Uncertainty evaluation of Climatol's adjustment algorithm applied to daily air temperature time series

Oleg Skrynyk[a,b *], Enric Aguilar[a], Jose Guijarro[c], Luc Yannick Andreas Randriamarolaza[a], Sergiy Bubin[d]

[a]*Center for Climate Change, C3, Geography Department, Universitat Rovira i Virgili, Tarragona, Spain*

[b]*Ukrainian Hydrometeorological Institute, Kyiv, Ukraine*

[c]*State Meteorological Agency (AEMET), Balearic Islands Office, Spain*

[d]*Department of Physics, Nazarbayev University, Nur-Sultan, Kazakhstan*

[*] Corresponding author, 15 C. Joanot Martorell, Vila-seca, 43480, Spain oleg.skrynyk@urv.cat

Abstract

The present study investigated the uncertainty associated with Climatol's adjustment algorithm applied to daily minimum and maximum air temperature. The uncertainty quantification was performed based on several numerical experiments and the benchmark data, that were created in the frame of the INDECIS project. Using a complex approach, the uncertainty was evaluated on different levels of detail (day-to-day evaluation through formalism of random functions and through six statistical metrics) and time resolution (daily and yearly). However, only the main source of potential residual errors was considered, namely station signals introduced into a raw data set to be homogenized/adjusted. Other influencing factors were removed from the analysis or kept almost unchanged.

According to our calculations, the Climatol's adjustment uncertainty, evaluated on the daily scale, varies in time. The width of the residual errors distribution in summer months is substantially less compared to wintertime. The slight seasonality is also observed in the means of the residual errors. The uncertainty evaluation based on the statistical metrics usually neglect such non-stationarity of the residual errors providing just averaged in time assessments. On the other hand, metrics provide detailed information regarding both types of the residual errors, systematic and scatter. The metrics values showed good capability of the Climatol software to remove the systematic errors related to jumps in the means, while the scatter errors are removed from the raw time series with less efficiency. On yearly scale, the uncertainty evaluation was performed for yearly temperature data and several climate extreme indices. The both types of the errors are removed well in yearly time series of the air temperature and the extreme indices. The metrics values also showed significant reduction of the adjustment uncertainty of Climatol's adjustment. Substantial decreasing of linear trend errors in yearly time series can also be reported.

Key words: uncertainty, homogenization adjustment, Climatol, minimum and maximum daily air temperature, INDECIS

1. Introduction

Detection of modern climate change and analysis of climate variability and extreme events on national, regional or even global scales are mainly performed based on a statistical analysis of time series of measured meteorological variables such as air temperature and precipitation (e.g. Alexander et al., 2006; Klein Tank et al., 2009; Hartmann et al., 2013). However, in order to extract accurate and reliable conclusions from the analysis it is necessary firstly to homogenize raw data sets due to many spurious artefacts (inhomogeneities) that are usually present in the data (Aguilar et al., 2003; Trewin, 2010). By performing homogenization, one tries to remove the inhomogeneities (abrupt shifts/jumps, gradual trends, outliers etc.) and in such way to approximate the data to the real climate signal, happened on some area. Usually the homogenization procedure allows to increase consistency of the data what is plainly seen after statistical comparison of the raw and homogenized time series (e.g. Mamara et al., 2014; Prohom et al., 2016; Osadchyi et al., 2018; Yosef et al., 2018; Skrynyk et al., 2019; Fioravanti et al., 2019; Dumitrescu et al., 2020). However, a question remains unclear: how far are the homogenized data from the true climate signal? Or in other words, what potential uncertainties could be still present in the data, homogenized by means of some homogenization algorithm or software? It is the important but still extremely complicated issue because the climate signal (clean data) is usually unknown and it is impossible to conduct direct quantitative comparison and evaluation of the homogenization results. Understanding of the uncertainties and their causes is vital to correctly interpret outputs of any predicting model (e.g. Iman and Helton, 1988), including homogenization software.

The problem of climate data homogenization can be divided into two sub-problems, namely detection of discontinuities (most probable dates of potential inhomogeneities) and adjustment of inhomogeneous data (some segments of raw time series) to homogeneous state. Both sub-problems might produce a certain part of common errors, which deviate the homogenized data from the true climate signal. An evaluation of efficiency of the detection algorithms has been performed in many works (e.g. Ducré-Robitaille et al., 2003; De Gaetano, 2006; Reeves et al., 2007; Domonkos, 2011; Kuglitsch et al., 2012; Venema et al., 2012; Willett et al., 2014; Killick, 2016; Yozgatligil and Yazici, 2016; Coll et al., 2020). On the other hand, an assessment of performance of adjustment methods has been addressed in papers (e.g. Della-Marta and Wanner, 2006; Mestre et al., 2011; Trewin, 2013; Squintu et al., 2020). In both cases, the evaluation was mainly performed in a relative form, that is, several homogenization algorithms are usually compared in order to define which one gives the best output and is most suitable for practical applications. Such relative comparison is usually performed based on some benchmark data. However, the quantification of uncertainties of homogenization procedures has been published just in several works (e.g. Lindau and Venema, 2016; Vincent et al., 2018; Trewin, 2018). Lindau and Venema (2016) studied uncertainty of the multiple breakpoint detection algorithms applied to yearly climate time series. To do so, they defined a probability distribution for possible shifts of the detected break from its true position based on a theoretical approach. According to their findings, the probability of the shifts or, in other words detection errors, can be described statistically by a Brownian motion with drift. Vincent et al. (2018) and Trewin (2018) evaluated uncertainty of homogenization adjustment algorithms applied to daily air temperature time series. In both works, parallel measurements of temperature were used in order to assess potential residual errors. However, the uncertainty of the adjustment was quantified using different methodology. In (Vincent et al., 2018) the remaining errors in corrected time series were evaluated through two statistical metrics, the root mean square error ($RMSE$) and the percentage of days within 0.5ºC ($POD05$) that were calculated based on daily data. As mentioned in the paper, $RMSE$ and $POD05$ were used to assess the uncertainty in the mean and extreme temperature values, respectively. In (Trewin, 2018) the uncertainty is also evaluated through some statistical indicators, but they were calculated on seasonal and annual scales. The uncertainty was defined as a standard deviation of the indicator values that were obtained by repeating calculations for slightly different adjustment conditions (changing a set of reference stations, their number etc.). Important to note that despite of intuitively clear meaning of the term 'uncertainty', which can be simply interpreted as a range or a distribution of possible residual errors, there is no unique methodology how it can be quantified for homogenization/adjustment of climate data.

The objective of this paper is to evaluate the uncertainty associated to the adjustment of daily maximum and minimum temperature series using Climatol (Guijarro, 2018). We constrain our work assuming a perfect detection to focus on Climatol's adjustment algorithm. It is also worth noting, that the problem of the uncertainty evaluation of homogenization adjustment is especially important when dealing with daily time series, since climate data with such time resolution is the basis for many modern climatological studies (e.g. monitoring, detection and attribution of changes in climate extremes). In order to achieve our goal we used benchmark data sets (Aguilar et al., 2018; Pérez-Zanón et al., 2018) specially elaborated in the frame of the European project INDECIS (Integrated approach for the development across Europe of user oriented climate indicators for GFCS high-priority sectors: agriculture, disaster risk reduction, energy, health, water and tourism) (INDECIS, 2018).

The methodology proposed in this paper and applied to Climatol can be generalized for other homogenization software, which are able to adjust daily time series of climatological variables in automatic mode with predefined break points. Our findings should also be helpful for developers of homogenization methods and software as well as for their potential users who ought to know what possible errors they still could expect after applying the homogenization adjustment.

2. Data and methods

2.1. The Climatol homogenization software

The R package Climatol is a homogenization software that has been widely used recently in order to remove inhomogeneities from collections of raw time series of different climate variables and different time resolution (e.g. Mamara et al., 2013; Sanchez-Lorenzo et al., 2015; Guijarro et al., 2018; Meseguer-Ruiz et al., 2018; Azorin-Molina et al., 2019; Dumitrescu et al., 2020; Coll et al., 2020). The effectiveness of the software has been evaluated during several benchmark tests (Venema et al., 2005; MULTITEST, 2015; Killik, 2016; Guijarro et al., 2017) where it showed good results, which are comparable to other high quality and well tested homogenization algorithms. According to the benchmarking, both part of the homogenization procedure in Climatol, namely detection and adjustment, work well allowing to remove different type of the artefacts and increase consistency of raw data sets. One of Climatol's characteristics is that it can be used automatically what significantly increases its objectivity and applicability to large data sets such as the European Climate Assessment and Dataset (ECA&D) (Klein Tank et al., 2002). Several versions of the software have been updated since its creation. In our work, we used Climatol 3.1.1., available through CRAN (https://cran.r-project.org/package=climatol).

The Climatol detection method (Guijarro, 2018) is based on the standard normalized homogeneity test (SNHT) (Alexandersson, 1986; Alexandersson and Moberg, 1997). For any candidate time series, Climatol uses data from neighbor stations to create only one composite reference series as their optionally weighted average.

Climatol first normalizes the data and infills missing values through an iterative process during which the main statistical properties of time series, namely means and standard deviations, are recalculated at every iteration until their stable values are obtained. Once the means become stable, all data are normalized and estimated (whether existing or missing, in all of the series) by means of respective value from the composite reference series, i.e. as a weighted average of a prescribed number of the nearest available data. From the statistical point of view, the approach used is equivalent to applying a type II linear regression model (Sokal and Rohlf, 1969), what is reasonable since all climatic time series from a network under study usually have similar errors. On the next step, the normalized original data and their estimates are used to create time series of anomalies (the estimated values are subtracted from the observed ones), which in turn are exploited to find and eliminate outliers and to detect inhomogeneities by applying SNHT. Since SNHT is a test originally devised for finding a single break point in a series, it is applied iteratively, splitting the candidate time series or its segment into two parts every cycle until no inhomogeneous segments are found. Moreover, during iterations, the test is applied twice: (1) to stepped overlapping temporal windows and after that (2) to complete series. Such two-stage procedure allows to minimize detection errors arisen when two or more shifts in the mean of similar size could mask its results. Finally, all homogeneous sub-periods originate complete reconstructed series by using new estimated values to fill all missing data in.

2.2. The INDECIS benchmark data sets

In the frame of the INDECIS project (see www.indecis.eu), two different collections of benchmark time series, which cover two regions in Europe with different climate (Southern Sweden and Slovenia) were created (Aguilar et al., 2018; Pérez-Zanón et al., 2018). Each collection contains daily series of nine essential climate variables (cloud cover, wind speed, relative humidity, sea level pressure, precipitation amount, snow depth, sunshine duration, maximum and minimum air temperature) over the period of 1950-2005. Each benchmark data set consists of clean data, extracted from the output of the Royal Netherlands Meteorological Institute (KNMI) Regional Atmospheric Climate Model (RACMO) version 2, driven by Hadley Global Environment Model 2 - Earth System (MOHC-HadGEM2-ES) (Collins et al., 2008), and inhomogeneous data, created by introducing realistic breaks and errors. Missing values and other quality problems (different from biases) were also added to generate other flavors of the perturbed benchmarks, however they were not used in our study. The RACMO model was chosen due to its high spatial resolution ($0.11^{o}\times0.11^{o}$) and the daily time step of the output provided: gridded time series of essential climate variables.

In our study, we used only the maximum (TX) and minimum (TN) air temperature benchmark data sets for the southern Sweden (Fig. 1 a). Both data sets contain 100 ‘stations’, a subset of the RACMO grid points chosen to imitate stations spatial distribution. Their geographical locations on the domain under study are shown in Fig. 1 b.

The introduction of biases in the homogeneous series was done by simulating relocations. First, closest pairs of the RACMO grid time series were used to build a database of differences (or ratios, depending on the variable) between nearby locations. Then, for every random sub-period to perturb in the homogeneous series, a difference (or a ratio) was randomly chosen, modified by a random factor, and applied to bias the sub-period. Total numbers of break points introduced into TN and TX clean time series are 258 and 280, respectively. That is, the mean break frequency was set to ~4/~5 (TN/TX) in 100 years, as it was found in previous studies on European series (e.g. Domonkos, 2011; Venema et al., 2012; Domonkos, 2017). Fig 2 represents the time distribution of the break points, while Fig 3 shows the distribution of the number of stations/time series with respect to the number of breaks in one time series.

Due to the daily time resolution and the way that was used to create the realistic, as much as possible, station signals (considered here as the time series of the introduced errors, see an example in Fig. 7 a below), they are characterized by intensive noise presence at each of homogeneous segments except the last ones. That makes it difficult to define precisely factors and amplitudes of the shifts at the break points. Nevertheless, we estimated such parameters by averaging respective sub-periods of the error time series. Thus, in our case the factors are mean values of errors at the homogeneous segments, while the amplitudes are differences between pairs of two consecutive factors: between the means at previous and next segments. As can be seen from Fig. 4, where histograms of the factors and amplitudes are presented, their range for TN, approximately from -6 to 6ºC (Fig. 4 a, c), is wider comparing to TX, (-3; 3) (ºC) (Fig.4 b, d). This was deliberately introduced into the benchmark to mimic real effects such as those related to larger local microclimate differences at nights comparing to daylight period of days (e.g. Brunet et al., 2008). Beside the factors and amplitudes, the homogeneous segments can also be characterized by standard deviations (SD) of errors. Fig. 5 shows their histograms for TN and TX time series. The mean and SD of the errors on the homogeneous segments can be combined in a single parameter called as signal to noise ratio. But in our work, we consider them separately.

The presented statistical properties of the break points and respective homogeneous segments in the introduced station signals are close to reality. Such conclusion is supported by many homogenization results of real data sets where similar statistical features of inhomogeneities have been found (e.g. Brunet et al., 2008; Trewin, 2018).

2.3. Methodology used to evaluate uncertainty of homogenization adjustment

In order to describe our approach to the evaluation of Climatol's adjustment uncertainty, we first introduce the formalism and present some graphical illustrations. Let

$$X^I,\ X^H, \text{ and } X^C \quad (1)$$

be inhomogeneous, homogenized, and clean daily data, respectively. X^I and X^C can be also referred to as raw and homogeneous data, correspondingly. All these data sets are collections of time series

$$X=\{x_{ij}\},\ i=1,\dots,M,\ j=1,\dots,N, \quad (2)$$

where M is the number of meteorological stations considered and N is the number of time steps/days. From mathematical point of view X is a rectangular matrix with dimension of $M\times N$. Let X_k, which is the k-th row in (2), denote the entire time series for the k-th station. The homogenization adjustment can be formally thought as mapping g that transform the input matrix X^I in to the output one X^H

$$X^I \underset{\rightarrow}{g} X^H. \quad (3)$$

X^C is the reference, etalon result for the outputs.

Based on the data available in (1), time series of real, E^R, detection, E^D, and homogenization, E^H, errors can be calculated:

$$E^R = X^I - X^C,\ \ E^D = X^I - X^H,\ \ E^H = X^H - X^C. \tag{4}$$

Specifically in our case, E^R is a collection of station signals (or, more precisely, station signals plus noise; but we will call them as station signals for simplicity) that were introduced into the clean data X^C. E^H is a dataset of residual errors that might be still present in the homogenized or adjusted series X^H. The error datasets E^R, E^D and E^H are also $M \times N$-matrices: $E = \{e_{ij}\}$, $i = 1, \ldots, M$, $j = 1, \ldots, N$.

Fig. 6 shows some typical examples of the time series associated with the same (k-th) station. They were extracted from the TN raw, homogenized by means of the Climatol software, and clean data sets. Fig. 7 shows the corresponding error time series (4), calculated from the data given in Fig. 6. All figures can be also interpreted as graphical representations of the k-th rows in the respective matrices. We will refer to both figures throughout this paper to illustrate the configuration and layout of our numerical experiments and results.

The main object of our study is the matrix E^H: we want to know how large could be the residual errors in the adjusted data, or in other words, how large could be the departure of the adjustment prediction X^H from the reference, etalon result X^C. According to (e.g., Walker et al., 2003), such departure is usually called as ‘uncertainty’. Typically, there exist multiple reasons, referred to as sources of the uncertainty (Jakeman et al., 2006), which may affect the adjustment performance and magnitude of the errors in E^H. Therefore, in order to evaluate the uncertainty of the homogenization adjustment we must consider all these sources - the whole credible range of every uncertain input and parameter of the adjustment software - and define the effective width of the corresponding probability distribution of the residual errors (Domonkos and Efthymiadis, 2013). The wider the error distribution, the more uncertain the software prediction X^H is.

The residual errors of the homogenization adjustment E^H should depend on the introduced errors E^R. The more complex station signals in E^R (e.g. the larger number of break points, the higher amplitudes of shifts, etc.), the larger residual errors should be expected. Thus, to clarify how wide the distribution of the potential remaining errors could be, we have to consider as many as possible different but real variants of E^R. Performing the homogenization adjustment for each of them provides a respective ensemble of Climatol's outputs, necessary for the uncertainty quantification.

The result of the homogenization adjustment should also depend on other factors, such as a mean correlation between candidate and reference time series (Szentimrey, 2008; Guijarro, 2011; Domonkos and Coll, 2017), the number of reference series (Trewin, 2018) etc. However, in the present study we focus only on the influence of the station signals on the adjustment result. That is, we try to quantify the adjustment uncertainty, which comes from only one source: errors introduced into the input data to be adjusted. The sensitivity of Climatol's adjustment to other possible factors will be addressed in our future works.

2.3.1. The concept of a random field/function applied to the residual errors E^H. The considerations presented above suggest an appropriate theoretical model for E^H that can provide a basis for further calculations and can make calculation results more statistically and theoretically solid. Since we are going to consider an ensemble of different realizations of E^H, it is natural to assume that E^H is a random field or, more generally, a random function, that is given at the limited number ($M \times N$) of discrete points in space and time domains, D and T, respectively. Therefore, in order to evaluate the homogenization adjustment and to quantify the adjustment uncertainty we have to define and study statistical properties of the random field E^H. According to the theory, a multidimensional ($M \times N$-dimensional) probability distribution function

$$f_{M \times N}\left(e_{11}^H, e_{12}^H, \ldots, e_{1N}^H, e_{21}^H, \ldots, e_{2N}^H, \ldots, e_{MN}^H\right) \tag{5}$$

provides complete and the most detailed description of E^H. Based on $f_{M \times N}$ it is possible to derive multidimensional probability distribution of the residual errors in any of M meteorological stations.

For instance, for k-th station we get $f_N\left(e_{k1}^H, e_{k2}^H, \ldots, e_{kN}^H\right)$. The f_N is obtained by integrating $f_{M\times N}$ with respect to its all arguments except $e_{k1}^H, e_{k2}^H, \ldots, e_{kN}^H$. Function $f_1\left(e_{kl}^H\right)$ defines probability distribution of the residual error in k-th meteorological station ($i=k$) and l-th day ($j=l$).

In the most general case, a random field might be non-stationary in time and heterogeneous in space. In this situation, the simplest statistical properties of the random field defined in a single point of the space-time domain, such as the mean or standard deviation, vary in the domain. On the contrary, when the field is stationary and homogeneous, these statistical moments are constant in time and space. Specifically to the homogenization adjustment, we can expect E^H to be non-stationary (e.g. due to seasonal cycle in temperature time series) and heterogeneous (e.g. due to possible different topography in D and, as a result, different local correlation between temperature time series). Such peculiarities of E^H, non-stationarity and spatial heterogeneity, make its analysis more difficult. In particular, that means we cannot use ergodic assumption in order to calculate statistical properties of E^H based on its only realization.

Let E^{Rq}, $q=1,\ldots,Q$ be Q different but real variants of the collection of the introduced station signals. Assume also that the same number of numerical experiments, the homogenization adjustments, were performed and corresponding number of realizations of E^H were obtained using a chain of the calculations

$$E^{Rq} + X^C = X^{Iq},\;\; X^{Iq} \underset{\rightarrow}{g} X^{Hq},\;\; X^{Hq} - X^C = E^{Hq},\;\; q=1,\ldots,Q, \tag{6}$$

Based on these realizations, it is theoretically possible to evaluate $f_{M\times N}$. However, such task is hardly feasible in practice due to extremely large number of dimensions to be considered. On the other hand, based on the statistical ensemble of Q individual realizations of E^H we can evaluate some of the moments of the residual error distribution (5). In the context of our objective, the most important of them are a mean value (m) and some parameter that can characterize a width of the distribution such as a standard deviation (σ) or a percentile range. The mean value provides information regarding a systematic bias of the homogenization adjustment, while the standard deviation or the percentile range characterize its uncertainty. Both statistics, m and σ, can vary in the space-time domain where E^H is defined and they can be evaluated based on formulas

$$m_{ij}=\frac{1}{Q}\sum_{q=1}^{Q} e_{ij}^{Hq}, \tag{7.1}$$

$$\sigma_{ij}=\left(\frac{1}{(Q-1)}\sum_{q=1}^{Q}\left(e_{ij}^{Hq}-m_{ij}\right)^2\right)^{\frac{1}{2}}, \tag{7.2}$$

$$i=1,\ldots,M,\ \ j=1,\ldots,N.$$

While the proposed approach to the evaluation of the adjustment uncertainty on the daily time scale appears attractive and theoretically rigorous, it can potentially lead to some problems that may limit its practical applicability. For instance, one of the limitations can be related to difficulties with a construction of the statistical ensemble for E^R with a sufficient number of its individual realizations in order to perform the calculations according to (6). Another example of limitations can be explained as follow: typically, at the end of the time domain T, all station signals in E^R contain undisturbed segments (see, for example, Fig. 7 a). Hence, a lot of zero values in E^H are usually obtained there. Such zero values have to be excluded from the analysis when evaluating homogenization adjustment since they do not mean ‘perfect’ adjustment. However, it is not very easy to do so, because individual station signals usually have undisturbed segments of different length.

Estimating the statistical properties of the random field of the residual error E^H is not the only way to evaluate the performance of the homogenization adjustment and to quantify its uncertainty on the daily time resolution. An alternative approach is to use specially elaborated statistical metrics or indicators (e.g. Vincent et al., 2018; Trewin, 2018). As noted in Coll et al. (2020), such metrics can provide useful indications in relation to the strengths and weaknesses of homogenization methods used.

2.3.2. Metrics for the adjustment evaluation on the daily time scale. The performance evaluation of an adjustment algorithm and the quantification of its uncertainty are slightly different tasks in several aspects. For instance, we can evaluate the performance even if there is only one realization of the adjustment output X^H. Whereas to define the uncertainty we usually should have the statistical ensemble of X^H ($X^{Hq}, q=1,\ldots,Q$) and the respective ensemble of E^H ($E^{Hq}, q=1,\ldots,Q$). As was mentioned above, a single realization of E^H can be used for the uncertainty quantification only if E^H satisfies the special conditions. The evaluation is usually performed by means of some metrics or statistical indicators. The metrics are computed for each individual station in the data set based on error data E_i^H ($i=1,\ldots,M$) or on comparison of the corresponding pair of time series X_i^H and X_i^C. Calculated for a single output of the homogenization adjustment X^H, they yield general (averaged in time) estimates of the systematic and random residual errors in this actual software run. The metrics values can be averaged over all stations, providing overall (for the whole space domain) evaluation. Some of such averaged metrics, however, can be also used in order to quantify the adjustment uncertainty.

Fig. 8 a shows a graphical comparison between homogenized X_k^H and clean X_k^C time series, presented in Fig. 6 b and c. Similar plot for inhomogeneous X_k^I and clean X_k^C data (Fig. 6 a and c) is presented in Fig.8 b for comparison. The solid bisecting line of black color, usually referred to as a line of true predictions, represents full agreement between respective time series. The perfect/ideal adjustment algorithm would yield corrected values, which would be completely the same as respective clean data. In this case, all dots depicting all pairs $\left(x_{kj}^C, x_{kj}^H\right)$, $j=1,\ldots,N$ would lie on the line of true predictions. The dots lying below the black line mean underestimation of the adjustment algorithm, while the above black line dots show overestimation. Other lines in the diagrams are explained later. The figures are used below for further explanations.

The discrepancy between the homogenized and clean time series (Fig. 8 a) is obviously reduced compared to the discrepancy between the inhomogeneous and clean data (Fig. 8 b). The residual disagreement in Fig. 8 a might be quantified by means of some statistical metrics. Due to the random nature of X_k^H and X_k^C, it is evident, that several metrics should be used because no sole one can provide complete information regarding the residual errors of both types, systematic and random.

Keeping in mind the daily resolution of our data, we applied six different metrics: bias (B), root mean square error ($RMSE$), factor of exceedance ($FOEX$), percentage of days within ±0.5/±2 °C margin ($POD05/POD2$), and difference in slopes ($SlopeD$). The metrics B, $FOEX$ and $SlopeD$ are intended to estimate the systematic errors, while other three, $RMSE$ and $POD05/POD2$, are used for evaluation of the random or scatter residual errors. In the context of the uncertainty evaluation, the two most important metric are B and $RMSE$, which averaged values can also provide information regarding the overall deviation of the adjustment prediction from the true climate signal and the range of the possible residual errors, respectively. Formulas for the majority of the metrics are standard and well known, however we include them for clarification. Note that all formulas are presented for individual pairs of time series, X_i^H and X_i^C, $i=1,\ldots,M$. Obviously, similar metrics can be calculated for inhomogeneous data by replacing X_i^H with X_i^I.

1) Bias

$$B_i=\frac{1}{N_i}\sum_{j=1}^{N_i}\left(x_{ij}^H-x_{ij}^C\right)=\frac{1}{N_i}\sum_{j=1}^{N_i}e_{ij}^H, \tag{8}$$

where N_i is a number of pairs $\left(x_{ij}^C, x_{ij}^H\right)$ in an adjusted segment/segments. The data from the last uncorrected segment are not used in calculations $\left(N_i<N\right)$. The bias can be positive or negative. Depending on its sign it shows average overestimation (+) or underestimation (-) of the adjusted data. However, the bias does not provide any information regarding whether overestimations are more frequent than underestimations or vice-versa. The 'perfect' homogenization algorithm would give 0 for this metric, while $B_i=0$ does not mean that all differences $x_{ij}^H-x_{ij}^C=e_{ij}^H$, $j=1,\ldots,N_i$ are zeros. In a case when the statistical ensemble of Q individual realizations of the adjustment outputs is available, B_i can be averaged over this statistical ensemble. By comparing (7.1) and (8) it

becomes clear that such averaged value can be considered as an estimate of the mean of the random field E^H for i-th station.

2) Root mean squared error

$$RMSE_i = \left(\frac{1}{N_i} \sum_{j=1}^{N_i} \left(x_{ij}^H - x_{ij}^C \right)^2 \right)^{\frac{1}{2}} = \left(\frac{1}{N_i} \sum_{j=1}^{N_i} \left(e_{ij}^H \right)^2 \right)^{\frac{1}{2}}. \tag{9}$$

$RMSE$ provides information about an average deviation of the adjusted data from the true climate signal. However, this metric can be also interpreted as a value that is proportional to the Euclidian distance between X_i^H and X_i^C in a multidimensional space. Consequently, such interpretation provides qualitative explanation why $RMSE_i$, averaged over the statistical ensemble of Q model runs, can characterize the width of possible residual error distribution for i-th station and, hence, can be used to characterize the homogenization adjustment uncertainty. Comparing (7.2) and (9), it can be also concluded, that such averaged value should be close to the standard deviation of the random field E^H for i-th station.

3) Factor of excedance

$$FOEX_i = \left(\frac{N_{\left(x_{ij}^H > x_{ij}^C\right)}}{N_i} - 0.5 \right) 100, \tag{10}$$

where $N_{\left(x_{ij}^H > x_{ij}^C\right)}$ is a number of pairs $\left(x_{ij}^C, x_{ij}^H\right)$ when $x_{ij}^H > x_{ij}^C$, i.e. a homogenized value is overestimated comparing to a respective value from a clean time series. The factor of excedance is measured in % and its values range from -50% to 50%. For instance, $FOEX = 50$ means that all homogenized data are overestimated with respect to true climate data. This measure is widely used in climate analysis and applied meteorology, e.g. Mosca et al. (1998).

4-5) Percentage of days within ±0.5/±2 ºC margin. In addition to the line of true values in Fig 8, other reference lines might be shown on a scatter diagram in order to facilitate the qualitative evaluation of adjustment performance. For instance, pairs of parallels, defined as

$$\left| X_i^H - X_i^C \right| = \Delta T, \tag{11}$$

where $||$ denotes an absolute value, ΔT is a certain threshold of temperature differences, can be drawn. In our study as the thresholds, we chose 0.5ºC following Vincent et al. (2018), and 2ºC by analogy with the Factor of 2 used in other fields of applied meteorology (e.g. Mosca et al., 1998). A pair of such reference lines when $\Delta T = 2$ are shown in red color in Fig. 8. Now metrics $POD05$ and $POD2$ can be simply explained as percentage of dots $\left(x_{ij}^{C}, x_{ij}^{H}\right)$, which lie in the area between respective reference lines (11). That is,

$$POD05_i = \frac{N_{\left|x_{ij}^{H} - x_{ij}^{C}\right| < 0.5}}{N_i} 100 \text{ and } POD2_i = \frac{N_{\left|x_{ij}^{H} - x_{ij}^{C}\right| < 2}}{N_i} 100, \qquad (12)$$

where $N_{\left|x_{ij}^{H} - x_{ij}^{C}\right| \le 0.5}$ and $N_{\left|x_{ij}^{H} - x_{ij}^{C}\right| \le 2}$ mean numbers of dots $\left(x_{ij}^{C}, x_{ij}^{H}\right)$, which lie in the areas inside respective lines (11). Such metrics show how large scatter of the adjusted values around the clean data is.

6) Difference in slopes

$$SlopeD_i = b_i - 1, \qquad (13)$$

where b_i is a slope of a linear regression model $X_i^H = a_i + b_i X_i^C$, built using the standard least-squares approach. The need to introduce such metric can be explained based on Fig. 8 a. As can be seen from this figure, neither B nor $FOEX$ can clearly capture the tendency of general simultaneous underestimation of positive temperatures and overestimation of negative ones (the opposite situation is also possible). The absolute values of the under/over-estimations depend of the temperature magnitude, and they are the largest for temperature extreme. In other words, the under/over-estimation should be reflected in the underestimation of an amplitude of the seasonal cycle showing less variability of the adjusted temperature values. We propose to evaluate such type of discrepancies (systematic error) between homogenized and clean data based on comparison of slopes of the true value line, which always equals to 1, and the linear regression built on the data (blue line in the Fig. 8). The metric is important when evaluating the adjustment of daily data, since the under/over-estimation of values from tails of the temperature distribution can influence calculating of some climate extremes indices. The best value for $SlopeD$ is 0. It worth noting that similar approach was used in (Della-Marta and Wanner, 2006), where comparison of a candidate series against a reference one through a scatter diagram was a part of a newly developed adjustment method. According to this paper, deviation of a slope of a line that fits the data from 1 indicates that daily temperatures at the candidate are less/more variable than those at the reference.

The set of the introduced metrics are capable to provide a fairly detailed description of the adjustment performance on the daily time resolution.

2.3.3. Quantifying discrepancies between homogenized and clean data on the yearly scale. As it was pointed out in the introduction, daily air temperature data are used in order to calculate climate extremes indices. Therefore, it is important to evaluate how accuracy of the adjustment algorithm for data with such temporal resolution is reflected in calculation of these indices and their regular tendencies (trends) (Trewin and Trevitt, 1996). To do so, we calculated yearly time series of the temperature data, TNy and TXy, and the following indices (Klein Tank et al., 2009; Zhang et al., 2011): FD (frost days), TR (tropical nights), TN10p (cold nights), TN90p (warm nights), ID (ice days), SU (summer days), TX10p (cold days), TX90p (warm days). However, due to peculiarities of the Southern Sweden climate (relatively cold) we slightly shifted the standard absolute thresholds in the respective climate extremes indices. That is, instead of 0 and 20°C for FD and TR, respectively, we used -10 and 10°C. Instead of 0 and 25°C for ID and SU, respectively, the thresholds of 5 and 20°C were used. Calculation of the indices was performed for raw, clean and homogenized data based on the RClimDex software (Zhang et al., 2018). After that, quantifying the discrepancies between the indices calculated based on the clean and homogenized data was performed by means of only two metrics, namely B and $RMSE$. Similarly to the daily time series, the metrics were calculated based on only adjusted segment/segments. In addition, we computed differences/errors in the indices linear trends (TrD), calculated for adjusted and clean data. The trends were evaluated over the whole time series (including undisturbed segments) through the least squares regression.

2.3.4. Ensemble of introduced station signals. As was noted above, the main source of the uncertainty for the homogenization adjustment is the station signals introduced into the raw time series. In other words, the results of the adjustment are sensitive to the input data and magnitude of errors contained there. It is natural to expect that the larger the deviation of raw time series from the clean ones, the larger the residual errors should be after the adjustment. In turn, the deviation of the raw time series from the clean data is controlled by the system of break points and corresponding statistical properties of homogeneous segments in the station signals E^R, such as the shift amplitudes/factors, signal to noise ratios etc. In real situation when homogenizing a some set of raw time series, such information is usually unknown. This is a reason why in order to estimate the adjustment uncertainty we have to use the benchmark data and consider all possible but real variants of the station signals or, in other words, consider their statistical ensemble$E^{Rq}, q=1,\ldots,Q$.

Such ensemble is preferred for further calculations, no matter what approach is used to quantify the adjustment uncertainty: the statistical metrics or the random field formalism. Our general idea regarding creating $E^{Rq}, q=1,\ldots,Q$ is to use the collections of the error time series, introduced in the benchmark, and apply to them replacements and/or permutations. As was shown in Section 2.2., the collection of the station signals E^R, that was created in the INDECIS project, possesses statistical properties, which are close to reality. Therefore, we should expect that a sufficient number of the replacements and/or permutations in the set of 94/96 (TN/TX, see Fig. 3) different station signals will provide enough number of individual realizations of E^H. Our methodology will be applied to two different case studies, with increasing complexity, which will be fully described in the Results section.

3. Results

3.1. Case study #1

This first case study considers ten stations (Fig. 9) and limits the length of the corresponding time series to the period of 1971-1980 (10 years). Nine time series (the references), belonging to the stations marked in black color in Fig. 9, are left clean, while the time series of the tenth station (the candidate), depicted in red, is assumed to be corrupted with only one break point dated to 01.01.1976. That is, the first half (1971-1975) of the period under study is intended to be corrupted. Using matrix notations similar to (2), these initial conditions can be written as follows

$$\{x_{ij}^{I}\} = \{x_{ij}^{C}\}, \text{ when } i=1,\ldots,9,\ \ j=1,\ldots,3653, \text{ or } i=10,\ \ j=1827,\ldots,3653; \quad (14.1)$$

$$\{x_{ij}^{I}\} \neq \{x_{ij}^{C}\}, \text{ when } i=10,\ \ j=1,\ldots,1826, \quad (14.2)$$

where 3653 is a total number of days in 1971-1980, 1826 is a number of days in 1971-1975.

An average distance between the candidate and the reference stations is ~34 *km*, while averaged Pearson's correlation coefficient between X_{10}^{C} and X_{i}^{C}, $i=1,\ldots,9$ is 0.96 for TN and 0.97 for TX data. Before the correlation calculation, the seasonal cycle was removed from every time series by using an approach similar to Vincent et al. (2018).

In order to construct the raw data with the corrupted 5-year sub-period ($\{x_{ij}^{I}\}$, $i=10$, $j=1,\ldots,1826$), we analyzed all station signals in E^{R}, that were initially introduced in the INDECIS benchmark, and defined homogeneous error segments, which length is more than 5 complete consecutive years (since January 1 until December 31). For instance, in the error time series shown in Fig. 7 a, all three homogeneous non-zero segments satisfy the stated above condition. The total numbers of such segments in TN and TX error data sets are 185 and 193, respectively. Then 185 for TN and 193 for TX different versions of the raw time series were constructed by shifting a 5-year period from each of the defined segments to 1971-1975 and adding them to the respective clean data $\{x_{ij}^{C}\}$, $i=10$, $j=1,\ldots,1826$. In such way (by performing such replacements), we obtained a statistical ensemble of individual realizations of the raw data set X^{Iq}, $q=1,\ldots,Q$, where $Q=185$ for TN and $Q=193$ for TX. The members of the ensemble differ from each other by only statistical properties of the disturbed segment in the tenth series (see (14.1) and (14.2)), which are well known (Fig. 4 and 5) and, hence, can be considered as controlled. Applying Climatol with the predefined break point to each member of the statistical ensemble, we obtained a sample of the respective number of the adjustment results, which were used for further calculations. It should be mentioned that the average correlation between X_{10}^{Iq}, $q=1,\dots,Q$ and the system of the reference series X_i^C, $i=1,\dots,9$ slightly varies for different q. For TN data the range of the correlation coefficient values is $(0.80, 0.95)$ with the mean around 0.89, while for TX data the range and the mean are $(0.81, 0.96)$ and 0.91, respectively. We believe that such variations are not substantially influencing on the adjustment results and, furthermore, they are unavoidable since they come from the variations of station signals in the statistical ensemble of the candidate time series.

The same corrupted period along with unchanged system of reference series allows to conduct statistically reliable and justified evaluation of the residual errors. Moreover, the approach, used in case study #1, provides an assessment of an almost pure effect of the introduced station signals on the adjustment uncertainty since any other reasons, which might have some influence on the homogenization adjustment, were kept approximately constant or removed.

Fig. 10 shows results of the adjustment uncertainty quantification on daily scale by applying the concept of a random field to the residual errors E^H. Since only one time series of the raw data set was corrupted on 1971-1975, E^H has non-zero values only for one point in the space domain (i.e. for tenth station) and only for the first half of the period under study. Therefore, statistical properties of E^H were defined only for these station and period. In Fig. 10, the mean values, 5$^{\text{th}}$ ($P05$) and 95$^{\text{th}}$ ($P95$) percentiles of empirical distributions of E^H, calculated for each day of 1971-1975, are shown. Figure (a) shows the calculations for TN, while (b) depicts the similar results for TX. The mean values were calculated based on formula (7.1), whereas the percentiles were evaluated based on the samples of Q (185 for TN and 193 for TX) values e_{10j}^{Hq}, $q=1,\dots,Q$ for each day ($j=1,\dots,1826$).

As can be seen from the figure, the calculated parameters, means and percentiles, vary in time. Beside noise, which is due to the limited number of individual realizations in the statistical ensemble, a regular one-year periodicity can be observed. Generally, the range of the residual error is less in summertime compared to winter months. Such non-stationary/periodic behavior of the widths of the residual error distributions can be obviously explained by the similar periodicity of the introduced errors E^R. The reason for the seasonality in E^R is significantly less local spatial variability of air temperature in a summer period compared to winter. Thus, we could expect that the adjusted values of air temperatures, both TN and TX, are closer to the true climate signal in summer than in winter.

The similar 1-year periodicity of the mean values of the residual error distributions implies periodic bias of the air temperature, adjusted by the Climatol software. For both climatic variables, the residual errors are slightly shifted to negative values during summertime, while in winter months the shift has opposite direction. Such bias periodicity means the average underestimation of temperature in summer, and the overestimation in winter and it should have some influence on the amplitude of the seasonal cycle of the adjusted minimum and maximum air temperature.

In order to provide additional evidences for the conclusions, stated after the qualitative analysis of the results presented in Fig. 10, we averaged the empirical error distributions over the whole period, and over January and July months separately (Fig. 11). Table 1 contains some of the parameters of these averaged distributions. Similar parameters for the introduced errors are presented in the table for comparison. The seasonality of the residual error distributions is seen in the figure for both variables and it is also supported by the table content.

In summer months, the percentile intervals of the residual errors, $(P05, P95)$, for the adjusted daily TN and TX air temperatures are $(-2.80, 1.70)$ (^oC) and $(-2.60, 1.90)$ (^oC), respectively. Note, that such quantitative assessments can be considered as one of possible measures of Climatol's adjustment uncertainty. The corresponding mean values of the error distributions are -0.41^oC and -0.22^oC. These results imply that in summer we could expect any adjusted temperature value x_{ij}^H to be slightly underestimated (on average) compared to a respective clean temperature x_{ij}^C by 0.41^oC for TN and $0.22\ ^oC$ for TX. Also, we could expect with 90% probability that for minimum air temperature the adjusted value x_{ij}^H lays in the interval $(x_{ij}^C - 2.80, x_{ij}^C + 1.70)$ (^oC), while for maximum air temperature the interval is $(x_{ij}^C - 2.60, x_{ij}^C + 1.90)$ (^oC). It is important to note a reduction by

~26/11% (TN/TX) in the percentile range length of the residual errors compared to the introduced ones. Such decreasing of the uncertainty is a quantitative assessment of the added value (Sturm and Engström, 2019) of the homogenization adjustment performed by the Climatol software on day-to-day level in a summer period.

In winter months, the ranges $(P05, P95)$, evaluated for the homogenization adjustment errors in TN and TX data are $(-3.60, 4.50)$ (^oC) and $(-2.00, 2.60)$ (^oC), respectively. The corresponding mean values of the error distributions are 0.40^oC for TN and 0.28^oC for TX. Thus, in winter we could expect any adjusted temperature value x_{ij}^H to be slightly overestimated (on average) by $0.40\,^oC$ for TN and 0.28^oC for TX relatively to the respective clean value x_{ij}^C and with 90% probability it lays in the interval $(x_{ij}^C - 3.60, x_{ij}^C + 4.50)$ (^oC) in case of TN air temperature and $(x_{ij}^C - 2.00, x_{ij}^C + 2.60)$ (^oC) in case of TX. Compared to summer months, there is noticeable difference between widths of $(P05, P95)$ intervals calculated for TN and TX winter residual errors. For minimum air temperature such interval is substantially larger (almost twice) meaning larger uncertainty in the adjusted values of TN in this period of the year. Similar to the summer period, the homogenization adjustment reduced the width of the introduced error distribution by15/13% (TN/TX).

The parameters of the empirical distribution of the residual errors, averaged over the whole 5-year period (see Table 1), can characterize only overall (time-averaged) Climatol performance and uncertainty. Some peculiarities of the errors time evolution are neglected. For instance, the shifts of the error mean values in the opposite directions during the winter and summer seasons compensate each other yielding perfect, almost unbiased Climatol's adjustment. The 5^{th} and 95^{th} percentile for TN and TX are between the respective summer and winter values, showing averaged uncertainty of the Climatol software. The standard deviations of the residual error distributions, which also can be used to characterize the adjustment uncertainty along with the percentile range, are 2.15^oC for TN and 1.64^oC for TX. These numbers are important because they can be compared later with averaged values of $RMSE$, which are also intended to show the general/overall uncertainty of the homogenization adjustment. It is worth noting, that parameters of the error distribution for the whole 5-year period can be also used in the evaluation of the adjustment uncertainty in spring and autumn, which can be considered as transitional periods between two limiting cases: summer and winter.

Thus, we can conclude that, if it is possible, the errors of the homogenization adjustment of daily air temperature time series should be evaluated on daily or, at least, seasonal scale. The overall time-averaged evaluation might omit some peculiarities of the residual errors.

Fig. 12 summaries evaluating results of Climatol's adjustment performance (including its uncertainty), which were obtained by applying the statistical metrics. It is important to keep in mind when interpreting these results that the metrics can provide only information regarding overall time-averaged performance of the software. As was pointed above, the six metrics that were used in the study yield detailed evaluation of Climatol's capability to remove systematic and random errors in each individual realization of the raw time series of the statistical ensemble. However, only averaged value of $RMSE$ (averaged over the statistical ensemble) can be considered as measure of the adjustment uncertainty, providing information regarding the width of empirical distribution of the potential residual errors. For each metric, 185/193 (TN/TX) values were calculated, that corresponds to the numbers of individual realizations in the statistical ensembles. These metric values are summarized as boxplots in the figure. Note, that the boxplots of the metrics, calculated for the respective raw data, are also shown for relative evaluation of the adjustment efficiency. Due to very short adjusted period (just 5 years) the climate extremes indices were not calculated and the evaluation of the Climatol software on the yearly scale was not performed in this series of numerical experiments.

As can be seen from the figure, the mean value of bias (B) and its interquartile range (IQR), which we use as a convenient measure of the metric distribution width directly shown on the boxplots, tend to zero for both variables, TN and TX. Similar tendencies are observed for $FOEX$. Here IQR is not zero, but it has relatively small magnitude, especially for TN. Both these metrics indicate the almost perfect performance of the Climatol software in removing systematic errors (shifts in the means). Such conclusion is plainly and brightly supported by a simple visual comparison with the same metrics in the raw data.

However, another type of the systematic residual errors associated with the seasonality of discrepancies between the homogenized and clean data (described by $SlopeD$) is not removed. Moreover, such type of errors seems to be slightly amplified by Climatol in a sense that almost all values of $SlopeD$ became negative compared to symmetric distribution of the metric values in the raw data. That means the simultaneous underestimation of summer temperatures and overestimation of winter ones, and as the result - the underestimation of an amplitude of seasonal cycle. Such conclusion is fully supported by the day-to-day evaluation provided above. The potential ability of the Climatol software to slightly alter seasonality was also pointed out by (Sturm and Engström, 2019).

The performance of the Climatol software in removing random errors is not so pronounced as the removing systematic ones. After adjusting, the means and IQRs of metrics $RMSE$, $POD05$ and $POD2$ for both variables, TN and TX, are slightly improved compared to similar values in the raw data. However, this improvement seems to be associated with the almost perfect removing of break point shifts in the means, and not directly related to the real Climatol's capability to cope with the scatter of errors. The mean value of $RMSE$, which yield the overall, time-averaged assessment of the adjustment uncertainty, is $2.06^{\circ}C$ for TN and $1.53^{\circ}C$ for TX. Such values are very close to the previously calculated standard deviations of the residual error distributions, calculated on the day-to-day level and averaged over 5-year period (see Table 1). The coincidence of the uncertainty estimates that were obtained by applying different approaches indicates robustness of the drawn conclusions and the quantitative assessments. In addition, our assessments of $RMSE$ for TN and TX adjusted data are close to similar estimates presented by Vincent et al. (2018).

It is worth noting again that the provided quantitative assessments of Climatol's performance and uncertainty (as well as those given in the following section) are valid only for cases when the correlation between candidate and reference series is quite high, ~ $(0.80, 0.95)$ for TN and $(0.81, 0.96)$ for Tx. As already mentioned, the uncertainty quantification in other situations, i.e. with other values of correlation ties between time series, will be performed in our future work.

According to (Vincent et al., 2018), adjustment algorithms, applied to daily air temperature data, might show worse ability to remove small size shifts compared to large ones. Thus, it would be interesting to define if there are some relationships between statistical characteristics of the introduced errors, such as their mean value (an amplitude of shift in the break point) and standard deviation (SD), and the corresponding values of the metrics, calculated after applying Climatol. The main purpose of the following calculations is to define what kind of errors (with small or large shift amplitude, with small or large noise component) is removed better. Because the statistical ensemble of Climatol runs contains 185 different individual realizations for TN data, the same numbers of different values of the error means and SDs were calculated and bound to corresponding values of the metrics (Fig. 13). Similar figure was created also for TX, but it is not included in the text. Note, that in Fig. 13 metrics calculated based on the raw data are also shown for comparison.

The relationships for B and $FOEX$ are trivial and they were expected due to the almost perfect performance of the Climatol software in removing jumps in the means. However, other metrics show more interesting dependencies on the error means and SDs. For instance, $SlopeD$ has negative values for any shift amplitude. However, the metric depends almost linearly on SD of the introduced errors. The larger the standard deviation, the larger negative value of $SlopeD$ should be expected, meaning the more intensive seasonality in the residual error time series. There are no any visible relations between the values of $RMSE$, $POD05$ and $POD2$ and the shift amplitudes from some interval around zero (shifts of small magnitudes). In this interval (approximately from -2 to 2 ^{o}C for TN and from -1 to $1^{o}C$ for TX), there are also no visible differences between the metric values computed based on the homogenized and raw data. It means that removing shifts of small magnitudes has small influence the random part of the residual errors. However, certain improvement of the metrics is observed for relatively large shifts. This conclusion is agreed well with the results by Vincent et al. (2018). Similar to $SlopeD$, the metrics $RMSE$, $POD05$ and $POD2$ show noticeable relationships with the standard deviations of the introduced errors. The larger magnitude of this statistical parameter, the larger random residual errors should be expected, what is indicated by the worse values of the metrics.

3.2. Case study #2

This case study is more complex since the raw time series can have more than one break point and their positions are not strictly fixed: they are different in different realizations of the experiment. Here, we used the same ten stations presented in Fig. 9 but considered them on the initially defined period of time 1950-2005. Similar to case study #1, nine time series (the references) are always kept clean, while constructing of the tenth disturbed or candidate series was slightly changed. Formally, these initial conditions can be stated in the following form

$$\{x_{ij}^I\}=\{x_{ij}^C\}, \text{ when } i=1,\ldots,9,\ \ j=1,\ldots,20454, \text{ or } i=10,\ \ j=N_{10}+1,\ldots,20454; \quad (15.1)$$

$$\{x_{ij}^I\}\neq\{x_{ij}^C\}, \text{ when } i=10,\ \ j=1,\ldots,N_{10}, \quad (15.2)$$

where 20454 is a total number of days in 1950-2005, N_{10} is a number of days in a disturbed segment/s of the candidate time series. N_{10} varies in different realizations of the numerical experiment.

In the INDECIS benchmark, 94 and 96 different non-zero station signals were created for TN and TX data, respectively (Fig. 3). By adding these error series to the clean data of the tenth station alternately, we created corresponding numbers of different realizations of raw data, which were used as inputs for the Climatol software. As in the previous case, each realization of this statistical ensemble consists of nine clean and one perturbed time series. By performing such replacement of the station signals, we do not change significantly the statistical properties of the introduced errors: the distributions of their means and standard deviations are almost the same as in case study #1. Besides, we do not change the system of reference stations. Pearson's correlation coefficients between X_{10}^C and X_i^C, $i=1,\ldots,9$ and between X_{10}^{Iq} $(q=1,\ldots,Q)$ and X_i^C, $i=1,\ldots,9$ are almost the same as in the previous case for both TN and TX data. But we change the structure and timing of break points, make it more difficult for Climatol to adjust different segments happened simultaneously in the raw time series. In addition, in this set of numerical experiments we can estimate Climatol's performance and its uncertainty on the yearly scale by defining the residual errors in the adjusted time series of climate extremes indices. Evaluation of the Climatol software in case study #2 on the daily scale was performed only through metrics, i.e. only overall, time-averaging evaluation was carried out. Day-to-day estimation of the residual error distributions, based on the concept of a random field, was not conducted. Such estimation is difficult to perform statistically correct in case study #2 since individual realizations of the raw candidate time series in the statistical ensemble have last undisturbed periods of different lengths. Consequently, for days in the end of 1950-2005 the calculations would operate with considerably less quantity of the non-zero error values compared to days in the beginning of 1950-2005.

Fig. 14 contains boxplots of the metrics that were calculated on the daily scale for the adjusted TN and TX data. Similar to the previous case, we provided also respective metric values for raw data in order to evaluate relative success of the adjustment algorithm.

As it can be seen from the figure, the distributions of the metric values are almost the same as in the previous case. That means good Climatol's performance in removing systematic errors (shifts in the means) and moderate improvement of the metrics showing removing of scatter/random residual errors. However, the seasonality of the residual errors and the related issue of the underestimation of the seasonal cycle amplitude is also preserved in this case study. Therefore, the number of break points in raw time series does not influence significantly the accuracy of Climatol's homogenization adjustment. If they are correctly defined during the detection process, the same (on average) adjustment results should be expected, no matter how many breaks were detected in each of raw time series.

The mean value of *RMSE* for the adjusted TN data is $2.07^{o}C$, while for the TX adjusted time series this parameter equals to $1.54^{o}C$. These values are very close to the similar estimates that were obtained in case study #1. Thus, the overall time-averaged uncertainty of Climatol's adjustment is not influenced significantly by including multiple break points in the raw time series.

The boxplots of the metrics calculated based on the adjusted yearly time series of air temperature data and the climate extremes indices are presented in Fig. 15. Similar results that were obtained based on raw yearly series are also presented in the figure for comparison. As can be seen in the figure, the averaging TN and TX daily data to the yearly scale almost completely remove both types of residual errors. Nearly zero values of *B* for adjusted TNy and TXy series are obvious, since Climatol removes very well systematic errors even in daily data. The mean value of *RMSE* for TNy is reduced after adjustment from 0.94ºC to 0.20ºC (by ~78%) while for TXy the reduction is slightly less: from 0.56ºC to 0.16ºC (by ~63%). Such substantial improvement of *RMSE* for both climatic variables can be explained by the fact that averaging data to yearly scale removes random/noisy part of the residual errors, seen on the daily scale. Note, that the mean values of *RMSE*, 0.20ºC for TNy and 0.16ºC for TXy, can be also considered as the measures of Climatol's adjustment uncertainty on the yearly time scale. In addition, as can be seen in the figure, Climatol removes most of the trend error in TNy and TXy data. The mean value and IQR of *TrD* are almost zeros (~0.00 and ~0.01ºC/decade, respectively) for both climatic variables.

Climatol removes well both types of errors also in the time series of all considered extreme indices. This is clearly seen in the figure, where empirical distributions of *B* and *RMSE*, calculated based on the adjusted data, can be compared with similar distributions, obtained for raw series. Both metrics for all indices indicate substantial improvement after applying Climatol's adjustment. The underestimation of the seasonal cycle amplitude in the adjusted data, seen on the daily time resolution, is not so noticeable in the indices time series, probably due to relatively small negative values of *SlopeD* (see Fig. 14). However, the means of *B* for all indices with fixed thresholds are slightly negative, meaning general slight underestimation of these indices in the adjusted data.

Below we focus mainly on trend evaluation in the time series of the extreme indices due to their critical importance in climatological applications. The empirical distributions of errors (differences) in trends, *TrD*, calculated for adjusted data are also presented in Fig. 15. Table 2 contains some of parameters of the empirical distributions of *TrD* values. The first noticeable qualitative conclusion that can be drawn from the figure is substantial decreasing of the trend errors in the adjusted data compared to the raw ones. Regular tendencies of all extreme indices, evaluated based on corrected data, are much closer to the real trends than evaluated based on the raw time series.

Based on the table content, quantitative assessments of Climatol's accuracy and uncertainty in the indices trend calculation can be derived. For instance, the mean value of the trend errors in the adjusted series of FD (frost days) is relatively small, 0.29 *days/decade* (2.9 *days/100years*). The uncertainty of the trend calculation in the adjusted FD data can be estimated by mean of the standard deviation (0.42 *days/decade*) or the percentile range $(P05, P95)$, which is $(-0.23, 0.94)$ (*days/decade*). Thus, we could expect, that a linear trend, calculated in the FD yearly time series that was corrected by the Climatol software, is slightly shifted (on average) on 0.29 *days/decade* relatively to the true climate trend (Tr^C), and with 90% probability it lies in the interval $(Tr^C - 0.23, Tr^C + 0.94)$(*days/decade*). It is worth noting, that the percentile range of the trend errors in the raw time series is significantly larger, $(-3.00, 2.92)$(*days/decade*), i.e. after applying Climatol, a 80% decrease of the uncertainty can be reported. Similar assessments can be obtained from Table 2 for other climate extreme indices. We also can conclude, that, in general, the trends can be estimated more accurately and with less uncertainty in the adjusted time series of the TX extreme climate indices than in TN extremes. One more important conclusion is that despite the substantial amount of the residual scatter/random errors which still remained in the adjusted daily time series, the linear trends calculated on the corrected yearly time series are reliable and close to real regular tendencies and they can be evaluated with significantly removed uncertainty.

4. Conclusion

In this study, the uncertainty quantification and the general performance evaluation of Climatol's adjustment algorithm, applied to daily minimum and maximum air temperature time series, are presented. We focused our attention only on the most influencing and important source of the uncertainty, namely introduced station signals into the raw data set to be adjusted. Other possible sources of the adjustment uncertainty were removed from the analysis or kept approximately constant. For instance, the mean correlation between candidate and reference series was around $(0.80, 0.95)$ for TN and $(0.81, 0.96)$ for Tx data. Therefore, our results are valid only for cases where the mentioned mean correlation can be observed. The sensitivity of the obtained quantitative assessments to other factors/sources will be addressed in our future work.

In order to evaluate the adjustment uncertainty, we used the INDECIS benchmark data and applied a complex approach, quantifying the uncertainty at different levels of detail and time resolution. According to our findings, Climatol's adjustment uncertainty, evaluated on day-to-day level, varies in time and depends on the season. In summer months, the residual errors in the adjusted daily TN and TX series are expected to belong to the intervals, $(P05, P95)$, $(-2.80, 1.70)$ (^{o}C) and $(-2.60, 1.90)$ (^{o}C), respectively. In winter months, the ranges of the possible remaining errors are larger: $(-3.60, 4.50)$ (^{o}C) for TN and $(-2.00, 2.60)$ (^{o}C) for TX. The overall adjustment uncertainty, averaged over all seasons, can be evaluated as the error range, $(P05, P95)$, $(-3.20, 3.20)$ (^{o}C) for TN and $(-2.50, 2.30)$ (^{o}C) for TX. In terms of standard deviations of the residual error distributions, the overall uncertainty can be evaluated as $2.15^{o}C$ for TN and $1.64^{o}C$ for TX data. These estimates agree well with the mean values of *RMSE*, which also can be used as a measure of the width of the empirical distribution of the residual errors. Besides 1-year periodicity in the width of the residual error distributions, their mean values are also slightly shifted periodically. For both climatic variables, the shift is toward negative values during summertime, while in winter months it has opposite direction. Such peculiarities of the residual errors can lead to the slight underestimation of the amplitude of the seasonal cycle of the adjusted TN and TX data. The calculations based on the specially introduced metric (*SlopeD*) provide additional evidence for such conclusion. Other metrics, used in the study, showed that Climatol removes extremely well systematic errors related to jumps in the mean and this Climatol's capability is valid for shifts of any magnitude and does not depend on the number of break points in the raw time series. The ability of Climatol to remove scatter/random errors in the daily raw time series is not so pronounced.

However, on the yearly time scale, both types of residual errors are removed well in adjusted time series. The adjusted yearly TN and TX temperature data are unbiased, and their uncertainty is reduced significantly: mean values of *RMSE* for TNy and TXy were decreased to 0.20°C (by ~78%) and 0.16°C (by ~63%), respectively. In addition, Climatol removes most of the trend error in TNy and TXy data, so trend analysis is more solid and better represents climate variations.

Similar conclusions are valid for the yearly time series of the considered climate extreme indices: both types of errors are removed well by Climatol. The underestimation of the seasonal cycle amplitude in the adjusted data, seen on the daily time resolution, is not clearly reflected in the indices time series. However, the mean values of bias (B) for all indices with fixed thresholds are slightly negative, meaning slight underestimation of these indices in the adjusted data. However, this does not have substantial influence on the linear trend calculations in the indices time series. The trends calculated in the adjusted time series are generally unbiased. The percentile $\left(P05, P95\right)$ ranges of the errors in the indices trends, calculated based on adjusted data, is reduced by ~70-80% compared to the trend errors in the corresponding raw time series. Despite the substantial amount of the residual scatter errors in daily time series, the linear trends calculated on the corrected yearly time series are close to real regular tendencies and they can be evaluated with significantly removed uncertainty.

Acknowledgement. The work was performed in the frame of the INDECIS project, that is a part of ERA4CS, an ERA-NET initiated by JPI Climate, and funded by FORMAS (SE), DLR (DE), BMWFW (AT), IFD (DK), MINECO (ES), ANR (FR) with co-funding by the European Union (Grant 690462). The work has been partially supported by the Ministry of Education and Science of Kazakhstan (Grant BR05236454) and Nazarbayev University (Grant 090118FD5345)

References

1. Aguilar E., Auer I., Brunet M., Peterson T.C., Wieringa J. 2003. *WMO Guidelines on climate metadata and homogenization*. WCDMP No.53, WMO-TD No. 1186, WMO, Geneva, Switzerland.
2. Aguilar E., van der Schrier G., Guijarro J.A., Stepanek P., Zahradnicek P., Sigro J., Coscarelli R., Engstrom E., Curley M., Caloiero T., Lledo L., Ramon J., Antonia Valente M. 2018. Quality control and homogenization benchmarking-based progress from the INDECIS Project. Vienna, Austria: General Assembly of the European Geosciences Union, 8–13 April 2018, EGU2018-16392
3. Alexander L.V., Zhang X., Peterson T.C., Caesar J., Gleason B., Klein Tank A.M.G., Haylock M., Collins D., Trewin B., Rahim F., Tagipour A., Kumar Kolli R., Revadekar J.V., Griffiths G., Vincent L., Stephenson D.B., Burn J., Aguilar E., Brunet M., Taylor M., New M., Zhai P., Rusticucci M., Vazquez Aguirre J.L. 2006. Global observed changes in daily climate extremes of temperature and precipitation. *J. Geophys. Res.*, 111, D05109. https://doi.org/10.1029/2005JD006290
4. Alexandersson H. 1986. A homogeneity test applied to precipitation data. *J. of Climatol.*, **6** (6), 661-675. https://doi.org/10.1002/joc.3370060607
5. Alexandersson H, Moberg A. 1997. Homogenization of Swedish temperature data. Part I: homogeneity test for linear trends. *Int. J. Climatol.*, **17** (1), 25–34. https://doi.org/10.1002/(SICI)1097-0088(199701)17:1<25::AID-JOC103>3.0.CO;2-J
6. Azorin-Molina C., Guijarro J.A., McVicar T.R., Trewin B.C., Frost A.J., Chen D. 2019. An approach to homogenize daily peak wind gusts: An application to the Australian series. *Int. J. Climatol*., **39** (4), 2260–2277. https://doi.org/10.1002/joc.5949
7. Brunet M., Saladié O., Jones P., Sigró J., Aguilar E., Moberg A., Lister D., Walther A., Almarza C. 2008. A case-study/guidance on the development of long-term daily adjusted

temperature datasets. WMO-TD No. 1425, WCDMP No. 66. World Meteorological Organization, Geneva

8. Coll J., Domonkos P., Guijarro J., Curley M., Rustemeier E., Aguilar E., Walsh S., Sweeney J. 2020. Application of homogenization methods for Ireland's monthly precipitation records: Comparison of break detection results. *Int. J. Climatol.*, 1–20. https://doi.org/10.1002/joc.6575
9. Collins W.J., Bellouin N., Doutriaux-Boucher M., Gedney N., Hinton T., Jones C. D., Liddicoat S., Martin G., O'Connor F., Rae J., Senior C., Totterdell I., Woodward S., Reichler T., Kim J. 2008. Evaluation of the HadGEM2 model. MetOffice Hadley Centre Technical Note 74, 47 pp.
10. DeGaetano A.T. 2006. Attributes of several methods for detecting discontinuities in mean temperature series. *J. Climate,* **19**, 838–853. https://doi.org/10.1175/JCLI3662.1
11. Della-Marta P., Wanner H. 2006. A method of homogenizing the extremes and mean daily temperature measurements. *J. Climate,* **19**, 4179–4197. https://doi.org/10.1175/JCLI3855.1
12. Domonkos P. 2011. Efficiency evaluation for detecting inhomogeneities by objective homogenization methods. *Theor. Appl. Climatol.,* **105**, 455-467. https://doi.org/10.1007/s00704-011-0399-7
13. Domonkos P., Efthymiadis D. 2013. Development and testing of homogenization methods: moving parameter experiments with ACMANT. *Adv. Sci. Res.*, **10**, 43–50, https://doi.org/10.5194/asr-10-43-2013
14. Domonkos P. 2017. Time series homogenization with optimal segmentation and ANOVA correction: past, present and future. *Proceeding of 9th Seminar for homogenization and quality control in climatological databases and 4th conference on spatial interpolation techniques in climatology and meteorology* (Budapest, April 3-7), WMO WCDMP-No.85, pp. 46-62

15. Domonkos P., Coll J. 2017. Time series homogenization of large observational datasets: impact of the number of partner series on efficiency. *Clim. Res.*, **74**, 31-42. https://doi.org/10.3354/cr01488

16. Ducré-Robitaille J.-F., Vincent L.A., Boulet G. 2003. Comparison of techniques for detection of discontinuities in temperature series. *Int. J. Climatol.*, **23**, 1087-1101. https://doi.org/10.1002/joc.924

17. Dumitrescu A., Cheval S., Guijarro J.A. 2020. Homogenization of a combined hourly air temperature dataset over Romania. *Int. J. Climatol.*, **40** (5), 2599-2608. https://doi.org/10.1002/joc.6353

18. Fioravanti G., Piervitali E., Desiato, F. 2019. A new homogenized daily data set for temperature variability assessment in Italy. *Int. J. Climatol.*, **39** (15), 5635-5654. https://doi.org/10.1002/joc.6177

19. Guijarro J.A. 2011. Influence of network density on homogenization performance. *Proceeding of 7th Seminar for Homogenization and Quality Control in Climatological Databases jointly organized with the Meeting of COST ES0601 (HOME) Action MC Meeting*. Budapest, Hungary, 24-27 October, WMO WCDMP-No. 78, pp. 11-18.

20. Guijarro J.A., López J.A., Aguilar E., Domonkos P., Venema V.K.C., Sigró J., Brunet M. 2017. Comparison of homogenization packages applied to monthly series of temperature and precipitation: The MULTITEST project. *Proceeding of 9th Seminar for homogenization and quality control in climatological databases and 4th conference on spatial interpolation techniques in climatology and meteorology*. Budapest, Hungary, 3-7 April 2017, WMO WCDMP-No.85, pp. 46-62.

21. Guijarro J.A. 2018. Homogenization of climatic series with Climatol. Version 3.1.1. Guide.

22. Guijarro J.A., Aguilar E., Caloiero T., Coscarelli R., Curley M., Pérez-Zanón N. 2018. Homogenization of daily Essential Climatic Variables with Climatol 3.1 within the INDECIS

project. Budapest, Hungary: European Conference for Applied Meteorology and Climatology, 3-7 September 2018, EMS2018-413.

23. Guijarro J.A., Aguilar E., Domoncos P., Sigró J., Štepánek P., Venema V., Zahradnícek P. 2019. Benchmarking results of the homogenization of daily Essential Climatic Variables within the INDECIS project. Vienna, Austria: General Assembly of the European Geosciences Union, 7–12 April 2019, EGU2019-10896-1.

24. Hartmann D. L., Klein Tank A.M.G., Rusticucci M., Alexander L.V., Brönnimann S., Charabi Y., Dentener F.J., Dlugokencky E.J., Easterling D.R., Kaplan A., Soden B.J., Thorne P.W., Wild M., Zhai P.M. 2013. Observations: atmosphere and surface. In: *Climate Change: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.* Cambridge University Press: Cambridge, UK and New York. NY.

25. INDECIS, 2018. Integrated approach for the development across Europe of user oriented climate indicators for GFCS high-priority sectors: agriculture, disaster risk reduction, energy, health, water and tourism. http://www.indecis.eu/ [Accessed January 10, 2020]

26. Iman R.L., Helton J.C. 1988. An Investigation of Uncertainty and Sensitivity Analysis Techniques for Computer Models. *Risk Analysis.*, **8** (1), 71-90. https://doi.org/10.1111/j.1539-6924.1988.tb01155.x

27. Jakeman A.J., Letcher R.A., Norton J.P. 2006. Ten iterative steps in development and evaluation of environmental models. *Environ. Model. Softw.*, **21** (5), 602–614. https://doi.org/10.1016/j.envsoft.2006.01.004

28. Killick R.E. 2016. Benchmarking the Performance of Homogenization Algorithms on Daily Temperature Data. PhD Thesis, University of Exeter, 249 pp. Available from https://ore.exeter.ac.uk/repository/handle/10871/23095 (November 2019).

29. 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., Antonio López J., 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** (12), 1441-1453. https://doi.org/10.1002/joc.773

30. Klein Tank A.M.G., Zwiers F.W., Zhang X. 2009. Guidelines on analysis of extremes in a changing climate in support of informed decisions for adaptation, climate data and monitoring WCDMP-No 72, WMO-TD No 1500, p 55.
31. Kuglitsch F.G., Auchmann R., Bleisch R., Broennigmann S., Martius O., Stewart M. 2012. Break detection of annual Swiss temperature series. *J. Geophys. Res.*, **117**, D13105. https://doi.org/10.1029/2012JD017729
32. Lindau R., Venema V. 2016. The uncertainty of break positions detected by homogenization algorithms in climate records. *Int. J. Climatol.*, **36**, 576-589. https://doi.org/10.1002/joc.4366
33. Mamara A., Argiriou A.A., Anadranistakis M. 2013. Homogenization of mean monthly temperature time series of Greece. *Int. J. Climatol.*, **33** (12), 2649–2666. https://doi.org/10.1002/joc.3614
34. Mamara A., Argiriou A.A., Anadranistakis M. 2014. Detection and correction of inhomogeneities in Greek climate temperature series. *Int. J. Climatol.*, **34** (10), 3024–3043. https://doi.org/10.1002/joc.3888
35. Meseguer-Ruiz O., Ponce-Philimon P.I., Quispe-Jofré A.S., Guijarro J.A., Sarricolea P. 2018. Spatial behavior of daily observed extreme temperatures in Northern Chile (1966–2015): data quality, warming trends, and its orographic and latitudinal effects. *Stoch. Environ. Res. Risk Assess.*, **32**, 3503–3523. https://doi.org/10.1007/s00477-018-1557-6

36. Mestre O., Gruber C., Prieur C., Caussinus H., Jourdain S. 2011. SPLIDHOM: a method for homogenization of daily temperature observations. *J. Appl. Meteorol. Climatol.*, **50**, 2343–2358. https://doi.org/10.1175/2011JAMC2641.1

37. Mosca S., Graziani G., Klug W., Bellasio R., Bianconi R. 1998. A statistical methodology for the evaluation of long-range dispersion models: an application to the ETEX exercise. *Atmos. Environ.*, **32** (24), 4307–4324. http://dx.doi.org/10.1016/S1352-2310(98)00179-4.

38. MULTITEST, 2015. http://www.climatol.eu/MULTITEST/ [Accessed January 10, 2020]

39. Osadchyi V., Skrynyk O.A., Radchenko R., Skrynyk O.Y. 2018. Homogenization of Ukrainian air temperature time series. *Int. J. Climatol.*, **38** (1), 497-505. https://doi.org/10.1002/joc.5191.

40. Pérez-Zanón N., Sigró J., Aguilar E., Guijarro J.A., van der Schrier G., Stepanek P., Zahradnicek P., Coscarelli R., Engström E., Curley M., Caloiero T., Lledó L., Ramon J., Valente M.A., Carvalho S. 2018. First Steps towards a Benchmarking Experiment in Quality Control and Homogenization of Observed Data. Budapest, Hungary: European Conference for Applied Meteorology and Climatology, 3-7 September 2018, EMS2018-465.

41. Prohom M., Barriendosb M., Sanchez-Lorenzod A. 2016. Reconstruction and homogenization of the longest instrumental precipitation series in the Iberian Peninsula (Barcelona, 1786–2014). *Int. J. Climatol.*, **36** (8), 3072–3087. https://doi.org/10.1002/joc.4537.

42. Reeves J., Chen J., Wang X.L., Lund R., Lu Q. 2007. A review and comparison of change points detection techniques for climate data. *J. Appl. Meteorol. Climatol.*, **46**, 900–915. https://doi.org/10.1175/JAM2493.1

43. Sanchez-Lorenzo A., Wild M., Brunetti M., Guijarro J.A., Hakuba M.Z., Calbó J., Mystakidis S., Bartok B. 2015. Reassessment and update of long-term trends in downward surface shortwave radiation over Europe (1939–2012). *J. Geophys. Res. Atmos.*, **120**, 9555–9569, https://doi.org/10.1002/2015JD023321.

44. Skrynyk O.Y., Aguilar E., Skrynyk O.A., Sidenko V., Boichuk D., Osadchyi V. 2019. Quality control and homogenization of monthly extreme air temperature of Ukraine. *Int. J. Climatol.*, **39** (4), 2071-2079. https://doi.org/10.1002/joc.5934

45. Sokal R.R., Rohlf P.J. 1969. Introduction to Biostatistics. 2nd edition, 363 pp, W.H. Freeman, New York. ISBN 978-0486469614

46. Sturm C., Engström E. 2019. Estimating the sensitivity and accuracy of homogenization: a case study with Climatol on temperature from the INDECIS benchmark. 12th EUMETNET Data Management Workshop, De Bilt, the Netherlands, 6-8 November 2019.

47. Szentimrey T. 2008. Methodological questions of series comparison. *Proceeding of 6th Seminar for Homogenization and Quality Control in Climatological Databases*. Budapest, Hungary, 26-30 May 2008, WMO WCDMP-No. 76, pp. 1−7.

48. Trewin B.C., Trevitt A.C.F. 1996. The development of composite temperature records. *Int. J. Climatol.*, **16** (11), 1227–1242. https://doi.org/10.1002/(SICI)1097-0088(199611)16:11<1227::AID-JOC82>3.0.CO;2-P

49. Trewin B. 2010. Exposure, instrumentation, and observing practice effects on land temperature measurements. *WIREs Clim. Change*, **1** (4), 490–506. https://doi.org/10.1002/wcc.46.

50. Trewin B. 2013. A daily homogenized temperature data set for Australia. *Int. J. Climatol.*, **33**, 1510–1529. https://doi.org/10.1002/joc.3530

51. Trewin B. 2018. The Australian Climate Observations Reference Network – Surface Air Temperature (ACORN-SAT).Version 2. Bureau Research Report No. 032. Available at: http://www.bom.gov.au/climate/change/acorn-sat/documents/BRR-032.pdf [Accessed April 2019].

52. Venema V, Mestre O, Aguilar E, Auer I, Guijarro JA, Domonkos P, Vertacnik G, Szentimrey T, Stepanek P, Zahradnicek P, Viarre J, Muller-Westermeier G, Lakatos M, Williams CN, Menne M, Lindau R, Rasol D, Rustemeier E, Kolokythas K, Marinova T, Andresen L,

Acquaotta F, Fratianni S, Cheval S, Klancar M, Brunetti M, Gruber C, Duran MP, Likso T, Esteban P, Brandsma T. 2012. Benchmarking monthly homogenization algorithms. *Clim. Past*, **8**, 89–115. https://doi.org/10.5194/cp-8-89-2012.

53. Vincent L.A., Milewska E.J., Wang X.L., Hartwell M.M. 2018. Uncertainty in homogenized daily temperatures and derived indices of extremes illustrated using parallel observations in Canada. *Int. J. Climatol.*, **38** (2). 692-707. https://doi.org/10.1002/joc.5203
54. Walker W.E., Harremoës P., Rotmans J., van der Sluijs J.P., van Asselt M.B.A., Janssen P., Krayer von Krauss M.P. 2003. Defining Uncertainty: A conceptual basis for uncertainty management in model-based decision support. *Integr. Assess.*, **4** (1), 5-17. https://doi.org/10.1076/iaij.4.1.5.16466
55. Willett K., Williams C., Jolliffe I.T., Lund R., Alexander L.V., Brönnimann S., Vincent L.A., Easterbrook S., Venema V.K.C., Berry D., Warren R.E., Lopardo G., Auchmann R., Aguilar E., Menne M.J., Gallagher C., Hausfather Z., Thorarinsdottir T., Thorne P.W. 2014. A framework for benchmarking of homogenization algorithm performance on the global scale. *Geosci. Instrum. Method. Data Syst.*, **3**, 187–200. https://doi.org/10.5194/gi-3-187-2014
56. Yosef Y., Aguilar E., Alpert P. 2018. Detecting and adjusting artificial biases of long-term temperature records in Israel. *Int. J. Climatol.*, **38** (8), 3273-3289. https://doi.org/10.1002/joc.5500.
57. Yozgatligil C., Yazici C. 2016. Comparison of homogeneity tests for temperature using a simulation study. *Int. J. Climatol.*, **36**, 62–81. https://doi.org/10.1002/joc.4329
58. Zhang X., Alexander L., Hegerl G.C., Jones P., Klein Tank A., Peterson T.C., Trewin B., Zwiers F.W. 2011. Indices for monitoring changes in extremes based on daily temperature and precipitation data. *WIRES Clim. Change*, **2**, 851–870. https://doi.org/10.1002/wcc.147
59. Zhang X., Feng Y., Chan R. 2018. Introduction to RClimDex v1.9. Guide. Climate research Division, Environment Canada, Downsview Ontario, Canada.

Tables

Table 1. Parameters of averaged empirical distributions of errors: homogenization/residual E^H and real/introduced E^R (all in ^oC)

		Year		January		July	
		E^H	E^R	E^H	E^R	E^H	E^R
TN	Mean	-0.03	-0.11	0.40	-0.08	-0.41	-0.13
	SD	2.15	2.53	2.56	2.97	1.39	1.85
	P05	-3.20	-4.00	-3.60	-4.90	-2.80	-3.20
	P95	3.20	3.70	4.50	4.60	1.70	2.90
	P95-P05	6.40	7.70	8.10	9.50	4.50	6.10
TX	Mean	-0.02	-0.00	0.28	-0.03	-0.22	0.04
	SD	1.64	1.84	1.58	1.78	1.48	1.67
	P05	-2.50	-2.70	-2.00	-2.70	-2.60	-2.50
	P95	2.30	2.60	2.60	2.60	1.90	2.50
	P95-P05	4.80	5.30	4.60	5.30	4.50	5.00

Table 2. Parameters of empirical probability distributions of *TrD* (errors/differences in linear trends), defined for yearly time series of climate extreme indices: (a) TN, (b) TX

a)	FD *days/decade*		TR *days/decade*		TN10p *%/decade*		TN90p *%/decade*	
	hom-cln	raw-cln	hom-cln	raw-cln	hom-cln	raw-cln	hom-cln	raw-cln
Mean	0.29	-0.26	0.64	-0.79	-0.35	-0.52	-0.29	-0.73
SD	0.42	1.83	0.74	3.59	0.42	1.25	0.34	1.27
P05	-0.23	-3.00	-0.42	-6.65	-1.02	-2.22	-0.79	-2.54
P95	0.94	2.92	2.05	2.55	0.32	1.44	0.31	0.28
P95-P05	1.17	5.92	2.47	9.20	1.34	3.66	1.10	2.82
b)	ID *days/decade*		SU *days/decade*		TX10p *%/decade*		TX90p *%/decade*	
	hom-cln	raw-cln	hom-cln	raw-cln	hom-cln	raw-cln	hom-cln	raw-cln
Mean	-0.05	-0.36	0.21	-0.56	-0.13	-0.13	-0.10	-0.36
SD	0.27	0.88	0.44	1.73	0.33	0.79	0.23	0.64
P05	-0.49	-1.88	-0.37	-3.41	-0.71	-1.47	-0.49	-1.40
P95	0.39	0.96	0.96	2.00	0.33	1.06	0.23	0.56
P95-P05	0.88	2.84	1.33	5.41	1.04	2.53	0.72	1.96

Figures

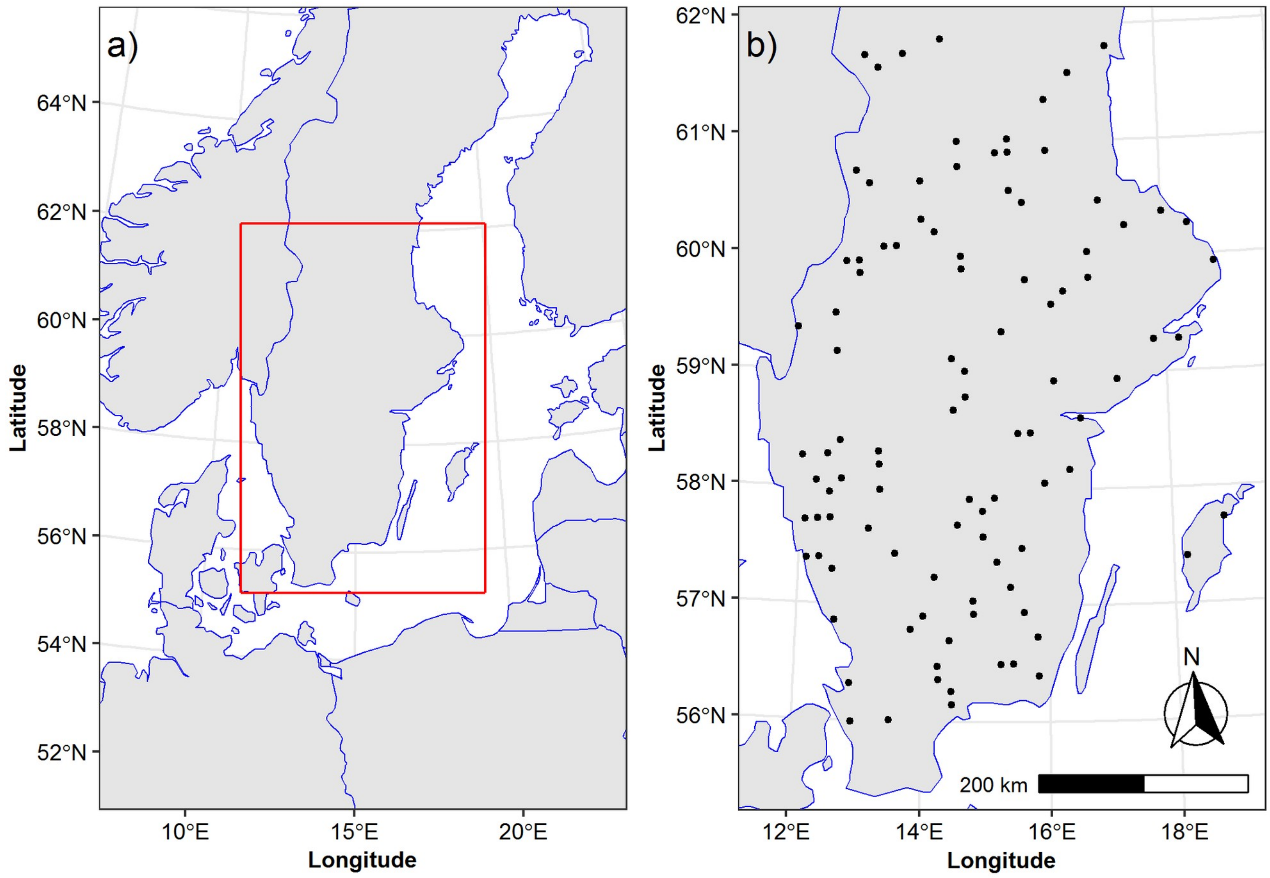

Fig. 1. (a) The domain of the Southern Sweden (inside of the red rectangular frame) and (b) locations of the 'stations' (the subset of the RACMO grid points, shown as black dots) on it

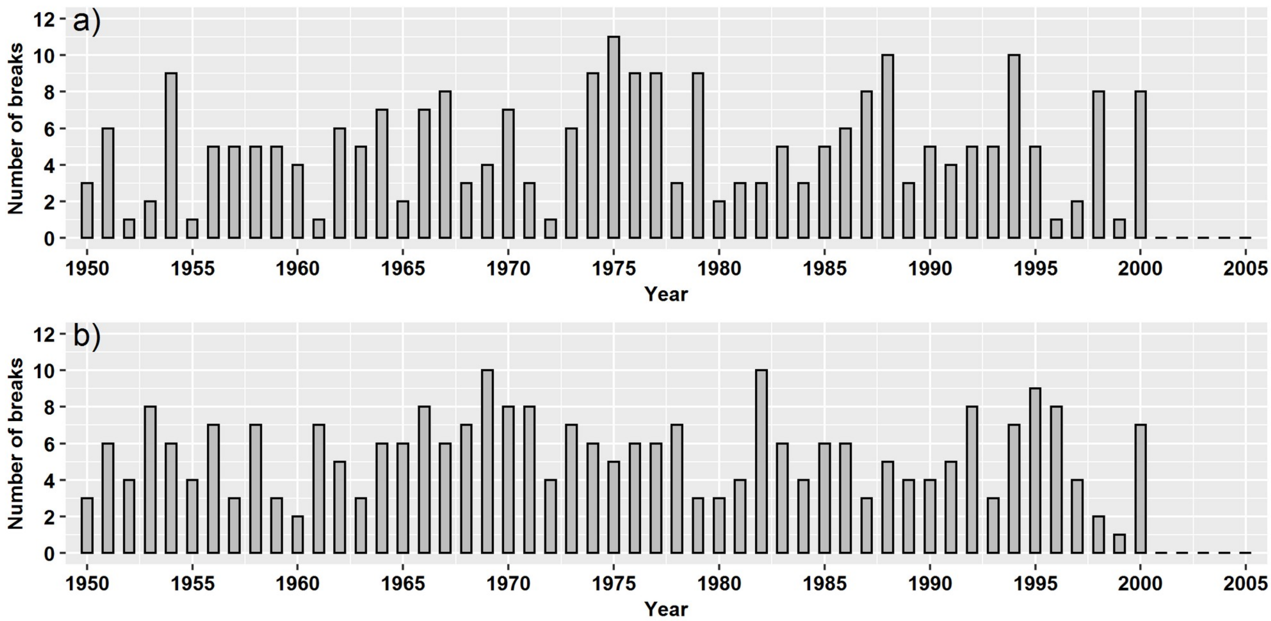

Fig. 2. Number of break points per year introduced to clean (a) TN and (b) TX air temperature time series

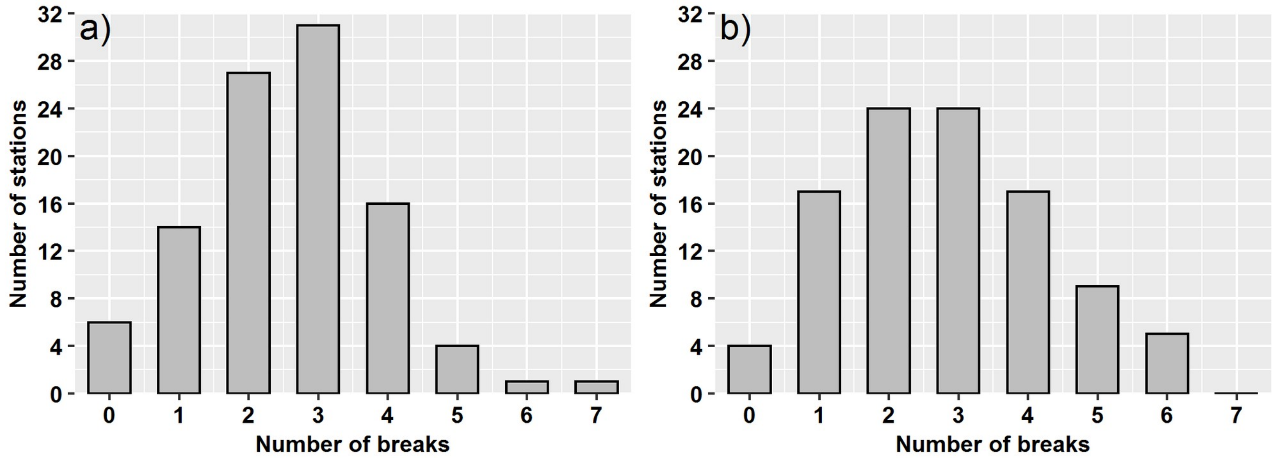

Fig. 3. Distribution of the number of stations/time series with respect to the number of break points in one time series: (a) TN, (b) TX

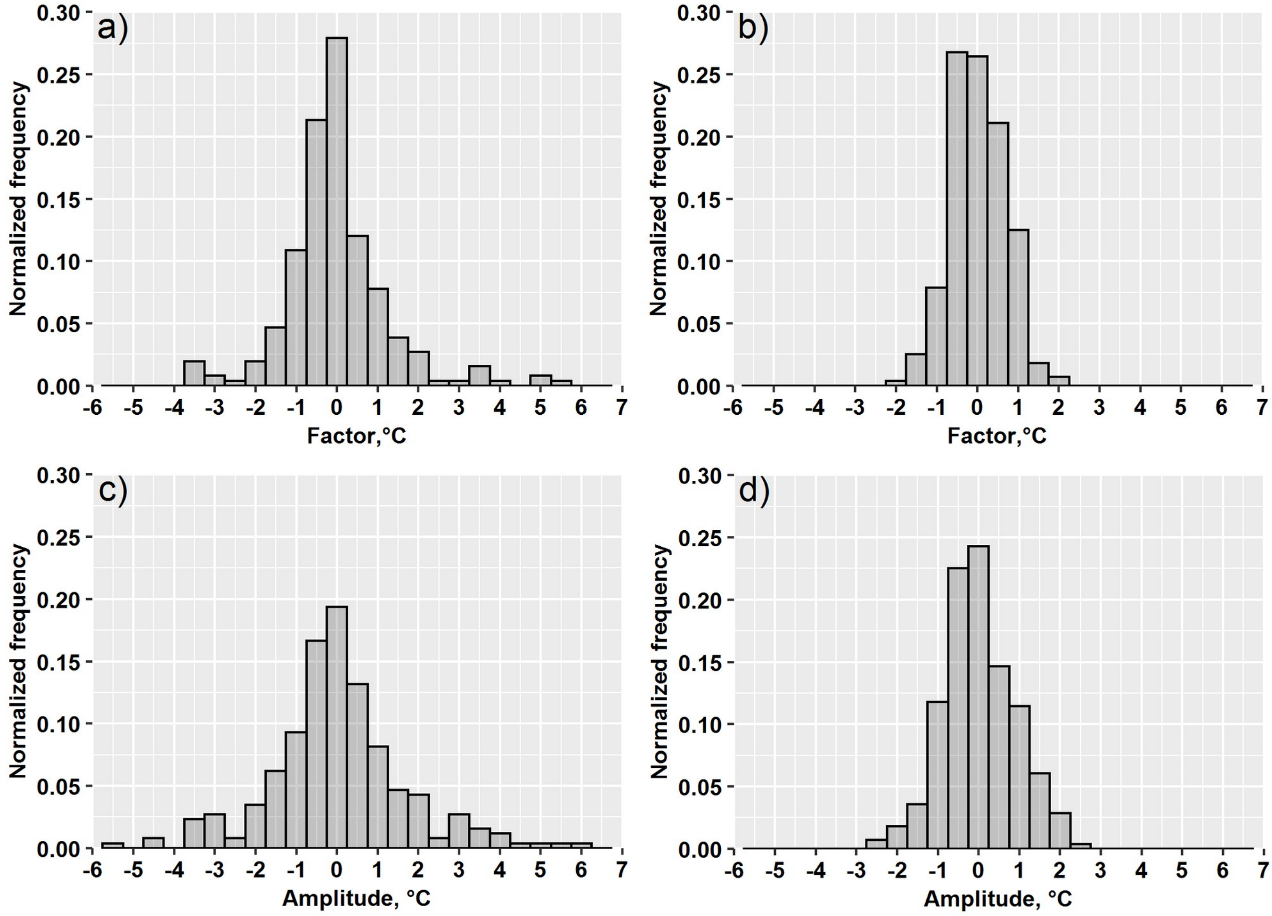

Fig. 4. Histograms of the factors (a, b) and amplitudes (c, d) of the shifts at break points, that were introduced to TN (a, c) and TX (b, d) clean data sets. The frequency/count was normalized by the total number of the breaks. The factors/amplitudes were estimated by averaging homogeneous segments in the time series of the introduced error

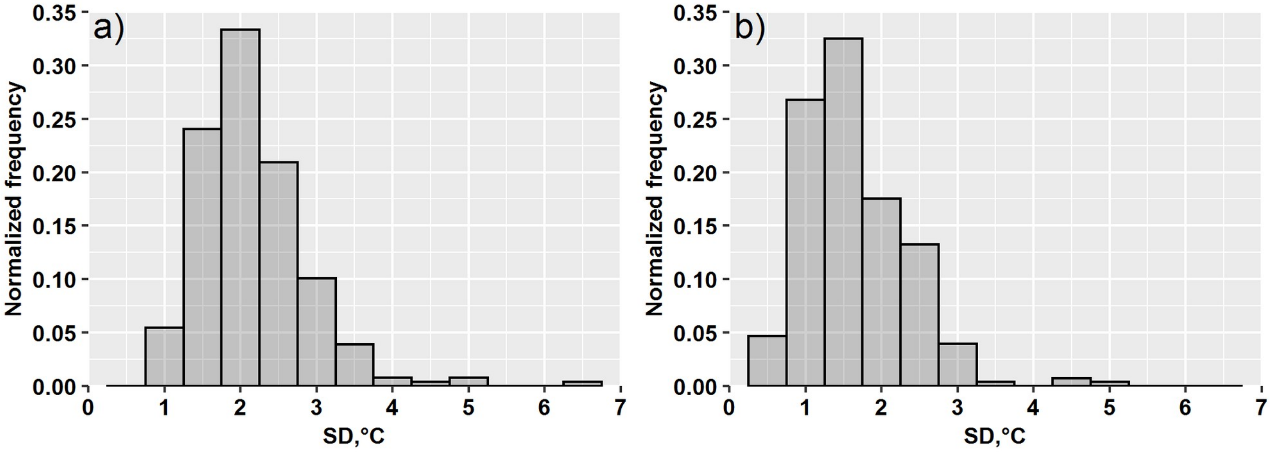

Fig. 5. Histograms of standard deviations (SD) of the introduced errors at the homogeneous segments: (a) TN, (b) TX. The frequency/count was normalized by the total number of the breaks.

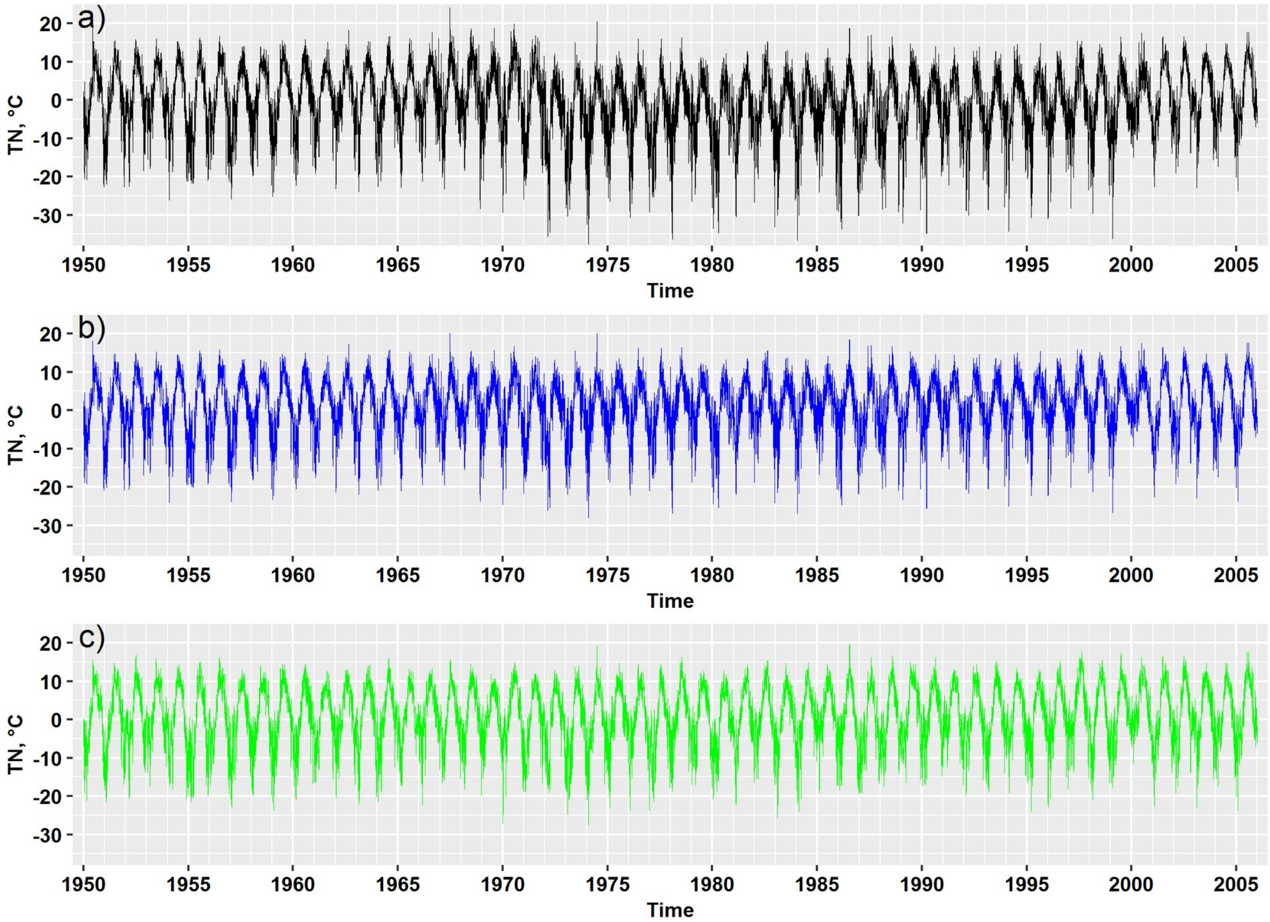

Fig. 6. Examples of TN time series belonging to the same (k-th) station extracted from the inhomogenious X^I (a), homogenized X^H (b) and clean X^C (c) data sets

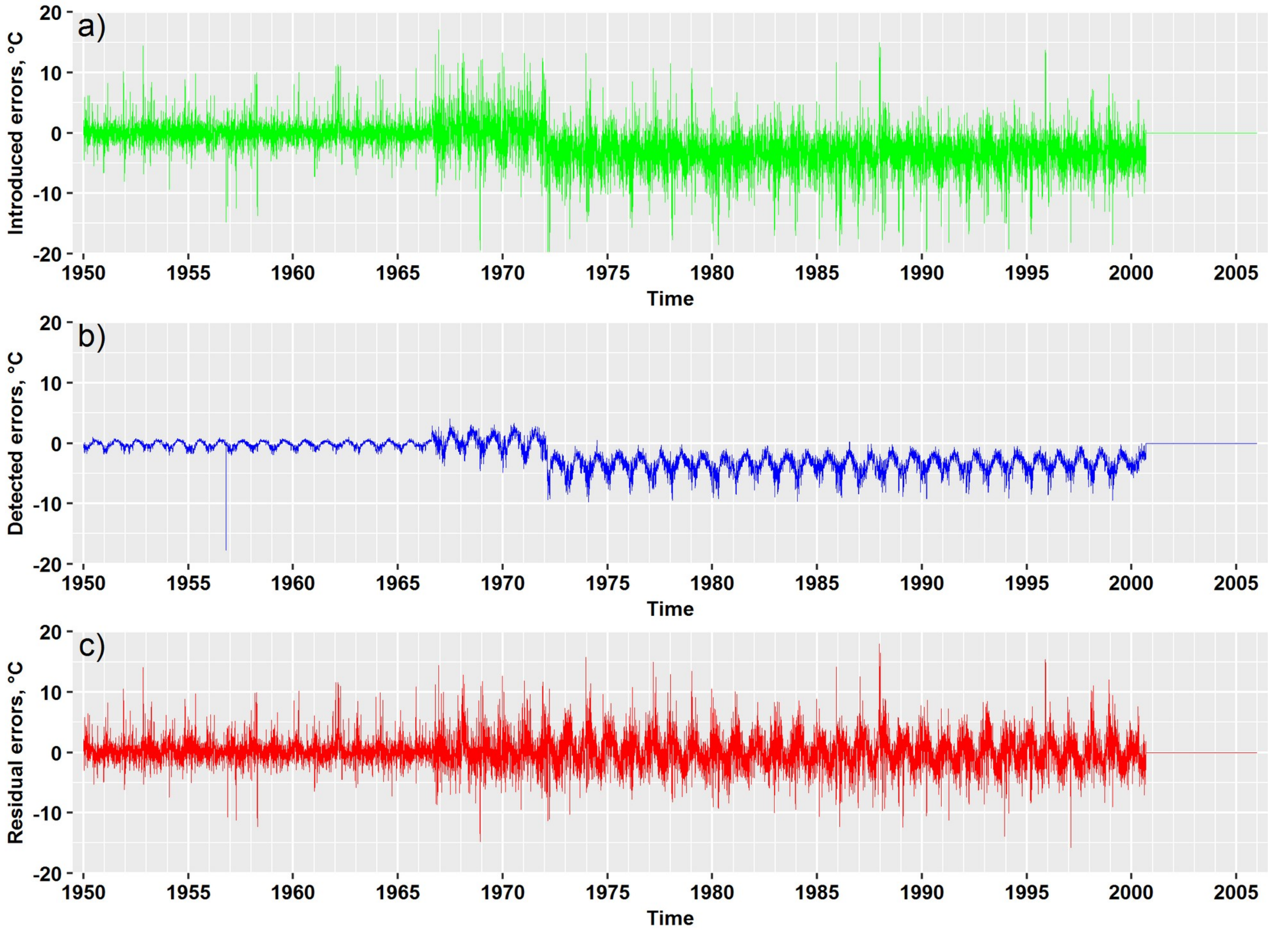

Fig. 7. Examples of time series of errors: real/introduced E_k^R (a), detected E_k^D (b) and residual E_k^H (c) calculated from the data presented in Fig. 6

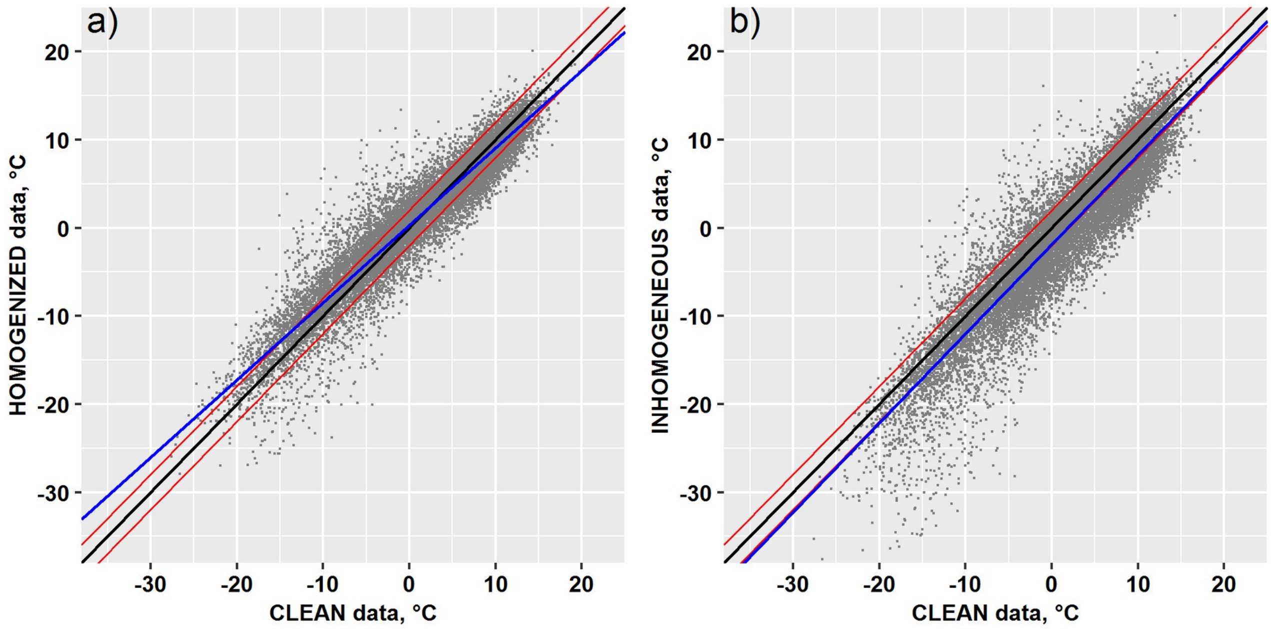

Fig. 8. Example of scatter diagrams. Homogenized X_k^H (a) and raw X_k^I (b) daily data are built against respective clean values X_k^C presented in Fig. 6

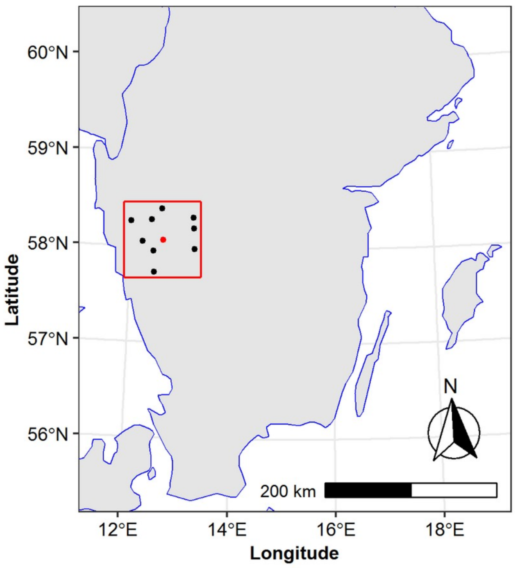

Fig. 9. The chosen set of meteorological stations in case study #1. Black dots show the stations whose time series were always clean, red dot is the station where inhomogeneities were introduced

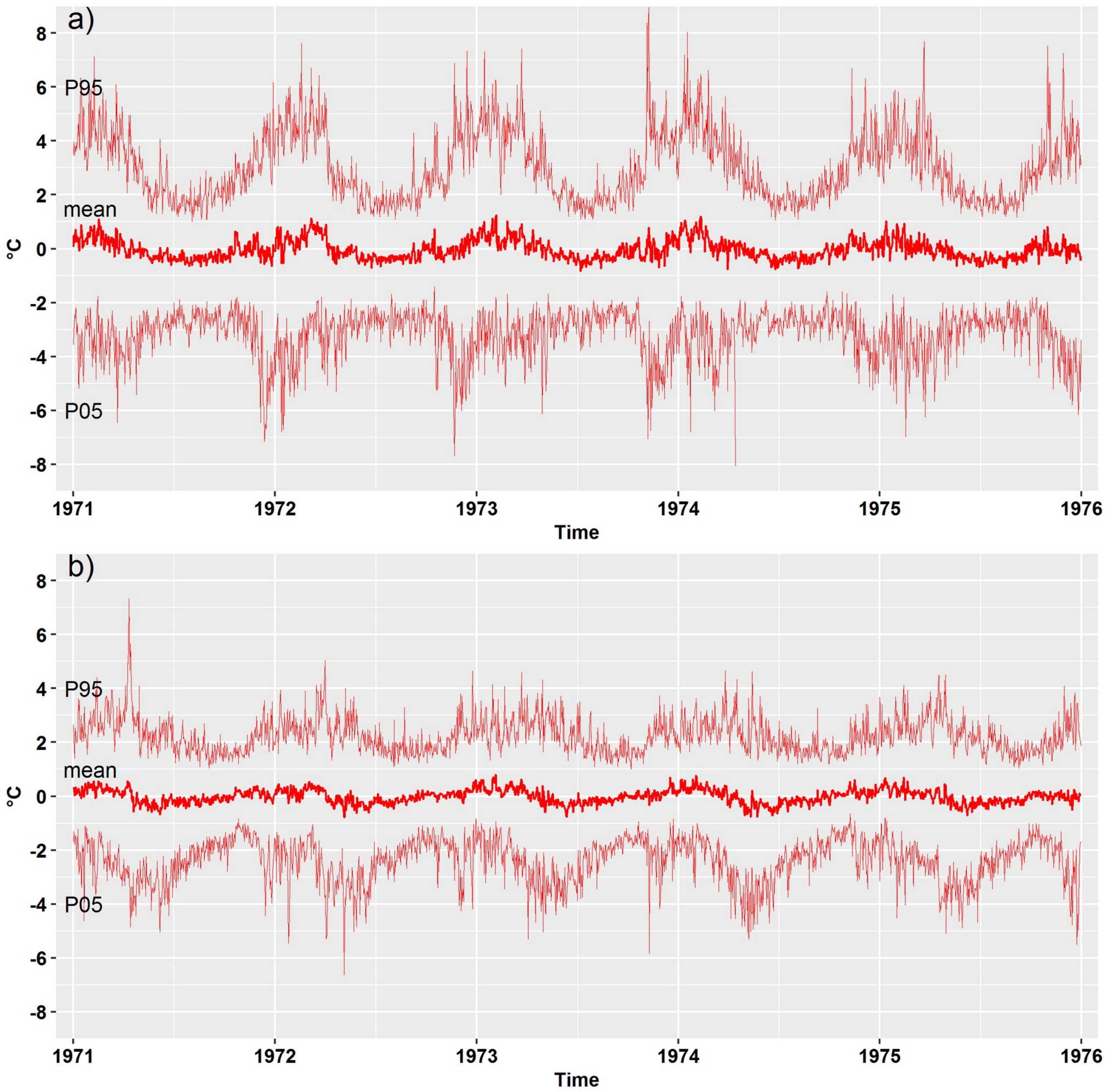

Fig. 10. Mean, 5th and 95th percentiles (P05 and P95) of empirical distributions of the residual errors, evaluated for each day of the corrupted segment: (a) TN, (b) TX.

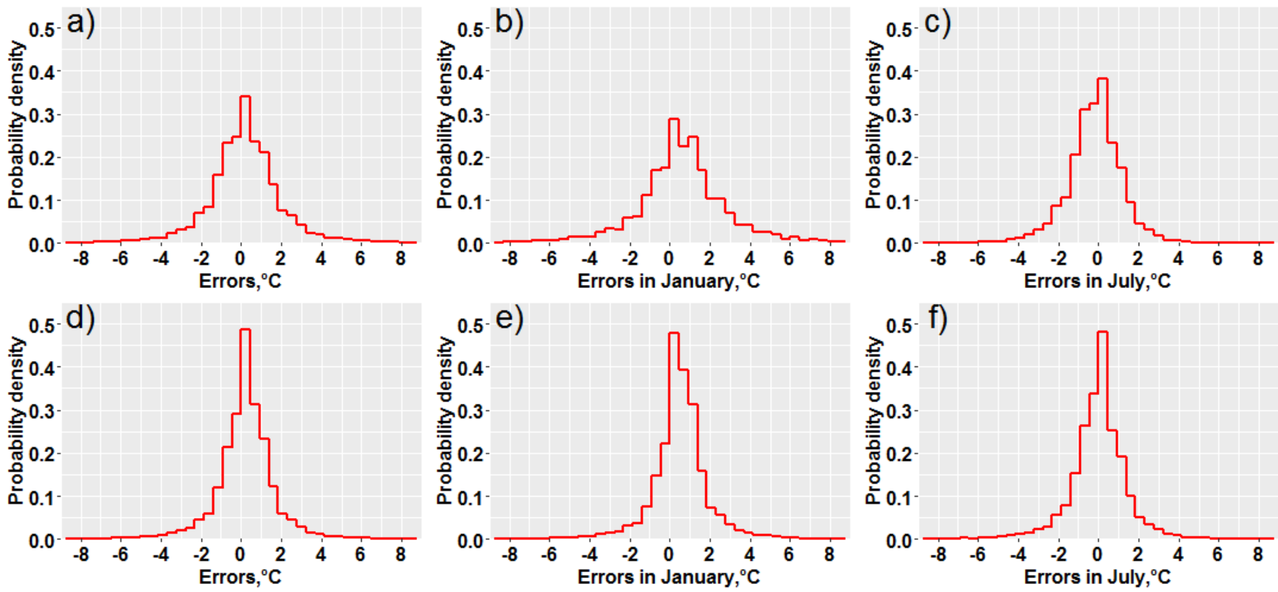

Fig. 11. Empirical distributions of the residual errors, averaged over (a, d) the whole 5-year period, (b, e) January months, (c, f) July months: (top panel) TN, (bottom panel) TX.

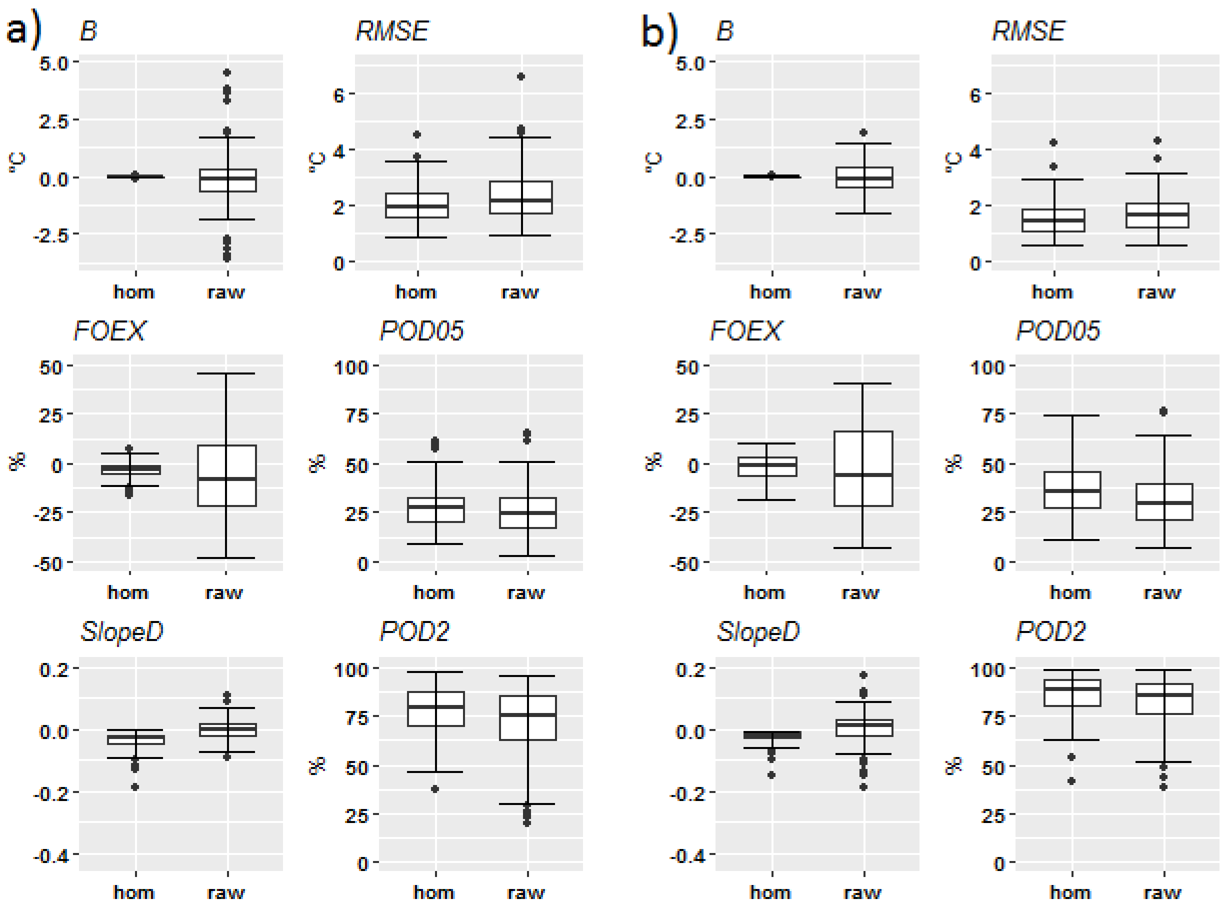

Fig. 12. Boxplots of the metrics, calculated in the set of numerical experiments #1: (a) TN, (b) TX.

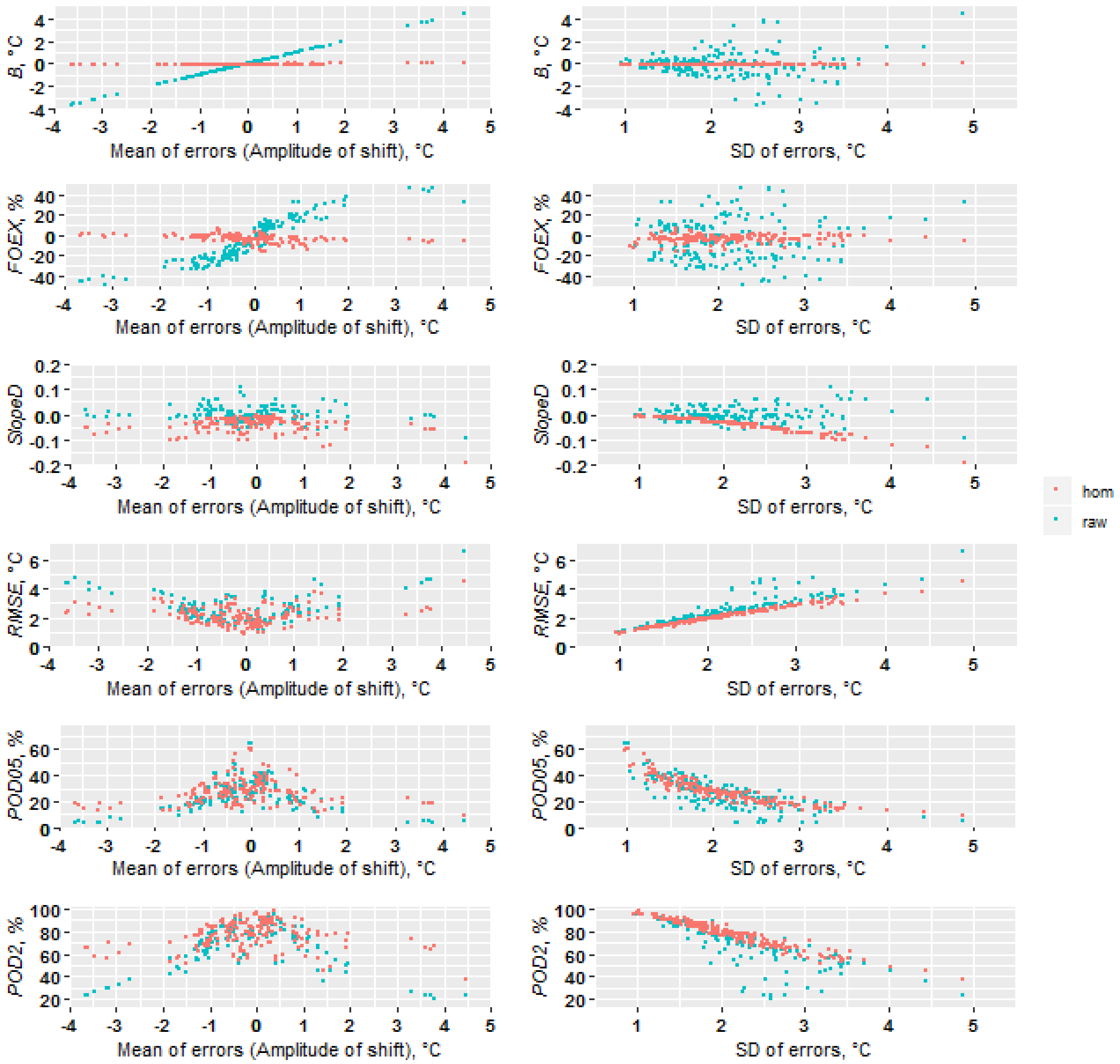

Fig. 13. Relationships between the metric values and the main statistical properties of corrupted segment in the station signals: means (left column) and standard deviations (right column). TN data.

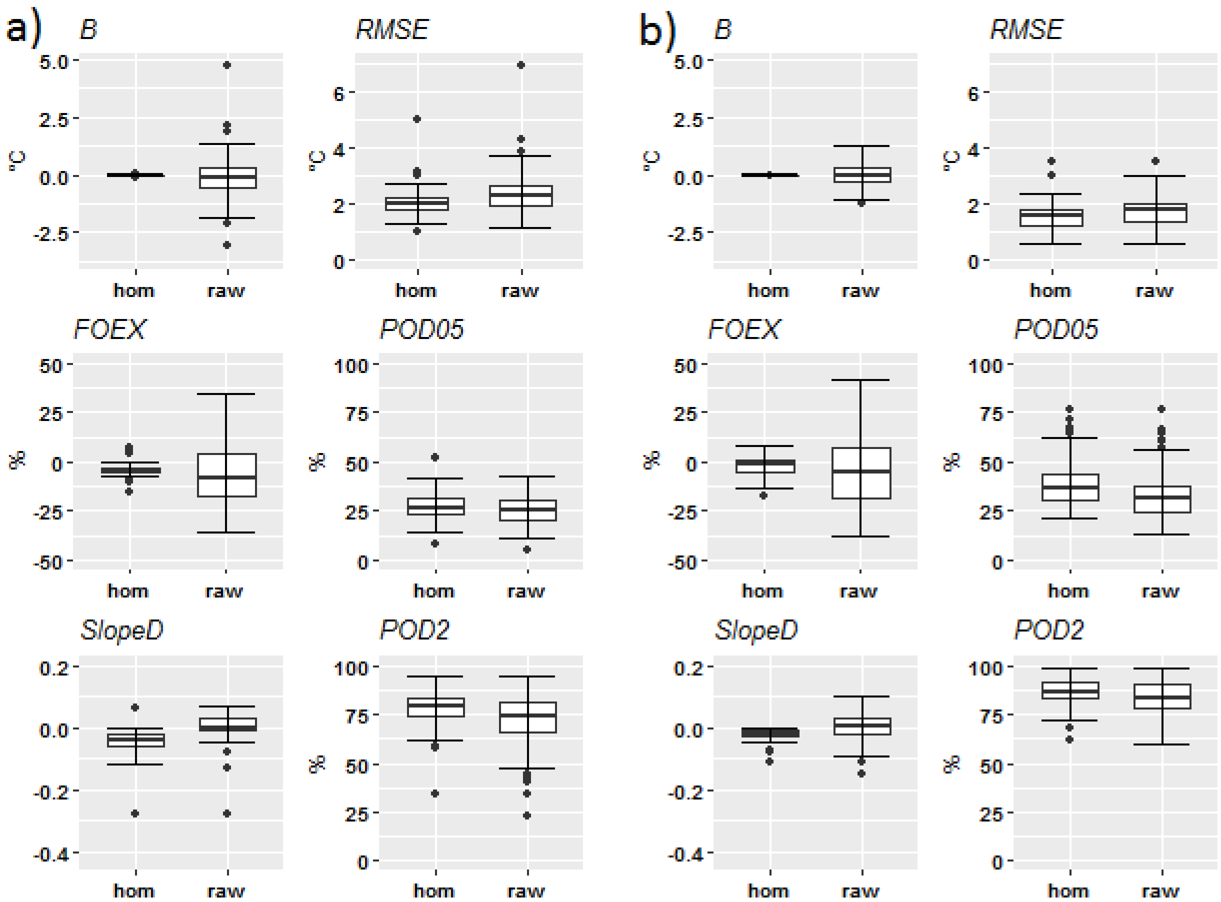

Fig. 14. Boxplots of the metrics calculated in the set of numerical experiments #2: (a) TN, (b) TX.

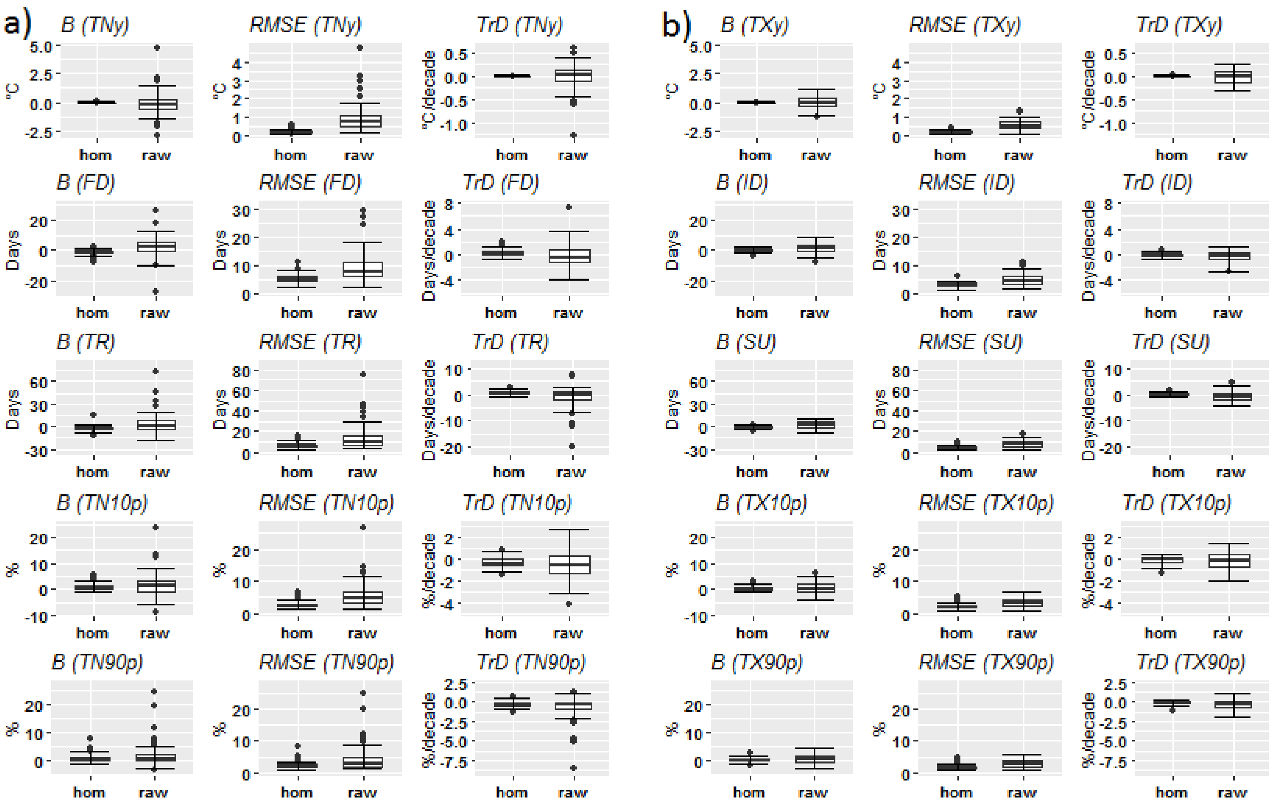

Fig. 15. Box-plots of the metrics calculated based on the yearly series of the climate extremes indices in the set of numerical experiments #2: (a) TN, (b) TX