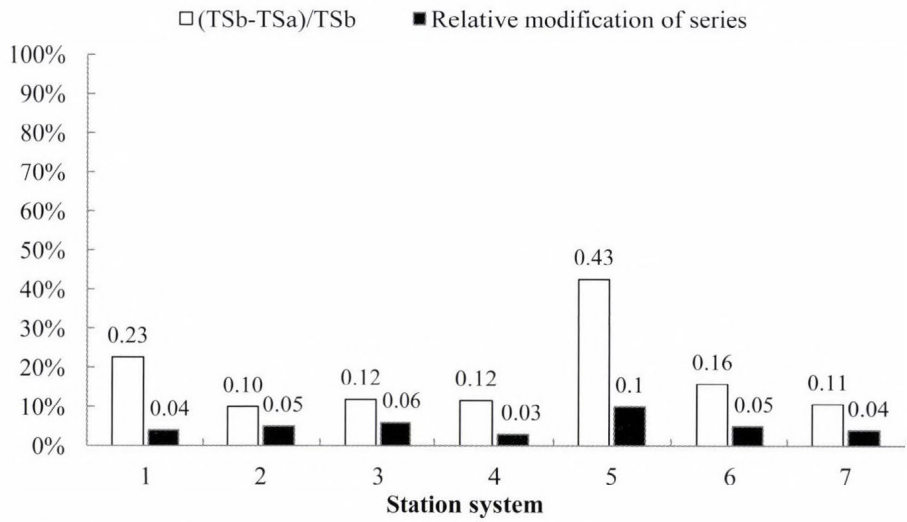

Fig. 4. Verification results for precipitation

Verification results for all the 12 elements can be followed up in the project deliverables related to the issues of the homogenization process *(D1.12)*. The data rescue and digitization activity in Module 1, and the data homogenization and QC performed by applying MASH procedure guarantee the availability of the high quality daily time series for the basic climate elements in the Carpathian region in the period of 1961–2010.

5. Analysis of the climate trends on the harmonized gridded dataset

The final outcome of the CarpatClim tender service is a ~10 × 10 km resolution gridded dataset on daily scale for elements listed in *Table 1*. Interpolation of the homogenized time series is carried out by applying the MISH (Meteorological Interpolation based on Surface Homogenized data basis; *Szentimrey* and *Bihari*, 2007) method. The MISH method was developed for interpolation of meteorological data, and an adequate mathematical background was also developed (*Szentimrey et al.*, 2011) for the purpose of efficient use of all the valuable meteorological and auxiliary model information. The main difference between MISH and the usual geostatistical interpolation methods is the application of the meteorological data series for modeling. In geostatistics (*Cressie*, 1991), the sample for modeling is only the predictor data, which is a single realization in time, while in meteorology there are data series, i.e., a sample in time and space as well.

5.1. Data harmonization with the homogenized data exchange

The cross border harmonization is essential in the project to avoid breaks at the boundaries on climate maps based on the gridded data. It can be ensured by the changes of the homogenized series across the borders as it was in case of the raw data exchange. The cross border harmonization is acceptable if some improvement appears in test statistics (*D2.5*). The gridding of the harmonized series was executed on national level by applying MISH, and the merging of the separate but harmonized grid parts followed up in the end.

5.2. Trend estimation based on the created dataset

Investigation of the climate extremes, observed trends, changes in frequency and intensity could contribute to the establishment of the adaptation strategies in the region. Climate indices are used in several projects on climate change as prevailing indicators of changes in extremes. Spatial interpolation of indices values for station locations is a difficult task, as the distribution functions of the several derived values are unknown. However, the basic variables, such as temperature and precipitation can be gridded by the knowledge of their statistical properties, thus higher quality gridded datasets can be constructed for

further analysis, as it was created in CarpatClim (*Lakatos et al.*, 2010). The gridded database produced in daily temporal resolution provides relevant outcomes for studying extremes.

One temperature and one precipitation index was chosen to show the first results of the trend analysis based on the high quality dataset covering the region. These are the number of hot days per year (daily maximum ≥ 30 °C) and the number of days with heavy rainfall (daily precipitation amount >20 mm).The changes obtained from the linear trend estimation are demonstrated on the grid defined in the specification (*JRC*, 2010). The maps indicate the changes in the examined period, i.e., the slope of the estimated linear trend multiplied with the length of the changing period.

Fig. 5 strengthens the warming trend in the entire region. The changes are in strong correspondence with the orography. The growth is less at higher mountains than at lower altitudes. More hot days occur in the basin, especially in the territory between Danube and Tisza rivers, by 18–22 days from 1961 to 2010. The Transylvanian basin shows fewer rises. The region is lying under the south and east Carpathians turned up the largest growing in the number of hot days (over 24) during the examined period.

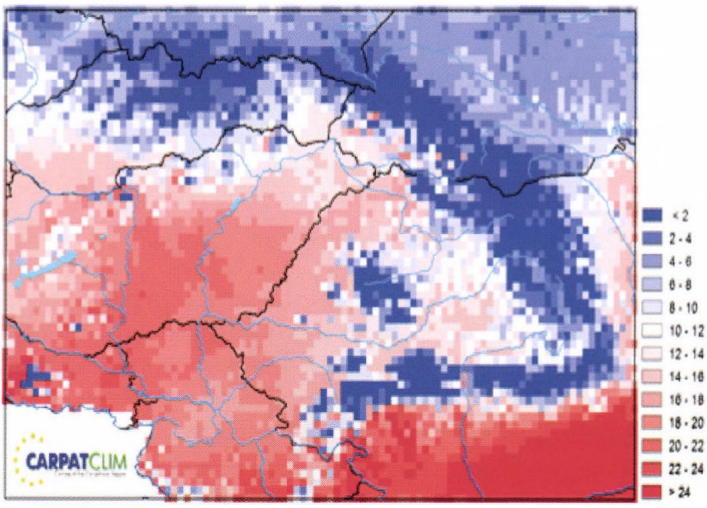

Fig. 5. Change in the number of hot days per year (daily maximum ≥ 30 °C) in the Carpathian region in the period of 1961–2010.

Fig. 6 visualizes the changes in the days above 20 mm precipitation during the whole 50 years period. The estimated changes indicated varied spatial distribution. The topographical effects are not so evident than in hot day's

changes. The changes are between –2 and 3 days in the extended area of the region. More intense decreasing or increasing was found mostly on small territories. The highest increase was indicated in the northeast Carpathians and the Bihor Mountains with 7 days.

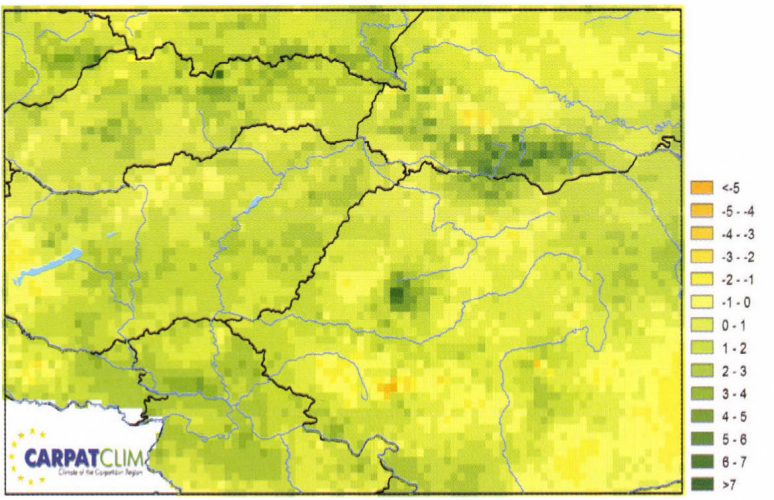

Fig. 6. Change in the number of days with heavy rainfall (daily sum > 20 mm) in the Carpathian region during the period from 1961 to 2010

6. Conclusion

The COST HOME Action had drawn the attention to the importance of data homogenization and recent methods. The monthly benchmark results of COST HOME denoted that MASH is one of the best monthly homogenization methods. The COST participants from the Carpathian region started the work with MASH during the STSMs supported by the COST. These STSMs established a common project for creating a homogenized dataset covering the region.

There are many advantageous attributes of MASH. Due to the automatic execution it allows performing the data homogenization, quality control, and data completion for the entire Carpathian region within a reasonable time. The MASH was used for numerous stations, 1319 climate and precipitation stations together, and 12 elements for a fifty-year long period in the Climate of the Carpathian Region tender service. The consortium members implemented the homogenization separately by the common procedure. The cross border harmonization was guaranteed by near border data exchange. The automatically

generated verification results presented in this paper confirm that the quality of the data highly improved during the homogenization and quality control procedure.

The Climate of the Carpathian Region Project contributes to the availability of a set of homogeneous and spatially representative data to prepare climate change studies relevant in the region. The warming trend is obvious on the harmonized, gridded data in the period of 1961–2010 as indicated from the preliminary trend analysis. The changes in the number of days with precipitation above 20 mm show significant decrease or increase only on small areas of the region in the examined 50-year long period.

Acknowledgements–This work was supported by the COST (European Cooperation in Science and Technology) Action ES0601 titled "Advances in homogenization methods of climate series: an integrated approach (HOME)" (2007–2011) and by JRC Desert Action in the framework of the "Climate of the Carpathian Region (CarpatClim)" Project. The authors take this opportunity to thank the following members of the CarpatClim Homogenization and Interpolation Group for data homoginization:
Austria: Ingeborg Auer, Johann Hiebl
Croatia: Janja Milković
Czech Republic: Pavel Zahradníček, Petr Štěpánek, Radim Tolasz
Hungary: Tamás Szentimrey, Zita Bihari, Mónika Lakatos, Tamás Kovács, Ákos Németh, Sándor Szalai
Poland: Piotr Kilar, Robert Pyrc, Danuta Limanowka
Romania: Sorin Cheval, Monica Matei
Serbia: Dragan Mihic, Predrag Petrovic, Tatjana Savic
Slovakia: Peter Kajaba, Gabriela Ivanakova, Oliver Bochnicek, Pavol Nejedlik, Pavel Šastný
Ukraine: Oleg Skrynyk, Yurii Nabyvanets, Natalia Gnatiuk,
and Annamari Marton for depection of maps.

References

Cressie, N., 1991: Statistics for Spatial Data. Wiley, New York.

D1.12: Final report on quality control and data homogenization measures applied per country, including QC protocols and measures to determine the achieved increase in data quality. http://www.carpatclim-eu.org/pages/deliverables/

D 2.5: Report with final results of the data harmonization procedures applied, including all protocols, per country. http://www.carpatclim-eu.org/pages/deliverables/

JRC, 2010: Climate of the Carpathian Region. Technical Specifications (Contract Notice OJEU 2010/S 110-166082 dated 9 June 2010). http://desert.jrc.ec.europa.eu/action/php/index.php?action=view&id=550

Lakatos, M., Szentimrey, T., and *Bihari, Z.*, 2010: Application of gridded daily data series for calculation of extreme temperature and precipitation indices in Hungary. *Időjárás 115*, 99–109.

Szentimrey, T., 1999: Multiple Analysis of Series for Homogenization (MASH). Proceedings of the Second Seminar for Homogenization of Surface Climatological Data, Budapest, Hungary; WMO, WCDMP-No. 41, 27–46.

Szentimrey, T. and *Bihari, Z.*, 2007: Mathematical background of the spatial interpolation methods and the software MISH (Meteorological Interpolation based on Surface Homogenized Data Basis). Proceedings from the Conference on Spatial Interpolation in Climatology and Meteorology, Budapest, Hungary, 2004, COST Action 719, COST Office, 17–27.

Szentimrey, T., 2008: Development of MASH homogenization procedure for daily data. Proceedings of the Fifth Seminar for Homogenization and Quality Control in Climatological Databases, Budapest, 2006; WCDMP-No. 71, WMO/TD-NO. 1493, 2008, 123–130.

Szentimrey, T., 2011: Manual of homogenization software MASHv3.03. Hungarian Meteorological Service.

Szentimrey, T., Bihari, Z., Lakatos, M., and *Szalai, S.*, 2011: Mathematical, methodological questions concerning the spatial interpolation of climate elements. Proceedings from the Second Conference on Spatial Interpolation in Climatology and Meteorology, Budapest, Hungary, 2009, Időjárás 115, 1–2, 1–11.

UNEP, 2007: Carpathians Environment Outlook. Geneva. http://www.grid.unep.ch.

Venema, V., Mestre, O., Aguilar, E., Auer, I., Guijarro, J.A., Domonkos, P., Vertacnik, G., Szentimrey, T., Štěpánek, P., Zahradnicek, P., Viarre, J., Müller-Westermeier, G., Lakatos, M., Williams, C.N., Menne, M., Lindau, R., Rasol, D., Rustemeier, E., Kolokythas, K., Marinova, T., Andresen, L., Acquaotta, F., Fratianni, S., Cheval, S., Klancar, M., Brunetti, M., Gruber, C., Duran, M.P., Likso, T., Esteban, P. and *Brandsma,T.*, 2012: Benchmarking monthly homogenization algorithms. *Climate of the Past 8*, 89–115.

Vincent, LA, Zhang, X, Bonsal, BR, and *Hogg, WD.*, 2002: Homogenization of daily temperatures over Canada. *J. Climate 15*, 1322–1334.

INSTRUCTIONS TO AUTHORS OF *IDŐJÁRÁS*

The purpose of the journal is to publish papers in any field of meteorology and atmosphere related scientific areas. These may be

- research papers on new results of scientific investigations,
- critical review articles summarizing the current state of art of a certain topic,
- short contributions dealing with a particular question.

Some issues contain "News" and "Book review", therefore, such contributions are also welcome. The papers must be in American English and should be checked by a native speaker if necessary.

Authors are requested to send their manuscripts to

Editor-in Chief of IDŐJÁRÁS
P.O. Box 38, H-1525 Budapest, Hungary
E-mail: journal.idojaras@met.hu

including all illustrations. MS Word format is preferred in electronic submission. Papers will then be reviewed normally by two independent referees, who remain unidentified for the author(s). The Editor-in-Chief will inform the author(s) whether or not the paper is acceptable for publication, and what modifications, if any, are necessary.

Please, follow the order given below when typing manuscripts.

Title page: should consist of the title, the name(s) of the author(s), their affiliation(s) including full postal and e-mail address(es). In case of more than one author, the corresponding author must be identified.

Abstract: should contain the purpose, the applied data and methods as well as the basic conclusion(s) of the paper.

Key-words: must be included (from 5 to 10) to help to classify the topic.

Text: has to be typed in single spacing on an A4 size paper using 14 pt Times New Roman font if possible. Use of S.I. units are expected, and the use of negative exponent is preferred to fractional sign. Mathematical formulae are expected to be as simple as possible and numbered in parentheses at the right margin.

All publications cited in the text should be presented in the *list of references*, arranged in alphabetical order. For an article: name(s) of author(s) in Italics, year, title of article, name of journal, volume, number (the latter two in Italics) and pages. E.g., *Nathan, K.K.,* 1986: A note on the relationship between photo-synthetically active radiation and cloud amount. *Időjárás 90,* 10-13. For a book: name(s) of author(s), year, title of the book (all in Italics except the year), publisher and place of publication. E.g., *Junge, C.E.,* 1963: *Air Chemistry and Radioactivity.* Academic Press, New York and London. Reference in the text should contain the name(s) of the author(s) in Italics and year of publication. E.g., in the case of one author: *Miller* (1989); in the case of two authors: *Gamov* and *Cleveland* (1973); and if there are more than two authors: *Smith et al.* (1990). If the name of the author cannot be fitted into the text: *(Miller,* 1989); etc. When referring papers published in the same year by the same author, letters a, b, c, etc. should follow the year of publication.

Tables should be marked by Arabic numbers and printed in separate sheets with their numbers and legends given below them. Avoid too lengthy or complicated tables, or tables duplicating results given in other form in the manuscript (e.g., graphs).

Figures should also be marked with Arabic numbers and printed in black and white or color (under special arrangement) in separate sheets with their numbers and captions given below them. JPG, TIF, GIF, BMP or PNG formats should be used for electronic artwork submission.

Reprints: authors receive 30 reprints free of charge. Additional reprints may be ordered at the authors' expense when sending back the proofs to the Editorial Office.

More information for authors is available: journal.idojaras@met.hu

Published by the Hungarian Meteorological Service

Budapest, Hungary

INDEX 26 361

HU ISSN 0324-6329