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.

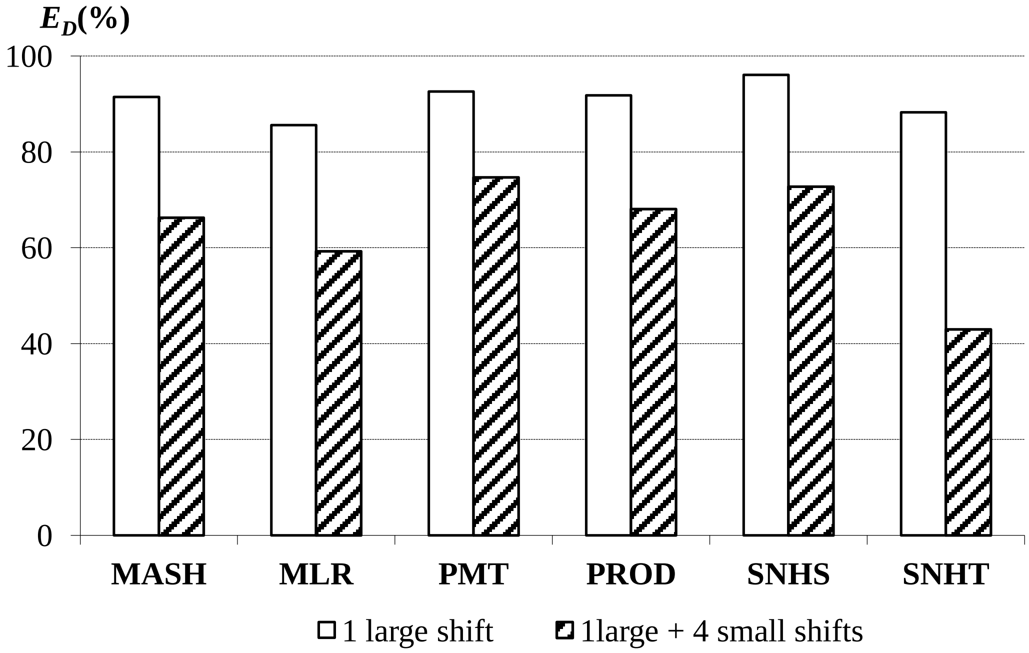

Fig. 1. Detection skill (E_D) of large shifts with the background of white noise and with that of white noise plus four small shifts. PMT – Penalized maximum t-test (*Wang et. al*., 2007); PROD – PRODIGE, SNHS – SNHT for shifts only (*Alexandersson*, 1986); the explanation of other denotations is in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

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

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

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

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).

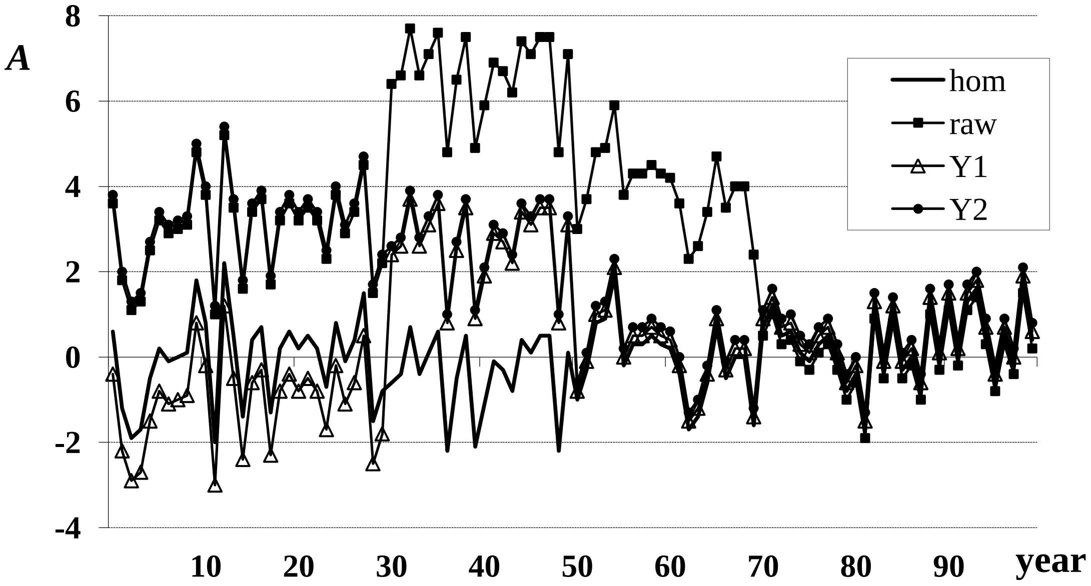

Fig. 2. Example of homogenization results. The raw time series is a 100 years long white noise with shifts in years 30, 50, and 70, whose magnitudes are +4, –3, –4, respectively. The unit of the values (*A*) is the standard deviation of white noise. hom = homogenous data, raw = "raw" (simulated) data, 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.

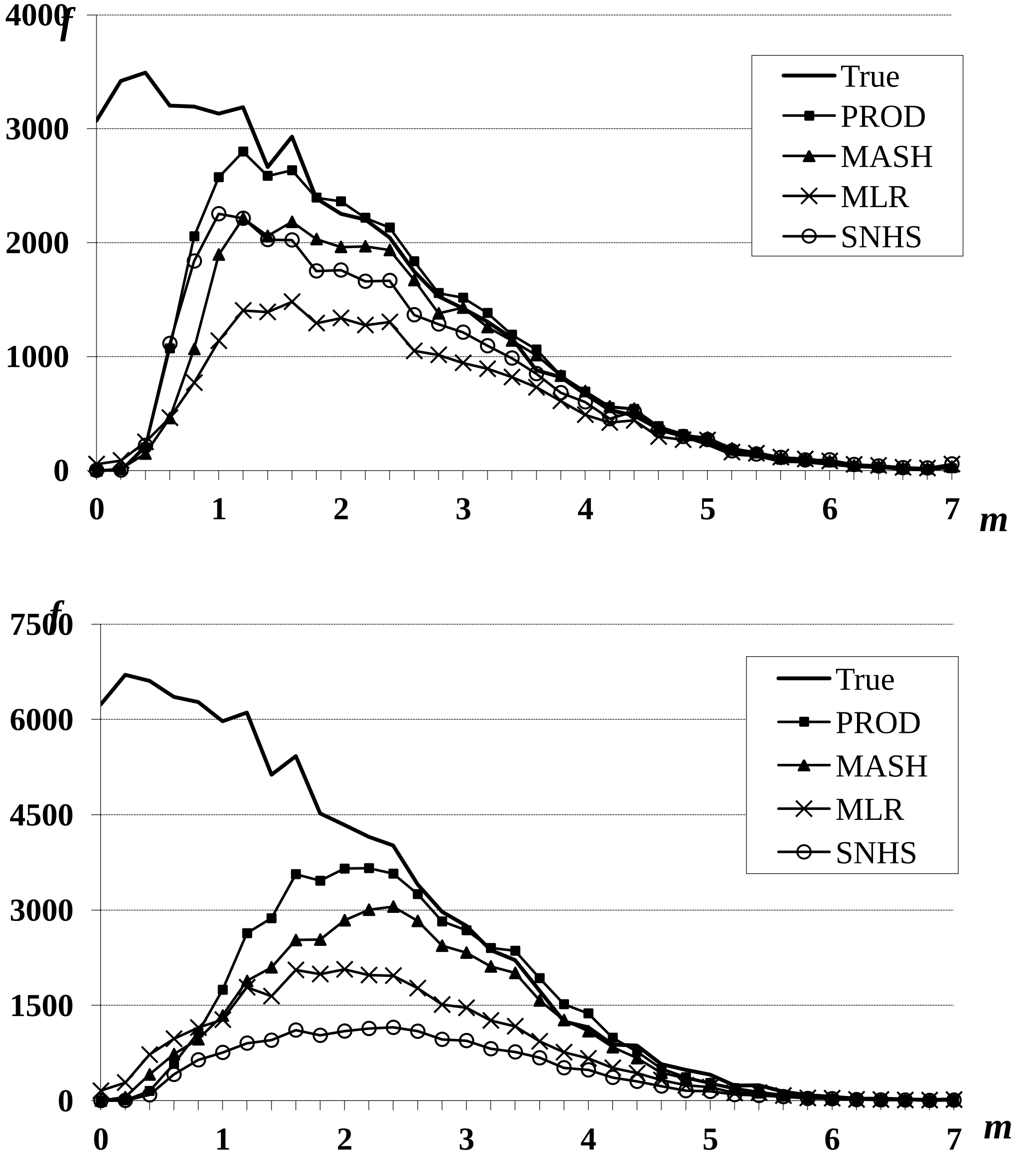

Fig. 3. Frequency (*f*) of detected change-points as a function of shift-magnitudes (*m*), when 5 shifts with random positions (top panel) and 5 platforms (pair of shifts with the same *m* and opposite directions, bottom panel) with random positions are inserted into 100 years long white noise process. 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*.