Received: 5 September 2019 | Revised: 1 April 2020 | Accepted: 5 April 2020

DOI: 10.1002/joc.6597

RESEARCH ARTICLE

International Journal of Climatology RMetS

Analyzing the impact of automatization using parallel daily mean temperature series including breakpoint detection and homogenization

Lisa Hannak | **Karsten Friedrich** | **Florian Imbery** | **Frank Kaspar**

Deutscher Wetterdienst, National Climate Monitoring, Offenbach, Germany

Correspondence
Lisa Hannak, Deutscher Wetterdienst, National Climate Monitoring, Frankfurter Str. 135, 63067 Offenbach, Germany.
Email: lisa.hannak@dwd.de

Funding information
Deutscher Wetterdienst, Business Area Research and Development, research program "Innovation in Applied Research and Development" (IAFE)

Abstract

High-quality time series of meteorological observations are required for reliable assessments of climate trends. To analyze inhomogeneities in time series, parallel measurements can be used. Germany's national meteorological service DWD (Deutscher Wetterdienst) operates a network of climate reference stations. At these stations, manual and automatic observations have been taken in parallel. These parallel measurements therefore allow analyzing the impact of the transition on the homogeneity of time series of several meteorological parameters. Here, we present results for temperature. The differences between automatic and manual measurements are tested on breakpoints caused by instrumental defects or changes in the measurement conditions. The time series are highly correlated such that small breaks can be identified. The detected breakpoints are verified against metadata if available. In the case of no available metadata information, a procedure is suggested to identify the inhomogeneous time series (manual or automatic time series). Afterwards, the time series are homogenized. The homogenized time series are used to analyze the impact of changing the observing system from manual to automatic measurements on daily mean temperature.

KEYWORDS

automatization, breakpoint detection, climate observations, homogenization, parallel measurements, temperature series

1 | INTRODUCTION

Parallel measurements provide information on how changes in the observing system can affect time series. Furthermore, these measurements can determine uncertainties and can be used to control the quality of the data. If the behaviour of the differences changes significantly, the change can indicate a break in at least one time series and therefore a need for homogenization. Most homogenization methods require a reference series. In the case of parallel measurements, each of the time series can be used as reference series. Usually, parallel measurements are highly correlated which facilitates the breakpoint detection and homogenization.

Thermometer screens have an influence on the temperature measurements. Therefore, several studies have

been performed to compare thermometer screens (Brandsma and Van der Meulen, 2008; Brunet *et al*., 2011; Hoover and Yao, 2018) or changes from unscreened to screened measurement conditions (Böhm *et al*., 2010). The measurement arrangements or meteorological conditions such as wind speed, cloud cover, or radiation can influence temperature measurements. Large radiant flux can induce significant differences. The radiation effect on temperature measurements can be minimized by applying small-sized sensors (Erell *et al*., 2005). Auchmann and Brönnimann (2012) used parallel measurements to evaluate a physics-based correction model to homogenize temperature data. At German climate reference stations, manual and automatic measurement instruments are operated in parallel and can therefore be directly compared. Kaspar *et al*. (2016) analyzed temperature measurements with the result of only minor differences in the comparison of manual and automatic observations at the traditional observing times (06:30 UTC, 13:30 UTC and 20:30 UTC). The analysis of daily maximum temperature revealed an annual cycle in the time series of the differences with warmer automatically measured temperature maxima in summer at some stations. The main reason is a radiation effect on the shelter (LAM 630) used for automatic measurements. This error can be reduced by optimizing the position of the automatic instrument in the shelter (see Kaspar *et al*., 2016). Another reason for the annual cycle in the differences of daily maximum temperature in Germany is the different screen characteristics (e.g., shelter size and ventilation) between the modern and the historical screen. Parallel measurements of daily sunshine duration are analyzed in the study of Hannak *et al*. (2019) with the result of significant differences between manual and automatic daily sunshine duration measurements. To homogenize the daily sunshine duration data, a regression model (as introduced in their study) can be used to adapt the automatic measurements. Baciu *et al*. (2005) compared automatic and historical observations in Romania with the result of minor differences of daily mean temperature values but larger differences for daily minimum and maximum temperature values. Doesken (2005) analyzed the impact of automatization on temperature measurements at one station in the United States.

Usually, parallel measurements have a short temporal coverage. For this reason, nearby stations are often used to detect breaks and to homogenize the data ('called relative method'). In most cases, the correlation between nearby stations and the candidate time series is smaller than using parallel measurements such that small breaks are difficult to detect or to homogenize. Most homogenization procedures are applied on annual or monthly data like the HISTALP dataset (for the Alpine region). Their method includes relative homogeneity testing and metadata information (Auer *et al*., 2007). Monthly mean temperature and precipitation time series of Switzerland are homogenized by applying the software THOMAS (Begert *et al*., 2005). Israelian time series of temperature maxima and minima are homogenized by Yosef *et al*. (2018) and the homogenized data is used for trend analysis. Hannart *et al*. (2014) introduce a fully automatized breakpoint detection method using pairwise comparisons of the candidate series and neighbouring series, building groups of breakpoints and homogenize yearly time series in Argentina. In their study, the trends in long temperature series are stronger after homogenization. Peterson *et al*. (1998) introduce several breakpoint detection and homogenization methods used worldwide and discuss the limitation of homogenized data. An updated review of homogenization methods and breakpoint detection can be found in the study of Ribeiro *et al*. (2016). They conclude that relative methods (with reference series) are better than absolute methods and breakpoint detection methods which are able to detect multiple breakpoints are better than detection methods which can only detect one breakpoint and are run several times to detect multiple breakpoints.

Some studies focus on the homogenization of daily data. The breakpoint detection and homogenization of daily data is complicated by higher variability and autocorrelation compared to annual or monthly data. Breaks can affect the mean and higher-order moments which aggravates break detection and homogenization. To homogenize daily data the software SPLIne Daily HOMogenization (SPLIDHOM) can be used (coded in R [R Core Team, 2015]). SPLIDHOM uses an indirect nonlinear regression method which uses cubic smoothing splines and can adjust the mean and higher-order moments of the candidate series (Mestre *et al*., 2011). The method which is applied by Della-Marta and Wanner (2006) adjusts the mean and higher-order moments of daily temperature time series as well. Very similar to that Toreti *et al*. (2010) have enhanced that method to handle autocorrelation and uses an objective parameter estimation. Kuglitsch *et al*. (2009) homogenize daily maximum temperature series. In their study, the breaks are detected with nearby stations. To adjust the mean and higher-order moments of the candidate series a nonlinear regression method is used which requires a highly correlated reference series. Daily temperature data is homogenized by Hewaarachchi *et al*. (2017) using metadata information, a reference series and deals with the seasonal cycle and autocorrelation of the series. Lund *et al*. (2007) considered autocorrelation and periodic features in time series to detect breakpoints.

There also exist fully- or partly automatic homogenization software tools. The European project COST ES0601 (HOME) compared homogenization software tools. The software MASH, PRODIGE and ACMANT showed good results for temperature data. HOMER is a R-software combining features of several tested software tools and was developed after this project. It can be used with metadata in a semi-automatic mode and fully automatically (Mestre *et al*., 2013).

In this study, we use parallel measurements of temperature, aggregated to daily mean values, to detect breaks and to homogenize these time series. The parallel time series are highly correlated and can be used as reference series for each other (used for the breakpoint detection and homogenization step). Three different homogenization methods are compared to evaluate if they are able to homogenize the detected and identified breaks. The homogenized data are compared to the results of Kaspar *et al*. (2016). In the first part, the data and methods are introduced including the breakpoint detection, the identification of the inhomogeneous time series and the homogenization method. Afterwards the results of parallel measurements at 13 stations in Germany are summarized. The homogenized data is compared to the raw data in the next part. Finally the results are summarized.

2 | DATA AND METHODS

In Germany, historical measurements of air temperature were performed with a mercury-in-glass thermometer three times per day. Therefore, this setting is also used for manual measurements at climate reference stations (currently at 6:30 UTC, 13:30 UTC and 20:30 UTC). To calculate daily mean values these three observations are used with double weight on the evening value. The manual instrument is inside a wooden Stevenson screen. To directly compare daily mean temperature values of manual and automatic observations, the same equation was applied to the automatic measurements. Even though the temporal resolution of automatic measurements is higher, only values at 6:30 UTC, 13:30 UTC and 20:30 UTC are used for this comparison. The automatic instrument is a platinum resistance thermometer (PT100, manufacturer Ketterer). At most sites, the ventilated lamellar shelter 'LAM 630' (manufacturer Eigenbrodt) is used for automatic temperature instruments. Exceptions are the stations Brocken (at this station a shelter called 'Gießener Hütte' is used), Fichtelberg and Frankfurt airport (until October 2014) where the Stevenson screen is used for automatic and manual instruments. Figure 1 shows Frankfurt (airport) as one example of a German climate reference station. The geographical position of the stations and the time period of available parallel measurements are summarized in Table 1 (see Hannak *et al*., 2019). In Kaspar *et al*. (2016) more information about the instruments and characteristics of climate reference stations can be found.

FIGURE 1 Example of one climate reference station (station Frankfurt airport) [Colour figure can be viewed at wileyonlinelibrary.com]

To filter outliers and to control the data quality, differences greater than four times the pseudo standard deviations (SD) are excluded from both time series. After Lanzante (1996), the pseudo SD can be calculated by the interquartile range divided by 1.349. The pseudo SD is less influenced by outliers itself which is the reason for preferring the pseudo SD instead of the 'original' SD. For a Gaussian normal distribution, the pseudo SD and the SD are equal. This is a very strict outlier control but the results of the detection of breaks and the homogenization are improved by excluding outliers. The number of outliers is summarized in Table 1.

The breakpoints in the time series are compared to metadata information (modification history) of the instrument or shelter type. Examples of available metadata information are the date of a replacement or the date of a calibration.

2.1 | Detection of breaks

To detect breaks in time series, differences of automatic minus manual daily mean values (difference series) are used. The assumption is, that both time series have a similar climate signal, such that the difference series do not include climate features like annual cycle, trend, etc. The breakpoint detection is performed using the R-function 'uniseg'. The R-function 'uniseg' (part of the R package 'cghseg') was originally developed for

TABLE 1 Time range with parallel measurements; location and elevation (in meters), pairs of data, and number of outliers of each climate reference station (Hannak *et al.*, 2019)

WMO ID	Station name	Parallel measurements used	Latitude in degree	Longitude in degree	Elevation in m	Pairs of data	Number of outliers
10015	Helgoland	2006–2013	54.1750	7.8920	4	2,689	24
10035	Schleswig	2006–2017	54.5275	9.5486	43	4,006	152
10147	Hamburg Fuhlsbüttel	2008–2014	53.6332	9.9881	11	2,348	47
10379	Potsdam	2008–2017	52.3813	13.0622	81	3,193	9
10393	Lindenberg	2008–2017	52.2085	14.1180	98	3,324	43
10453	Brocken	2008–2017	51.7986	10.6183	1,134	3,361	57
10499	Görlitz	2008–2014	51.1622	14.9506	238	2,408	19
10501	Aachen	2008–2011	50.7827	6.0941	202	893	9
10505	Aachen-Orsbach	2011–2014	50.7982	6.0244	231	1,170	17
10578	Fichtelberg	2008–2014	50.4283	12.9535	1,213	2,400	28
10637	Frankfurt main (airport)	2008–2017	50.0259	8.5213	100	3,343	146
10929	Konstanz	2007–2012	47.6774	9.1901	443	1953	48
10962	Hohenpeißenberg	2008–2017	47.8009	11.0109	977	3,154	108

comparative genomic hybridization (CGH) data, but works for difference series of climate data as well (Picard *et al.*, 2016). The identification of the positions of breakpoints is based on a dynamic programming algorithm for joint segmentation and uses a maximum likelihood criterion to find the best number of segments and the best position of these breakpoints. More information about the method and algorithm can be found in Picard *et al.* (2011).

The first step in the breakpoint detection procedure is the calculation of *monthly* mean differences between automatic and manual measurements. Then, the R-function 'uniseg' is used to detect breaks in the *monthly* difference series. The next step is to use the *daily* difference series. Within a time range of plus/minus 2 months around the break detected using *monthly* data, 'uniseg' is used with *daily* data to get a more precise break date. If 'uniseg' is not able to detect a break in this time range using *daily* data, the break date based on the *monthly* data is used for further steps. Figure 2 shows an example for the results of the method 'uniseg' with *monthly* and *daily* data.

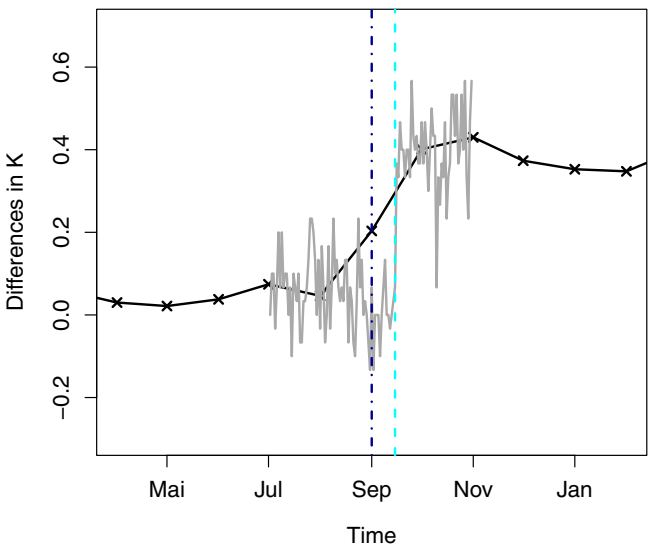

FIGURE 2 Example for the breakpoint detection with differences (automatic minus manual observations) of monthly data (black line) and daily data (grey line); The vertical lines represent the results of the breakpoint detection method (here the R-function 'uniseg') for different temporal resolution (daily: cyan line, monthly: blue line) [Colour figure can be viewed at wileyonlinelibrary.com]

2.2 | Identification of time series with breaks

Differences facilitate breakpoint detection but do not provide information about the time series responsible for the break in the difference series. To identify the inhomogeneous time series, different comparisons are made, for example, with metadata information or nearby stations.

For first comparison, metadata information of the manual and the automatic instrument is used. In a given

time range around the break date, metadata information for each instrument is counted ('metadata score'). The metadata information with the smallest time lag (in days) between the break date and the metadata information has an extra weight (+0.5 to the total 'metadata score'). The time range used for the comparison depends on the signal-to-noise ratio (SNR) and the probability to miss a break (Lindau and Venema, 2016). The SNR is calculated after SNR = $|D/2|/\sigma$, where D is the difference of the mean value (daily data) before the break and after the break (Lindau and Venema, 2016). When there are multiple breaks in the time series, D is calculated with the mean values of two subsequent segments and the SD σ of the first segment is used.

For the second comparison, a reference series is calculated with the help of nearby stations. The stations are weighted with their correlation coefficients between the day-to-day changes of daily mean temperature of the automatic time series and the day-to-day changes of the neighbouring station. A minimal correlation is set to 0.9. Only stations with a higher or equal correlation coefficient are used for the estimation of the reference series with the following equation (Alexandersson and Moberg, 1997):

$$x_{\mathrm{ref}}(t) = \sum \mathrm{cor}_j \cdot \left(x_j(t) - x_{j,\mathrm{mean}} + y_{\mathrm{mean}}\right) / \sum \mathrm{cor}_j, \quad (1)$$

where x_j stands for the different time series of the nearby stations, y_{mean} is the mean value of the automatically measured time series and cor_j are the correlation coefficients of each station. In a given time range around the breaks, all breaks detected by 'uniseg' (using the differences of automatic/manual observations minus the reference time series) are counted.

The third comparison is based on related parameters. One automatic instrument is used to measure daily mean, daily maximum and daily minimum temperature. For manual measurements three different thermometers are used. So if there is a break in more than one difference series of different parameters, it is likely that the automatic instrument is causing the break.

Finally, the presumed inhomogeneous time series can be derived from the three comparison. The total score of the first comparison is weighted four times, the total score of the second comparison is weighted twice and the total score of the third comparison is weighted once. If the sum of these scores for the automatic instrument is larger than for the manual instrument, then it is likely that the automatic instrument is causing the break and the automatic time series has to be homogenized. If the score of the automatic and the manual instrument is equal, it is not possible to draw a conclusion which instrument is responsible for the break in the difference series and therefore no homogenization can be done. Figure 3 summarizes the procedure of breakpoint detection and identification of the inhomogeneous time series.

A final comparison is carried out at the end of the procedure to identify a break date that is as accurate as possible. First, the instrument with the highest total score is identified. Afterwards, it is checked whether there is a metadata information of the instrument in the given time range. If there is metadata information, the metadata information with the smallest time lag in days (within the given time range around the break) is used as break date instead of the break date detected by 'uniseg'.

2.3 | Homogenization

With the detected breakpoints and the information about the inhomogeneous time series (manual or automatic), the data can be homogenized. The first step is to divide the time series into segments. The breakpoints define the segment areas. The most recent segment is used as training period and the other segments are adjusted to that segment. For each time series, different segments are used (dependent on their break dates) but the training period is the same for both series. The segments are adjusted with the oldest segment first. At the end, all segments are adjusted to the training period by using the difference series of automatic minus manual observations.

To homogenize the data, three different methods are used. The first method is called Linear Scaling (similar to Vincent *et al*., 2002). For this method, monthly correction factors are estimated to homogenize the data. The monthly correction factors are determined by the differences of the mean differences (candidate minus reference) between the training period and the break period.

For example, data are available from January 1, 2008 to January 1, 2015 and the automatic time series has a break on May 1, 2010. To calculate a correction factor for January, the mean value of the difference series of all January values in the period January 1, 2011 to January 1, 2015 (training period) is calculated. This value is compared to the mean difference of all January values in the period January 1, 2008 to January 31, 2010 (break period). The difference of these two mean values is the correction factor for January. To correct January values in the period January 1, 2008 to January 31, 2010, the January correction factor is subtracted from the automatic observations. The same approach is repeated for each month.

With these monthly factors, *monthly* data can be corrected. To homogenize *daily* data, the monthly factors are smoothed using a spline. With this method, every day

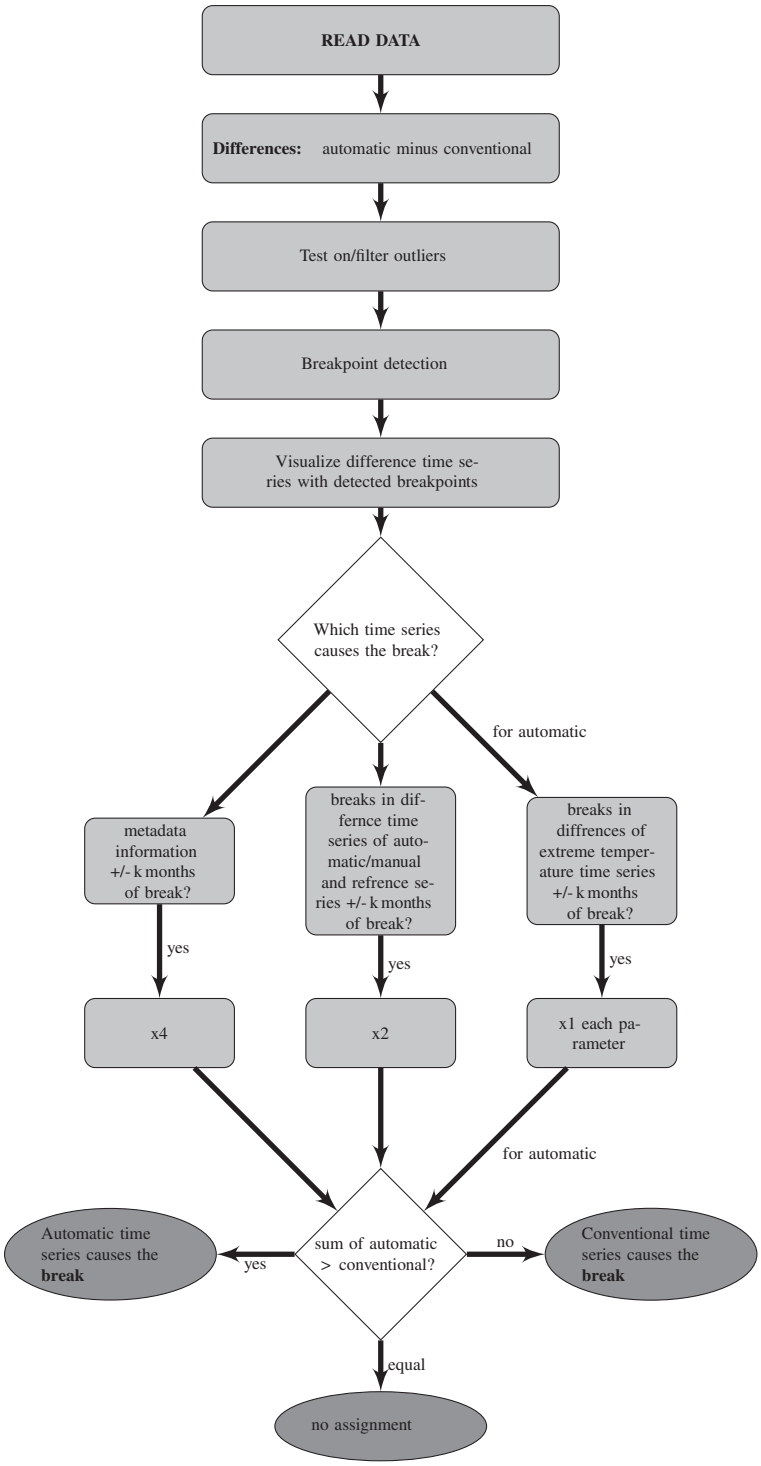

FIGURE 3 Flowchart for a suggested procedure for breakpoint detection and identification of time series including detected breakpoints

of the year has an own correction factor. This method only corrects the data in the mean value, not in the higher order moments. If the break also affects the SD of the time series, the method is not able to correct this feature.

The second method to homogenize *daily* data was suggested by Della-Marta and Wanner (2006) (called HOM). This method uses quantile mapping to adjust the data. The distribution of the differences (candidate minus reference series) during the break period is compared to the training period and adjusted such that after the correction the distributions are more consistent to each other. The adjustments are applied separately for each season and segment for example, to adjust winter values from the break segment only winter data of the training and break period are used.

The last method is based on SPLIDHOM. With an indirect nonlinear regression method and cubic smoothing splines the data of the break period is adjusted to the

training period (Mestre *et al*., 2011). The adjustments are done separately for each month and each segment (similar to the training and break periods described for the method Linear Scaling).

These three methods are examples for homogenization methods which can be used for daily data. There exist also other homogenization procedures but in most cases these methods are comparable to one of the three methods described here.

3 | EVALUATION

3.1 | Results of breakpoint detection

Table 2 summarizes the results of the breakpoint detection. At four of the 13 stations, no breaks are detected. At nine stations, at least two breaks are detected. Usually, the break size (in term of differences in the mean value) is small. The mean SNR is 0.46 and in most cases the automatically measured time series is inhomogeneous. Potential reasons for the breaks are replacements of the automatic instrument or modification of the instrument position inside the lamellar shelter (type: LAM 630). A replacement of an instrument (done in regular intervals) can have impacts on the homogeneity of the time series. For example, the uncertainty of the automatic instrument is 0.1 K (checked in the calibration laboratory). Accordingly, the combined calibration uncertainty of two instruments is 0.14 K (JCGM J, 2008).

The first break at Fichtelberg is caused by a calibration of the manual instrument and the other breaks can be related to modifications of the Stevenson shelter (not specified in details). The best identification method is the

TABLE 2 Station name, date of breakpoint (detected by 'uniseg'), signal-to-noise-ratio (SNR), total number of the first comparison (with metadata), total number of the second comparison (with reference series), total number of third comparison (with related parameter, daily maximum and minimum temperature), and total score for each instrument (manual or automatic)

			Metadata score		Nearby stations		T_{min}/T_{max}	Total score	
Station	**Break date**	**SNR**	**Manual**	**Auto**	**Manual**	**Auto**	**Auto**	**Manual**	**Auto**
Helgoland	May 1, 2008	0.62	2.5	0	—	—	1	*10*	1
	May 18, 2011	0.63	0	1.5	—	—	1	0	*7*
Schleswig	March 30, 2012	0.24	2	3	0	0	1	8	*13*
	August 1, 2015	0.23	0	0	0	0	0	0	0
Hamburg	January 14, 2012	0.22	0	0	0	0	0	0	0
	March 20, 2013	0.31	0	1.5	0	1	0	0	*8*
Potsdam	June 18, 2013	0.35	1	1.5	0	0	0	4	*6*
	March 15, 2016	0.29	0	2.5	0	0	0	0	*10*
Lindenberg	No breaks								
Brocken	October 30, 2010	0.94	0	0	—	—	1	0	*1*
	June 7, 2011	1.11	1	1.5	—	—	1	4	*7*
Görlitz	April 1, 2009	0.29	0	0	0	0	2	0	*2*
	November 6, 2013	0.22	0	2.5	0	0	1	0	*11*
Aachen	No breaks								
Aachen-Orsbach	No breaks								
Fichtelberg	August 25, 2009	0.37	1.5	0	0	0	3	*6*	3
	July 12, 2010	0.37	2.5	1	0	0	1	*10*	5
	November 11, 2013	0.24	1.5	0	0	0	1	*6*	1
Frankfurt	No breaks								
Konstanz	No breaks								
Hohen-peißenberg	September 11, 2013	0.83	1	2.5	—	—	2	4	*12*
	October 22, 2014	0.64	0	2.5	—	—	2	0	*12*

Note: Italic values represent the time series (manual or automatic) with the highest score of each station.

comparison with metadata information ('metadata score', first comparison) and the comparison of related parameters (third comparison). The comparison with nearby stations (second comparison) is less successful. Probably, the break size is too small and the difference series of manual/automatic minus reference time series is too noisy resulting in a small SNR. At the station Brocken, Helgoland and Hohenpeißenberg no reference series can be calculated. The correlation coefficients between manual/automatic time series and the series of nearby stations are too small. These three stations are located on a mountain top (station Brocken and Hohenpeißenberg) or on an island (station Helgoland). Only in one case (station Hamburg), the break in the difference series (automatic minus manual) can also be found in the differences of automatic minus reference series (second comparison).

In two cases (station Schleswig and Hamburg), it is not possible to identify the inhomogeneous time series using the three comparisons.

- The difference series and the break (detected and identified) for the station Schleswig is shown in Figure 4. At March 16, 2012 the PT100 instrument was replaced by a new one. After the detected break, the difference series has a linear trend. One reason for a trend in the difference series can be a drift in the instrument. This trend period affects the results of the breakpoint detection method ('uniseg'). The second break (detected but not identified) can be an artefact of the detection method dealing with the linear trend. This can be the reason why no metadata information is available for that time period.
- At the station Hamburg, the two detected breaks are small (inside the uncertainty of the instruments). There is only metadata information for the second break. The first break has no metadata information (first comparison), there is no break in the difference series between the measurements of Hamburg and nearby stations (comparison two), and no breaks are detected in the difference series of the related parameter daily maximum temperature and daily minimum temperature (comparison three). The total score of the manual and the automatic instrument is zero. For that reason, the detected break can not be assigned to one instrument (manual or automatic) and no homogenization is done.

The series of Hohenpeißenberg has breaks with large SNR (compared to the other breaks in Table 2) and metadata information is available (see Figure 5, top). Additionally, the break can be detected in the difference series (automatic minus manual observations) of the parameters daily mean temperature, daily maximum temperature and daily minimum temperature (third comparison). This indicates a break in the automatically measured time series.

3.2 | Results of homogenization

All homogenization methods used here are based on the idea of using a training period (including the most recent measurements) and adjust the data of the break period to the training period. Table 3 summarizes mean and SD of the differences between automatic and manual observations for the complete time series, for the training period and the complete time series after homogenization using Linear Scaling, SPLIDHOM or HOM. Differences between the homogenization methods are small. All methods are able to adjust the data to the training period.

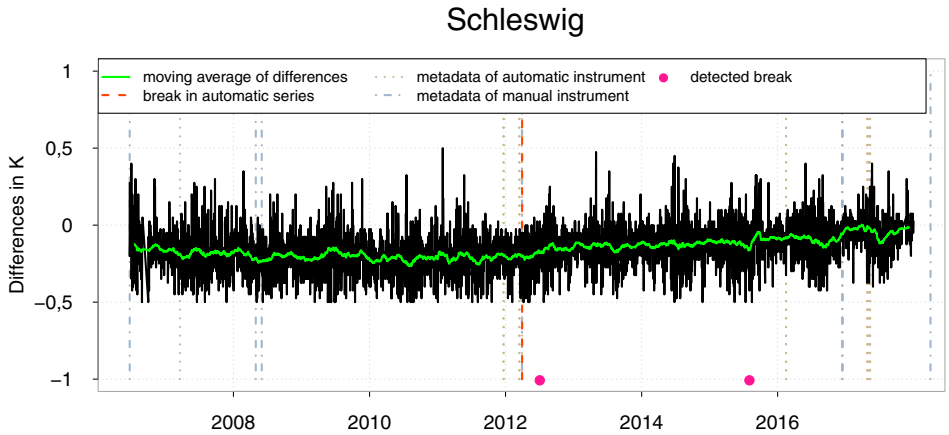

FIGURE 4 Differences of daily mean temperature in K: (station Schleswig) automatic minus manual measurements (black line), moving average (green line), and classification of detected break (orange: automatic instrument). Outliers are filtered. The pink dots in the bottom part of the plot show the detected breaks with 'uniseg'. The vertical lines show the dates of metadata information and the detected and identified breaks [Colour figure can be viewed at wileyonlinelibrary.com]

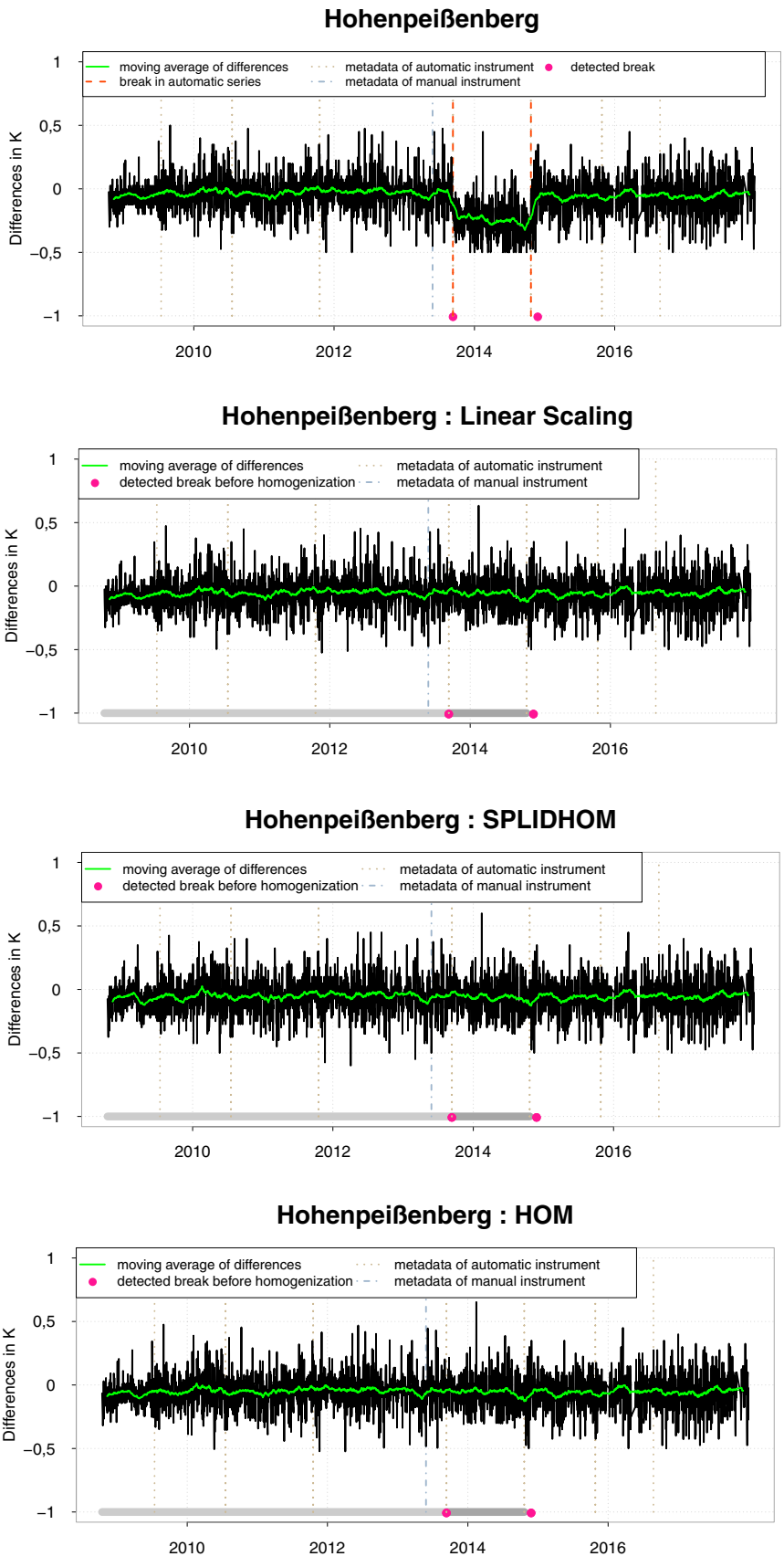

FIGURE 5 Difference series of raw data (top), homogenized data with Linear Scaling (second), homogenized with SPLIDHOM (third), and homogenized with HOM (bottom) at station Hohenpeißenberg. The grey area in the bottom part of the plots represent the different segments of the time series separated by the detected and identified breaks. The pink points in the bottom part of the plot show the detected breaks with 'uniseg'. The vertical lines show the dates of metadata information and the detected and identified breaks [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 3 Mean values and standard deviation of difference series (automatic minus manual observations) before homogenization, in training period and after homogenization using the methods Linear Scaling, SPLIDHOM, and HOM

	Before homogenization				Linear Scaling		SPLIDHOM		HOM	
	(All data)		(Training)		(All data)		(All data)		(All data)	
Station	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean
Helgoland	0.11	−0.03	0.1	*0.01*	0.11	*0.04*	0.11	*0.04*	0.11	*0.04*
Schleswig	0.14	−0.15	0.13	*−0.11*	0.14	*−0.11*	0.14	*−0.11*	0.14	−0.1
Hamburg	0.13	0.05	0.12	*−0.04*	0.12	*−0.04*	0.13	*−0.04*	0.12	*−0.04*
Potsdam	0.12	0.08	0.12	*0.07*	0.12	*0.07*	0.12	0.06	0.12	*0.07*
Brocken	0.11	−0.1	0.08	*−0.14*	0.09	*−0.14*	0.09	*−0.14*	0.09	−0.13
Görlitz	0.11	0.03	0.1	*−0.01*	0.11	*−0.01*	0.11	0	0.11	0
Fichtelberg	0.07	0.03	0.07	*0.01*	0.07	0	0.08	0	0.07	*0.01*
Hohenpeißenberg	0.15	−0.07	0.14	*−0.06*	0.13	*−0.06*	0.14	*−0.06*	0.13	−0.05

Note: The grey background represents the smallest differences between the standard deviation of the training period and of the time series after homogenization using the different homogenization methods.

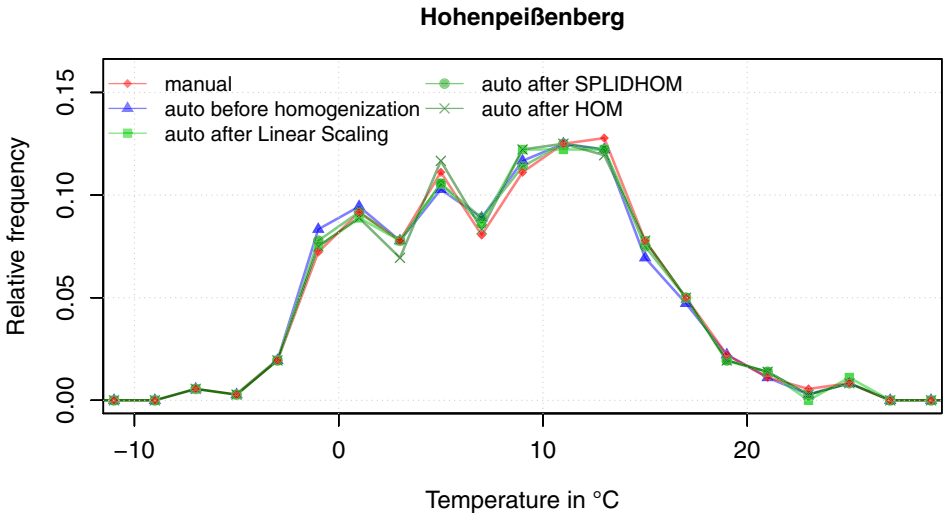

FIGURE 6 Histogram of manual and automatic daily mean values at station Hohenpeißenberg before homogenization (red for manual and blue for automatic measurements), after homogenization using Linear Scaling (light green), after homogenization using SPLIDHOM (medium green), and after homogenization using HOM (dark green) for the break period September 11, 2013 to November 23, 2014 [Colour figure can be viewed at wileyonlinelibrary.com]

The reason for the small differences between the homogenization method is, that in most cases the breaks in this study affect the mean but not the SD of the individual segments of the difference series.

Figure 5 shows the difference series of automatic minus manual observations before and after the homogenization with all three methods for the station Hohenpeißenberg. Differences are very small between the methods. After homogenization no further breaks are detected (i.e., the homogenization was successful). Changes in the distribution before and after homogenization are small (see Figure 6). The distribution of the automatic measurements is shifted to the right towards the manual distribution.

In a few cases, the procedure of breakpoint detection, identification and homogenization failed. For the series of Hamburg, Helgoland and Schleswig, breaks are detected in the difference series after the homogenization. At these stations, the result is independent of the homogenization method.

- In Helgoland, the detected break has a time lag to the metadata information (Figure 7, first row). One possible explanation is that the metadata information is not related to the break and the wrong time series is adjusted. Another possible explanation is that the metadata information has a wrong or shifted date. If metadata is available, the date of the metadata information is used instead of the detected break date. Therefore, an incorrect date in the metadata will result in an incorrect break date and the homogenization is influenced.

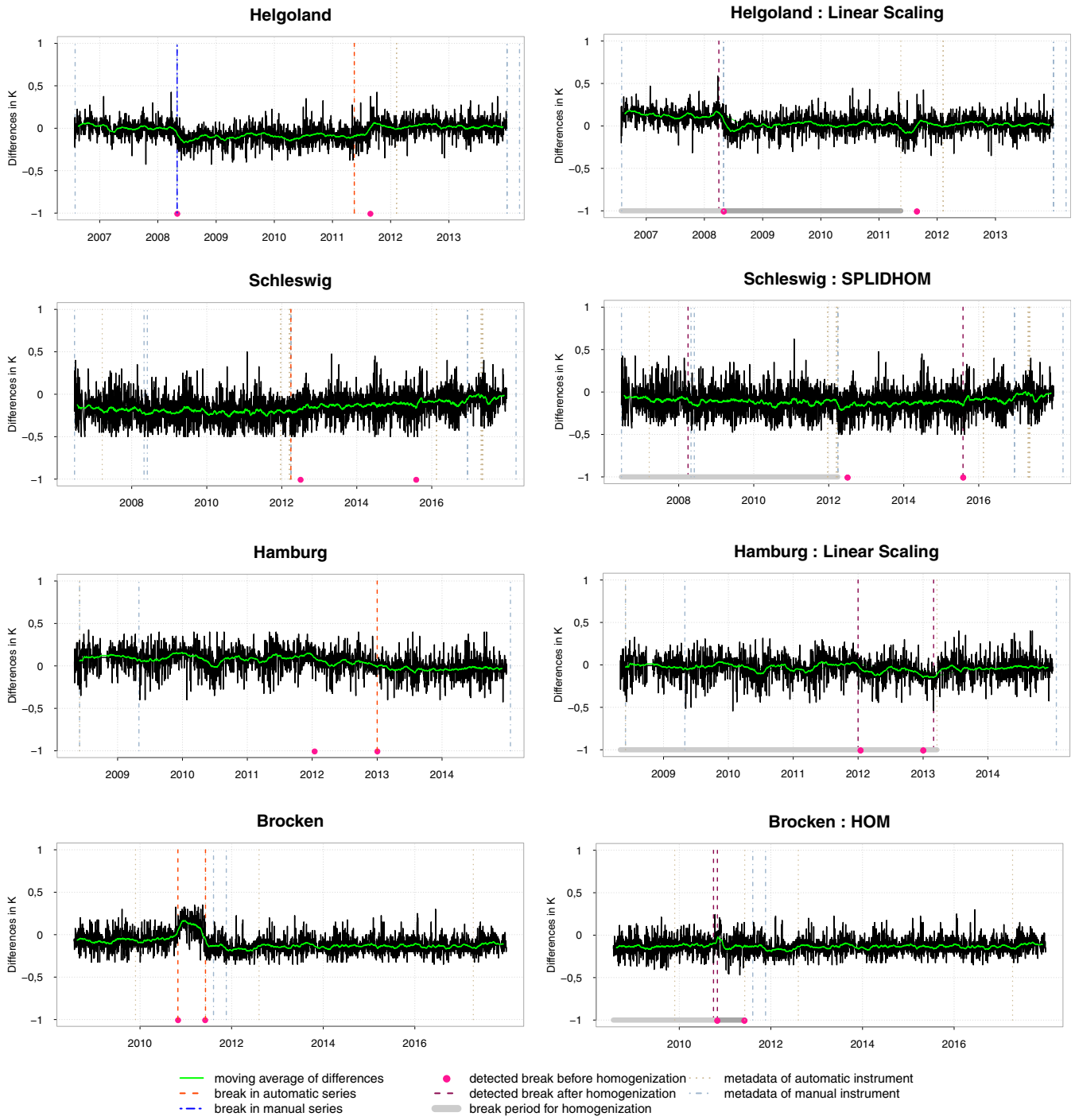

FIGURE 7 Difference series of raw data (left) and homogenized data (right) with Linear Scaling at station Helgoland (top row), difference series of homogenized data with SPLIDHOM at station Schleswig (second row), difference series of homogenized data with Linear Scaling at station Hamburg (third row), and difference series of homogenized data with HOM at station Brocken (forth row). The grey area in the bottom part of the plots on the right represent the different segments of the time series separated by the detected and identified breaks. The pink points in the bottom part of the plot show the detected breaks with 'uniseg'. The vertical lines show the dates of metadata information, the detected and identified breaks (orange for automatic and blue for manual) and the detected breaks in the difference series after homogenization (dark pink line) [Colour figure can be viewed at wileyonlinelibrary.com]

- In Schleswig, the last period with the linear trend in the difference series is used as training period. The homogenization methods have problems with this linear trend (Figure 7, second row).
- In Hamburg, the breakpoint is detected at the same position as before the homogenization because it was not possible to identify the inhomogeneous time series (automatic or manual). No homogenization is done for

that break/segment (Figure 7, third row). This affected the homogenization results of the complete time series.

- After the homogenization using HOM a break can be detected at a similar position as in the raw data for the series Brocken (Figure 7, bottom). The break in the difference series (with homogenized data) is smaller than before (with raw data) so there is an improvement. Using the other two homogenization methods (SPLIDHOM and Linear Scaling) no breaks are detected.

3.3 | Comparison of raw data and homogenized data

After the homogenization, the differences between homogenized data and raw data (only been controlled for outliers) are analyzed. Figure 8 shows the histograms of the differences between automatic and manual measurements without outliers of the original data, and the data after homogenization with Linear Scaling, SPLIDHOM or HOM. The mean value of all differences remains almost identical (−0.03 K), that is, breaks in the time series compensate each other. Some breaks are related to higher temperature values for the automatic instrument and some are connected to smaller temperature values for the automatic instruments. On average breaks have no effect on the mean differences between the automatic and manual daily mean temperature values. The mean difference between the two measurement systems (manual and automatic) is close to zero. No break is expected in long time series of daily mean temperature in Germany related to the automatization (at least for stations with the same measurement

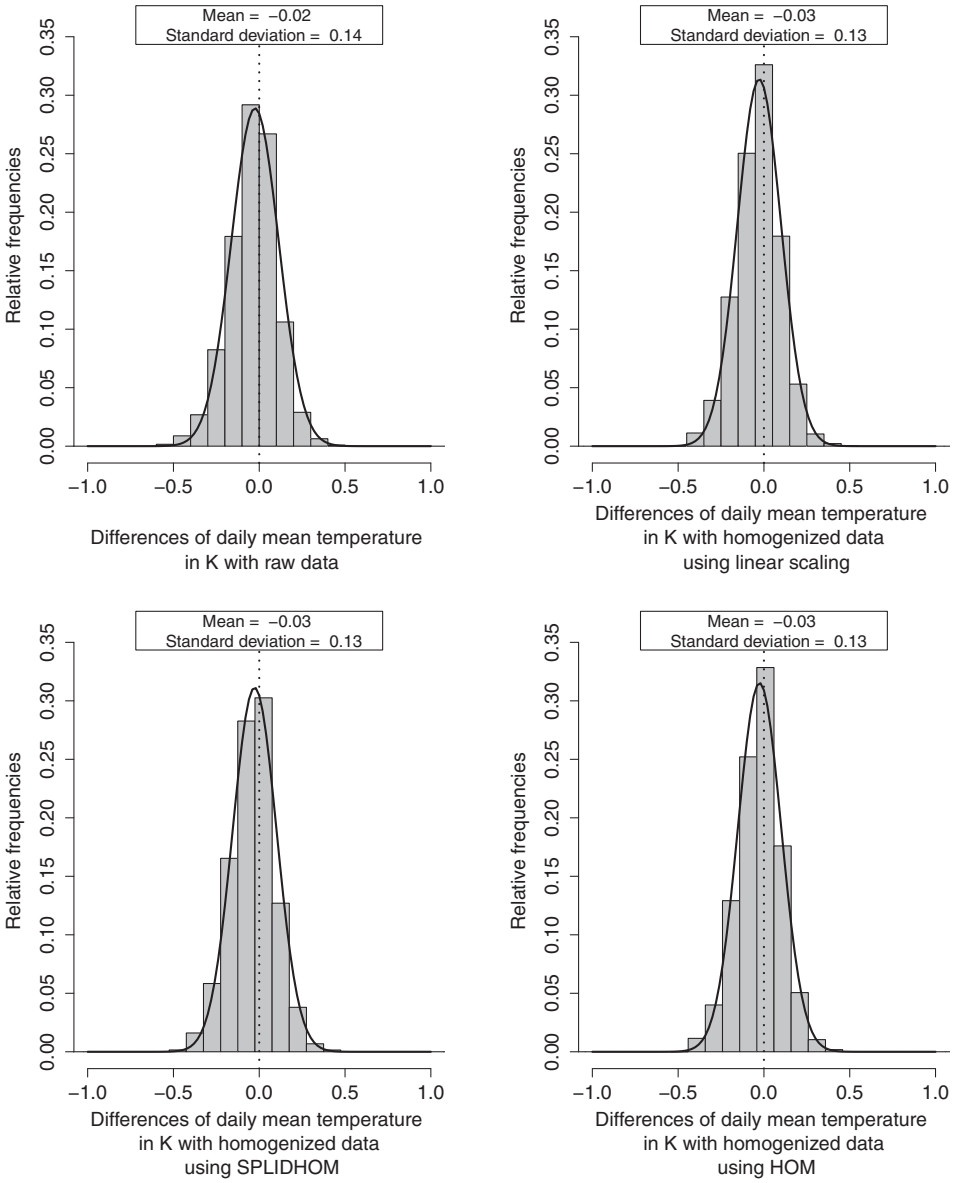

FIGURE 8 Histogram of manual and automatic daily mean values before homogenization (top, left), after homogenization using Linear Scaling (top, right), after homogenization using SPLIDHOM (bottom, left), and after homogenization using HOM (bottom, right)

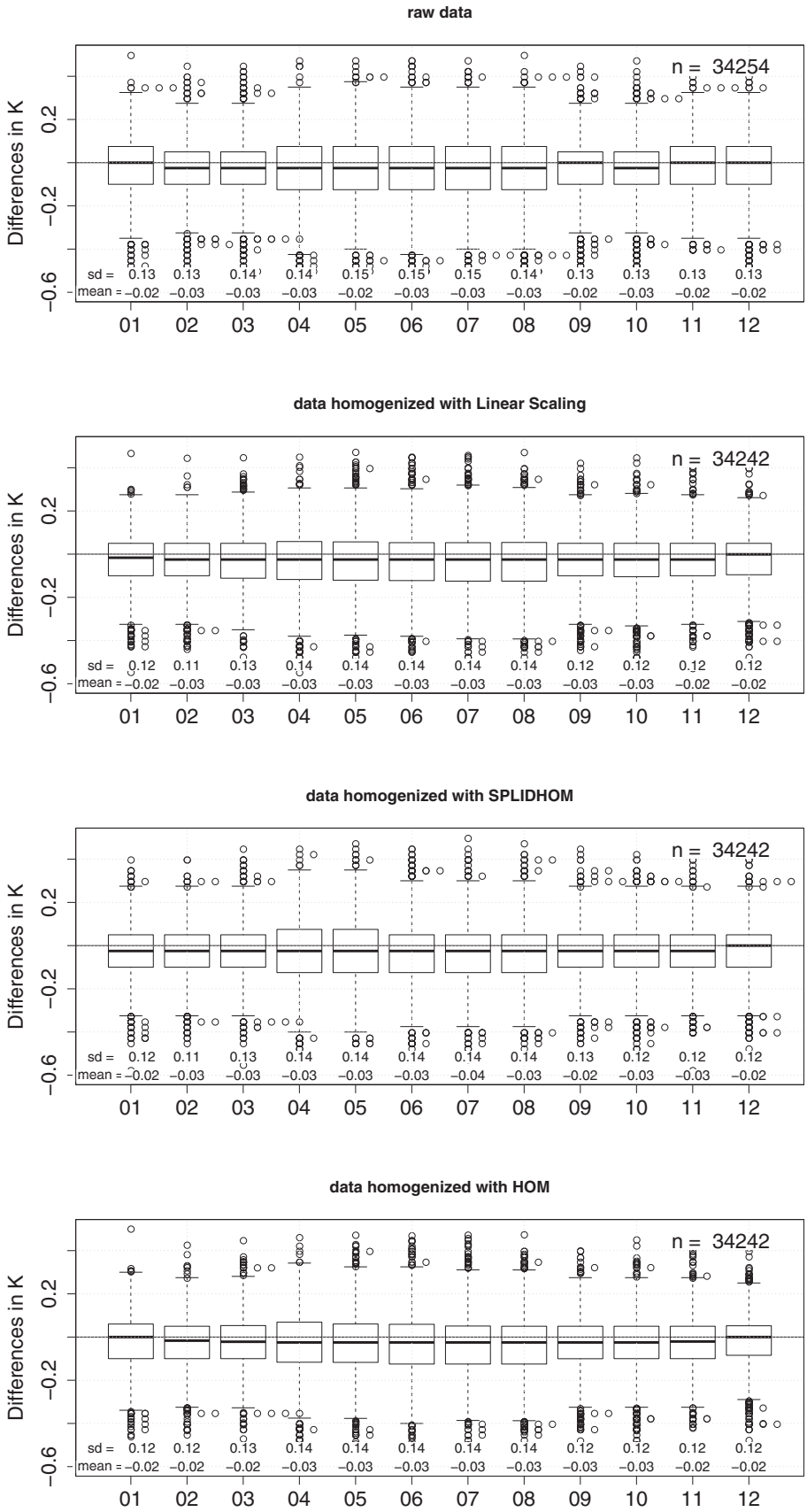

FIGURE 9 Monthly boxplots of differences in K. Top row: based on raw data, second row: homogenizes data (with Linear Scaling), third row: homogenizes data (with SPLIDHOM), fourth row: homogenizes data (with HOM). The mean difference and the standard deviation (SD) of the differences are presented below the monthly boxplots

conditions as they are present at German climate reference stations). Station relocation or environmental changes may have a stronger influence on long time series than the automatization.

The SD of the differences between automatic and manual measurements after the homogenization only differ in a small range (0.01 K) compared to the original data. As generally expected, the SD is smaller after homogenization because the SD of the original data is increased by the breaks.

As shown in Figure 9, there is no annual cycle in the differences of manual and automatic daily mean temperature observations of the raw data and after homogenization.

4 | SUMMARY, CONCLUSIONS AND OUTLOOK

In this study, differences between automatic and manual daily mean temperature measurements of 13 stations are analyzed including an outlier control, a breakpoint detection and the homogenization of time series. The mean (−0.02 K) and the SD (0.14 K) of the differences between automatic and manual measurements of daily mean temperature before the homogenization are small. This finding is in agreement with the results of Baciu *et al*. (2005). With these values no break is expected in long time series of daily mean temperature (calculated with the traditional equation) caused by the transition from manual to automatic measurement instruments. To study the effects of breaks in time series, the time series are analyzed on breakpoints and the data is homogenized (with three homogenization methods: Linear Scaling, SPLIDHOM and HOM). Afterwards the differences are analyzed again. The mean difference of the two observing techniques (manual and automatic) remain almost constant. In most cases, the break size is below the instrument calibration uncertainty. The homogenization of the time series only has a small effect but after the homogenization the SD of the differences is even smaller than before. The largest breaks were found for the automatic instrument at the station Brocken and Hohenpeißenberg. Here, the break size is larger than the instrument calibration uncertainty.

The analysis of German climate reference stations has shown that for Germany the homogenization of the data is only of minor relevance in the context of analyzing the impact of the automatization on long time series of daily mean temperature values. In this case, the breaks in the time series (during the time period of parallel measurements) are small and compensate each other. In some cases, replacement of PT100 instruments causes small breaks caused by the instrument calibration uncertainty (resulting in a potential offset inside a range of 0.14 K). The maintenance intervals of the instrument are short enough to detect problems of the instrument sufficiently early to ensure that the quality of the data is not strongly affected by breaks. Replacements of instruments or calibration dates are well documented such that breaks can be identified easily (in most cases).

The results of this study can be summarized as follows:

- The mean differences between manual and automatic daily mean temperature values are small. Therefore, it can be concluded that the automatization of temperature measurements did not cause relevant breaks in the German time series of daily mean temperature.
- The differences between the results of the three homogenization methods SPLIDHOM, HOM and Linear Scaling are small. All three methods are able to homogenize the breaks as for example the breaks in the time series of Hohenpeißenberg.
- The detected breaks in the time series of daily mean temperature (within the time period of parallel measurements) are small indicating consistent data quality and sufficiently short maintenance intervals.

At German climate reference stations also parallel measurements of other meteorological parameters are performed (e.g., precipitation, daily sunshine duration, relative humidity, and wind speed). The analysis of the impact of changing measurement systems on the homogeneity of long time series of these parameters will be subject of future studies.

ACKNOWLEDGEMENTS

We thank one anonymous reviewer and Alba Gilabert Gallart for their helpful comments on the manuscript. We also want to thank our colleagues at the observatories and from the technical infrastructure for helpful discussions. This research was supported by the Deutscher Wetterdienst research program "Innovation in Applied Research and Development" (IAFE).

ORCID

Lisa Hannak https://orcid.org/0000-0002-9376-2383
Karsten Friedrich https://orcid.org/0000-0003-2135-1750
Florian Imbery https://orcid.org/0000-0002-9616-1874
Frank Kaspar https://orcid.org/0000-0001-8819-8450

REFERENCES

Alexandersson, H. and Moberg, A. (1997) Homogenization of Swedish temperature data. Part I: Homogeneity test for linear trends. *International Journal of Climatology*, 17(1), 25–34.

Auchmann, R., and Brönnimann, S. (2012) A physics-based correction model for homogenizing sub-daily temperature series. *Journal of Geophysical Research: Atmospheres*, 117, D17119. https://doi.org/10.1029/2012JD018067.

Auer, I., Böhm, R., Jurkovic, A., Lipa, W., Orlik, A., Potzmann, R., Schöner, W., Ungersböck, M., Matulla, C., Briffa, K., Jones, P., Efthymiadis, D., Brunetti, M., Nanni, T., Maugeri, M., Mercalli, L., Mestre, O., Moisselin, J.M., Begert, M., Müller-Westermeier, G., Kveton, V., Bochnicek, O., Stastny, P., Lapin, M., Szalai, S., Szentimrey, T., Cegnar, T., Dolinar, M., Gajic-Capka, M., Zaninovic, K., Majstorovic, Z. and Nieplova, E. (2007) HISTALP—historical instrumental climatological surface time series of the Greater Alpine Region. *International Journal of Climatology*, 27(1), 17–46.

Baciu, M., Copaciu, V., Breza, T., Cheval, S., and Pescaru, I. V. (2005). Preliminary results obtained following the intercomparison of the meteorological parameters provided by automatic and classical stations in Romania. In WMO Technical Conference on Meteorological and Environmental Instruments and Methods of Observation (TECO-2005).

Begert, M., Schlegel, T. and Kirchhofer, W. (2005) Homogeneous temperature and precipitation series of Switzerland from 1864 to 2000. *International Journal of Climatology*, 25(1), 65–80.

Böhm, R., Jones, P.D., Hiebl, J., Frank, D., Brunetti, M. and Maugeri, M. (2010) The early instrumental warm-bias: a solution for long central European temperature series 1760–2007. *Climatic Change*, 101(1–2), 41–67.

Brandsma, T. and Van der Meulen, J. (2008) Thermometer screen intercomparison in De Bilt (The Netherlands) – Part II: Description and modeling of mean temperature differences and extremes. *International Journal of Climatology*, 28(3), 389–400.

Brunet, M., Asin, J., Sigró, J., Bañón, M., García, F., Aguilar, E., Palenzuela, J.E., Peterson, T.C. and Jones, P. (2011) The minimization of the screen bias from ancient Western Mediterranean air temperature records: an exploratory statistical analysis. *International Journal of Climatology*, 31(12), 1879–1895.

Della-Marta, P. and Wanner, H. (2006) A method of homogenizing the extremes and mean of daily temperature measurements. *Journal of Climate*, 19(17), 4179–4197.

Doesken, N. J. (2005). The National Weather Service MMTS (Maximum-Minimum Temperature System)–20 years after. In *15th Conference on Applied Climatology*, *Abstract JPI.26*, *5pp*.

Erell, E., Leal, V. and Maldonado, E. (2005) Measurement of air temperature in the presence of a large radiant flux: an assessment of passively ventilated thermometer screens. *Boundary-Layer Meteorology*, 114(1), 205–231.

Hannak, L., Friedrich, K., Imbery, F. and Kaspar, F. (2019) Comparison of manual and automatic daily sunshine duration measurements at German climate reference stations. *Advances in Science and Research*, 16, 175–183.

Hannart, A., Mestre, O. and Naveau, P. (2014) An automatized homogenization procedure via pairwise comparisons with application to Argentinean temperature series. *International Journal of Climatology*, 34(13), 3528–3545.

Hewaarachchi, A.P., Li, Y., Lund, R. and Rennie, J. (2017) Homogenization of Daily Temperature Data. *Journal of Climate*, 30(3), 985–999.

Hoover, J. and Yao, L. (2018) Aspirated and non-aspirated automatic weather station Stevenson screen intercomparison. *International Journal of Climatology*, 38(6), 2686–2700.

JCGM. (2008) Evaluation of measurement data-guide to the expression of uncertainty in measurement (GUM), BIPM Report 100: 2008, Sévres-BIPM Joint Committee for Guides in Metrology WG1.

Kaspar, F., Hannak, L. and Schreiber, K.-J. (2016) Climate reference stations in Germany: Status, parallel measurements and homogeneity of temperature time series. *Advances in Science and Research*, 13, 163–171.

Kuglitsch, F.G., Toreti, A., Xoplaki, E., Della-Marta, P.M., Luterbacher, J. and Wanner, H. (2009) Homogenization of daily maximum temperature series in the Mediterranean. *Journal of Geophysical Research: Atmospheres*, 114, D15108. https://doi.org/10.1029/2008JD011606.

Lanzante, J.R. (1996) Resistant, robust and non-parametric techniques for the analysis of climate data: Theory and examples, including applications to historical radiosonde station data. *International Journal of Climatology*, 16(11), 1197–1226.

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.

Lund, R., Wang, X.L., Lu, Q.Q., Reeves, J., Gallagher, C. and Feng, Y. (2007) Changepoint detection in periodic and autocorrelated time series. *Journal of Climate*, 20(20), 5178–5190.

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. *IDOJÁRÁS Quarterly Journal of the Hungarian Meteorological Service*, 117, 47–67.

Mestre, O., Gruber, C., Prieur, C., Caussinus, H. and Jourdain, S. (2011) SPLIDHOM: A Method for Homogenization of Daily Temperature Observations. *Journal of Applied Meteorology and Climatology*, 50(11), 2343–2358.

Peterson, T.C., Easterling, D.R., Karl, T.R., Groisman, P., Nicholls, N., Plummer, N., Torok, S., Auer, I., Boehm, R., Gullett, D., Vincent, L., Heino, R., Tuomenvirta, H., Mestre, O., Szentimrey, T., Salinger, J., Førland, E.J., Hanssen-Bauer, I., Alexandersson, H., Jones, P. and Parker, D. (1998) Homogeneity adjustments of in situ atmospheric climate data: a review. *International Journal of Climatology*, 18(13), 1493–1517.

Picard, F., Hoebeke, M., Lebarbier, E., Miele, V., Rigaill, G., and Robin, S. (2016). *cghseg: Segmentation Methods for Array CGH Analysis*. R package version 1.0.2-1.

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

Core Team, R. (2015) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ribeiro, S., Caineta, J. and Costa, A. (2016) Review and discussion of homogenisation methods for climate data. *Physics and Chemistry of the Earth, Parts A/B/C*, 94, 167–179.

Toreti, A., Kuglitsch, F.G., Xoplaki, E., Luterbacher, J. and Wanner, H. (2010) A novel method for the homogenization of

daily temperature series and its relevance for climate change analysis. *Journal of Climate*, 23(19), 5325–5331.

Vincent, L.A., Zhang, X., Bonsal, B. and Hogg, W. (2002) Homogenization of daily temperatures over Canada. *Journal of Climate*, 15(11), 1322–1334.

Yosef, Y., Aguilar, E. and Alpert, P. (2018) Detecting and adjusting artificial biases of long-term temperature records in Israel. *International Journal of Climatology*, 38(8), 3273–3289.

How to cite this article: Hannak L, Friedrich K, Imbery F, Kaspar F. Analyzing the impact of automatization using parallel daily mean temperature series including breakpoint detection and homogenization. *Int J Climatol*. 2020;1–16. https://doi.org/10.1002/joc.6597