Indecis

Sectorial Climate Services

Integrated approach for the **d**evelopment across **E**urope of user oriented **c**limate **i**ndicators for GFCS high-priority **s**ectors: Agriculture, disaster risk reduction, energy, health, water and tourism

Work Package 3

Deliverable 3.5

Report on the uncertainty of the homogenization process

E. Aguilar[1], T. Caloiero[2], G.N. Caroletti[3], R. Coscarelli[3], J. Guijarro[4], L.Y.A.Randriamarolaza[1], S.M. Vicente-Serrano[5], O. Skrynyk[1]

[1]Centre for Climate Change (C3), Rovira i Virgili University (URV), Vila-seca, Spain
[2]Institute for Agricultural and Forest Systems in the Mediterranean, National Research Council of Italy, Rende (CS), Italy
[3]Research Institute for Geo-hydrological Protection, National Research Council of Italy, Rende (CS), Italy
[4]State Meteorological Agency (AEMET), Balearic Islands Office, Spain
[5]Instituto Pirenaico de Ecología, Spanish National Research Council (IPE-CSIC), Zaragoza, Spain

This report arises from the Project INDECIS which is 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's Horizon 2020 research and innovation programme

TABLE OF CONTENTS

1. Introduction

Homogenization of climatological data on the daily time scale is attempted through the detection of inhomogeneities at annual, seasonal or monthly means level, and then adjusting the corresponding daily values through various techniques (e.g. Aguilar et al., 2003; Trewin, 2010; Vincent et al., 2017). The detection of break points, where an inhomogeneity takes place (e.g. an instrument is replaced, an extraction method is changed, a ground station is moved), is of major importance in this process (Kuglitsch et al., 2012). However, the adjustment of inhomogeneous data (some segments of raw time series) to homogeneous state is also very important since both parts of the homogenization procedure might produce a certain number of common errors, which deviate the homogenized data from the true climate signal.

By performing a homogenization, one aims to remove the detected inhomogeneities (abrupt shifts/jumps, gradual trends, outliers etc.) and approximate the data to the real climate signal, that took place in some area. Usually the homogenization procedure allows to improve the consistency of the data, which can be seen in the process of a statistical comparison of the raw and homogenized time series. However, the question that may remain unclear is: how far are the homogenized data from the true climate signal? Or, in other words, what potential uncertainties could still be present in the data homogenized by means of some homogenization algorithm or software? It is a very important yet largely overlooked issue, because the climate signal (clean data) is essentially unknown and it is impossible to conduct a direct quantitative comparison and evaluation of the homogenization results. At the same time, understanding the uncertainties and their causes is vital for the correct interpretation of outputs of any predicting model, including homogenization software.

It is important to note that in spite 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 the homogenization (detection and/or adjustment) of climate data (Lindau and Venema, 2016; Trevin, 2018; Vincent et al., 2018).

2. INDECIS benchmark data sets

In the scope of the INDECIS project (www.indecis.eu), two different collections of benchmark time series were created (Aguilar et al., 2018), which cover two regions in Europe with different climate, namely southern Sweden and Slovenia (Figure 1). Each collection contains the 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. The RACMO model was chosen due to its high spatial resolution (0.11°×0.11°) and the daily time step of the output provided: gridded time series of essential climate variables.

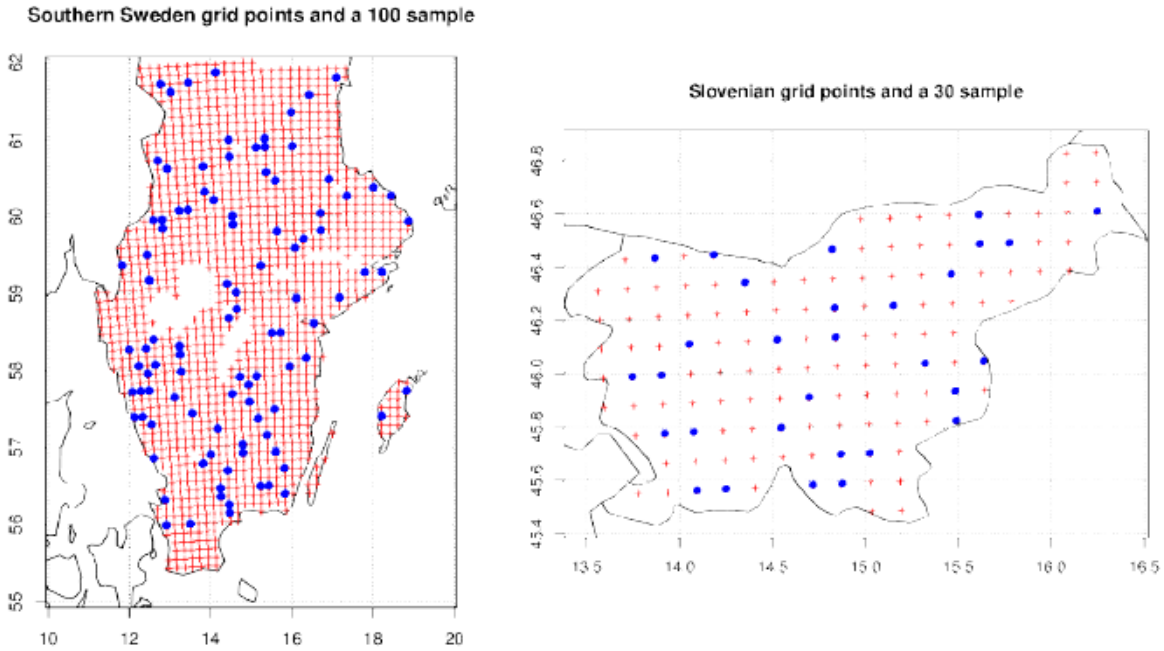

Figure 1. Two European domains, Southern Sweden (left panel) and Slovenia (right panel), with chosen subsets of the Regional climate model grid points (shown as blue dots) to imitate ground station spatial distribution.

The benchmark data set for southern Sweden contain 100 'stations', a subset of the RACMO grid points chosen to imitate stations spatial distribution, while only 30 'stations' were selected in the Slovenia domain. Their geographical locations are shown in Figure 1.

The introduction of biases (break points) in the homogeneous series was done by simulating relocations. First, the 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 to enhance the lower variability of modeled series, and applied to bias the sub-period.

As an example, Figures 2–5 show in the graphical form statistical properties of station signals introduced into minimum (TN) and maximum (TX) air temperature clean time series from the southern Sweden domain. Figure 2 represents the time distribution of the break points. Figure 3 shows the distribution of the number of stations/time series with respect to the number of breaks in one time series. Figure 4 contains the histograms of the factors and amplitudes (defined here according to the HOMER notations) of jumps in the break points. Beside the factors and amplitudes, the homogeneous segments in the introduced errors time series (station signals) can also be characterized by standard deviations (SD) of errors. Figure 5 shows their histograms.

The 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).

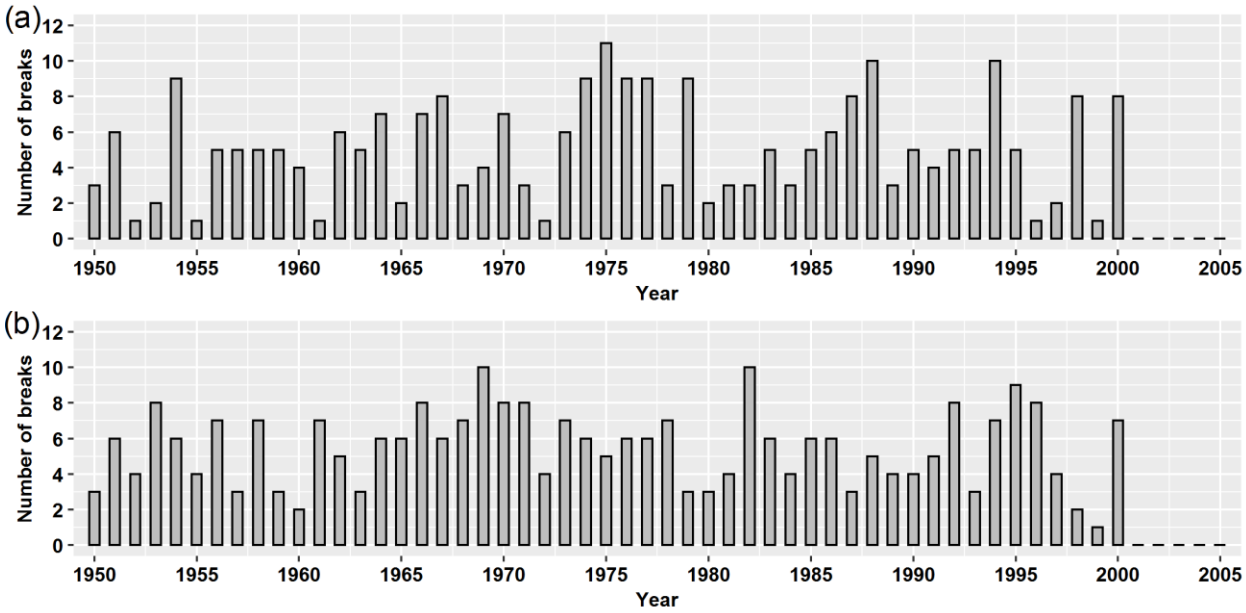

Figure 2. Number of break points per year introduced to clean (a) TN and (b) TX air temperature time series. The southern Sweden domain.

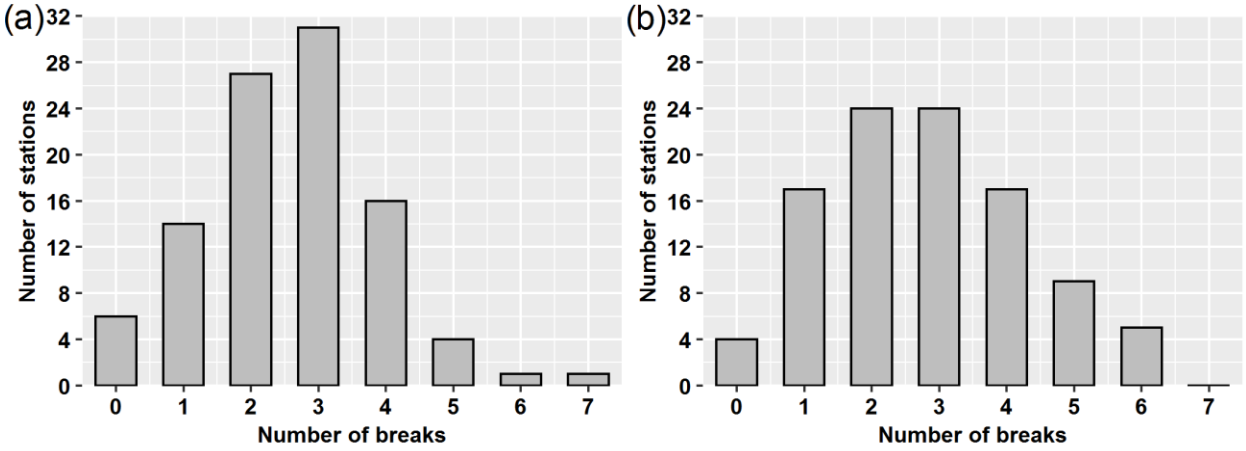

Figure 3. Distribution of the number of stations/time series with respect to the number of break points in one time series: (a) TN, (b) TX. The southern Sweden domain.

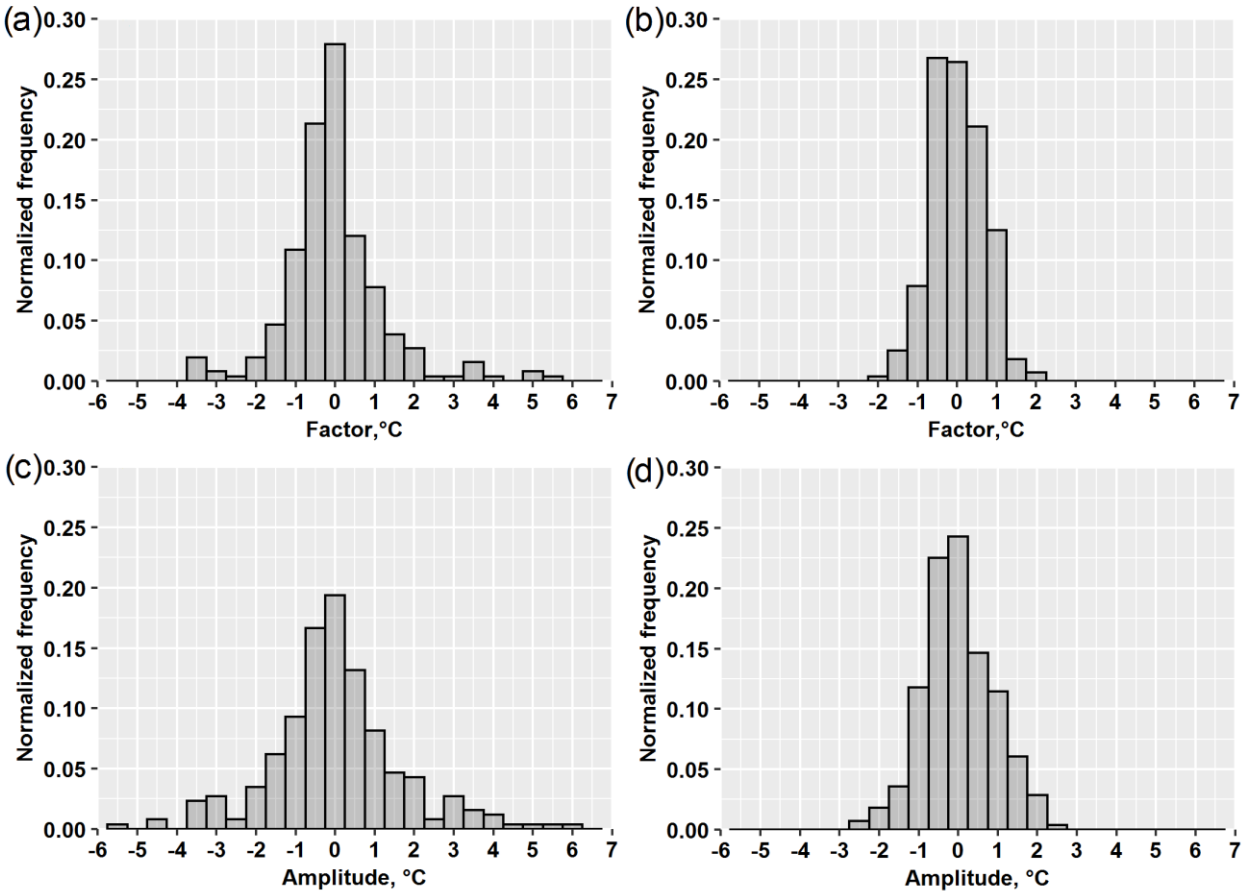

Figure 4. Histograms of the factors (a, b) and amplitudes (c, d) of the shifts at break points, that were introduced to TN (a, c) and TX (b, d) clean data sets. The frequency/count was normalized by the total number of the breaks. The factors/amplitudes were estimated by averaging homogeneous segments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

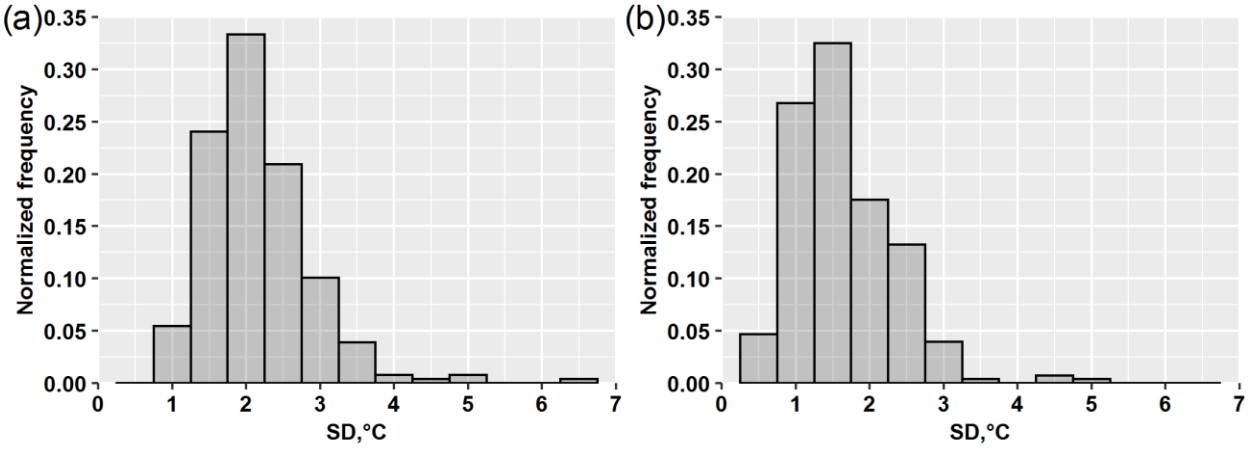

Figure 5. Histograms of SDs of the introduced errors at the homogeneous segments: (a) TN, (b) TX. The frequency/count was normalized by the total number of the breaks. The southern Sweden domain.

3. Homogenization software analyzed

4.1. Climatol

The R package Climatol (Guijarro, 2018; http://climatol.eu/) is a homogenization software that has been selected as the main homogenization tool in the INDECIS project. The effectiveness of the software has been evaluated in several benchmark tests where it demonstrated good results, which are comparable in terms of accuracy to other well established and tested homogenization algorithms. One of Climatol's key feature characteristics is that it can be used automatically, which significantly increases its 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 released since its creation. In the scope of the INDECIS project, Climatol 3.1.1. is being used, available through CRAN (https://cran.r-project.org/package=climatol).

The Climatol detection method (Guijarro, 2018) is based on the standard normal homogeneity test (SNHT) (Alexandersson, 1986). For any candidate time series, Climatol uses data from neighboring stations to create a single composite reference series as their optionally weighted average. This composite series is used further to create time series of anomalies (in order to detect breaks) and to estimate all missing data and all sub-periods/segments after break point detected. From the statistical point of view, the approach employed in the estimation process is equivalent to applying a type II linear regression model.

4.2. HOMER

The HOMER (HOMogenization software in R) software package (Mestre et al., 2013) was developed under the umbrella of the COST Action ES0601 and integrates the parts of different homogenization algorithms such as PRODIGE (Caussinus and Mestre, 2004), ACMANT (Domonkos, 2011) and Climatol (Guijarro, 2018), which were tested and validated via benchmarking (Venema et al., 2012). HOMER is applied to monthly time series, usually through an interactive procedure.

In order to detect potential break points, HOMER uses three different approaches: (**i**) pairwise comparison, (**ii**) joint segmentation, (**iii**) bivariate detection on annual and seasonal changes. In the interactive mode, however, the final decision regarding breaks are made by software users. The correction of detected inhomogeneities are performed by means of ANOVA two factors model.

4.3. HOMER (SMHI version)

Due to the interactive nature of the HOMER software, its application is time consuming and quite limited to relatively small datasets. The homogenization with HOMER of temperature observations at SMHI has previously been performed with a set of criteria for the confirmation of a suggested homogeneity break (Joelsson et al., 2020). These criteria have been implemented in the HOMER (interactive mode) source code by assigning the break signals from the methods different weights and applying a threshold for the sum of the weighted break signals each year for the confirmation of a break year. The user can choose to adjust these threshold and weights to fit their needs. All user interactions

are removed to enable batch processing. This automatic mode of HOMER will be indicated in the Report as SMHI-HOMER from now on.

4.4. ACMANT

The ACMANT (Adapted Caussinus-Mestre algorithm for networks of temperature series) homogenization software (Domonkos, 2011), as it follows from its full name, is a further development of the Caussinus-Mestre method (Caussinus and Mestre, 2004). ACMANT treats in a special way the seasonal changes of inhomogeneity sizes in temperature time series and applies a bivariate test for searching the timings of breaks. The two variables are the annual mean temperature and the amplitude of seasonal temperature-cycle.

4. Methodology of the uncertainty quantification and performance evaluation for homogenization software

As it was mentioned above, unfortunately there is no commonly used methodology for uncertainty quantification of homogenization procedures. Besides, it is worth noting that the performance evaluation of a homogenization algorithm and the quantification of its uncertainty are slightly different tasks in several aspects.

Let

$$\boldsymbol{X}^I, \boldsymbol{X}^H, \text{ and } \boldsymbol{X}^C \tag{1}$$

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

$$\boldsymbol{X} = \{x_{ij}\},\ i = 1, \dots, M,\ j = 1, \dots, N, \tag{2}$$

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

$$\boldsymbol{X}^I \xrightarrow{g} \boldsymbol{X}^H. \tag{3}$$

$\boldsymbol{X}^C$ is the reference, etalon result for the outputs.

Based on the data available in (1), time series of real, $\boldsymbol{E}^R$, detected, $\boldsymbol{E}^D$, and homogenization, $\boldsymbol{E}^H$, errors can be calculated:

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

Specifically in the considered case, $\boldsymbol{E}^R$ is a collection of station signals (or, more precisely, station signals plus noise; but for simplicity they will be referred further as station signals) that were introduced into

the clean data $\boldsymbol{X^C}$. $\boldsymbol{E^H}$ is a dataset of residual errors that might be still present in the homogenized series $\boldsymbol{X^H}$. The error datasets $\boldsymbol{E^R}$, $\boldsymbol{E^D}$ and $\boldsymbol{E^H}$ are also $M \times N$-matrices: $\boldsymbol{E} = \{e_{ij}\}$, $i = 1, \dots, M$, $j = 1, \dots, N$.

Figure 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 (the southern Sweden domain). Figure 7 shows the corresponding error time series (4), calculated from the data given in Figure 6. All figures can be also interpreted as graphical representations of the *k*-th rows in the respective matrices.

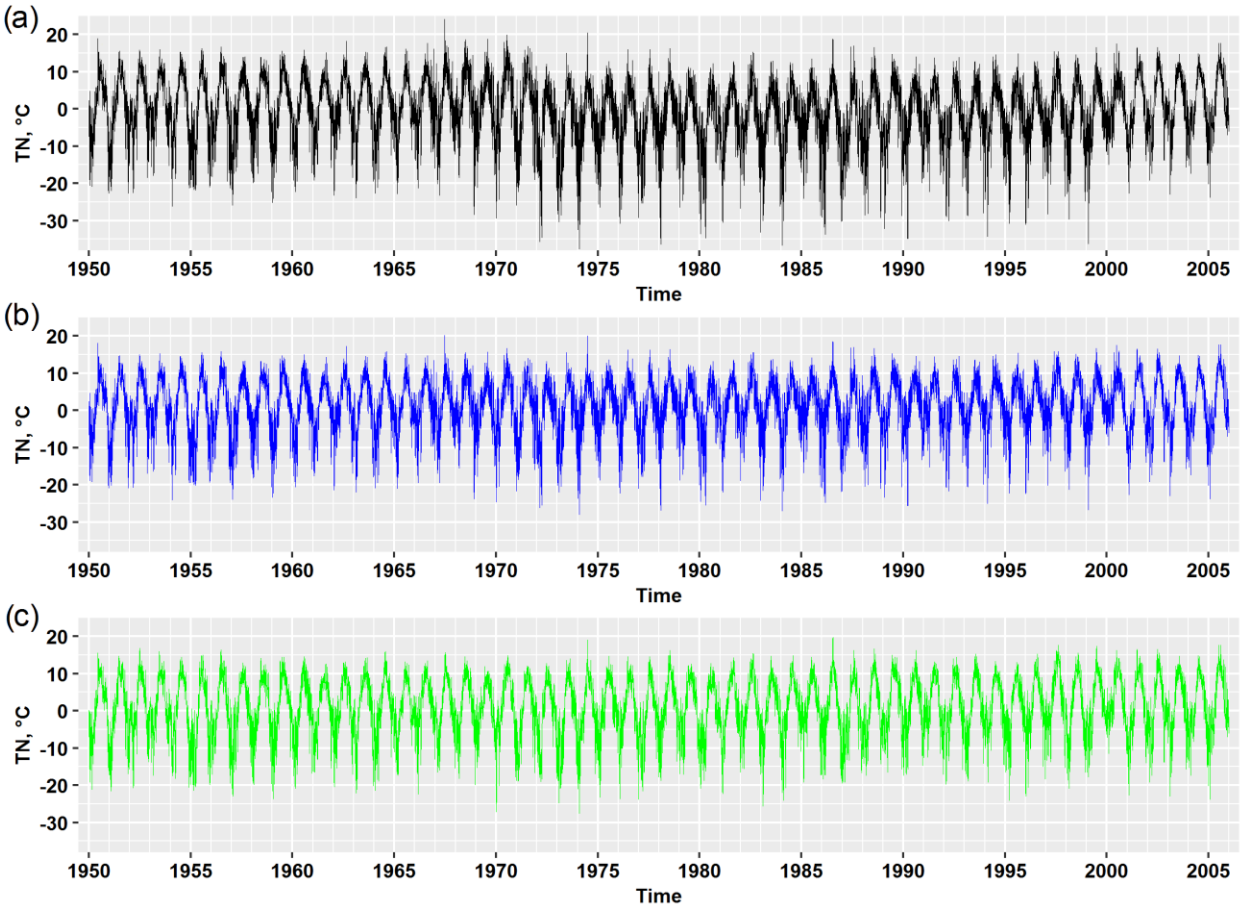

Figure 6. Examples of TN time series (the southern Sweden domain) belonging to the same (*k*-th) station extracted from the inhomogeneious $\boldsymbol{X^I}$ (a), homogenized $\boldsymbol{X^H}$ (b) and clean $\boldsymbol{X^C}$ (c) data sets

The main object of the uncertainty quantification study is the matrix $\boldsymbol{E^H}$: it is necessary to define how large the residual errors in the adjusted data could be or, in other words, how large the departure of the adjustment prediction $\boldsymbol{X^H}$ from the reference result $\boldsymbol{X^C}$ could be. According to e.g. Walker et al. (2003), such departure is usually called 'uncertainty'. Typically, there exist multiple reasons, referred to as sources of the uncertainty, which may affect the homogenization performance and magnitude of the errors in $\boldsymbol{E^H}$. Therefore, in order to evaluate the uncertainty of the homogenization, all these sources -

the whole credible range of every uncertain input and parameter of the homogenization software – must be considered and the effective width of the corresponding probability distribution of the residual errors should be defined (Domonkos and Efthymiadis, 2013). The wider the error distribution, the more uncertain the software prediction $\boldsymbol{X^H}$ is.

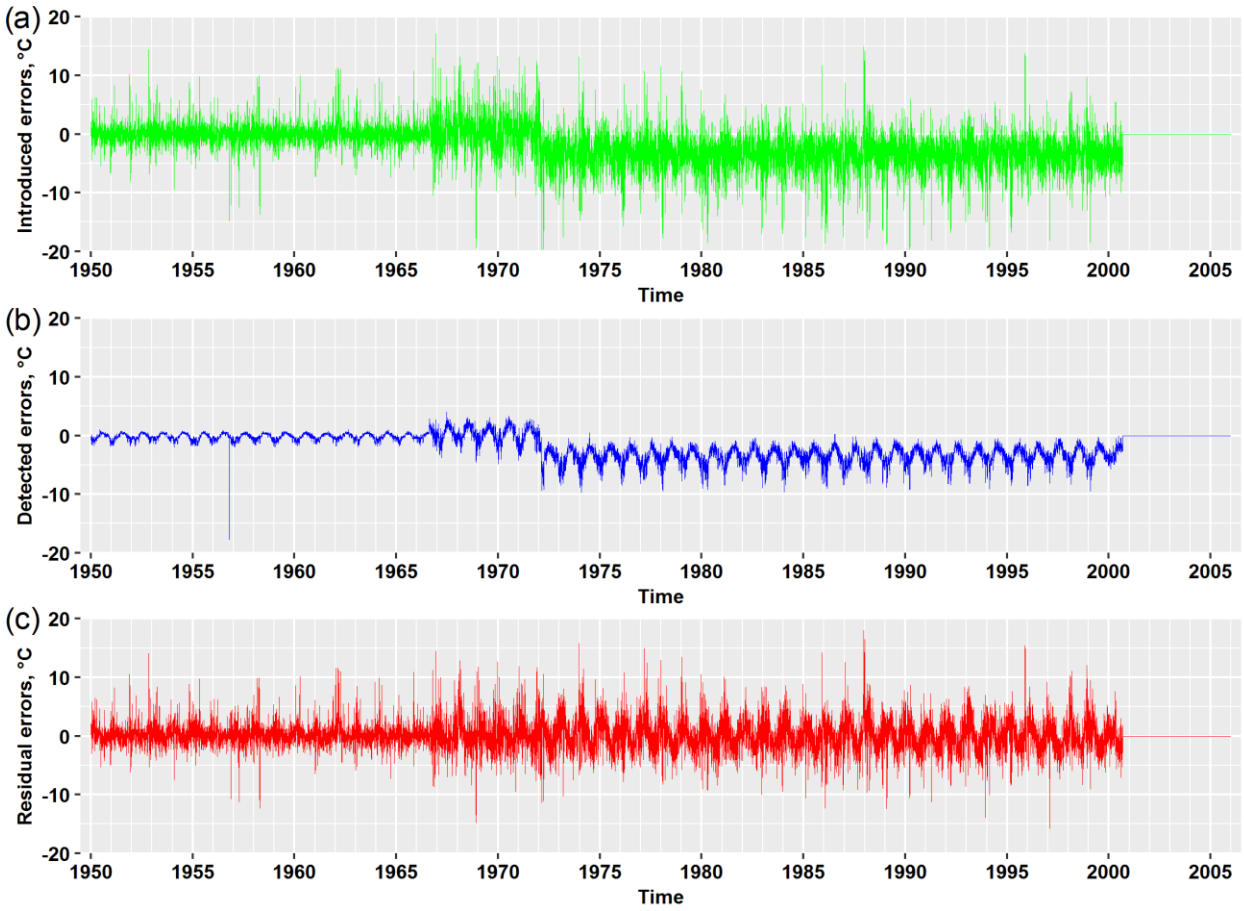

Figure 7. Examples of time series of errors: real/introduced $\boldsymbol{E_k^R}$ (a), detected $\boldsymbol{E_k^D}$ (b) and residual/homogenization $\boldsymbol{E_k^H}$ (c) calculated from the data presented in Figure 6

The residual errors of the homogenization $\boldsymbol{E^H}$ should depend on the introduced errors $\boldsymbol{E^R}$. The more complex station signals in $\boldsymbol{E^R}$ (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, a large number of different yet real variants of $\boldsymbol{E^R}$ has to be considered. Performing the homogenization adjustment for each of them provides a respective ensemble of a homogenization software's outputs, necessary for the uncertainty quantification. The result of the homogenization should also depend on other factors, such as the mean correlation between candidate and reference time series (e.g. Guijarro, 2011), number of reference series (e.g. Trewin, 2018) etc.

4.1. The concept of a random field/function applied to the residual errors

The considerations presented above suggest an appropriate theoretical model for $\boldsymbol{E}^{\boldsymbol{H}}$ that can provide a basis for further calculations and can make calculation results more solid, both statistically and theoretically. Since it is necessary to consider an ensemble of different realizations of $\boldsymbol{E}^{\boldsymbol{H}}$, it is natural to assume that $\boldsymbol{E}^{\boldsymbol{H}}$ 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 and to quantify the homogenization uncertainty it is necessary to define and study statistical properties of the random field $\boldsymbol{E}^{\boldsymbol{H}}$. According to the theory, a multidimensional ($M \times N$ -dimensional) probability distribution function

$$f_{M\times N}(e_{11}^{H}, e_{12}^{H}, \dots, e_{1N}^{H}, e_{21}^{H}, \dots, e_{2N}^{H}, \dots, e_{M\times N}^{H}) \quad (5)$$

provides the most detailed and complete description of $\boldsymbol{E}^{\boldsymbol{H}}$. 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 the function $f_N(e_{k1}^{H}, e_{k2}^{H}, \dots, e_{kN}^{H})$ can be obtained by integrating $f_{M\times N}$ with respect to all its arguments except $e_{k1}^{H}, e_{k2}^{H}, \dots, e_{kN}^{H}$. Function $f_1(e_{kl}^{H})$ 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 procedure, it can be expected that $\boldsymbol{E}^{\boldsymbol{H}}$ is 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 $\boldsymbol{E}^{\boldsymbol{H}}$, namely non-stationarity and spatial heterogeneity, make its analysis more difficult. In particular, that means that the ergodic assumption cannot be used in order to calculate statistical properties of $\boldsymbol{E}^{\boldsymbol{H}}$ based on its only realization.

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

$$\boldsymbol{E}^{\boldsymbol{Rq}} + \boldsymbol{X}^{\boldsymbol{C}} = \boldsymbol{X}^{\boldsymbol{Iq}}, \ \boldsymbol{X}^{\boldsymbol{Iq}} \xrightarrow{g} \boldsymbol{X}^{\boldsymbol{Hq}}, \ \boldsymbol{X}^{\boldsymbol{Hq}} - \boldsymbol{X}^{\boldsymbol{C}} = \boldsymbol{E}^{\boldsymbol{Hq}}, \ q = 1, \dots, Q \quad (6)$$

Based on these realizations, it is theoretically possible to evaluate $f_{M\times N}$. However, such task is hardly feasible in practice due to the extremely large number of dimensions to be considered. On the other hand, based on the statistical ensemble of Q individual realizations of $\boldsymbol{E}^{\boldsymbol{H}}$ some of the moments of the residual error distribution (5) can be evaluated. In the context of the homogenization uncertainty quantification, the most important of them are a mean value (m) and some parameter that can

characterize a width of the distribution such as the standard deviation (σ) or the percentile range. The mean value provides information regarding the systematic bias of the homogenization, while the standard deviation or the percentile range characterize its uncertainty. Both statistics, m and σ, can vary in the space-time domain where $\boldsymbol{E}^{\boldsymbol{H}}$ is defined and they can be evaluated using the following formulas

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

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

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

While the proposed approach to the evaluation of the homogenization 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 constructing a statistical ensemble for $\boldsymbol{E}^{\boldsymbol{R}}$ with a sufficient number of its individual realizations in order to perform the calculations according to (6). Another example of possible limitations can be explained as follow: typically, at the end of the time domain *T*, all station signals in $\boldsymbol{E}^{\boldsymbol{R}}$ contain undisturbed segments (see, for example, Figure 7a). Hence, a lot of zero values in $\boldsymbol{E}^{\boldsymbol{H}}$ are usually obtained there. Such zero values have to be excluded from the analysis when evaluating the homogenization uncertainty since they do not mean the 'perfect' homogenization. However, it is not very easy to do so, because individual station signals usually have undisturbed segments of different length.

4.2. Verification/validation statistical metrics

Estimating the statistical properties of the random field of the residual error $\boldsymbol{E}^{\boldsymbol{H}}$ is not the only way to evaluate the performance of the homogenization and to quantify its uncertainty on the daily or monthly 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.

As it was mentioned above, the performance evaluation of a homogenization 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 a single realization of the adjustment output $\boldsymbol{X}^{\boldsymbol{H}}$. Whereas to define the uncertainty we normally should have a statistical ensemble of $\boldsymbol{X}^{\boldsymbol{H}}$ ($\boldsymbol{X}^{\boldsymbol{Hq}}$, $q = 1, \dots, Q$) and the corresponding ensemble of $\boldsymbol{E}^{\boldsymbol{H}}$ ($\boldsymbol{E}^{\boldsymbol{Hq}}$, $q = 1, \dots, Q$). As it was already mentioned, a single realization of $\boldsymbol{E}^{\boldsymbol{H}}$ can be used for the uncertainty quantification only if $\boldsymbol{E}^{\boldsymbol{H}}$ satisfies 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 $\boldsymbol{E}_{\boldsymbol{i}}^{\boldsymbol{H}}$ ($i = 1, \dots, M$) or on comparison of the corresponding pair of time series $\boldsymbol{X}_{\boldsymbol{i}}^{\boldsymbol{H}}$ and $\boldsymbol{X}_{\boldsymbol{i}}^{\boldsymbol{C}}$. Calculated for a single output of the homogenization adjustment $\boldsymbol{X}^{\boldsymbol{H}}$, 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 these averaged metrics, however, can also be used to quantify the homogenization uncertainty.

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

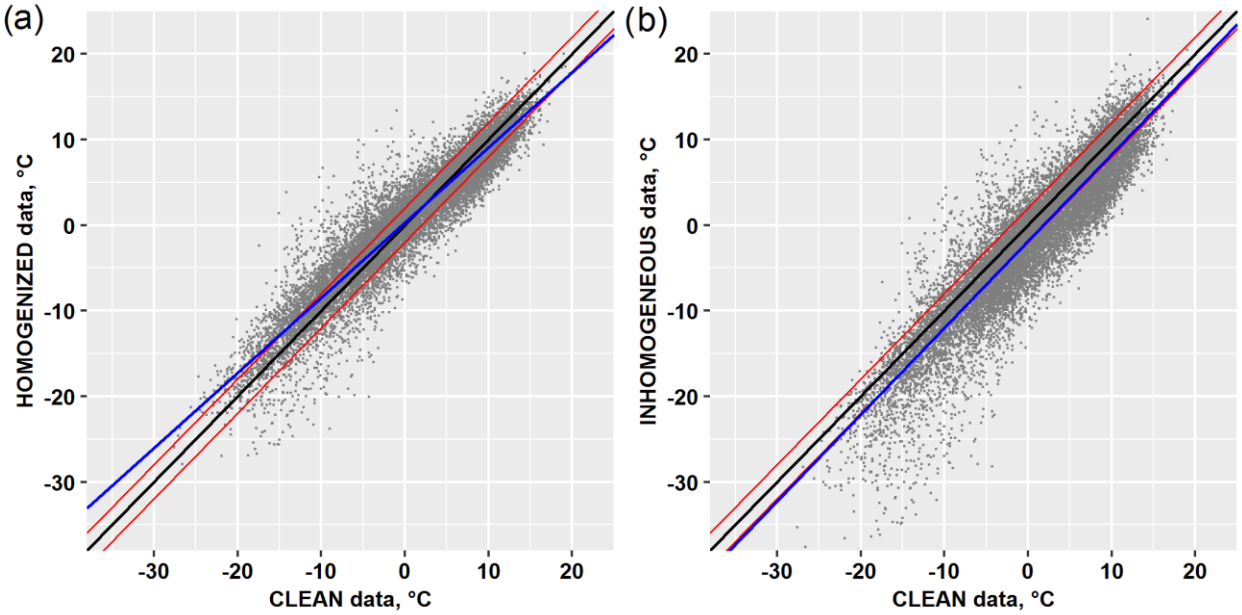

Figure 8. Example of scatter diagrams. Homogenized $\boldsymbol{X}_k^H$ (a) and raw $\boldsymbol{X}_k^I$ (b) daily data are built against respective clean values $\boldsymbol{X}_k^C$ presented in Figure 6

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

Eight different metrics were applied in the report: the bias (B), the absolute bias (B^{abs}), root mean square error ($RMSE$), factor of exceedance ($FOEX$), percentage of days within ±0.5/±2 °C margin

($POD05/POD2$), Pearson's correlation coefficient (CC) and difference in slopes ($SlopeD$). The use of metrics B, $FOEX$ and $SlopeD$ is intended for estimating the systematic errors, while the other four, $RMSE$, B^{abs} 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 homogenization prediction from the true climate signal and the range of the possible residual errors, respectively. Formulas for most of the metrics are standard and well known. However, they were included in the report for completeness. Note that all formulas are presented for individual pairs of time series, $\boldsymbol{X}_i^H$ and $\boldsymbol{X}_i^C$, $i = 1, \dots, M$. Obviously, similar metrics can be calculated for inhomogeneous data by replacing $\boldsymbol{X}_i^H$ with $\boldsymbol{X}_i^I$ (in case of no missing values in the raw data).

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 corresponding time series. Depending on its sign it shows average overestimation (+) or underestimation (-) of the homogenized data. However, the bias does not provide any information 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 $x_{ij}^H - x_{ij}^C = e_{ij}^H$, $j = 1, \dots, N_i$ are zeros. In other words, $B_i = 0$ is a necessary, but not sufficient, condition for having a perfect model or algorithm. In the case when a 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 an estimate of the mean of the random field $\boldsymbol{E}^H$ for i-th station.

2) Absolute bias

$$B_i^{abs} = \frac{1}{N_i}\sum_{j=1}^{N_i}\left|x_{ij}^H - x_{ij}^C\right| = \frac{1}{N_i}\sum_{j=1}^{N_i}\left|e_{ij}^H\right|. \tag{9}$$

Absolute bias is used to provide an effective measure of the difference between the validated series and the validation set. In this case, $B_i^{abs} = 0$ is a necessary and sufficient condition for having a perfect model or algorithm. However, there is no information regarding the average sign of the difference (overestimation/underestimation).

3) Root mean squared error

$$RMSE_i = \left(\frac{1}{N_i}\sum_{j=1}^{N_i}\left(x_{ij}^H - x_{ij}^C\right)^2\right)^{\frac{1}{2}} = \left(\frac{1}{N_i}\sum_{j=1}^{N_i}\left(e_{ij}^H\right)^2\right)^{\frac{1}{2}}. \tag{10}$$

$RMSE_i$ provides information about the average deviation of the homogenized data from the true climate signal. However, this metric can be also interpreted as a value that is proportional to the Euclidian distance between $\boldsymbol{X}_i^H$ and $\boldsymbol{X}_i^C$ in a multidimensional space. Consequently, such an 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 the i-th station and, hence, can be used to characterize the homogenization adjustment uncertainty. Comparing (7.2) and (10), it can be concluded that such averaged value should be close to the standard deviation of the random field $\boldsymbol{E}^H$ for the i-th station.

4) Factor of exceedance

$$FOEX_i = \left(\frac{N_{\left(x_{ij}^H > x_{ij}^C\right)}}{N_i} - 0.5 \right) 100, \tag{11}$$

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 in comparison with the respective value from a clean time series. The factor of exceedance 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 the true climate data. This measure is widely used in climate analysis and applied meteorology.

5-6) Percentage of days within ±0.5/±2 °C margin. In addition to the line of true values in Figure 8, other reference lines might be shown on a scatter diagram in order to facilitate the qualitative evaluation of a homogenization performance. For instance, pairs of parallels can be drawn that are defined as

$$\left| \boldsymbol{X}_i^H - \boldsymbol{X}_i^C \right| = \Delta T, \tag{12}$$

where $|\ldots|$ denotes an absolute value, ΔT is a certain threshold of temperature differences, $\boldsymbol{X}_i^H$ and $\boldsymbol{X}_i^C$ denote here just dependent and independent variables shown in Figure 8a. Following Vincent et al. (2018), the thresholds of 0.5°C can be used or 2°C by analogy with the factor of 2 used in other fields of applied meteorology. A pair of such reference lines when $\Delta T = 2$°C are shown in red color in Figure 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 (12). That is,

$$POD05_i = \frac{N_{\left|x_{ij}^H - x_{ij}^C\right| < 0.5}}{N_i} 100 \text{ and } POD2_i = \frac{N_{\left|x_{ij}^H - x_{ij}^C\right| < 2}}{N_i} 100, \tag{13}$$

where $N_{\left|x_{ij}^H - x_{ij}^C\right| \le 0.5}$ and $N_{\left|x_{ij}^H - x_{ij}^C\right| \le 2}$ stand for the numbers of dots $\left(x_{ij}^C,\ x_{ij}^H\right)$, which lie in the areas inside respective lines (12). Such metrics characterize the magnitude of the scatter of the adjusted values around the clean data.

7) Pearson's correlation coefficient

$$CC_i = \frac{\sum_{j=1}^{N_i} \left(x_{ij}^H - \bar{x}_i^H\right)\left(x_{ij}^C - \bar{x}_i^C\right)}{\sqrt{\left(x_{ij}^H - \bar{x}_i^H\right)^2} \sqrt{\left(x_{ij}^C - \bar{x}_i^C\right)^2}} \tag{14}$$

where $\bar{x}_i^H$ and $\bar{x}_i^C$ are calculated as

$$\bar{x}_i^H = \frac{1}{N_i} \sum_{j=1}^{N_i} x_{ij}^H \text{ and } \bar{x}_i^C = \frac{1}{N_i} \sum_{j=1}^{N_i} x_{ij}^C.$$

Pearson's correlation coefficient is a measure of the linear relationship between two variables or datasets. CC_i can vary from -1 (perfect anticorrelation or negative correlation) and +1 (perfect

correlation or positive correlation), while 0 means there is no correlation at all. In a case of homogenization evaluation, the best result is the perfect positive correlation between $\boldsymbol{X}_i^H$ and $\boldsymbol{X}_i^C$, i.e. $CC_i = 1$. In literature, generally values from 0.4 to 0.6 are said to yield weak correlation, values from 0.6 to 0.8 to yield moderate correlation, and values above 0.8 to yield strong correlation.

8) Difference in slopes

$$SlopeD_i = b_i - 1, \tag{15}$$

where b_i is the slope of a linear regression model $\boldsymbol{X}_i^H = a_i + b_i \boldsymbol{X}_i^C$, which is built using the standard least-squares approach. The need to introduce such metric can be explained based on Figure 8a. 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 on 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 the amplitude of the seasonal cycle showing less variability of the homogenized/adjusted temperature values. Such type of discrepancies (systematic error) between homogenized and clean data can be evaluated 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 Figure 8). The metric is important when evaluating the adjustment of the daily data, since the under/over-estimation of values from tails of the temperature distribution can affect the calculations of some climate extremes indices. The best value for $SlopeD$ is 0. It is worth noting that a similar approach was used in (Della-Marta and Wanner, 2006), where a comparison of the candidate and reference series by means of a scatter diagram was part of the proposed adjustment method. According to that work, deviation of the 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 is capable of providing a fairly detailed description of the homogenization performance on the daily or monthly time resolution.

5. Results

5.1. Verification of the homogenization software on the monthly scale

The main questions considered in this section were the following: (**i**) what metrics can be meaningfully used to validate the best-performing homogenization technique for a temperature record in a region? (**ii**) does temperature homogenization techniques' performance depend on physical features of a station like its geographical position, i.e., latitude, altitude, distance from the sea? (**iii**) does temperature homogenization techniques' performance depend on the nature of the inhomogeneities, i.e., the number of break points and missing data?

In order to clarify the stated questions monthly time series of TN and TX have been calculated from the raw data for both domains (see Figure 1). Then the homogenized monthly data have been obtained through three homogenization techniques: ACMANT, and two versions of HOMER: the standard one and the manual mode setup performed by partners at the Swedish Meteorological and Hydrological Institute, SMHI (SMHI-HOMER). Metrics used for the evaluation of the homogenization techniques are bias, absolute bias, root mean square error, Pearson correlation and factor of exceedance. The metrics have been calculated for each pair of homogenized-clean time series for both domains separately and then averaged over the corresponding number of stations (100 for southern Sweden and 30 for Slovenia). The same metrics were also applied to study the differences and errors between the inhomogeneous (raw) datasets and the clean data. In this way, it was possible to evaluate quantitatively the improvements obtained by the use of homogenization techniques. Tables 1 and 2 show the results of the regional mean metrics calculated for the corrupted dataset and the three homogenized datasets, for southern Sweden and Slovenia respectively.

Table 1: 100-station average of five metrics of corrupted dataset and homogenization techniques for the southern Sweden compared to the clean data.

SOUTHERN SWEDEN (100 stations)					
Metric	**Variable**	**Corrupted**	**Standard HOMER**	**SMHI-HOMER**	**ACMANT**
CC	**TX**	0.9959	0.9979	0.9979	0.9981
	TN	0.9908	0.9959	0.9959	0.9966
***RMSE* (°C)**	**TX**	0.7094	0.4986	0.5082	0.4570
	TN	0.8908	0.6049	0.6059	0.5351
***B* (°C)**	**TX**	-0.0286	-0.0344	-0.0399	-0.0095
	TN	-0.0434	0.0047	-0.0341	-0.0089
***B^{abs}* (°C)**	**TX**	0.3797	0.3053	0.3112	0.2490
	TN	0.4786	0.3581	0.3577	0.2833
***FOEX* (%)**	**TX**	-28.3854	-21.4211	-22.1607	-23.1310
	TN	-28.9896	-19.5327	-22.6176	-23.8854

Table 2: 30-station average of five metrics of corrupted dataset and homogenization techniques for Slovenia compared to the clean data.

SLOVENIA (30 stations)					
Metric	**Variable**	**Corrupted**	**Standard HOMER**	**SMHI-HOMER**	**ACMANT**
CC	**TX**	0.9730	0.9951	0.9955	0.9953
	TN	0.9856	0.9930	0.9938	0.9942
***RMSE* (°C)**	**TX**	2.1440	0.8040	0.7876	0.7673
	TN	1.3062	0.8743	0.8316	0.7584
***B* (°C)**	**TX**	0.1370	0.1462	0.0125	0.0210
	TN	0.0927	0.0526	0.0168	-0.0315
***B^{abs}* (°C)**	**TX**	1.3452	0.4944	0.4832	0.4190
	TN	0.6656	0.4816	0.4818	0.3732
***FOEX* (%)**	**TX**	-21.3442	-7.3512	-13.9831	-20.3175
	TN	-29.3552	-19.6875	-20.0992	-28.3829

It is clear from the metrics results that homogenization improves the correspondence of the dataset to the real data on all accounts except for the bias. *RMSE*, B^{abs} and *FOEX* all allow to evaluate quantitatively meaningful improvements in the homogenized datasets. For instance, for maximum temperature, mean *RMSE* in Sweden is reduced from 0.71°*C* (corrupted dataset) to 0.50°*C* (standard HOMER), 0.51°*C* (SMHI-HOMER) and 0.46°*C* (ACMANT); for minimum temperature, mean *RMSE* was reduced from 0.9 to 0.6 (°*C*), 0.61 and 0.53 (°*C*) respectively. In both cases, there was an improvement of about 30% with respect to the corrupted dataset. In Slovenia, improvements are even bigger in absolute terms, as the Slovenian corrupted dataset has much worse *RMSE* to start with: *RMSE* is 2.1°*C* for maximum temperature and 1.31°*C* for minimum temperature, while the homogenized *RMSE*s are respectively 0.8 and 0.87 (standard HOMER), 0.79 and 0.83 (SMHI-HOMER), 0.77 and 0.76 (ACMANT) (all in °*C*).

However, the Pearson correlation coefficient didn't improve much in either region or for either variable. The reason for this is that, even though artificially manipulated, the corrupted data still show a very high linear correlation with the real one, as is expected in the case of inhomogeneities to the instrumental sensitivity or the re-positioning of an instrument that are simulated by the introduction of artificial break points. Even the weakest correlation between corrupted data and original data, i.e., 0.9730 in Slovenia for maximum temperature, is so high that differences between it and the perfect correlation 1 are not significant. On the other hand, it is important to point out that homogenization always improves even this parameter.

For what concerns to the bias, it is true that homogenization not always improves this metric: for example, the southern Swedish maximum temperature corrupted dataset has a mean bias of -0.0286°*C*, while standard HOMER's bias is -0.0344°*C* and SMHI-HOMER's -0.0399°*C*. However, taking into account both the bias and the absolute bias (which shows reductions from 0.38 to 0.30 and 0.31 (°*C*) respectively), it is clear that the biases in the maximum temperature cancel out and the value goes towards zero, but this masks the true signal of the error. Validating the techniques through the bias, thus, can be used to assess if homogenization changes the sign of the bias, but it is not really suggested as a way to assess quantitatively whether there are improvements in the quality of the data.

According to the results, ACMANT is the best performer with regard to Pearson correlation, *RMSE* and absolute bias for both regions and both variables; in these instances, the two HOMER techniques are almost equivalent, with very small differences for each of these metrics. For what regards the factor of exceedance, the two HOMER techniques perform best, with standard HOMER being slightly the better one (TX Sweden: -21.4% vs -22.2%; TN Sweden: -19.5% vs -22.6%; TX Slovenia: -7.3% vs -14.0%; TN Slovenia -19.7% vs -20.1%).

Comparing results in southern Sweden and Slovenia, it is also clear that the homogenization produces different outcomes depending on the variable and on the region. For example, while in Sweden the factor of exceedance goes from the corrupted dataset *FOEX*=-28.4% for TX and *FOEX*=-29.0% for TN to the homogenized values -21.4% and -19.5% respectively (standard HOMER), in Slovenia the factor of exceedance of the corrupted dataset is -21.3% for TX and *FOEX*=-29.3% for TN and the homogenized

values with standard Homer are -7.3% and -19.7% respectively, so that in Slovenia the homogenization improves much more than in southern Sweden according to this metric.

In order to understand the differences in performance of the homogenization techniques, the relationship between physical features, inhomogeneities and the performance was investigated through the use of Pearson's correlation coefficients, with the goal of highlighting any possible linear correlation between them.

The metrics of the stations of each regional dataset have been compared to two different types of variables: (**i**) physical features of the stations ((**a**) latitude, (**b**) distance from the sea, (**c**) altitude (a.s.l.)); (**ii**) features of the corrupted station data ((**a**) the number of breaks introduced and (**b**) the introduction of missing data). It must be noted, though, that distance from the sea and altitude data were available only for the Swedish stations, so there was no analysis on these points on Slovenian stations.

5.1.1. Latitude

No relevant correlation was found between the latitude of the stations and the five metrics for TX in Sweden. There is a huge difference in the values of the five metrics for each station belonging to the two regional sets.

There is some significant correlation for TN in the Pearson correlation metrics; however, that was established to be the least interesting metric, as the Pearson correlation coefficient didn't improve much in either region or for either variable (see above). The weak correlations of the Bias metric (*B*) with latitude are also not relevant, considering the diminished role of the *B* metric determined in the analysis above.

For what regards Slovenia, the only relevant results seem to be those for *RMSE* in Standard HOMER, where we found *CC*=-0.38 for TX and *CC*=0.37 for TN. As the difference in latitude range is much smaller than in Sweden, it is difficult to establish whether this is truly a standout result.

In general, there seems to be no major discernible pattern in the correlation coefficients depending on variable (TX or TN) or region (southern Sweden or Slovenia).

Table 3. Correlation coefficients between the five metrics of homogenized maximum temperature datasets and the station latitude, for the 100 southern Sweden stations. Boxes highlighted in grey mean the value is at least 95% significant.

Southern SWEDEN	**Standard HOMER**		**SMHI-HOMER**		**ACMANT**	
TX	***CC***	**significance**	***CC***	**significance**	***CC***	**significance**
CC	**0.16**	0.10	**0.17**	0.08	**0.18**	0.08
RMSE	**0.04**	0.72	**0.09**	0.38	**0.01**	0.93
B	**-0.11**	0.25	**-0.16**	0.12	**0.05**	0.61
B^{abs}	**0.06**	0.54	**0.09**	0.36	**0.03**	0.78
FOEX	**0.04**	0.96	**-0.07**	0.48	**-0.05**	0.65

Table 4. Correlation coefficients between the five metrics of homogenized minimum temperature datasets and the station latitude, for the 100 southern Sweden stations. Boxes highlighted in grey mean the value is at least 95% significant.

Southern SWEDEN	Standard HOMER		SMHI-HOMER		ACMANT	
TN	***CC***	**significance**	***CC***	**significance**	***CC***	**significance**
CC	**0.32**	<0.01	**0.27**	<0.01	**0.25**	0.01
RMSE	**-0.12**	0.23	**-0.05**	0.60	**-0.06**	0.56
B	**-0.28**	<0.01	**-0.19**	0.06	**-0.20**	0.04
B^{abs}	**-0.08**	0.42	**<0.01**	0.96	**<0.01**	0.97
FOEX	**-0.12**	0.22	**-0.05**	0.59	**-0.02**	0.84

Table 5. Correlation coefficients between the five metrics of homogenized maximum temperature datasets and the station latitude, for the 30 Slovenia stations. Boxes highlighted in grey mean the value is at least 95% significant.

SLOVENIA	Standard HOMER		SMHI-HOMER		ACMANT	
TX	***CC***	**significance**	***CC***	**significance**	***CC***	**significance**
CC	**0.45**	0.01	**0.24**	0.21	**0.35**	0.06
RMSE	**-0.38**	0.04	**-0.16**	0.40	**-0.28**	0.13
B	**-0.11**	0.57	**0.30**	0.11	**0.51**	<0.01
B^{abs}	**-0.35**	0.06	**-0.10**	0.59	**-0.17**	0.38
FOEX	**-0.07**	0.72	**0.02**	0.90	**0.29**	0.11

Table 6. Correlation coefficients between the five metrics of homogenized maximum temperature datasets and the station latitude, for the 30 Slovenia stations. Boxes highlighted in grey mean the value is at least 95% significant.

SLOVENIA	Standard HOMER		SMHI-HOMER		ACMANT	
TN	***CC***	**significance**	***CC***	**significance**	***CC***	**significance**
CC	**-0.30**	0.11	**-0.22**	0.24	**-0.11**	0.56
RMSE	**0.37**	0.04	**0.27**	0.15	**0.20**	0.29
B	**0.12**	0.52	**-0.21**	0.27	**-0.14**	0.47
B^{abs}	**0.32**	0.08	**0.21**	0.26	**0.12**	0.54
F	**0.15**	0.43	**<0.01**	>0.99	**0.06**	0.76

5.1.2. Distance from the Sea

Very weak correlations were found between the distance from the sea of the Swedish stations and the five metrics. Most are found in the Pearson correlation metrics, that was established to be the least interesting metric, as the Pearson correlation coefficient didn't improve much in either region or for either variable (see above). The weak correlations of the Bias metric (*B*) with distance from the sea are also not relevant, considering the diminished role of the *B* metric determined in the analysis above.

On the other hand, of more interest is the weak but significant negative correlation between Factor of exceedance (*FOEX*) and the minimum temperature in Sweden for the HOMER homogenization techniques. As all three methods were found to underestimate the values of the validated series, from

these results it seems that increasing the station distance from the sea, the number of underestimated data increases slightly as well.

Table 7. Correlation coefficients between the five metrics of homogenized maximum temperature datasets and the station distance from the sea, for the 100 southern Sweden stations. Boxes highlighted in grey mean the value is at least 95% significant.

Southern SWEDEN	Standard HOMER		SMHI-HOMER		ACMANT	
TX	***CC***	**significance**	***CC***	**significance**	***CC***	**significance**
CC	**0.23**	0.02	**0.21**	0.04	**0.18**	0.07
RMSE	**-0.09**	0.35	**-0.05**	0.63	**-0.06**	0.53
B	**0.09**	0.37	**-0.02**	0.87	**0.22**	0.03
B^{abs}	**-0.06**	0.54	**-0.01**	0.92	**-0.02**	0.85
FOEX	**0.07**	0.50	**0.01**	0.93	**0.07**	0.50

Table 8. Correlation coefficients between the five metrics of homogenized minimum temperature datasets and the station distance from the sea, for the 100 southern Sweden stations. Boxes highlighted in grey mean the value is at least 95% significant.

Southern SWEDEN	Standard HOMER		SMHI-HOMER		ACMANT	
TN	***CC***	**significance**	***CC***	**significance**	***CC***	**significance**
CC	**0.25**	0.01	**0.27**	<0.01	**0.18**	0.07
RMSE	**-0.16**	0.11	**-0.15**	0.12	**-0.07**	0.50
B	**-0.30**	<0.01	**-0.30**	<0.01	**-0.23**	0.02
B^{abs}	**-0.14**	0.17	**-0.11**	0.30	**-0.04**	0.70
FOEX	**-0.24**	0.02	**-0.23**	0.02	**-0.10**	0.30

5.1.3. Station altitude (a.s.l.)

Like in the case of latitude, no correlation was found between station altitude and homogenization metrics for maximum temperature. On the other hand, a signal emerged linking altitude and minimum temperature for Pearson correlation, *RMSE* and factor of exceedance.

Table 9. Correlation coefficients between the five metrics of homogenized maximum temperature datasets and the station altitude, for the 100 southern Sweden stations. Boxes highlighted in grey mean the value is at least 95% significant.

Southern SWEDEN	Standard HOMER		SMHI-HOMER		ACMANT	
TX	***CC***	**significance**	***CC***	**significance**	***CC***	**significance**
CC	**0.16**	0.12	**0.16**	0.11	**0.14**	0.16
RMSE	**-0.04**	0.67	**-0.01**	0.92	**-0.02**	0.85
B	**0.05**	0.58	**-0.06**	0.52	**0.15**	0.13
B^{abs}	**-0.02**	0.83	**0.01**	0.91	**0.00**	0.96
FOEX	**0.06**	0.58	**-0.04**	0.70	**0.05**	0.61

Although we are not interested in the Pearson correlation much, for reasons already specified in this report, the negative correlations in *RMSE* and *FOEX* that show up in the HOMER homogenizations of minimum temperature suggest that the temperature at stations with higher altitude might be slightly underestimated than that of stations at lower altitude. Since this effect is weaker both in magnitude and significance in ACMANT homogenization results, this might indicate that the latter technique is more apt to correctly infer minimum temperature data in southern Sweden.

Table 10. Correlation coefficients between the five metrics of homogenized minimum temperature datasets and the station altitude, for the 100 southern Sweden stations. Boxes highlighted in grey mean the value is at least 95% significant.

Southern SWEDEN	**Standard HOMER**		**SMHI-HOMER**		**ACMANT**	
TN	***CC***	**significance**	***CC***	**significance**	***CC***	**significance**
CC	**0.26**	<0.01	**0.28**	<0.01	**0.21**	0.04
RMSE	**-0.21**	0.04	**-0.20**	<0.05	**-0.13**	0.19
B	**-0.30**	<0.01	**-0.31**	<0.01	**-0.24**	0.01
B^{abs}	**-0.18**	0.07	**-0.15**	0.13	**-0.10**	0.31
FOEX	**-0.29**	<0.01	**-0.30**	<0.01	**-0.18**	0.07

5.1.4. Number of breaks

Results show that there is a moderate correlation between the number of breaks and the skill of the homogenization techniques (Tables 11-14). It must be noted that in all these instances, the best performing technique will be the one where the relationship is least relevant. In the case of Pearson correlation, where we have anticorrelation, i.e., the more breaks and the further the homogenized dataset strays from the clean data, the best performing technique will be the one with the lowest negative correlation. In the case of RMSE and Babs, a positive correlation means that the magnitude of errors increases with the number of breaks. In the case of the exceedance factor, a strong correlation, whether negative or positive, will mean that with the increase of break points, underestimation or overestimation increase too, respectively.

Last but not least, the bias does not show any correlation: this is probably related to the intrinsic nature of the bias metric, as it is not adjusted for magnitude like the absolute bias. Absolute bias results prove that there is in fact a correlation between bias and number of breaks, but that correlation does not show when the sign of the bias is not accounted for.

The metric that shows the strongest correlation with the number of breaks is the exceedance factor *FOEX*, ranging from 0.30 (ACMANT maximum temperature in southern Sweden) to 0.57 (standard HOMER minimum temperature in Slovenia).

There are some slight differences between the metrics in southern Sweden and Slovenia in both variables: as the results for maximum temperature in Sweden are greater in magnitude and more robust statistically for *CC*, *RMSE* and B^{abs}, probably the different size of the sampling (100 stations against 30) means that the correlation is highlighted as we increase the number of stations in the regional dataset.

On the other hand, the correlation between maximum temperature exceedance factor and the number of breaks is stronger in the Slovenian case than in the southern Swedish one (in Slovenia 0.49 for standard HOMER, 0.44 for SMHI-HOMER and 0.39 for ACMANT, versus 0.42, 0.33 and 0.30 respectively in southern Sweden), so maybe in the former instance the correlation might be overestimated, again because of the difference of sampling size.

Table 11. Correlation coefficients between the five metrics of homogenized maximum temperature datasets and the number of breaks introduced in the corrupted set, for the 100 southern Sweden stations. Boxes highlighted in grey mean the value is at least 95% significant.

Southern SWEDEN	Standard HOMER		SMHI-HOMER		ACMANT	
TX	***CC***	**significance**	**CC**	**significance**	**CC**	**significance**
CC	**-0.39**	<0.01	**-0.37**	<0.01	**-0.38**	<0.01
RMSE	**0.49**	<0.01	**0.49**	<0.01	**0.49**	<0.01
B	**-0.04**	0.70	**-0.12**	0.21	**-0.07**	0.51
B^{abs}	**0.52**	<0.01	**0.50**	<0.01	**0.46**	<0.01
FOEX	**0.42**	<0.01	**0.33**	<0.01	**0.30**	<0.01

Table 12. Correlation coefficients between the five metrics of homogenized maximum temperature datasets and the number of breaks introduced in the corrupted set, for the 30 Slovenia stations. Boxes highlighted in grey mean the value is at least 95% significant.

SLOVENIA	Standard HOMERr		SMHI-HOMER		ACMANT	
TX	**CC**	**significance**	**CC**	**significance**	**CC**	**significance**
CC	**-0.34**	<0.01	**-0.29**	<0.01	**-0.25**	0.01
RMSE	**0.37**	<0.01	**0.33**	<0.01	**0.33**	<0.01
B	**0.18**	0.08	**-0.00**	0.98	**-0.08**	0.45
B^{abs}	**0.46**	<0.01	**0.39**	<0.01	**0.35**	<0.01
FOEX	**0.49**	<0.01	**0.44**	<0.01	**0.39**	<0.01

Table 13. Correlation coefficients between the five metrics of homogenized minimum temperature datasets and the number of breaks introduced in the corrupted set, for the 100 southern Sweden stations. Boxes highlighted in grey mean the value is at least 95% significant.

Southern SWEDEN	Standard HOMER		SMHI-HOMER		ACMANT	
TN	CC	**significance**	**CC**	**significance**	**CC**	**significance**
CC	**-0.31**	<0.01	**-0.34**	<0.01	**-0.29**	<0.01
RMSE	**0.46**	<0.01	**0.50**	<0.01	**0.47**	<0.01
B	**0.06**	0.55	**-0.01**	0.90	**0.05**	0.60
B^{abs}	**0.53**	<0.01	**0.54**	<0.01	**0.46**	<0.01
FOEX	**0.57**	<0.01	**0.50**	<0.01	**0.52**	<0.01

On the other hand, there are some differences between maximum temperature homogenization (Tables 11 and 12) and minimum temperature homogenization (Tables 13 and 14). While in the case of *RMSE* and B^{abs} the increase in number of stations from Slovenia to Sweden results once again in stronger correlation, i.e., the increase in number of breaks yields worse results, this also happens for *FOEX*,

contrarily to the maximum temperature case. Moreover, the correlation between homogenized minimum temperature and the clean dataset is less influenced by the number of breaks in southern Sweden (-0.31, -0.34 and -0.29) than in Slovenia (-0.39, -0.40 and -0.38), contrarily to results with maximum temperature.

Table 14. Correlation coefficients between the five metrics of homogenized minimum temperature datasets and the number of breaks introduced in the corrupted set, for the 30 Slovenia stations. Boxes highlighted in grey mean the value is at least 95% significant.

SLOVENIA	Standard HOMER		SMHI-HOMER		ACMANT	
TN	**CC**	**significance**	**CC**	**significance**	**CC**	**significance**
CC	**-0.39**	<0.01	**-0.40**	<0.01	**-0.38**	<0.01
RMSE	**0.42**	<0.01	**0.45**	<0.01	**0.46**	<0.01
B	**0.03**	0.80	**-0.04**	0.7253	**-0.14**	0.16
B^{abs}	**0.49**	<0.01	**0.47**	<0.01	**0.48**	<0.01
FOEX	**0.50**	<0.01	**0.44**	<0.01	**0.42**	<0.01

5.1.5. Impact of missing data

With regard to missing data, it is important to note that HOMER and ACMANT have different approaches. HOMER reduces the number of missing data much more drastically than ACMANT (see Table 15): for instance, for maximum temperature, there are on average 111 missing data in southern Sweden stations and 98 missing data in Slovenia stations. These missing data are on average completely replaced in the HOMER homogenization, while with ACMANT 63 and 53 missing data respectively remain on average per station.

Table 15. Missing data in S Sweden and Slovenia corrupted and homogenized datasets: maximum and mean number of missing data in the stations set.

Missing Data	Corrupted Dataset		Standard HOMER		SMHI-HOMER		ACMANT	
TX	**S Sweden**	**Slovenia**	**S Sweden**	**Slovenia**	**S Sweden**	**Slovenia**	**S Sweden**	**Slovenia**
Max	**256**	**229**	**19**	**9**	**19**	**9**	**167**	**156**
Mean	**111**	**98**	**0**	**0**	**0**	**0**	**63**	**53**
TN	**S Sweden**	**Slovenia**	**S Sweden**	**Slovenia**	**S Sweden**	**Slovenia**	**S Sweden**	**Slovenia**
Max	**250**	**214**	**97**	**28**	**97**	**28**	**168**	**168**
Mean	**104**	**117**	**2**	**4**	**2**	**4**	**63**	**71**

Tables 16-19 show the results of the correlation between the missing data for each station and each metric used in this study. Since Homer replaces the missing values almost entirely, it is clear that, as can be expected, the number of missing data is not significant for the metrics. It might be that, with much more missing data, the skill of the homogenization method to repair the dataset could break down, but it might happen at a number of missing data so big to make the actual dataset de facto useless.

On the other hand, for what regards the ACMANT technique, since much less missing data are replaced, the number of missing data bears an impact on the skill of the homogenization. Although there is no significant worsening of Pearson correlation results (given the already high correlation between the corrupted dataset and the clean one), especially B^{abs} and *FOEX* are significantly affected, for both regional datasets in both variables.

Table 16. Correlation coefficients between the five metrics of homogenized maximum temperature datasets and the number of missing data introduced in the corrupted set, for the 100 southern Sweden stations. Boxes highlighted in grey mean the value is at least 95% significant.

Southern SWEDEN	Standard HOMER		SMHI-HOMER		ACMANT	
TX	**CC**	**significance**	**CC**	**significance**	**CC**	**significance**
CC	**0.01**	0.89	**0.06**	0.58	**0.16**	0.10
RMSE	**-0.02**	0.84	**-0.08**	0.45	**-021**	0.04
B	**0.13**	0.21	**0.05**	0.59	**0.20**	0.04
B^{abs}	**-0.06**	0.57	**-0.11**	0.27	**-0.38**	<0.01
FOEX	**0.01**	0.94	**-0.05**	0.59	**-0.25**	0.01

Table 17. Correlation coefficients between the five metrics of homogenized maximum temperature datasets and the number of missing data introduced in the corrupted set, for the 30 Slovenia stations. Boxes highlighted in grey mean the value is at least 95% significant.

SLOVENIA	Standard HOMER		SMHI-HOMER		ACMANT	
TX	**CC**	**significance**	**CC**	**significance**	**CC**	**significance**
CC	**-0.04**	0.71	**-0.04**	0.71	**0.13**	0.20
RMSE	**0.08**	0.41	**0.06**	0.57	**-0.13**	0.21
B	**0.11**	0.28	**0.07**	0.48	**-0.03**	0.78
B^{abs}	**0.06**	0.52	**0.03**	0.77	**-0.28**	<0.01
FOEX	**0.08**	0.41	**0.06**	0.53	**-0.33**	<0.01

Table 18 Correlation coefficients between the five metrics of homogenized minimum temperature datasets and the number of missing data introduced in the corrupted set, for the 100 southern Sweden stations. Boxes highlighted in grey mean the value is at least 95% significant.

Southern SWEDEN	Standard HOMER		SMHI-HOMER		ACMANT	
TN	**CC**	**significance**	**CC**	**significance**	**CC**	**significance**
CC	**0.05**	0.60	**0.05**	0.63	**0.16**	0.12
RMSE	**0.02**	0.85	**0.03**	0.76	**-0.14**	0.18
B	**0.09**	0.38	**0.09**	0.36	**0.06**	0.53
B^{abs}	**0.04**	0.70	**0.04**	0.66	**-0.27**	<0.01
FOEX	**0.12**	0.22	**0.13**	0.20	**-0.20**	<0.05

The one exception to this pattern happens for minimum temperature in Slovenia. Here, there is no significant relationship between missing values and ACMANT performance, while we can see some significant, albeit weak, correlation between Homer and missing data, especially for the Standard

HOMER technique. It is very difficult to pinpoint to an explanation for this difference, as the magnitude in the number of missing data remains the same.

Table 19. Correlation coefficients between the five metrics of homogenized minimum temperature datasets and the number of missing data introduced in the corrupted set, for the 30 Slovenia stations. Boxes highlighted in grey mean the value is at least 95% significant.

SLOVENIA	**Standard HOMER**		**SMHI-HOMER**		**ACMANT**	
TN	**CC**	**significance**	**CC**	**significance**	**CC**	**significance**
CC	**-0.17**	0.08	**-0.14**	0.16	**-0.02**	0.83
RMSE	**0.23**	0.02	**0.19**	0.05	**0.04**	0.72
B	**0.20**	0.04	**0.14**	0.16	**0.06**	0.52
B^{abs}	**0.24**	0.01	**0.17**	0.09	**-0.10**	0.30
FOEX	**0.24**	0.02	**0.19**	0.06	**-0.18**	0.06

As the conclusion for this section, it can be noted that the results showed that *RMSE*, absolute bias and factor of exceedance are the most useful metrics for evaluating homogenization techniques' performance.

Very weak, significant negative correlations are detected between station distance from the sea and factor of exceedance (*FOEX*) and between station altitude and both *RMSE* and *FOEX* for minimum temperature homogenization results obtained with the two HOMER techniques. This suggests that temperature at stations further from the sea and at higher altitude might be very slightly underestimated when homogenized with Homer rather than with ACMANT. Latitude of the stations do not seem to have an impact on how well a technique homogenizes temperature data, although significant results were achieved for *RMSE* in Standard HOMER, where the increase in latitude seems to correlate with an increase in error in Slovenia.

Regardless of the technique used, the quality of homogenization anti-correlates meaningfully to the number of breaks. Missing data do not seem to have any impact on HOMER homogenization in southern Sweden for both variables, and for maximum temperature in Slovenia, while a very weak, albeit significant, negative impact emerges between standard setup HOMER performance and number of missing data for minimum temperature in Slovenia. The reverse is true about ACMANT: the number of missing data significantly affects homogenization performance in a negative way, with the exception of minimum temperature homogenization for the Slovenia dataset.

In general, the nature of the datasets (i.e., number of breaks and missing data) seems to have a more important role in yielding good homogenization results than physical parameters associated to the stations (i.e., latitude, elevation and distance from the sea). Even though from this point of view, the skill of HOMER to replace most missing data give it the upper hand over ACMANT, the actual metrics show that ACMANT still performs better for these variables in these regions for what concerns *RMSE* and absolute error B^{abs}, while HOMER performs better with regard to the factor of exceedance *FOEX*.

5.2. Uncertainty quantification of the Climatol adjustment on the daily scale

In this section, the uncertainty associated with Climatol's adjustment algorithm applied to daily minimum and maximum air temperature is investigated (i.e., perfect detection was assumed). The uncertainty quantification was performed based on several numerical experiments and the INDECIS benchmark data from the southern Sweden domain (Skrynyk et al., 2020). Using a complex approach, the adjustment uncertainty was evaluated at different levels of detail (day-to-day evaluation through formalism of random functions and 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, such as the averaged correlation between a candidate and references, number of reference stations etc., were removed from the analysis or kept almost unchanged.

5.2.1. Case study #1

This first case study considers ten stations (Figure 9) and limits the length of the corresponding time series to the period of 1971-1980 (10 years, similar to Vincent et al. (2018)). Nine time series (the references), belonging to the stations marked in black color in Figure 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 the same matrix notations as in (2), these initial conditions can be written as follows

$$\{x_{ij}^I\} = \{x_{ij}^C\}, \text{ when } i = 1, \dots, 9, j = 1, \dots, 3653, \text{ or } i = 10, j = 1827, \dots, 3653; \quad (16.1)$$

$$\{x_{ij}^I\} \neq \{x_{ij}^C\}, \text{ when } i = 10,\ j = 1, \dots, 1826, \quad (16.2)$$

where 3653 is the total number of days in the time interval 1971-1980, while 1826 is the number of days in the interval 1971-1975.

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

In order to construct the raw data with the corrupted 5-year sub-period ($\{x_{ij}^I\}$, $i = 10$, $j = 1, \dots, 1826$), we analyzed all station signals in $\boldsymbol{E}^R$, that were initially introduced in the INDECIS benchmark, and defined homogeneous error segments that have the length of more than 5 complete consecutive years (since January 1 until December 31). For instance, in the error time series shown in Figure 7a, all three homogeneous non-zero segments, i.e. 01.01.1950-13.08.1966, 14.08.1966-19.02.1972, 20.02.1972-08.09.2000, satisfy this 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 (translating along the time axis) a 5-year period from each of the defined segments to 1971-1975 and adding them to the respective clean data $\{x_{ij}^C\}$, $i = 10$,

$j = 1, \dots, 1826$. This way (by performing such replacements), we obtained a statistical ensemble of individual realizations of the raw data set $\boldsymbol{X}^{Iq}$, $q = 1, \dots, Q$, 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 (16.1) and (16.2)), which are well known (Figure 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 $\boldsymbol{X}_{10}^{Iq}$, $q = 1, \dots, Q$ and the system of the reference series $\boldsymbol{X}_i^C$, $i = 1, \dots, 9$ 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.

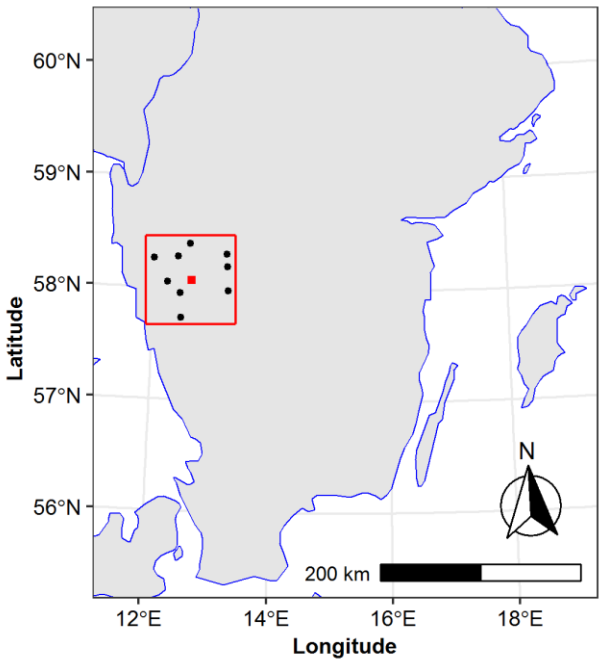

Figure 9. The chosen set of meteorological stations in case study #1. Black dots show the stations whose time series were always clean, red square is the station where inhomogeneities were introduced

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 a nearly pure effect of the introduced station signals on the adjustment uncertainty. This is because any other factors, which might have some effect on the homogenization adjustment, were kept approximately constant or removed.

Figure 10 shows the results of the adjustment uncertainty quantification on the daily scale by applying the concept of a random field to the residual errors $\boldsymbol{E}^H$. Since only a single time series of the raw data

set was corrupted on 1971-1975, $\boldsymbol{E}^{H}$ has non-zero values only for one point in the space domain (i.e. for tenth station) and just for the first half of the period under study. Therefore, the statistical properties of $\boldsymbol{E}^{H}$ were defined only for this station and period. In Figure 10, the mean values, 5th ($P05$) and 95th ($P95$) percentiles of empirical distributions of $\boldsymbol{E}^{H}$, 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 using the samples of Q (185 for TN and 193 for TX) values $e_{10\,j}^{Hq}$, $q = 1, \ldots, Q$ for each day ($j = 1, \ldots, 1826$).

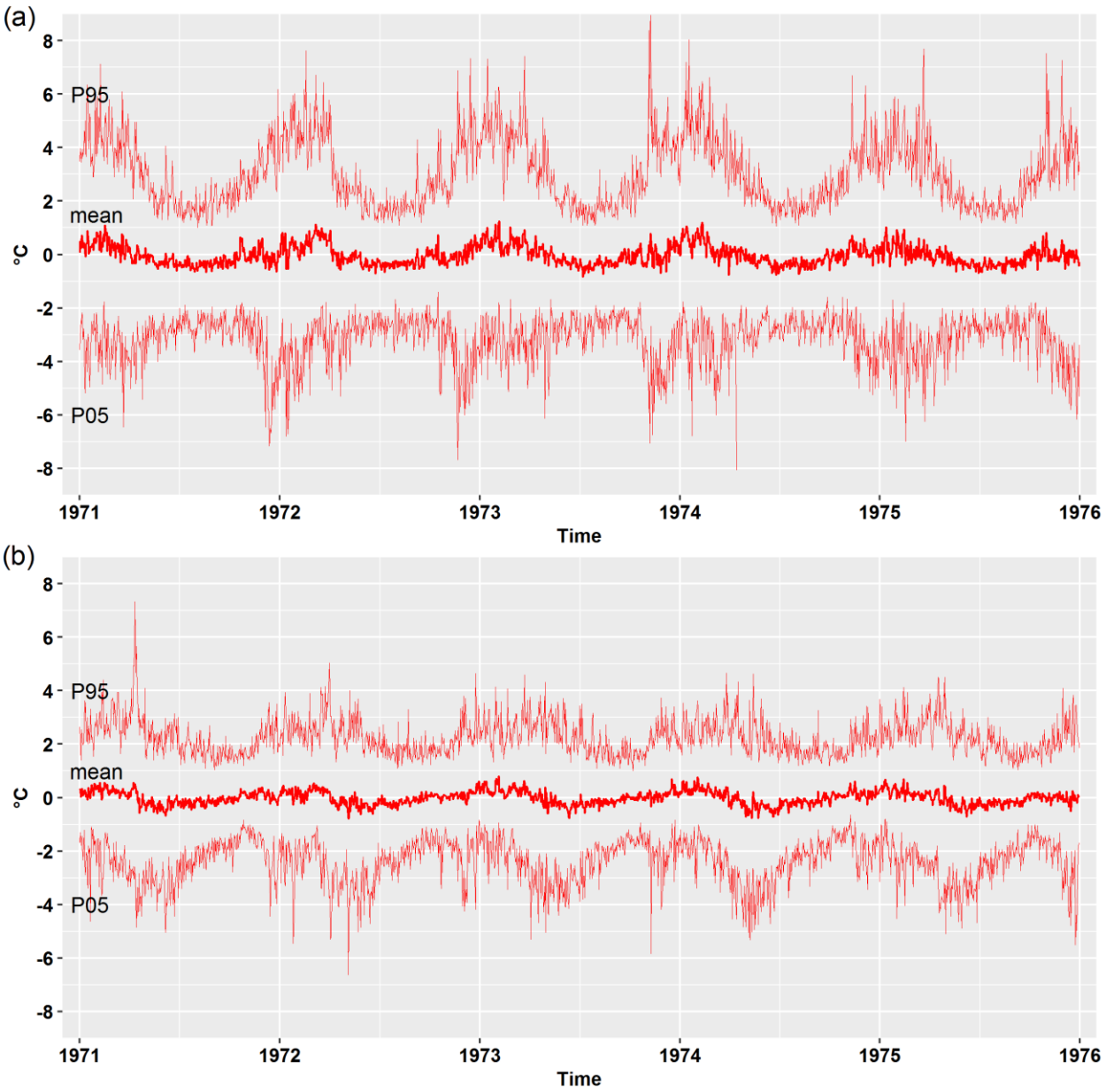

Figure 10. Mean, 5th and 95th percentiles ($P05$ and $P95$) of empirical distributions of the residual errors, evaluated for each day of the corrupted segment: (a) TN, (b) TX

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 attributed to the similar periodicity of the introduced errors $\boldsymbol{E}^R$. The reason for the seasonality in $\boldsymbol{E}^R$ 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 Figure 10, we averaged the empirical error distributions over the whole period, and over January and July months separately (Figure 11). Table 20 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.

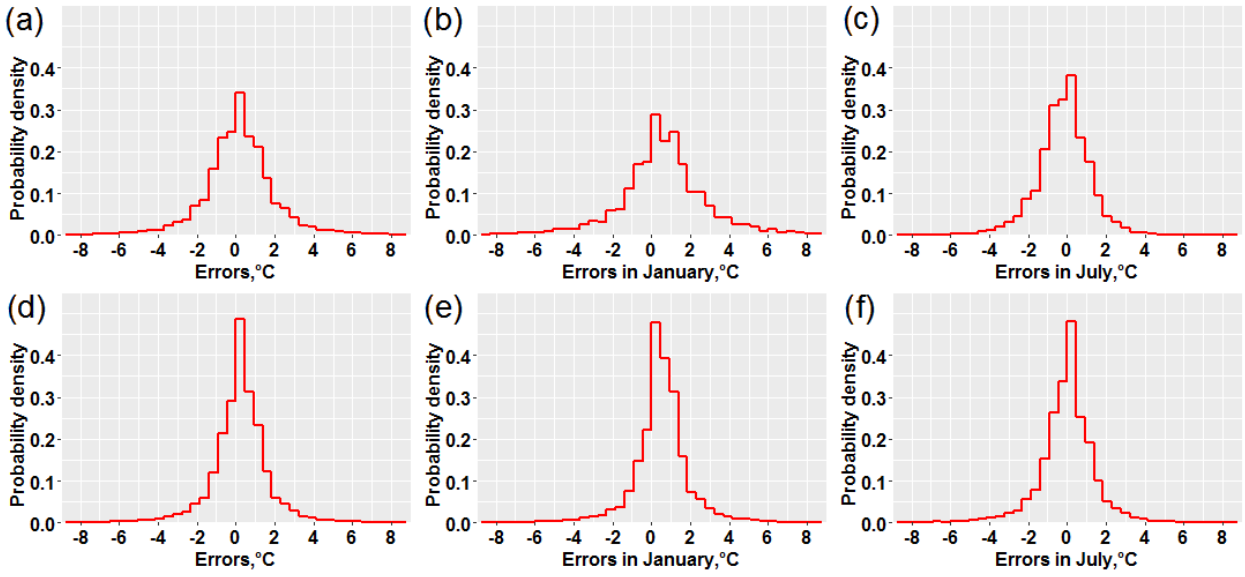

Figure 11. Empirical distributions of the residual errors, averaged over (a, d) the whole 5-year period, (b, e) January months, (c, f) July months: (top panel) TN, (bottom panel) TX

In summer months, the percentile intervals of the residual errors, $(P05, P95)$, for the adjusted daily TN and TX air temperatures are $(-2.80, 1.70)$ (°*C*) and $(-2.60, 1.90)$ (°*C*), 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 $-0.41°C$ and $-0.22°C$. 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 $0.41°C$ for TN and 0.22 $°C$ for TX. Also, we could expect with 90% probability that for minimum air temperature the adjusted value x_{ij}^{H} lays in the interval $(x_{ij}^{C} - 2.80, x_{ij}^{C} + 1.70)$ ($°C$), while for maximum air temperature the interval is $(x_{ij}^{C} - 2.60, x_{ij}^{C} + 1.90)$ ($°C$). 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.

Table 20. Parameters of averaged empirical distributions of errors: homogenization/residual $\boldsymbol{E^H}$ and real/introduced $\boldsymbol{E^R}$ (all in $°C$)

		Year		January		July	
		$\boldsymbol{E^H}$	$\boldsymbol{E^R}$	$\boldsymbol{E^H}$	$\boldsymbol{E^R}$	$\boldsymbol{E^H}$	$\boldsymbol{E^R}$
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

In winter months, the ranges $(P05, P95)$, evaluated for the homogenization adjustment errors in TN and TX data are $(-3.60, 4.50)$ ($°C$) and $(-2.00, 2.60)$ ($°C$), respectively. The corresponding mean values of the error distributions are $0.40°C$ for TN and $0.28°C$ for TX. Thus, in winter we could expect any adjusted temperature value x_{ij}^{H} to be slightly overestimated (on average) by 0.40 $°C$ for TN and $0.28°C$ for TX relatively to the respective clean value x_{ij}^{C} and with 90% probability it lays in the interval $(x_{ij}^{C} - 3.60, x_{ij}^{C} + 4.50)$ ($°C$) in case of TN air temperature and $(x_{ij}^{C} - 2.00, x_{ij}^{C} + 2.60)$ ($°C$) 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 by 15/13% (TN/TX).

The parameters of the empirical distribution of the residual errors, averaged over the whole 5-year period (see Table 20), 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 5th and 95th 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.

Figure 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 of removing systematic and random errors in each individual realization of the raw time series of a statistical ensemble. However, only averaged value of $RMSE$ (averaged over a statistical ensemble) can be considered as a 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 in 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 the 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 the 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 of coping with the scatter of errors. The mean value of $RMSE$, which yields the overall, time-averaged assessment of the adjustment uncertainty, is $2.06°C$ for TN and $1.53°C$ 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 20). 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).

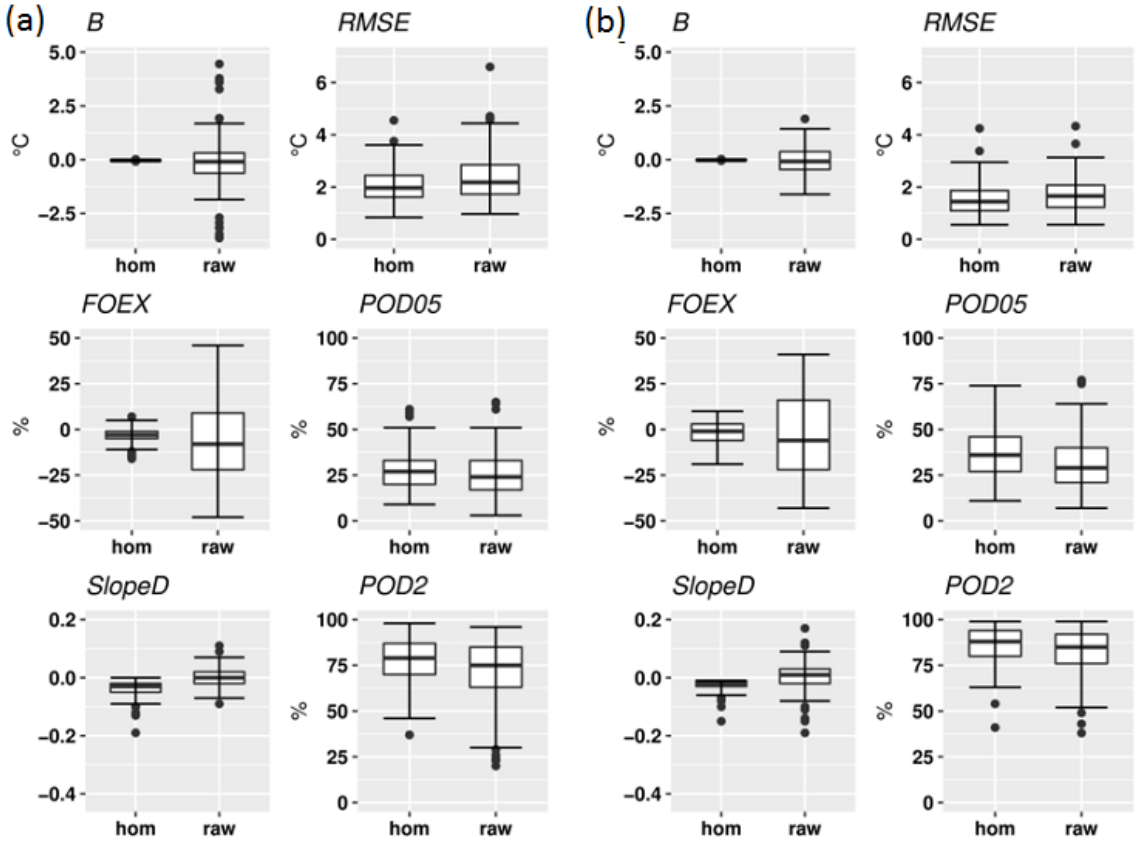

Figure 12. Boxplots of the metrics, calculated in the set of numerical experiments #1: (a) TN, (b) TX

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 it was 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.

5.2.2. 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 Figure 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

$$\{x_{ij}^I\} = \{x_{ij}^C\}, \text{ when } i = 1, \dots, 9, j = 1, \dots, 20454, \text{ or } i = 10, j = N_{10} + 1, \dots, 20454; \quad (17.1)$$

$$\{x_{ij}^I\} \neq \{x_{ij}^C\}, \text{ when } i = 10,\ j = 1, \dots, N_{10}, \quad (17.2)$$

where 20454 is the total number of days in the time interval 1950-2005, while N_{10} is the 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 for the southern Sweden domain, 94 and 96 different non-zero station signals were created for TN and TX data, respectively (Figure 3). By adding these error series to the clean data of the tenth station alternately, we created corresponding numbers of different realizations of the 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 $\boldsymbol{X}_{10}^C$ and $\boldsymbol{X}_i^C$, $i = 1, \dots, 9$ and between $\boldsymbol{X}_{10}^{Iq}$ $(q = 1, \dots, Q)$ and $\boldsymbol{X}_i^C$, $i = 1, \dots, 9$ 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 (which positions are predefined during Climatol calculations), make it more difficult for the software 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 the 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 calculations would operate with considerably less quantity of non-zero error values compared to days in the beginning of 1950-2005.

Figure 13 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 corresponding 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 residual errors and the related issue of the underestimation of the seasonal cycle amplitude is also preserved in this case study. Therefore, a number of break points in the 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.

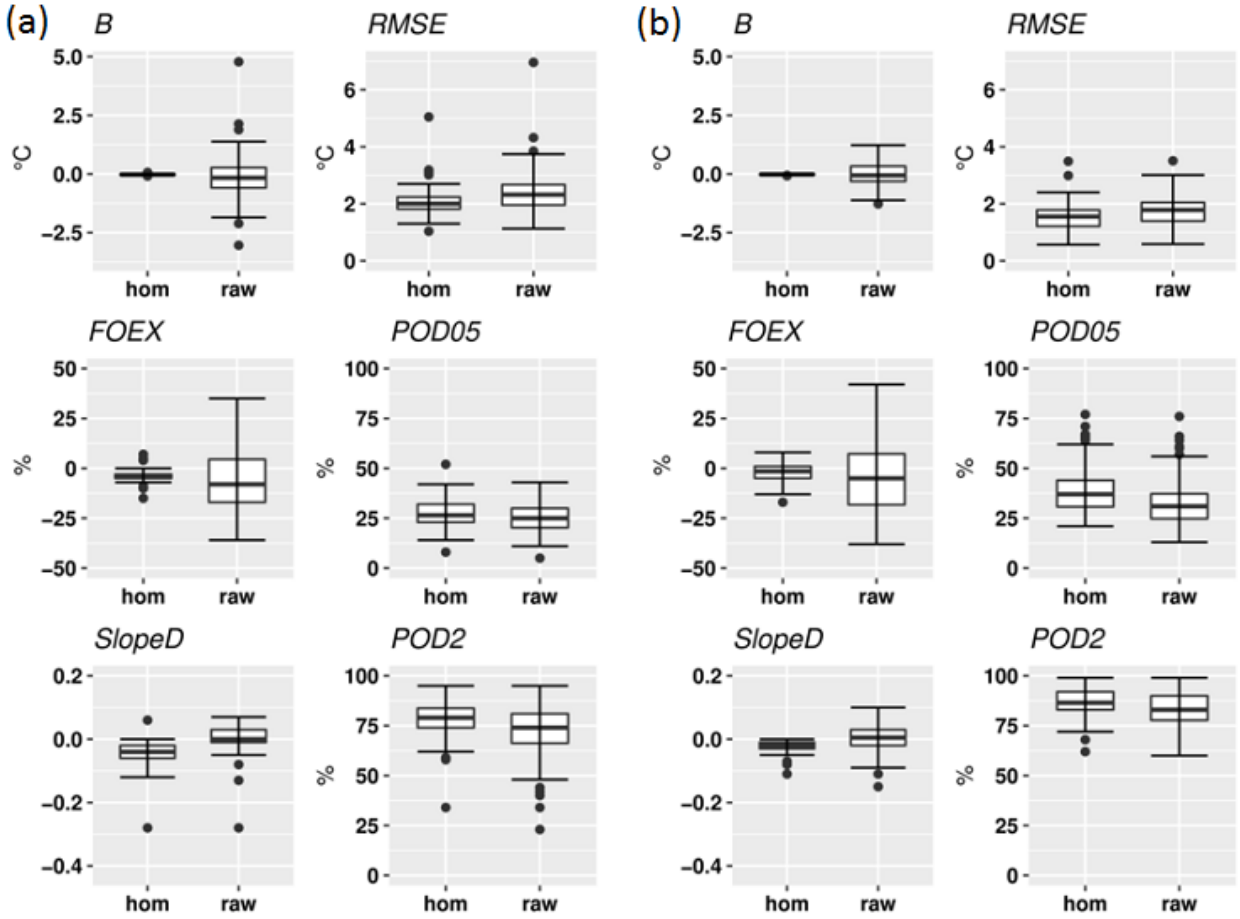

Figure 13. Boxplots of the metrics calculated in the set of numerical experiments #2: (a) TN, (b) TX

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.

It is important to evaluate how the accuracy of the adjustment algorithm for data with the daily temporal resolution is reflected in calculation of climate extreme indices and their regular tendencies (trends) (e.g. Trewin and Trevitt, 1996). To do so, we calculated the yearly time series of the temperature data, TNy and TXy, and the following indices (Klein Tank et al., 2009): 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. In order to indicate these changes in the calculating algorithms of the indices clearly, we will denote them as FD-10, TR10, ID5 and SU20. 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 using the adjusted segment/segments only. In addition, we computed differences/errors in the indices linear trends (TrD), calculated for the adjusted and clean data. The trends were evaluated over the whole time series (including undisturbed segments) through the least squares regression.

The boxplots of the metrics calculated based on the adjusted yearly time series of the air temperature data and the climate extremes indices are presented in Figure 14. Similar results that were obtained based on the 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 the both types of residual errors. Nearly zero values of B for adjusted TNy and TXy series is a trivial result, since Climatol removes very well systematic errors even in daily data. The mean value of $RMSE$ for TNy is reduced after adjustment from 0.94°C to 0.20°C (by ~78%) while for TXy the reduction is slightly less: from 0.56°C to 0.16°C (by ~63%). Such substantial improvement of $RMSE$ for both climatic variables can be explained by the fact that averaging data to the yearly scale removes random/noisy part of the residual errors, seen on the daily scale. Note, that the mean values of $RMSE$, 0.20°C for TNy and 0.16°C 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 ~0.01°C/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 the 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 Figure 13). 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 the adjusted data are also presented in Figure 14. Table 21 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 the corrected data, are much closer to real trends than evaluated based on the raw time series.

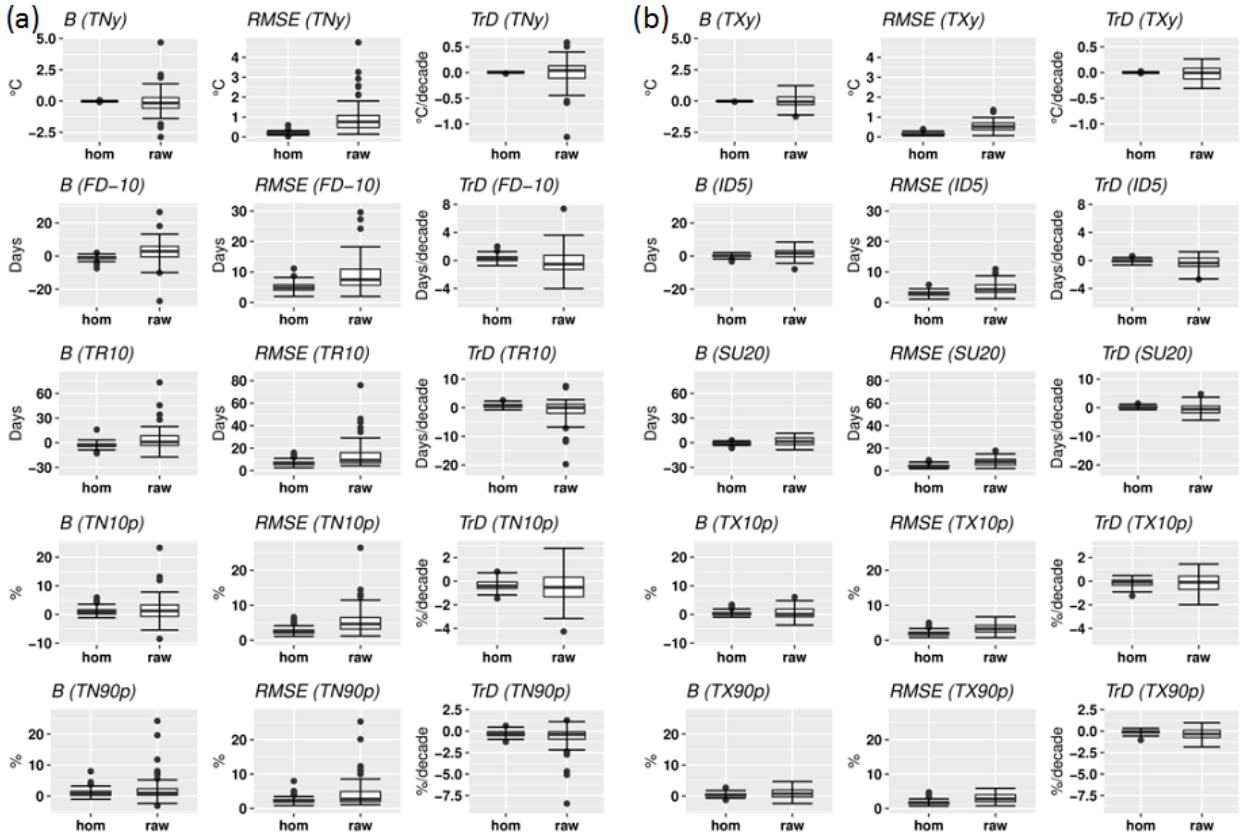

Figure 14. Boxplots of the metrics calculated based on the yearly series of the climate extremes indices in the set of numerical experiments #2: (a) TN, (b) TX

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-10 (frost days) is relatively small, 0.29 *days/decade* (2.9 *days/100years*). The uncertainty of the trend calculation in the adjusted FD-10 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-10 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 lie 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 21 for other climate extreme indices. We also can conclude, that, in general, 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 remain 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.

Table 21. 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-10** ***days/decade***		**TR10** ***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)	**ID5** ***days/decade***		**SU20** ***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

6. Conclusions

The INDECIS project has provided the excellent benchmark data sets which can be used for verification/validation/evaluation/uncertainty quantification of homogenization software. In the Report, the benchmark data were used to evaluate performance of several homogenization methods/techniques (HOMER, SMHI-HOMER and ACMANT) on the monthly time scale and to quantify uncertainty of the Climatol software on the daily scale.

Based on a set of calculated statistical metrics, the evaluation of the software HOMER, SMHI-HOMER and ACMANT on the monthly scale was focused on potential dependences of the homogenization results on physical features of a station (i.e. latitude, altitude, distance from the sea) and the nature of

the inhomogeneities (i.e. the number of break points and missing data). In general, the nature of the datasets (i.e., number of breaks and missing data) seems to have a more important role in yielding good homogenization results than physical parameters associated to the stations (i.e., latitude, elevation and distance from the sea).

For the Climatol software, the quantification of uncertainty of its adjustment algorithm was performed for daily air temperature time series. The residual errors were evaluated using complex approach which allowed to perform the uncertainty evaluation on the day-to-day scale as well as overall (averaged over time) assessment. On the yearly scale, the uncertainty was evaluated mainly based on calculation of climate extremes indices. As a general conclusion, it can be stated that Climatol removes very well systematic errors related to jumps in the means. Scatter errors in the daily raw time series are removed less efficiently. Besides, Climatol's adjustment uncertainty, evaluated on the daily scale, varies over time. The width of the residual errors distribution in summer months is substantially less compared to wintertime. In addition, both types of errors are removed well in the yearly time series of the air temperature and the extreme indices. Substantial decrease of the linear trend errors in the yearly time series can also be reported.

References

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

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. and 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.

Alexandersson, H. (1986) A homogeneity test applied to precipitation data. *Journal of Climatology*, 6(6), 661-675. https://doi.org/10.1002/joc.3370060607.

Brunet, M., Saladié, O., Jones, P., Sigró, J., Aguilar, E., Moberg, A., Lister, D., Walther, A.and 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.

Caussinus, H. and Mestre, O. (2004) Detection and correction of artificial shifts in climate series. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 53, 405-425. https://doi.org/10.1111/j.1467-9876.2004.05155.x.

Coll, J., Domonkos, P., Guijarro, J., Curley, M., Rustemeier, E., Aguilar, E., Walsh, S. and Sweeney, J. (2020) Application of homogenization methods for Ireland's monthly precipitation records: Comparison of break detection results. *International Journal of Climatology*, 1–20. https://doi.org/10.1002/joc.6575

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. and Kim, J. (2008) Evaluation of the HadGEM2 model. MetOffice Hadley Centre Technical Note 74, 47 pp.

Della-Marta, P. and Wanner, H. (2006) A method of homogenizing the extremes and mean daily temperature measurements. Journal of Climate, 19(17), 4179–4197. https://doi.org/10.1175/JCLI3855.1.

Domonkos, P. (2011) Adapted Caussinus-Mestre algorithm for networks of temperature series (ACMANT). *International Journal of Geosciences*, 2(3), 293–309. https://doi.org/10.4236/ijg.2011.23032.

Domonkos, P. and Efthymiadis, D. (2013) Development and testing of homogenization methods: moving parameter experiments with ACMANT. *Advances in Science and Research*, 10, 43–50. https://doi.org/10.5194/asr-10-43-2013.

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.

Guijarro, J.A. (2018) Homogenization of climatic series with Climatol. Version 3.1.1. Guide

Joelsson, M., Slättberg, N., Carnebring, A., Sturm, C., and Engström, E. (2020) Automation of the interactive mode of the homogenisation software HOMER for climatological applications , EGU General Assembly, Online, 4–8 May 2020, EGU2020-1397, https://doi.org/10.5194/egusphere-egu2020-1397.

Klein Tank, A.M.G., Wijngaard, J.B., Können, G.P. *et al.* (2002) Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. *International Journal of Climatology*, 22(12), 1441-1453. https://doi.org/10.1002/joc.773.

Klein Tank, A.M.G., Zwiers, F.W. and 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.

Kuglitsch, F.G., Auchmann, R., Bleisch, R., Broennigmann, S., Martius, O., and Stewart, M. (2012) Break detection of annual Swiss temperature series. *Journal of Geophysical Researcch. Atmosphere*. 117. D13105. https://doi.org/10.1029/2012JD017729.

Lindau, R. and Venema, V. (2016) The uncertainty of break positions detected by homogenization algorithms in climate records. International Journal of Climatology, 36(2), 576-589. https://doi.org/10.1002/joc.4366.

Mestre, O., Domonkos, P., Picard, F., Auer, I., Robin, S., Lebarbier, E., Bohm, R., Aguilar, E., Guijarro, J., Vertachnik, G., Klancar, M., Dubuisson, B. and Stepanek, P. (2013) HOMER: a homogenization software—methods and applications. *Idojaras* (*Quarterly Journal of Hungarian Meteorological Service*), 117(1), 47–67.

Skrynyk, O., Aguilar, E., Guijarro, J., Randriamarolaza, L.Y.A. and Bubin. S. (2020) Uncertainty evaluation of Climatol's adjustment algorithm applied to daily air temperature time series. *International Journal of Climatology*. https://doi.org/10.1002/joc.6854.

Sturm, C. and 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.

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

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

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.

Venema, V., Mestre, O., Aguilar, E., Auer, I., Guijarro, J.A., Domonkos, P., Vertacnik, G., Szentimrey, T., Stepanek, P., Zahradnicek, P., Viarre, J., Muller-Westermeier, G., Lakatos, M., Williams, C.N., Menne, M., Lindau, R., Rasol, D., Rustemeier, E., Kolokythas, K., Marinova, T., Andresen, L., Acquaotta, F., Fratianni, S., Cheval, S., Klancar, M., Brunetti, M., Gruber, C., Duran, M.P., Likso, T., Esteban, P. and Brandsma, T. (2012) Benchmarking monthly homogenization algorithms. *Climate of the Past*, 8, 89–115. https://doi.org/10.5194/cp-8-89-2012.

Vincent, L.A., Milewska, E.J., Wang, X.L. and Hartwell, M.M. (2018) Uncertainty in homogenized daily temperatures and derived indices of extremes illustrated using parallel observations in Canada. International Journal of Climatology, 38(2). 692-707. https://doi.org/10.1002/joc.5203.

Walker, W.E., Harremoës, P., Rotmans, J., van der Sluijs, J.P., van Asselt, M.B.A., Janssen, P. and Krayer von Krauss, M.P. (2003) Defining uncertainty: a conceptual basis for uncertainty management in model-based decision support. *Integrated Assessment*, 4(1), 5-17. https://doi.org/10.1076/iaij.4.1.5.16466.

Zhang, X., Feng, Y., Chan, R. (2018) Introduction to RClimDex v1.9. Guide. Climate research Division, Environment Canada, Downsview Ontario, Canada.