

## RESEARCH ARTICLE

# Automation and evaluation of the interactive homogenization tool HOMER

L. Magnus T. Joelsson  | Christophe Sturm | Johan Södling |  
Erik Engström  | Erik Kjellström 

Swedish Meteorological and Hydrological  
Institute, Norrköping, Sweden

**Correspondence**

L. Magnus T. Joelsson, Swedish  
Meteorological and Hydrological Institute,  
Norrköping, 601 76, Sweden.  
Email: magnus.joelsson@smhi.se

**Abstract**

The interactive HOMER is automated by the use of a set of parameters. This approach retains the skill and flexibility of an interactive method, but adds the speed and reproducibility of an automatic method. The automation of the interactive HOMER also enables systematic testing. Its performance is evaluated by the homogenization of the Indecis homogenization benchmark datasets. The overall performance of the interactive HOMER compares well with the methods using the homogenization tools Climatol and ACMANT and surpass the performance of the standard automatic HOMER. All the homogenization methods reduce the initial error. The average residual error and all error percentiles below and including the 99th error percentile do not differ more than 0.3°C between interactive HOMER and the other methods. Interactive HOMER and Climatol report fewer homogeneity breaks than the true number, while standard automatic HOMER and ACMANT report more homogeneity breaks. Across the methods, a higher number of reported homogeneity breaks renders a higher share of the true homogeneity breaks to be detected, but also a higher share of the reported breaks to be false positives. On average the differences in the corrected times series are small between the methods implying that the choice of homogenization method is a matter of preference.

**KEYWORDS**

ACMANT, Climatol, HOMER, homogenization, Indecis

**Abbreviations:** CRMSE, centred root mean square error; POD, probability of detection; POH, probability of hit; RMSE, root mean square error; Se, Sweden; Si, Slovenian; tn, daily minimum temperature; tx, maximum daily temperature.

## 1 | INTRODUCTION

Time series of meteorological observables might have nonclimatological shifts (henceforth, homogeneity breaks or simply breaks). These homogeneity breaks can be caused by changes in, for example, observation times, site or instruments. A true climate signal can be distorted by

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2021 The Authors. *International Journal of Climatology* published by John Wiley & Sons Ltd on behalf of Royal Meteorological Society.

such homogeneity breaks. Climatological studies require observational data to be homogenized (World Meteorological Organization, 2017a; 2017b). There is an array of homogenization software tools available (Venema *et al.*, 2020), both interactive and automatic.

Within homogenization methods using interactive tools, an operator investigates outputs from the tool and selects the time of homogeneity breaks. The data is then homogenized accounting for the selected homogeneity breaks. Interactive methods make use of the operator's skill and knowledge of the station network. If several large data sets are to be frequently homogenized, interactive methods might be unappealing as they are time-consuming. It is also difficult to quantitatively evaluate an interactive method as such evaluation is practically limited to smaller data sets and the results depend on subjectivity of the operator. Methods using tools that automatically select the time of homogeneity breaks and perform the homogenization can be used in a more time efficient way. Naturally, a problem with such methods is that they may not take skill and knowledge of experienced operators into account, which may lower the quality of the homogenization. However, studies have shown that such methods can perform as well as interactive methods (Venema *et al.*, 2012).

The Swedish Meteorological and Hydrological Institute (SMHI) has homogenized data sets of monthly averages of daily mean temperature and monthly precipitation observations from the Swedish network of weather stations using the Standard Normal Homogeneity Test (SNHT) (Alexandersson and Moberg, 1997). SNHT is a reliable but labour intensive method. A new homogenization protocol is to be established at SMHI, including the selection of a homogenization tool. A candidate is the well-established HOMogenizaton softwarE in R (HOMER) (Mestre *et al.*, 2013) that has been initially tested at SMHI. HOMER has been used by several meteorological institutes (e.g., Coll *et al.*, 2014; Vertačnik *et al.*, 2015; Kuya *et al.*, 2020) and in numerous scientific studies (e.g., Morán-Tejeda *et al.*, 2016; Kivinen *et al.*, 2017; Vicente-Serrano *et al.*, 2017). HOMER can be run in either interactive or automatic mode. In the interactive mode, HOMER suggests a number of homogeneity breaks for an operator to reject or confirm. In the automatic mode, HOMER automatically confirms or rejects homogeneity breaks.

A few issues have been reported with the use of HOMER in operational use. The automatic mode of HOMER has been found to deliver biased corrections of temperature time series of a network of stations in Switzerland (Gubler *et al.*, 2017). The error could potentially be linked to a shift in temperatures over large parts of western Europe around 1987 (De Laat and

Crok, 2013). Gubler *et al.* (2017) dissuaded the use of HOMER in automatic mode. A similar problem was also found by Mestre *et al.* (2013) who states 'However, the automatic joint-detection is not perfect'. Furthermore, the automatic mode of HOMER is not actually fully automatic and can therefore not be run in batch mode (Guijarro *et al.*, 2017).

Pérez-Zanón *et al.* (2015) used HOMER in interactive mode when comparing HOMER and the homogenization tool Adapted Caussinus-Mestre Algorithm for Networks of Temperature series (ACMANT) (Domonkos *et al.*, 2011). They conclude: 'While HOMER detects more breaks supported by metadata, this method is also more dependent on the user skill and thus sensitive to subjective errors'. Vertačnik *et al.* (2015), who used HOMER in an ensemble type operation with up to six experts independently homogenizing a smaller data set of the Slovenian station network, concluded that 'This semi-automatic homogenization approach based on metadata gave more reliable homogenization results than a fully automatic approach without metadata'. Even though the problem of subjectivity thereby is avoided, such operation is impractical as it requires the participation of several experts.

The current study has two objectives:

1. Develop a homogenization method that combines the speed of HOMER's automatic mode with the skill and flexibility of HOMER's interactive mode
2. Enable quantitative evaluation of HOMER's interactive mode

To achieve these objectives, HOMER's interactive mode is automated using a number of parameters. The method is evaluated by homogenizing monthly temperature time series of the synthetic Indecis benchmark data set (Indecis Project, 2017). The data set includes daily maximum ( $T_{\max}$ ) and daily minimum temperature ( $T_{\min}$ ) of two separate networks. The two networks reflect different real world like conditions. The evaluation of HOMER is made without facilitative measures, such as splitting of the network (Pérez-Zanón *et al.*, 2015; Kuya *et al.*, 2020) and without the influence of the operator skill. The results are compared with the results of the homogenization by the well-established homogenization tools Climatol (Guijarro, 2018) and ACMANT. The selection of reference series and the number of homogenization iterations are investigated in more detail.

The rest of the article is organized as follows. In Section 2 the tool HOMER, and how it is typically used, is described (Section 2.1), along with a shorter description of the benchmark tools (Section 2.2) and the evaluation data set (Section 2.3). In Section 3 the automation of

HOMER (Section 3.1), how the tests are conducted (Section 3.2) and evaluated (Section 3.3) are described. The results are presented in Section 4 and discussed in Section 5. The article is rounded up with conclusions in Section 6 and outlook for future studies and development in Section 7.

## 2 | DATA AND TOOLS

In the following sections, HOMER (Section 2.1), the other homogenization tools (Section 2.2), and the input data (Section 2.3) are described. 'Tool' here denotes a homogenization software package (e.g., ACMANT, Climatol and HOMER), 'method' denotes the application of a tool with a fixed protocol and set of parameter values (e.g., ACMANT and Climatol with their respective default parameter settings, and HOMER-auto and HOMER-inter) and 'function' denotes a procedure inside a tool (e.g., detection, gap filling and correction methods of the different software packages).

### 2.1 | HOMER

#### 2.1.1 | Description

HOMER uses a combination of functions for detection of homogeneity breaks (see Table 1; Figure 1). The detection functions all use a subset of the network to detect inhomogeneities in each time series. The time series in focus is henceforth referred to as the candidate series. The subset

of series which the candidate series is compared with are called references. The references are chosen either on the basis of geographical proximity or correlation. A threshold value for the longest acceptable distance or the lowest acceptable correlation and the minimum number of references are set. In the current study the correlation threshold is 0.95 and the minimum number of reference series is eight (Kuya *et al.*, 2020). There are four detection functions. First, the pairwise detection function of PRODIGE (Caussinus and Mestre, 2004) compares annual averaged data of the candidate series with each of the reference series one by one. Second, the pairwise detection function is used on seasonal (winter and summer) data. Third, a joint detection function from Picard *et al.* (2011) compares the candidate series with all the reference series simultaneously. Fourth and final, a detection function called 'ACMANT' in Mestre *et al.* (2013) (as it is borrowed from the ACMANT homogenization tool) includes detection of seasonal cycle range inhomogeneities. In order to avoid confusion between the ACMANT detection function and the ACMANT homogenization tool, the detection function is henceforth called the 'ACMANT-style detection function' and the homogenization tool is simple referred to as 'ACMANT'. The functions uses dynamic programming (Bellman, 1954; Fisher, 1958) to fit a step function with a given number of change points. The positions of the change points are set to minimize the internal variance, that is, the variance of the levels. A level here denotes the time between two breaks or between a break and an end point. The Caussinus and Lyazhri-criterion (Caussinus and Lyazhri, 1997), or in case of the joint-detection function, a modified Bayesian Information Criterion (BIC) (Picard *et al.*, 2011), is used to find the optimal number of change points. The functions are further described in Mestre *et al.* (2013). The results of the detection functions are combined in a number of figures for examination by the operator to reject or confirm suggested homogeneity breaks station by station. The HOMER tool includes the ANOVA-based (analysis of variance) correction and gap fill function from Caussinus and Mestre (2004). ANOVA is further explained in Lindau and Venema (2018).

Even though the input data has monthly resolution, the detection functions detect breaks on annual basis. HOMER is setup such that the breaks are assigned to December of the year in consideration. To assign the break its optimal placement on a monthly scale, a function borrowed from ACMANT is implemented. The so-called 'change month' function, see Figure 1, can move the homogeneity break  $\pm 24$  months from the initial placement (Mestre *et al.*, 2013). Note that the intention of this study is not to improve the core HOMER method; therefore the seemingly arbitrary value of  $\pm 24$  months for the change month function is not modified. The

TABLE 1 The functions of HOMER/Bart

Function	Description
Pairwise-detection, annual	Mean level homogeneity break detection function from PRODIGE on annual average data
Pairwise-detection, seasonal	Mean level homogeneity break detection function from PRODIGE on seasonal average data
Joint-detection	Mean level homogeneity break detection from the <i>cghseg</i> package Picard <i>et al.</i> , (2011)
ACMANT-style detection	Seasonal cycle homogeneity break detection function from ACMANT
Correction	ANOVA based correction function including gap filling
Change month	Finds the optimal change point on monthly scale from change points on annual scale

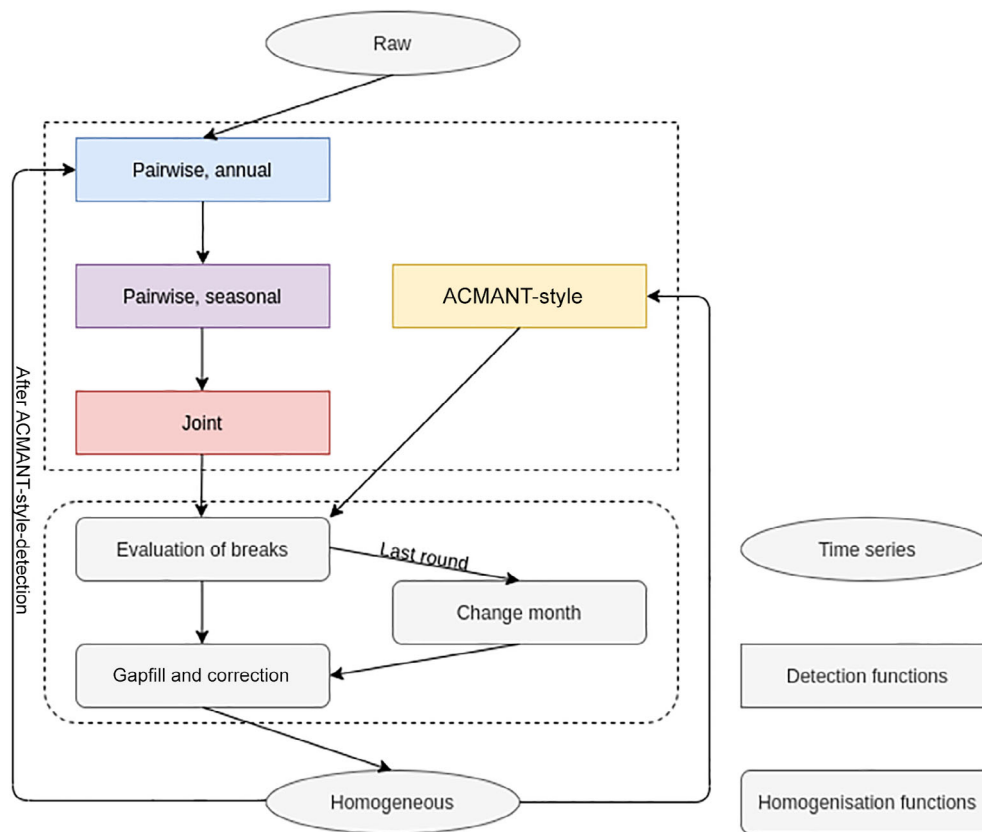


FIGURE 1 Flow chart of HOMER [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

change month function can also merge multiple breaks into a single break or split up a single break to multiple breaks within this time frame.

### 2.1.2 | Procedure

The suggested work flow (Mestre and Aguilar, 2011; Vertačnik *et al.*, 2015; Coll *et al.*, 2018; Kuya *et al.*, 2020) is depicted as a flow chart in Figure 1. First homogeneity breaks are detected by applying the detection function in the sequence pairwise annual-style, pairwise seasonal-style, joint-style and ACMANT-style detection. Since the ACMANT-style detection function needs pre-homogenized data, the correction function is applied both before and after the ACMANT-style detection function. The process should be repeated (at least) once. Before the final correction of the data series, it is recommended to apply the change month function.

### 2.1.3 | Confirmation of homogeneity breaks

When HOMER was previously tested at SMHI a 'traditional' procedure for the confirmation or rejection of possible homogeneity breaks was adopted. The procedure is as follows: For each year in each candidate series the number

of breaks reported by the pairwise-detection functions are counted. The pairwise-detection functions compare the candidate series with its reference series. For each year in each candidate series, there can maximum be as many breaks reported as the number of reference series. The number of breaks reported by the seasonal function is multiplied by  $\frac{1}{3}$ . Breaks reported by the joint-detection and the ACMANT-style detection functions are added to the sum. If the sum adds up to four or more, the break is confirmed. If the sum adds up to three, metadata is consulted. If the considered break is supported by metadata the break is confirmed. If metadata does not support the considered break, the break is noted in a notebook. If the same break is repeatedly noted, the break is confirmed. Reported breaks in the previous or following year adds to the candidate break, but with a penalty of one break signal. Breaks cannot be confirmed in consecutive years in the same time series, which is a common restriction for homogenization methods (Venema *et al.*, 2020).

## 2.2 | Benchmark tools

### 2.2.1 | Climatol

The Climatol version used here is Climatol v3.1.2. Climatol constructs standardized data series by using a

Reduced Major Axis linear regression (Clarke, 1980) on the reference series, parallel to the candidate series (Guijarro, 2018). The selection of references is based on proximity only such that all series, even short series with little or no temporal overlap with the candidate series, can be included. The difference between the constructed and the candidate series are used for detecting outliers and homogeneity breaks. The homogeneity break detection is based on the SNHT method. The detection is performed iteratively until no breaks larger than a certain threshold is found.

### 2.2.2 | ACMANT

The ACMANT version used here is ACMANT v4.3 (Domonkos, 2019). As HOMER, ACMANT is based on the PRODIGE homogenization method (Causinus and Mestre, 2004). Like Climatol and HOMER's joint detection function, ACMANT construct a composite reference series from the selected references of each candidate series but does not use the pairwise technique of HOMER. The functions work on deseasonalized data, the seasonal cycle is then added in the end of the process in ACMANT. The selection of references is based of Spearman correlation and can vary from year to year. The homogenization is performed in three iterations with increasing sensitivity. ACMANT has numerous special features: Detection of changes in seasonal cycles, ensemble homogenization and weighted ANOVA model besides the ordinary ANOVA correction model.

### 2.2.3 | Previous comparisons

A comprehensive evaluation of homogenization tools is the 'European Cooperation in Science and Technology Action ES0601: advances in homogenization methods of climate series: an integrated approach' (COST-HOME) project (Venema *et al.*, 2012). Two of the evaluated tools were Climatol and ACMANT. ACMANT had lower residual error and higher Probability of Detection (POD) than Climatol. Climatol had higher Heidke Skill Score (HSS) (Heidke, 1926) than ACMANT. HSS is designed to be more sensitive to false positives than other skill scores such as the Peirce Skill Score (Peirce, 1884). COST-HOME was followed by the 'Multiple verification of automatic software homogenizing monthly temperature and precipitation series' (MULTITEST) project (Guijarro *et al.*, 2017). A series of tests with varying degree of complexity was performed. ACMANT produced the lowest RMSE for most of the tests. For the more simple tests the RMSE of Climatol and ACMANT did not differ

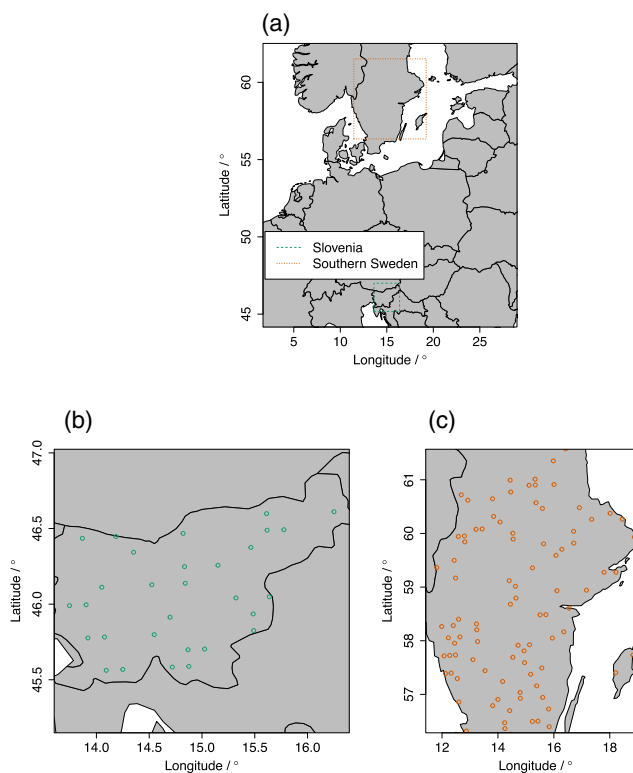
significantly from each other. The automatic mode of HOMER received as low RMSE as ACMANT for the simpler tests but higher RMSE for the more complex tests.

From these two studies it can be concluded that ACMANT's break detection is generally more sensitive than Climatol's as ACMANT finds higher number of both true and false breaks.

## 2.3 | Indecis benchmark data set

To quantitatively test the performance of HOMER and compare it with the performance of Climatol and ACMANT, a benchmark data set is required. In the present study the Indecis benchmark  $T_{\min}$  and  $T_{\max}$  data sets are used (Indecis Project, 2017; Guijarro, 2019). The Indecis benchmark data sets consists of synthetic time series with daily temporal resolution between 1950 and 2005 compiled from the KNMI's regional climate model RACMO v2 (Indecis Project, 2017; Skrynyk *et al.*, 2020) to mimic the records of observational networks. In the Indecis data set, there are two networks with virtual weather stations (see Figure 2): A Slovenian network with 30 time series (see Figure 2b) and a southern Swedish network with 100 time series (see Figure 2c). Since the aim of this study is to evaluate the general skill of HOMER and the other tools (not only their skill with Swedish climate) both available temperature data sets are used. The data are distorted by inserting homogeneity breaks and data gaps. The unhomogenized data with inserted homogeneity breaks and data gaps is henceforth denoted 'raw' data. 'True' data denotes the corresponding data without inserted homogeneity breaks and data gaps. The inserted breaks are referred to as 'actual breaks' where they need to be distinguished from breaks suggested by a method ('reported breaks'). Note that there are no actual breaks in the sense of breaks in the true data as the true data is supposedly completely homogeneous. Most stations have two or three actual breaks. The maximum number of breaks in one series is seven. The Indecis benchmark data sets used in the current study allows us to study four temperature cases: Slovenian  $T_{\max}$ , Swedish  $T_{\max}$ , Slovenian  $T_{\min}$  and Swedish  $T_{\min}$ .

The two networks have distinct characteristics: The Slovenian network is less dense, have lower temporal correlation and larger average error than the Swedish network. The average number of breaks per time series are about the same for all four cases. The mean variance of the annual data is larger for the Slovenian ( $\sigma_{\text{tx}}^2 = 4.6^\circ\text{C}^2$  and  $\sigma_{\text{tn}}^2 = 1.7^\circ\text{C}^2$ ) than the Swedish network ( $\sigma_{\text{tx}}^2 = 1.0^\circ\text{C}^2$  and  $\sigma_{\text{tn}}^2 = 1.5^\circ\text{C}^2$ ), especially for  $T_{\max}$ . In the Swedish data sets, the variance in the annual averaged data is slightly higher for  $T_{\min}$ .



**FIGURE 2** The distribution of the stations in the two networks. (a) The Indecis networks. (b) The Slovenian network. (c) The southern Swedish network [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

### 3 | APPROACHES AND EVALUATION

In the following sections, the development of the new automated version of HOMER is described (Section 3.1), followed by descriptions of the tests (Section 3.2) and metrics (Section 3.3) that are used in the current study to evaluate the different homogenization methods and tools.

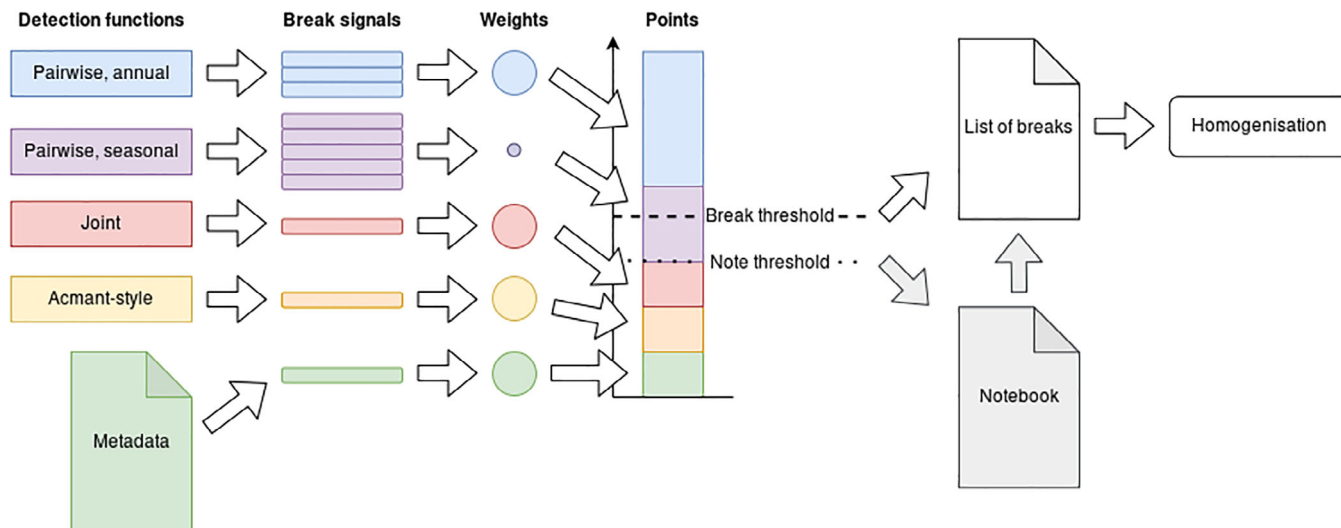
#### 3.1 | Automation of the interactive mode of HOMER

The interactive mode of HOMER is automated in a script called Bart. The source code and a user manual of Bart is included in the Appendix S1. The traditional procedure described in Section 2.1.3 is translated into a point system where each function is ascribed a weight factor (see Figure 3). The weights and the thresholds are controlled by parameter values. Decreasing threshold values or increasing weights of the detection functions makes the method more sensitive to homogeneity breaks. Breaks described in the metadata records are treated like breaks reported by the detection functions: The metadata breaks are also ascribed a weight controlled by a parameter

value. The role of metadata in the homogenization process can thus be chosen according to the quality of the metadata (Venema *et al.*, 2020). There are examples of such differences in treatment of metadata in the literature; Kuya *et al.* (2020) restricts their validation of homogeneity breaks to those supported by metadata, while Gubler *et al.* (2017) argues that metadata in their case with a sparse network only should decide the exact placement in time of statistically confirmed homogeneity breaks. Alternatively to the traditional use of two iterations, Bart can let HOMER's homogenization process converge by repeating the process until no additional breaks are reported.

This gives the user a number of input parameters, as tabulated in Table 2. The use of inputs enables the Bart script to run HOMER in batch mode, which has not been possible previously (Guijarro *et al.*, 2017).

A recurring problem when working with HOMER on incomplete data series and when breaks are confirmed in later stages with gapfilled data, is the occurrence of levels with no original data. In such cases, HOMER has, until now, required the interaction of the operator to remove one of the breaks, also in the automatic mode. This is automated in the Bart script: A break after a level with no original data is removed such that the level is merged with a later level. If the



**FIGURE 3** Depiction of the evaluation of a fictional potential homogeneity break in a certain year in a time series where the joint detection function and the Acmant-style detection function reports a break, the pairwise detection function reports breaks for three reference time series on annual basis and for five reference time series on seasonal basis, and the break is supported in the metadata. The sum of break points exceeds the break threshold and, consequently the break is added to the list of homogeneity breaks [Colour figure can be viewed at wileyonlinelibrary.com]

**TABLE 2** The input parameters and their set of values for the HOMER-inter (Inter.), the HOMER-auto (Auto.) homogenizations, the homogenization with constant number of references (Const. Ref.) and the convergence homogenization (Conv.)

Parameter	Inter.	Auto.	Const. ref	Conv.
Correlation threshold	0.95	0.95	1.00	0.95
Minimum number of reference series	8	8	8	8
Number of iterations	2	2	2	–
Convergence	Off	Off	Off	On
Pairwise-detection annual weight	3	0	3	3
Pairwise-detection seasonal weight	1	0	1	1
Joint-detection weight	3	1	3	3
ACMANT-style weight	3	1	3	3
Adjacent penalty	3	∞	3	3
Break validation threshold	12	1	12	12
Note threshold	9	–	9	9
Restricted time	1	0	1	1

last level does not have original data, the level is merged with the previous level.

### 3.2 | Homogenization of benchmark data sets

#### 3.2.1 | Preparation

The monthly data series used in this study are obtained from the daily data values of the Indecis benchmark data sets. In the Indecis benchmark data sets, parts of the data is removed to mimic missing data. To fill the gaps in the

daily data linear temporal interpolation is used. If more than 5 days are missing, the monthly value is considered missing. Interpolation is not possible if data is missing in the start or the end of the data series, or if the missing data is part of a longer gap. If interpolation is not possible, the mean is calculated on the available days.

#### 3.2.2 | Method comparison

The Indecis benchmark data sets for the four cases ( $T_{max}$  and  $T_{min}$  for Slovenia and Sweden) are homogenized by HOMER using the Bart script with settings corresponding

to the traditional use of the interactive mode (HOMER-inter) and with settings corresponding to the automatic mode (HOMER-auto). The settings for the two HOMER methods are tabulated in Table 2. The data sets are also homogenized by Climatol and ACMANT. Furthermore, HOMER's correction function is applied on the raw data with actual break positions and the default reference selection settings (0.95 correlation threshold, minimum eight references). These results can be considered as an upper limit for the score of HOMER with the current settings. To assess the overall skill of the four homogenization methods a number of metrics are calculated by comparing the homogenized data with the true data or by comparing the reported breaks with the actual breaks (see Section 3.3).

### 3.2.3 | Reference selection and homogenization iterations

The use of the correlation threshold for the selection of reference series means that the number of references can vary between stations. The sensitivity of especially the pairwise detection function could potentially vary between stations within a network. Since homogeneity breaks disturb the correlation between series, series with fewer breaks require less climatological similarity between them to be well correlated and can therefore be ascribed more references. To assess the effect of the number of references for the homogenization of each station, an alternative homogenization is conducted where the references are the eight most highly correlated series for each candidate (constant method).

Under the current settings of the minimum reference correlation and minimum number of references, the number of references varies between eight and 53 for the stations in the Swedish data sets, see Figure 4a. In the further analysis the stations are binned according to their number of references in the default selection of the Swedish  $T_{\max}$  data set. In the Slovenian data sets almost all stations have eight references. The Swedish  $T_{\max}$  case has a wider spread of references than the  $T_{\min}$  case, hence the sole focus on the  $T_{\max}$  case. To illustrate how the reference selection is influenced by the homogeneity of the time series, the number of actual breaks in series is plotted against the number of references as a bar plot in Figure 4b. The average number of actual breaks generally decreases for bins with increasing number of references.

To investigate the effect of the number of iterations in HOMER's homogenization process, the Indecis data sets are homogenized by repeating the process until no further homogeneity breaks are reported by the detection functions (convergence). The results are compared with the results of

the operational method, which includes two homogenization rounds. Skill scores (see Section 3.3.2) are then calculated after each correction of the process. Note that the correction function is applied twice for each iteration.

The settings for the two alternative HOMER methods (constant number of references and convergence) are tabulated in Table 2.

## 3.3 | Metrics

Two groups of metrics are used in the current study: Time series based metrics and skill scores. Time series based metrics compare the output corrected series with the true and the input raw time series. The skill scores compare the lists of breaks reported by the method with the list of actual breaks.

### 3.3.1 | Time series based metrics

In the following section  $x_{i,j,m}$  and  $y_{i,j}$  represents the homogenized and the true value respectively. The subscript  $i$  refer to the time  $t_i$ . The time vector spans from  $t_0$  to  $t_{(n-1)}$ . The subscript  $j$  refer to the identity of the time series (i.e., weather station). The subscript  $m$  refer to homogenization method.  $\bar{x}_{j,m}$  and  $\bar{y}_j$  represents the temporal averages of the homogenized and true values of the variable, respectively.  $m = \text{raw}$  refers to unhomogenized data. The root mean square error (RMSE) is defined as:

$$\text{RMSE}_{j,m} \equiv \sqrt{\frac{1}{n} \sum_{(i=0)}^{n-1} (x_{i,j,m} - y_{i,j})^2}. \quad (1)$$

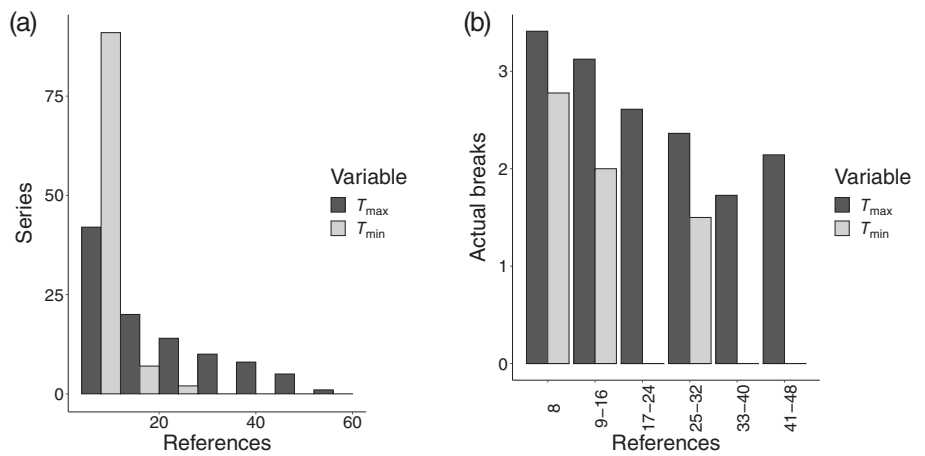
For the evaluation of homogenization methods, the Centred Root Mean Square Error (CRMSE) is recommended over the ordinary RMSE, since 'the main aim of homogenization is not to improve the absolute values but rather the temporal consistency' (Venema *et al.*, 2012). CRMSE for the homogenization method  $m$  is defined as:

$$\text{CRMSE}_{j,m} \equiv \sqrt{\frac{1}{n} \sum_{(i=0)}^{n-1} \left( (x_{i,j,m} - \bar{x}_{j,m}) - (y_{i,j} - \bar{y}_j) \right)^2} \quad (2)$$

In addition, the bias  $B$  (mean departure from true data) is calculated for both raw and homogenized data.

$$B_{j,m} \equiv \frac{1}{n} \sum_{(i=0)}^{n-1} (x_{i,j,m} - y_{i,j}). \quad (3)$$

**FIGURE 4** Statistics of number of references and number of actual breaks per time series, in the Indecis Swedish network data set homogenized by the operational interactive mode of HOMER. (a) Number of time series binned according to their number of references, Swedish network. (b) Average number of actual breaks per series versus number of references of the series, Swedish network



To evaluate the residual error relative to the raw error, the measure of efficiency  $E$  is used (Domonkos, 2013b; Gubler *et al.*, 2017):

$$E_{j,m} \equiv \frac{\text{RMSE}_{j,\text{raw}} - \text{RMSE}_{j,m}}{\text{RMSE}_{j,\text{raw}}} \quad (4)$$

Where the error is to be evaluated station by station, the raw error is in some cases small or even zero. In such cases the efficiency measure is inapplicable and instead error reduction is used:

$$\text{Error reduction}_{j,m} \equiv \text{RMSE}_{j,\text{raw}} - \text{RMSE}_{j,m}. \quad (5)$$

When homogenizing real observational data, the correct solution is not known. Every correction of raw data aims to bring the time series closer to the true values, but risks of adding error to the series. It is, therefore, preferable to achieve as low residual error as possible with the smallest correction possible. The mean absolute correction is calculated:

$$\text{Mean absolute correction}_{j,m} \equiv \frac{1}{n} \sum_{(i=0)}^{n-1} (\text{abs}(x_{i,j,\text{raw}} - x_{i,j,m})) \quad (6)$$

The time series metrics are given as averages over a network with  $k$  time series:

$$A_m = \frac{1}{k} \sum_{j=1}^k A_{j,m} \quad (7)$$

where,  $A$  is any metric. The time series metrics are calculated on the nongap filled data such that all methods and the raw data are easily compared and such that all metrics are calculated on the same data.

### 3.3.2 | Skill scores

A detection function has two tasks: Find as many true homogeneity breaks as possible and report as few false positives as possible (Van Malderen *et al.*, 2020). Different skill scores have been adopted to capture both these skills in a single score (Menne and Williams Jr, 2005, 2009; Venema *et al.*, 2012; Domonkos, 2013a). In the current study, the two skills are kept separate in two metrics: Probability of Detection (POD) and Probability of Hit (POH). The two metrics clearly describes the characteristics of the different methods. The two metrics are calculated from three measures: True positives,  $a$ , refer to the number of years where actual breaks are detected. False positives,  $b$ , refer to the number of years where breaks are falsely reported. False negatives,  $c$ , refer to the number of years where actual breaks are undetected. In this study a break is considered to be detected if it is  $\pm 2$  years from an actual break (Menne and Williams Jr, 2009). This margin is consistent with the HOMER change month function. POD is defined as:

$$\text{POD} \equiv \frac{\sum \text{true positive}}{\sum \text{year with actual break}} = \frac{a}{a+c} \quad (8)$$

The POH is defined as:

$$\text{POH} \equiv \frac{\sum \text{true positive}}{\sum \text{year with reported break}} = \frac{a}{a+b}. \quad (9)$$

Within a conservative approach to the homogenization of observational data, it is most important to correct the breaks with large amplitude. To test how well the methods detect the largest breaks, the probability of detection of the largest breaks is calculated.  $\text{POD}_{\text{large}}$  describe the probability of detection considering only the 25% largest actual breaks.

### 3.3.3 | Confidence intervals

95% confidence intervals for all the metrics are calculated with the bootstrapping method (Efron, 1992). The bootstrapping method generates a set of values for a metric and calculates the 95th percentile of the set. The set is generated by drawing a large number of random samples from the input data and calculate the metric from each sample. The size of the samples are equal to the size of the input data. There will be a number of duplicates in each sample, which gives the set its variation. In this study, the input data is the combined time series from all the stations of a case. In the skill score metrics the input data is the  $a$ ,  $b$  and  $c$  time series, which all will have a binomial value per year.

Since the Slovenian and the Swedish networks differ in size, the averages over the four cases are calculated with uncertainty propagation (Kircher, 2001). The uncertainty of the average  $\bar{\sigma}$ , is calculated over the cases  $i \in [1, N]$  as:

$$\bar{\sigma} = \sqrt{\frac{\sum_{i=1}^N \sigma_i^2}{N}} \quad (10)$$

where,  $\sigma_i$  is the uncertainty of case  $i$  for the metric in question.

## 4 | RESULTS

### 4.1 | Comparison of methods

The time series based metrics of four homogenized cases by the different methods are presented as bar plots in Figure 5 and tabulated in Table 3. The four cases are the monthly average  $T_{\max}$  and  $T_{\min}$  of the Slovenian and the Swedish Indecis benchmark network's data sets. The correction of the actual breaks by HOMER with the default settings are also included. Nongap filled data are used.

The CRMSE scores of ACMANT, HOMER-inter and HOMER-auto are equivalent within the 95% confidence interval for all the cases except the Swedish  $T_{\min}$  where HOMER-auto has a higher CRMSE than the other two methods. Averaged over the four cases, the CRMSE scores of ACMANT and HOMER-inter are  $0.69^\circ\text{C} \pm 0.01^\circ\text{C}$ , the average CRMSE score of HOMER-auto is  $0.70^\circ\text{C} \pm 0.01^\circ\text{C}$ . Climatol has the highest average CRMSE ( $0.73^\circ\text{C} \pm 0.01^\circ\text{C}$ ), but scores equivalent to most of the other methods for all cases except the Swedish  $T_{\max}$ . Correction of actual breaks in HOMER result in an average CRMSE score of  $0.65^\circ\text{C} \pm 0.01^\circ\text{C}$ , which can be considered an upper limit for the skill of HOMER methods under the current settings. All four methods

show CRMSE scores slightly higher than that implying that there is potential for some further refinement in the detection algorithms.

On average, the absolute values of the biases of the homogenizations are  $0.02^\circ\text{C} \pm 0.01^\circ\text{C}$  or smaller, with the exception of HOMER-auto ( $+0.04^\circ\text{C} \pm 0.00^\circ\text{C}$ ). The bias of the raw data is  $+0.05^\circ\text{C} \pm 0.01^\circ\text{C}$ . The average bias of HOMER-auto includes the highest bias of all the methods and cases (the Slovenian  $T_{\max}$  case:  $+0.13^\circ\text{C} \pm 0.01^\circ\text{C}$ ).

Climatol that has the smallest mean absolute correction with  $0.71^\circ\text{C} \pm 0.01^\circ\text{C}$ , followed by HOMER-inter with  $0.73^\circ\text{C} \pm 0.01^\circ\text{C}$ . ACMANT has the mean absolute correction  $0.75^\circ\text{C} \pm 0.01^\circ\text{C}$  and HOMER-auto  $0.77^\circ\text{C} \pm 0.01^\circ\text{C}$ , which is the largest mean absolute correction. The absolute corrections are larger for the Slovenian network than for the Swedish network. The extent of corrections follow the number of breaks as shown below.

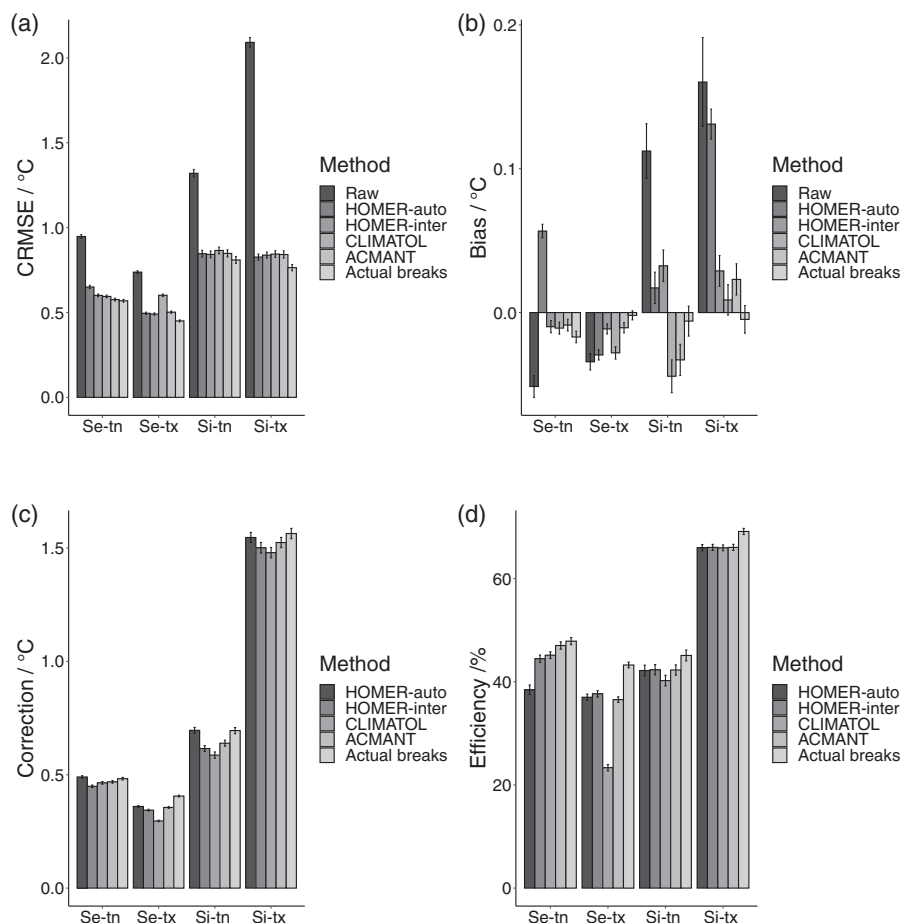
The ranking of the methods efficiency varies between the cases, but on average ACMANT and HOMER-inter has the highest efficiency ( $48.0 \pm 0.04\%$  and  $47.6 \pm 0.04\%$ , respectively), followed by HOMER-auto ( $45.9 \pm 0.04\%$ ). Climatol has the lowest efficiency in all cases except for Swedish  $T_{\min}$ , albeit not always significantly within the 95% confidence interval. Climatol's average efficiency of  $43.7 \pm 0.04\%$  includes the lowest efficiency value of all methods and cases (the Swedish  $T_{\max}$  case:  $23.3 \pm 0.06\%$ ).

For a measure of the distribution of error beyond the averages, the integer absolute error percentiles between 0 and 100 for the methods are plotted against the corresponding integer absolute error percentiles of the raw data for the four cases in Figure 6. The absolute error is defined as the absolute differences between the homogenized and the true values. None of the methods could perfectly recreate the true data. The absolute difference in error percentiles of all methods are smaller than  $0.3^\circ\text{C}$  for all percentiles below, and including, the 99th percentile.

POD, POH,  $\text{POD}_{\text{large}}$  and the number of reported breaks relative to the number of true breaks are presented as barplots in Figure 7 and tabulated in Table 4. POD and POH as a function of relative number of reported breaks are plotted as scatterplots in Figure 8.

The methods group in a pair reporting less homogeneity breaks (Climatol and HOMER-inter), and a pair which reports more (ACMANT and HOMER-auto). Furthermore, it can be concluded that higher number of reported breaks raise POD, but lower POH, see Figure 8. A method that reports more breaks probably finds both more true breaks and more false breaks. Climatol consistently reports both the lowest number of breaks and the highest POH (on average  $0.95 \pm 0.02$ ). ACMANT

**FIGURE 5** The time series based metrics in the homogenization of the monthly averages of the daily maximum (tx) and minimum temperatures (tn) of the Slovenian (Si) and the Swedish (Se) Indecis benchmark network's data sets by HOMER-auto and HOMER-inter, Climatol and ACMANT. Also included is the HOMER correction function applied on the actual breaks. The uncertainty ranges are the 95% confidence interval calculated with the bootstrap method. Only nonmissing data are included. (a) CRMSE; (b) bias; (c) Absolute correction; and (d) Efficiency



consistently reports the highest POD ( $0.74 \pm 0.03$ ), and in two of the four cases, also the highest number of breaks. HOMER-auto has the lowest POH ( $0.53 \pm 0.03$ ) and reports on average twice the number of breaks reported by Climatol and similar number of breaks as ACMANT for the four cases. HOMER-inter has the second highest POH ( $0.75 \pm 0.04$ ). HOMER-inter also has the second highest POD ( $0.62 \pm 0.04$ ) together with HOMER-auto ( $0.64 \pm 0.04$ ).

The extent of correction follows the number of breaks reported: Climatol has the smallest average correction and the smallest total number of breaks over the four cases (387 breaks, not shown), followed by HOMER-inter which has the second smallest average correction and the second smallest number of total breaks (570). ACMANT, which has the second largest average correction, has the second highest total number of breaks (726). HOMER-auto which has the largest average correction also has the largest number of total breaks (773).

The pattern is similar for the large breaks: ACMANT has the highest  $POD_{large}$  ( $0.85 \pm 0.05$ ), followed by HOMER-auto and HOMER-inter ( $0.76 \pm 0.06$  and  $0.74 \pm 0.06$ , respectively). Climatol has the lowest  $POD_{large}$  ( $0.67 \pm 0.06$ ). In relative terms, the methods

finds equal number of large breaks within the confidence intervals.

## 4.2 | Reference selection

The number of actual breaks reported by the operational method and the constant method is presented in Figure 9a. The operational method reports more breaks than the number of actual breaks for series with up to two actual breaks. The constant method reports no breaks for series with no actual break and on average one break for series with one actual break. On average, the operational method reports more breaks than the constant method for series with any number of actual breaks.

The average relative number of breaks (reported breaks per actual break) in series binned according to the number of references (in the operational method) is presented in Figure 9b for the methods applied on the Swedish  $T_{max}$  case. On average, the operational method reports more than two breaks per actual break in series with more than 35 references, whereas in the series with eight references the operational method reports less than

Method	CRMSE/0.1°C	Bias/0.1°C	Corr/0.1°C	E/%
<i>T<sub>max</sub></i> , Slovenia				
Raw	20.9 ± 0.3	1.6 ± 0.3		
HOMER-auto	8.3 ± 0.2	1.3 ± 0.1	15.5 ± 0.2	66.0 ± 0.6
HOMER-inter	8.4 ± 0.2	0.3 ± 0.1	15.0 ± 0.2	66.0 ± 0.6
CLIMATOL	8.4 ± 0.2	0.1 ± 0.1	14.8 ± 0.2	65.9 ± 0.6
ACMANT	8.4 ± 0.2	0.2 ± 0.1	15.2 ± 0.2	66.0 ± 0.6
True-breaks	7.6 ± 0.2	-0.0 ± 0.1	15.6 ± 0.2	69.1 ± 0.6
<i>T<sub>max</sub></i> , Sweden				
Raw	7.4 ± 0.1	-0.3 ± 0.1		
HOMER-auto	5.0 ± 0.1	-0.3 ± 0.0	3.6 ± 0.0	37.0 ± 0.6
HOMER-inter	4.9 ± 0.1	-0.1 ± 0.0	3.4 ± 0.0	37.7 ± 0.6
CLIMATOL	6.0 ± 0.1	-0.3 ± 0.0	3.0 ± 0.0	23.3 ± 0.6
ACMANT	5.0 ± 0.1	-0.1 ± 0.0	3.6 ± 0.0	36.5 ± 0.5
True-breaks	4.5 ± 0.1	-0.0 ± 0.0	4.1 ± 0.0	43.2 ± 0.6
<i>T<sub>min</sub></i> , Slovenia				
Raw	13.2 ± 0.2	1.1 ± 0.2		
HOMER-auto	8.5 ± 0.2	0.2 ± 0.1	7.0 ± 0.1	42.2 ± 1.0
HOMER-inter	8.4 ± 0.2	0.3 ± 0.1	6.2 ± 0.1	42.3 ± 1.0
CLIMATOL	8.7 ± 0.2	-0.4 ± 0.1	5.9 ± 0.1	40.2 ± 1.0
ACMANT	8.5 ± 0.2	-0.3 ± 0.1	6.4 ± 0.1	42.3 ± 1.0
True-breaks	8.1 ± 0.2	-0.1 ± 0.1	6.9 ± 0.1	45.1 ± 1.1
<i>T<sub>min</sub></i> , Sweden				
Raw	9.5 ± 0.1	-0.5 ± 0.1		
HOMER-auto	6.5 ± 0.1	0.6 ± 0.0	4.9 ± 0.1	38.4 ± 0.9
HOMER-inter	6.0 ± 0.1	-0.1 ± 0.0	4.5 ± 0.1	44.5 ± 0.7
CLIMATOL	5.9 ± 0.1	-0.1 ± 0.0	4.6 ± 0.1	45.1 ± 0.6
ACMANT	5.8 ± 0.1	-0.1 ± 0.0	4.7 ± 0.1	47.0 ± 0.7
True-breaks	5.7 ± 0.1	-0.2 ± 0.0	4.8 ± 0.1	47.9 ± 0.7

Note: Also included is the HOMER correction function applied on the actual breaks. The uncertainty ranges are the 95% confidence interval calculated with the bootstrap method. Only non-missing data are included.

one break per actual break. In the constant method the same series have about the same number of reported breaks per actual break.

The error reduction, correction, POH and POD versus the number of actual breaks in each station of the Swedish *T<sub>max</sub>* data set as homogenized by the operational and constant methods are presented as bar plots in Figure 10. The error reduction and correction are not significantly different between the two methods for series with any number of actual breaks, except for the series with no actual break. The operational method has negative, significantly nonzero, error reduction for the series with no actual break and significant nonzero correction, while the constant method has no error reduction or correction. The operational method shows decreasing POH for series with decreasing number of actual breaks.

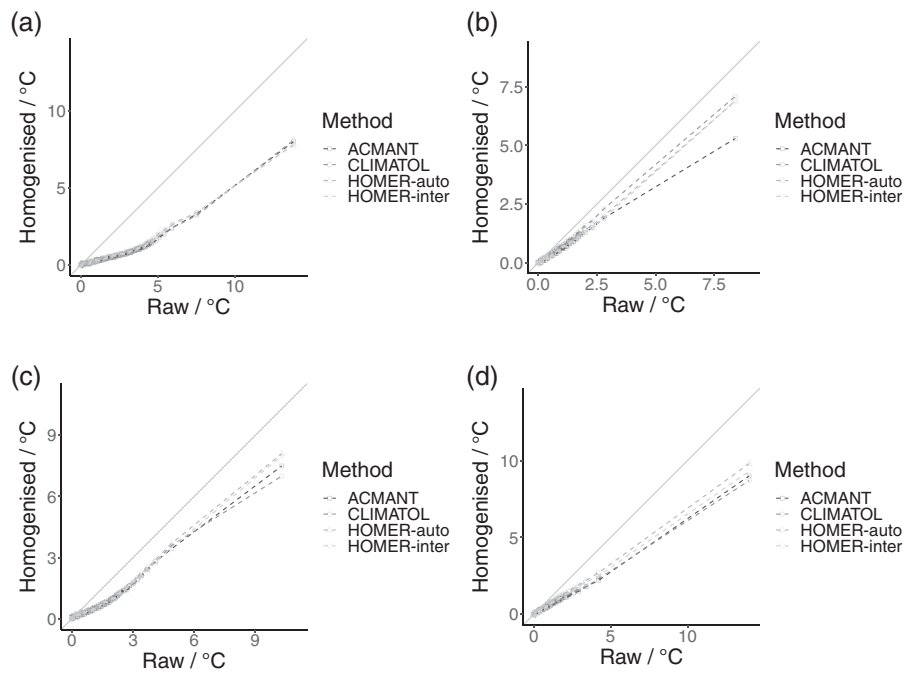
The constant method shows no significant trend of POH for series with decreasing number of actual breaks. For series with more than three actual breaks, there is no significant difference between the methods. POD is higher for the operational method for series with most number of actual breaks, albeit not significantly within the 95% confidence interval.

### 4.3 | Number of iterations in HOMER

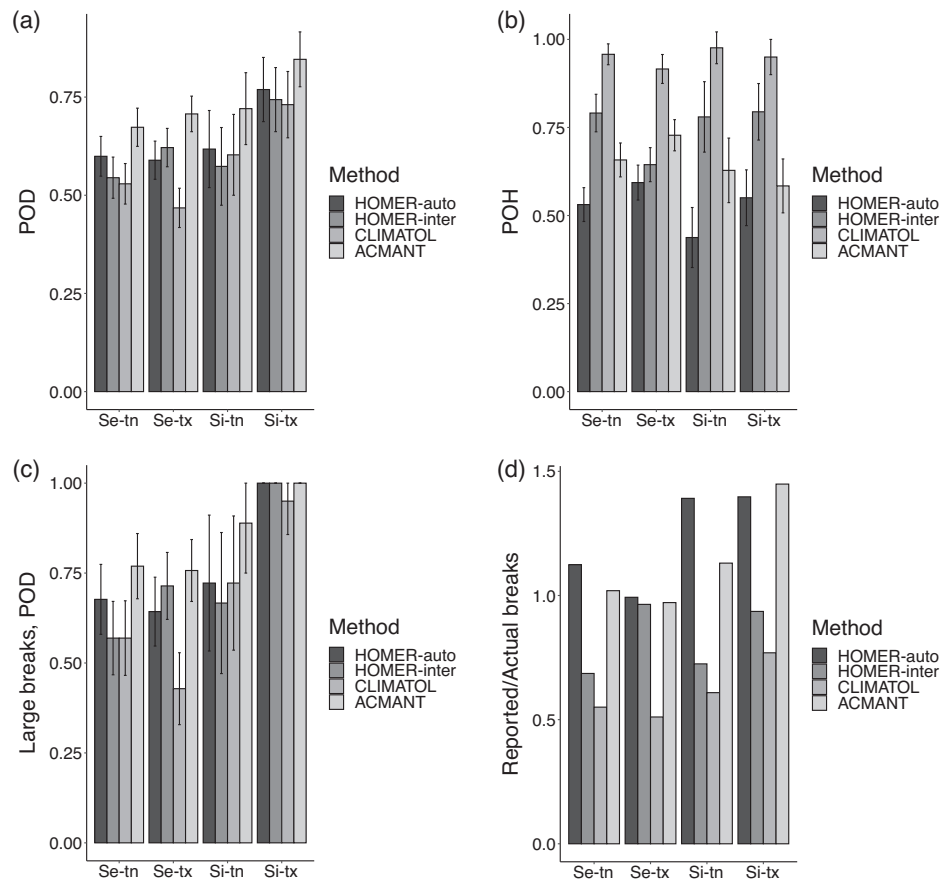
The methods' skill scores after each correction in the homogenization processes with and without the convergence option are plotted in Figure 11. The applications of the change month function in the end of the process are indicated in the figures.

TABLE 3 CRMSE, biases, mean absolute corrections and efficiency in the homogenization of the monthly averages of the daily maximum (tx) and minimum temperatures (tn) of the Slovenian (Si) and the Swedish (Se) Indecis benchmark network's data sets by HOMER-auto, HOMER-inter, Climatol and ACMANT

**FIGURE 6** Q–Q plot of the integer absolute error percentiles (0–100) of the homogenized monthly mean of daily maximum and minimum temperatures data of the Slovenian and Swedish Indecis data sets by HOMER-auto, HOMER-inter, Climatol and ACMANT against the integer raw error percentiles (0–100), no gap filled data included, the grey line indicate the 1:1 line (raw error percentiles), the points are connected by dashed lines. (a) The daily maximum temperature, Slovenian networks. (b) The daily maximum temperature, Swedish network. (c) The daily minimum temperature, Slovenian network. (d) The daily minimum temperature, Swedish network



**FIGURE 7** The skill scores in the homogenization of the monthly averages of the daily maximum (tx) and minimum temperatures (tn) of the Slovenian (Si) and the Swedish (Se) Indecis benchmark network’s data sets by HOMER-auto, HOMER-inter, Climatol and ACMANT. The uncertainty ranges are the 95% confidence interval calculated with the bootstrap method. (a) POD; (b) POH; (c)  $POD_{large}$ ; (d) number of reported breaks relative to the number of actual breaks



Method	POH	POD	POD <sub>large</sub>	Brks
<i>T</i> <sub>max</sub> , Slovenia				
HOMER-auto	0.55 ± 0.08	0.77 ± 0.08	1.00 ± 0.00	109
HOMER-inter	0.79 ± 0.08	0.74 ± 0.08	1.00 ± 0.00	73
CLIMATOL	0.95 ± 0.05	0.73 ± 0.08	0.95 ± 0.09	60
ACMANT	0.58 ± 0.08	0.85 ± 0.07	1.00 ± 0.00	113
<i>T</i> <sub>max</sub> , Sweden				
HOMER-auto	0.59 ± 0.05	0.59 ± 0.05	0.64 ± 0.10	278
HOMER-inter	0.64 ± 0.05	0.62 ± 0.05	0.71 ± 0.09	270
CLIMATOL	0.92 ± 0.04	0.47 ± 0.05	0.43 ± 0.10	143
ACMANT	0.73 ± 0.04	0.71 ± 0.05	0.76 ± 0.09	272
<i>T</i> <sub>min</sub> , Slovenia				
HOMER-auto	0.44 ± 0.09	0.62 ± 0.10	0.72 ± 0.19	96
HOMER-inter	0.78 ± 0.10	0.57 ± 0.10	0.67 ± 0.20	50
CLIMATOL	0.98 ± 0.05	0.60 ± 0.10	0.72 ± 0.19	42
ACMANT	0.63 ± 0.09	0.72 ± 0.09	0.89 ± 0.14	78
<i>T</i> <sub>min</sub> , Sweden				
HOMER-auto	0.53 ± 0.05	0.60 ± 0.05	0.68 ± 0.10	290
HOMER-inter	0.79 ± 0.05	0.54 ± 0.05	0.57 ± 0.10	177
CLIMATOL	0.96 ± 0.03	0.53 ± 0.05	0.57 ± 0.10	142
ACMANT	0.66 ± 0.05	0.67 ± 0.05	0.77 ± 0.09	263

Note: The uncertainty ranges are the 95% confidence interval calculated with the bootstrap method.

TABLE 4 POH, POD, the probability of detection of the breaks in upper quartile (POD<sub>large</sub>) and the number of reported breaks (Brks) in the homogenization of the monthly averages of the daily maximum (tx) and minimum temperatures (tn) of the Slovenian (Si) and the Swedish (Se) Indecis benchmark network's data sets by HOMER-auto, HOMER-inter, Climatol and ACMANT

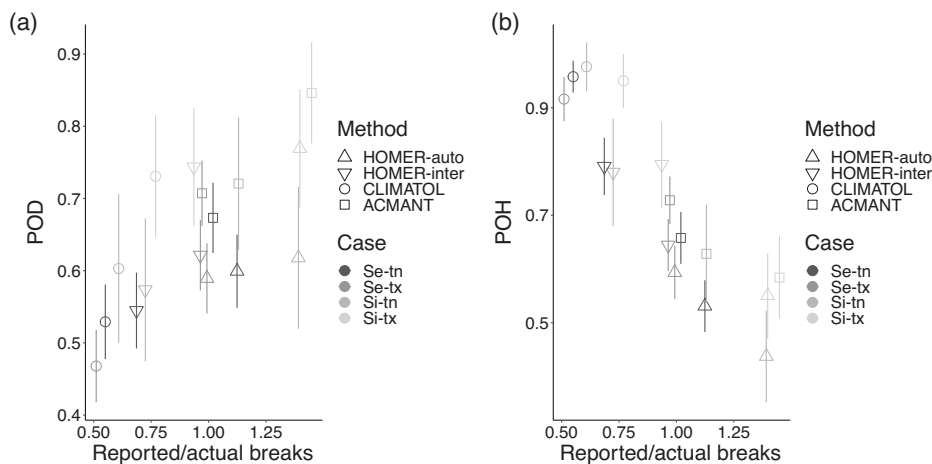


FIGURE 8 The skill scores against the relative number of breaks reported in the homogenization of the monthly averages of the daily maximum (tx) and minimum temperatures (tn) of the Slovenian (Si) and the Swedish (Se) Indecis benchmark network's data sets by HOMER-auto, HOMER-inter, Climatol and ACMANT. The uncertainty ranges are the 95% confidence interval calculated with the bootstrap method. (a) POD; (b) POH

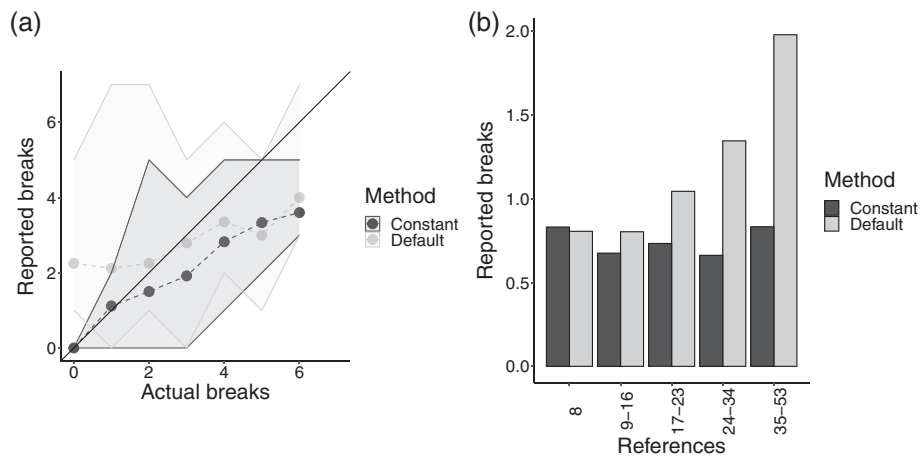
The differences between the operational and convergence run are not statistically significant within the 95% confidence interval. POD and number of reported breaks increase when the process is allowed to converge. POH, on the other hand, decrease. POD usually increases more than POH decreases. The change month function increases POH and affects only POD in one case. The effect of the change month function is indicated by the difference between the dashed and the solid lines at correction round 4, where the function is applied on the operational run, but not the convergence run.

## 5 | DISCUSSION

### 5.1 | Data and tools

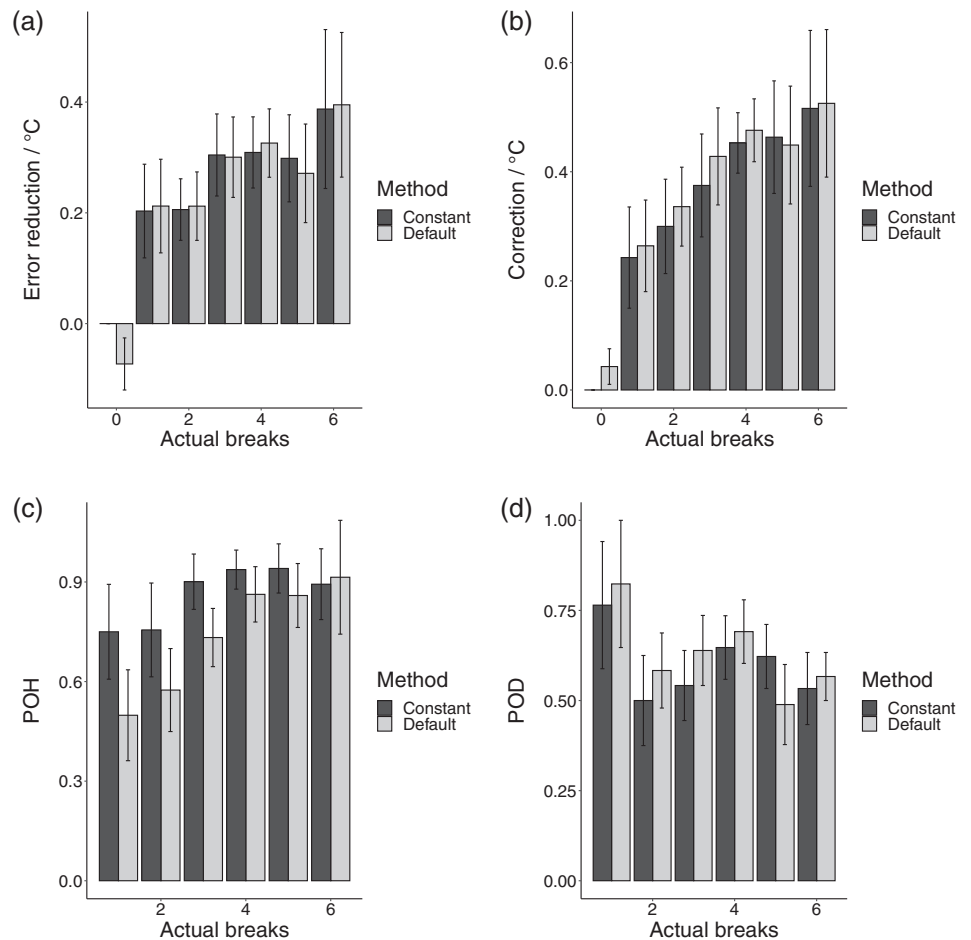
#### 5.1.1 | Benchmark data set

The usefulness of synthetic benchmark data sets for the evaluation of homogenization methods is an open question. It is, for example, uncertain how closely the synthetic homogeneity breaks resemble



**FIGURE 9** The number of reported breaks versus the number of actual breaks and number of references in the average daily maximum temperatures data of the Swedish Indecis data set homogenized by two versions of the interactive mode of HOMER: Default reference selection setting and the eight most highly correlated series as references (constant), in the left panel the shaded area indicate the full range of the number of reported breaks, the black line indicate the 1:1 line. (a) Average number of actual breaks. (b) Number of references

**FIGURE 10** Error reduction (a), mean absolute correction (b), POH (c) and POD (d) for each station versus the number of actual breaks in each series of the monthly average daily maximum temperatures data of the Swedish Indecis data set homogenized by two versions of the interactive mode of HOMER: Default reference selection setting and the eight most highly correlated series as references (constant), the lines indicate the 95% confidence interval



real ones. Recent studies have used benchmark data sets based on real observational data (e.g., Squintu *et al.*, 2020). The existence of true homogeneous solutions makes synthetic benchmark data sets to still be a good option.

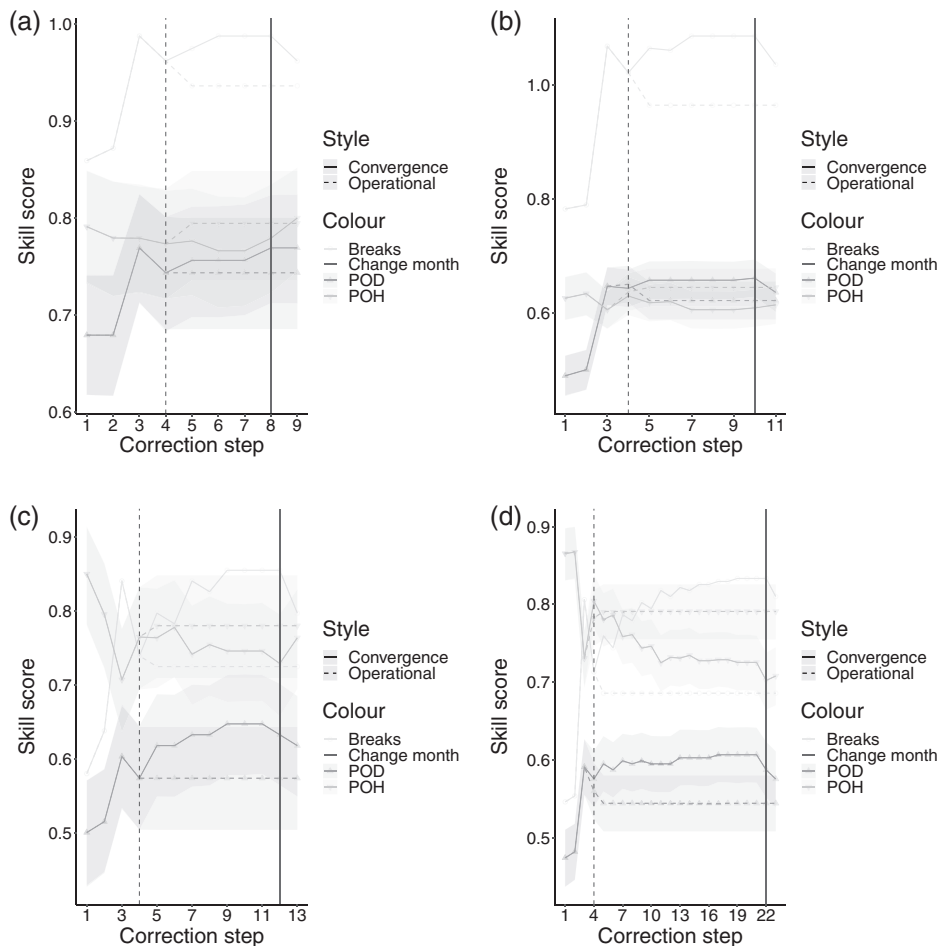
The results are, however, not necessarily directly transferable to observational data sets. The characteristics of the benchmark data sets and the observational data set must be considered: Spatial density and the fraction of missing data are two important characteristics. The temporal heterogeneity of data availability is another important characteristic (i.e., how much of the data that is missing early or late in the series and how long typical gaps are).

Any possible influence of differences of climate in Sweden and Slovenia is probably subdued with the use of annual data. There is no correlation between variance of the annual averaged data and the detection skill. Pearson's correlation coefficients between skill scores and variance on station level are close to zero ( $\rho^{\text{POD}} = 0.05$  and  $\rho^{\text{POH}} = -0.1$ ) for all the cases combined.

## 5.2 | Test results

### 5.2.1 | Method comparison

On average, the four methods produce similar homogenized data series. The average CRMSE of all the homogenized series are small. The difference between ACMANT and Climatol CRMSE is smaller than in the results of Venema *et al.* (2012). Especially ACMANT has a higher CRMSE in the results of the current study than in Venema *et al.* (2012). Our results compares qualitatively well with Guijarro *et al.* (2017) who reported small differences in RMSE between automatic HOMER, Climatol and ACMANT for tests resembling the Indecis benchmark data sets. Gubler *et al.* (2017) found in their study interactive HOMER to have lower CRMSE than automatic HOMER both for a dense and a sparse network of Swiss observational time series. The implications of these results for the results of the current study are however difficult to evaluate since interactive HOMER applied on the dense network, with metadata support, was defined as the true solution.



**FIGURE 11** POH, POD and number of breaks found as a fraction of the number of actual breaks (breaks) of the homogenized mean of daily maximum and minimum temperatures data of the Slovenian and Swedish Indecis data sets at each point of correction for HOMER-inter with the default settings (two full homogenization rounds) and the convergence option turned on, the shaded areas present the 95% confidence interval. Note that each homogenization round includes two correction steps. The point of the change month function is indicated with black vertical lines. (a) The daily maximum temperature, Slovenian network. (b) The daily maximum temperature, Swedish network. (c) The daily minimum temperature, Slovenian network. (d) The daily minimum temperature, Swedish network

For most cases, most of the methods successfully correct also the time series with the highest residual errors (see Figure 6). The correction of the methods correlates with the error in the raw data: large error in the raw data results in large corrections. Also the measure of efficiency follows the pattern of error in the raw data for the different cases; methods reduce the error more effectively in cases where the raw data error is large.

POD follows the number of reported breaks in most cases, while POH decreases for increasing number of breaks. POD values are higher than in the results of Venema *et al.* (2012) for both Climatol and ACMANT. However, the definition of true positive (i.e., the tolerance time) is not explicitly stated in Venema *et al.* (2012). Other studies have used shorter tolerance time than what is used in the current study (Vertačnik *et al.*, 2015; Van Malderen *et al.*, 2020).

Similar to the results of the current study, Pérez-Zanón *et al.* (2015) concluded that ACMANT, compared with the interactive mode of HOMER, detected more homogeneity breaks in 90% of the 44 temperature time series used in their study. On average, ACMANT detected about 50% more breaks than HOMER. In the current study, ACMANT detects on average 40% more breaks than HOMER-inter. Furthermore, Pérez-Zanón *et al.* (2015) concluded that HOMER detected more of the homogeneity breaks listed in metadata than ACMANT. The average POD (based on metadata breaks) of HOMER was 0.82, as defined by Equation (8). ACMANT's average POD was 0.35.

In Vertačnik *et al.* (2015), HOMER in automatic mode generally found more homogeneity breaks than HOMER in interactive mode, which agrees well with the results of the current study. About two-thirds of the breaks detected by HOMER in automatic mode was supported by metadata, corresponding to a POD (based on metadata breaks) between 0.6 and 0.7. The runs with HOMER in interactive mode resulted in POD-values ranging from 0.86 to 1.0.

The high POD for HOMER in interactive mode relative to ACMANT in Pérez-Zanón *et al.* (2015) and HOMER in automatic mode in Vertačnik *et al.* (2015) compared with the results of the current study can possibly be explained by the fact that the operator of HOMER in both studies had a priori knowledge of the metadata breaks. In Bart, there is a possibility to include information from metadata, but it is not utilized in the current study. Moreover, a metadata break does not necessarily cause a homogeneity break, which might subdue POD of the automatic methods.

HOMER-inter reports more breaks in  $T_{\max}$  data than in  $T_{\min}$  data. This is reflected in POD and POH skill scores. The difference is not as pronounced in the other

methods, though all methods report more breaks for  $T_{\max}$  than  $T_{\min}$  in the Slovenian data. If this is a systematic problem in HOMER-inter is not possible to determine from the current results alone.

The time series based metrics vary more between the cases than between the methods, while the skill scores vary significantly between the methods (see Figure 7). HOMER-inter can be considered quite close to the upper limit of the skill of HOMER methods, such that any further substantial average error reduction requires more than tuning the detection functions parameters. The skill can probably be enhanced in specific situations.

The evaluation of homogenization methods should not be limited to the ability to minimize the error of the raw data. Homogenization can be considered to be an asymmetric problem: It is most desirable to have the data correctly homogenized, but if this is not achieved, it is more desirable to keep the data unmodified than to have a homogenization that is even slightly false. Similarly, if data is improved on average, it still can be deteriorated locally. Modifications that bring the data further from the true values cannot simply be offset by modifications that bring the data closer to the true values somewhere else. When correcting observational data, the true values are not known. All modifications of data might bring the data further from the true values, and hence risk of adding noise to the data. The risk is larger when extensive modifications are done. Therefore can a conservative method (such as Climatol) arguably be preferred over a method more prone to modify the data (such as ACMANT), even if it's residual error in the homogenization of benchmark data is larger.

## 5.2.2 | Reference selection

The traditional use of HOMER includes the use of a correlation based criterion for selection of reference series. The number of references can vary substantially between the series within a network. Figure 4b reveals a connection between the number of references and the number of actual breaks. There is a risk of overhomogenizing relatively clean series, since there are more reported breaks than actual breaks for the homogenization of stations with many references, see Figures 9 and 10. When the number of references are set to be constant, POH and POD are less sensitive to the number of actual breaks in the data.

There is no significant effect of varying the number of references on the error reduction and correction, except for the series with no actual breaks. For most reference bins, the use of a constant number of references does not discard vital information, since the error reduction and

POD is not significantly higher in the operational method. The error in the series with higher number of references is not reduced significantly more than in series with fewer references, contrary to what is expected. The error reduction instead drops for higher number of references, indicating that the effect of higher correlation between relatively clean stations dominates over the detection functions benefiting from the use of more references. This is also supported by the decreasing correction for higher number of references and the small differences between the selection methods.

It remains unclear why also POD drops for increasing number of references. An option introduced in the Bart script is to let the pairwise detection function normalize the number of references for each candidate series, where the number of references exceeds a certain threshold. Tests conducted with this option (not shown) gives on average equivalent results to the operational method. However, for series with a large number of references, the too generous confirmation of breaks in the operational method is overcompensated, such that POD drops more than in the operational and constant methods.

Another option is to use a proximity-based criterion for the selection of references (Coll *et al.*, 2018), where all stations within a certain radius are selected as references. Tests conducted show the average CRMSE to be slightly reduced, especially for the Slovenian network (not shown). The skill scores are, on the other hand, slightly decreased. Reference selection with the proximity-based criterion requires the network to be climatologically homogeneous. It might not be suitable for networks including, for example, mountainous or archipelagic regions. The optimal settings for the selection of references need to be further investigated.

## 6 | CONCLUSION

A new automated version of the well-established homogenization tool HOMER has been developed. As opposed to the standard automatic version of HOMER, which only uses a part of the HOMER tool, the new version uses the full HOMER tool and can be customized to the user's needs. This development simplifies operational use and enables systematic testing of the full HOMER tool. Homogenization methods using the new automated version of HOMER has been evaluated by assessing different datasets for different variables. The methods represents the interactive mode (HOMER-inter) and the automatic mode of HOMER (HOMER-auto). Results have also been compared with results of other homogenization methods (using the tools Climatol and ACMANT).

The most important findings are:

- HOMER-inter produces homogenized time series with lower (or comparable) residual error compared with HOMER-auto, but with less modification of the input raw time series.
- The overall performance of HOMER-inter, compares well with Climatol and ACMANT.
- All the homogenization methods produce homogenized time series with significantly smaller deviation from the true data values compared with the raw time series.
- Regarding modification of data, HOMER-inter can be seen as a compromise between ACMANT, that modifies the data more, and Climatol, which is more conservative.
- The number of references and the reference selection method should be carefully considered when using HOMER.
- The practice of running two homogenizations rounds with HOMER is supported in this study as the probability of detection is not significantly increased with further rounds.

The results of the current study suggest that the choice of homogenization method is a matter of preference. It should, however, be stressed that both the new automation of HOMER and Climatol has extensive possibilities to change parameters which very well might change their respectively profiles. Such investigations are beyond the scope of this study.

## 7 | OUTLOOK

The Bart-script is still under development. A number of new features (including parallel computing and an improved reference selection function) are implemented in versions following the version used in the current study.

Homogenization of the Swedish observational network's monthly average temperature data set is currently ongoing. The results will constitute a future publication, including a comparison with the previous homogenization performed with SNHT.

### ORCID

L. Magnus T. Joelsson  <https://orcid.org/0000-0002-0287-4962>

Erik Engström  <https://orcid.org/0000-0002-6207-6460>

Erik Kjellström  <https://orcid.org/0000-0002-6495-1038>

### REFERENCES

Alexandersson, H. and Moberg, A. (1997) Homogenization of Swedish temperature data. Part i: homogeneity test for linear trends.

- International Journal of Climatology: A Journal of the Royal Meteorological Society*, 17, 25–34.
- Bellman, R. (1954) The theory of dynamic programming. *Bulletin of the American Mathematical Society*, 60, 503–515.
- Caussinus, H. and Lyazrhi, F. (1997) Choosing a linear model with a random number of change-points and outliers. *Annals of the Institute of Statistical Mathematics*, 49, 761–775.
- 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.
- Clarke, M. (1980) The reduced major axis of a bivariate sample. *Biometrika*, 67, 441–446.
- Coll, J., Curley, C., Walsh, S. and Sweeney, J. (2014) Ireland with homer. In Proceedings of the 8th seminar for homogenisation and quality control in climatological databases and 3rd conference on spatial interpolation in climatology and meteorology, climate data and monitoring WCDMP no. vol. 84, pp. 23–45.
- Coll, J., Curley, M., Walsh, S. and Sweeney, J. (2018) Homerun: relative homogenisation of the irish precipitation network. EPA Research Report 2012-CCRP-FS.11 Report.
- De Laat, A. and Crok, M. (2013) A late 20th century european climate shift: fingerprint of regional brightening? *Atmospheric and Climate Sciences*, 3, 291–300.
- Domonkos, P. (2013a) Efficiencies of inhomogeneity-detection algorithms: comparison of different detection methods and efficiency measures. *Journal of Climatology*, 2013, 1–15.
- Domonkos, P. (2013b) Measuring performances of homogenization methods. *Quarterly Journal of the Hungarian Meteorological Service*, 117, 91–112.
- Domonkos, P. (2019) ACMANT homogenization software: manual v4.3. 1–20.
- Domonkos, P., Sigró, J. and Poza, R. (2011) Adapted Caussinus-Mestre algorithm for networks of temperature series (acmant). *International Journal of Geosciences*, 2, 293–309.
- Efron, B. (1992) Bootstrap methods: another look at the jackknife. In: Kotz, S. and Johnson, N.L. (Eds.) *Breakthroughs in Statistics*. New York, NY: Springer, pp. 569–593.
- Fisher, W.D. (1958) On grouping for maximum homogeneity. *Journal of the American Statistical Association*, 53, 789–798.
- Gubler, S., Hunziker, S., Begert, M., Croci-Maspoli, M., Konzelmann, T., Brönnimann, S., Schwierz, C., Oria, C. and Rosas, G. (2017) The influence of station density on climate data homogenization. *International Journal of Climatology*, 37, 4670–4683.
- Guijarro, J. A. (2018) Homogenization of climatic series with climatol. Version 3.1.1. Available at: <https://CRAN.R-project.org/package=climatol> [Accessed 28th September 2021]
- Guijarro, J. A. (2019) Recommended Homogenization Techniques Based on Benchmarking Results.
- Guijarro, J. A., López Díaz, J. A., Aguilar, E., Domonkos, P., Venema, V. K., Sigró, J. and Brunet, M. (2017) Comparison of homogenization packages applied to monthly series of temperature and precipitation: the MULTITEST project. In 9th Seminar for homogenization and quality control in climatological databases and 4th conference on spatial interpolation techniques in climatology and meteorology (Budapest, 3–7 April 2017).
- Heidke, P. (1926) Berechnung des erfolges und der güte der windstärkevorhersagen im sturmwarnungsdienst. *Geografiska Annaler*, 8, 301–349.
- Indecis Project (2017) Homogenisation Benchmark [homepage on the Internet]. Available from: <http://www.indecis.eu/benchmarking.php> [Accessed 12th August 2020].
- Kircher, J. (2001) Data analysis toolkit# 5: Uncertainty analysis and error propagation. University of California Berkeley Seismological Laboratory. Available from: [http://seismo.berkeley.edu/kirchner/eps\\_120/Toolkits/Toolkit\\_05.pdf](http://seismo.berkeley.edu/kirchner/eps_120/Toolkits/Toolkit_05.pdf).
- Kivinen, S., Rasmus, S., Jylhä, K. and Laapas, M. (2017) Long-term climate trends and extreme events in northern fennoscandia (1914–2013). *Climate*, 5, 16.
- Kuya, E.K., Gjeltén, H.M. and Tveito, O.E. (2020) Homogenization of Norway's mean monthly. *Network*, 2, 1–95.
- Lindau, R. and Venema, V. (2018) On the reduction of trend errors by the anova joint correction scheme used in homogenization of climate station records. *International Journal of Climatology*, 38, 5255–5271.
- Menne, M.J. and Williams, C.N., Jr. (2005) Detection of undocumented change-points using multiple test statistics and composite reference series. *Journal of Climate*, 18, 4271–4286.
- Menne, M.J. and Williams, C.N., Jr. (2009) Homogenization of temperature series via pairwise comparisons. *Journal of Climate*, 22, 1700–1717.
- Mestre, O. and Aguilar, E. (2011) HOME\_R. Fast documentation, HOMER Training School [homepage on the Internet]. Available from: [http://www.c3.urv.cat/data/HOME\\_R.pdf](http://www.c3.urv.cat/data/HOME_R.pdf) [Accessed 16th June 16].
- Mestre, O., Domonkos, P., Picard, F., Auer, I., Robin, S., Lebarbier, E., Böhm, R., Aguilar, E., Guijarro, J.A., Vertacnik, G., et al. (2013) Homer: a homogenization software-methods and applications. *Időjárás*, 117, 47–67.
- Morán-Tejeda, E., Bazo, J., López-Moreno, J.I., Aguilar, E., Azorín-Molina, C., Sanchez-Lorenzo, A., Martínez, R., Nieto, J.J., Mejía, R., Martín-Hernández, N. and Vicente-Serrano, S.M. (2016) Climate trends and variability in Ecuador (1966–2011). *International Journal of Climatology*, 36, 3839–3855.
- Peirce, C.S. (1884) The numerical measure of the success of predictions. *Science*, 4, 453–454.
- Pérez-Zanón, N., Sigró, J., Domonkos, P. and Ashcroft, L. (2015) Comparison of homer and acmant homogenization methods using a central pyrenees temperature dataset. *Advances in Science and Research*, 12, 111–119.
- Picard, F., Lebarbier, E., Hoebeke, M., Rigail, G., Thiam, B. and Robin, S. (2011) Joint segmentation, calling, and normalization of multiple cgh profiles. *Biostatistics*, 12, 413–428.
- Skrynyk, O., Aguilar, E., Guijarro, J. A. and Bubín, S. (2020) Uncertainty of climatol adjustment algorithm for daily time series of additive climate variables. In *EGU General Assembly Conference Abstracts*, 5365.
- Squintu, A.A., van der Schrier, G., Štěpánek, P., Zahradníček, P. and Tank, A.K. (2020) Comparison of homogenization methods for daily temperature series against an observation-based benchmark dataset. *Theoretical and Applied Climatology*, 140, 1–17.
- Van Malderen, R., Pottiaux, E., Klos, A., Domonkos, P., Elias, M., Ning, T., Bock, O., Guijarro, J., Alshawaf, F., Hoseini, M., et al. (2020) Homogenizing gps integrated water vapor time series: benchmarking break detection methods on synthetic data sets. *Earth and Space Science*, 7, EA001121.
- Venema, V., Trewin, B., Wang, X., Szentimrey, T., Lakatos, M., Aguilar, E., Auer, I., Guijarro, J., Menne, M., Oria, C., Louamba, W. and Rasul, G. (2020) *Guidelines on Homogenization, 2020 Edition*. Geneva: World Meteorological Organization.

- Venema, V.K., Mestre, O., Aguilar, E., Auer, I., Guijarro, J.A., Domonkos, P., Vertacnik, G., Szentimrey, T., Stepanek, P., Zahradnicek, P., et al. (2012) Benchmarking homogenization algorithms for monthly data. *Climate of the Past*, 8, 89–115.
- Vertačnik, G., Dolinar, M., Bertalanič, R., Klančar, M., Dvoršek, D. and Nadbath, M. (2015) Ensemble homogenization of slovenian monthly air temperature series. *International Journal of Climatology*, 35, 4015–4026.
- Vicente-Serrano, S.M., Aguilar, E., Martínez, R., Martín-Hernández, N., Azorin-Molina, C., Sánchez-Lorenzo, A., El Kenawy, A., Tomás-Burguera, M., Moran-Tejeda, E., López-Moreno, J.I., et al. (2017) The complex influence of enso on droughts in Ecuador. *Climate Dynamics*, 48, 405–427.
- World Meteorological Organization. (2017a) *WMO Guidelines on Generating a Defined Set of National Climate Monitoring Products*. Geneva, Switzerland: WMO.

World Meteorological Organization. (2017b) *WMO Guidelines on the Calculation of climate Normals*. Geneva, Switzerland: WMO.

## SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

**How to cite this article:** Joelsson, L. M. T., Sturm, C., Södling, J., Engström, E., & Kjellström, E. (2021). Automation and evaluation of the interactive homogenization tool HOMER. *International Journal of Climatology*, 1–20. <https://doi.org/10.1002/joc.7394>