INTERNATIONAL JOURNAL OF CLIMATOLOGY
Int. J. Climatol. **37**: 4670–4683 (2017)
Published online 6 June 2017 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/joc.5114

RMetS
Royal Meteorological Society

The influence of station density on climate data homogenization

S. Gubler,[a]* S. Hunziker,[b,c] M. Begert,[a] M. Croci-Maspoli,[a] T. Konzelmann,[a] S. Brönnimann,[b,c] C. Schwierz,[a] C. Oria[d] and G. Rosas[d]

[a] *Federal Office of Meteorology and Climatology MeteoSwiss, Zürich, Switzerland*
[b] *Institute of Geography, University of Bern, Bern, Switzerland*
[c] *Oeschger Centre for Climate Change Research, University of Bern, Bern, Switzerland*
[d] *Servicio Nacional de Meteorología e Hidrología del Perú, SENAMHI, Lima, Peru*

ABSTRACT: Relative homogenization methods assume that measurements of nearby stations experience similar climate signals and rely therefore on dense station networks with high-temporal correlations. In developing countries such as Peru, however, networks often suffer from low-station density. The aim of this study is to quantify the influence of network density on homogenization. To this end, the homogenization method HOMER was applied to an artificially thinned Swiss network.

Four homogenization experiments, reflecting different homogenization approaches, were examined. Such approaches include diverse levels of interaction of the homogenization operators with HOMER, and different application of metadata. To evaluate the performance of HOMER in the sparse networks, a reference series was built by applying HOMER under the best possible conditions.

Applied in completely automatic mode, HOMER decreases the reliability of temperature records. Therefore, automatic use of HOMER is not recommended. If HOMER is applied in interactive mode, the reliability of temperature and precipitation data may be increased in sparse networks. However, breakpoints must be inserted conservatively. Information from metadata should be used only to determine the exact timing of statistically detected breaks. Insertion of additional breakpoints based solely on metadata may lead to harmful corrections due to the high noise in sparse networks.

KEY WORDS homogenization; station density; HOMER; metadata; temporal consistency, trend accuracy

Received 7 April 2016; Revised 30 March 2017; Accepted 31 March 2017

1. Introduction

Long-term and high-quality climate data are essential for monitoring and studying climate variability and change. Measurements are often affected by non-climatic influences such as station relocations, observer changes, changes in the station environment, changes of instruments, and station maintenance. To remove such inhomogeneities and to obtain more reliable climate data, time series must be homogenized (e.g. Peterson *et al.*, 1998; Aguilar *et al.*, 2003; Trewin, 2010). Besides affecting single station measurements, inhomogeneities may bias the network average if they have a tendency in one direction during a certain period (Venema *et al.*, 2012). In the Swiss station network for instance, a temperature trend analysis for the period 1864–2000 using raw data underestimates the temperature trend compared to homogenized data by 0.4 °C/100 years (Begert *et al.*, 2005), which amounts to around 40–50% of the reported trend. Nevertheless, many studies still use raw data for their analyses, often leading to misinterpretation of the results (Cao and Yan, 2012). Hence, conclusions of such research should be interpreted with care, because trustworthy trend assessments must rely on high-quality homogenous data (e.g. Begert *et al.*, 2005; Reeves *et al.*, 2007; Toreti *et al.*, 2010; Venema *et al.*, 2012).

Relative homogenization presumes that nearby stations have the same climate signals. The performance of homogenization methods is therefore often tested using highly correlated data. In many regions of the world station densities are low however. For instance, in the Peruvian Andes the station density is around ten times lower than in Switzerland (one temperature station per roughly 5000 km^2 in Peru compared to one per 475 km^2 in Switzerland). Additionally, data quality in such regions can be low (e.g. measurement errors and missing data), leading to a substantial fraction of stations which is not suitable for climate studies. The combination of strong climate gradients, the sparse network, and the exclusion of time series due to quality problems all contribute to weak correlations. This may impact the efficiency of the homogenization process (Caussinus and Mestre, 2004; Domonkos, 2013). However, only one publication comparing the homogenization of a dense and a thinned network is known to the authors: Vertačnik *et al.* (2015). They found that a reduction of the Slovenian station network from 60 to 44 (i.e. a reduction from one station per 307 km^2 to one station per 461 km^2) does not substantially influence the

* Correspondence to: S. Gubler, Federal Office of Meteorology and Climatology MeteoSwiss, Operation Center 1, P.O. Box 257, 8058 Zurich-Airport, Switzerland. E-mail: stefanie.gubler@meteoswiss.ch

homogenization results. However, the thinned network in Slovenia is still about ten times denser than the network in Peru.

Analysis and assessment of climate change and the implementation of climate services is of great importance for regions that are especially vulnerable to climate change impacts (Brooks and Adger, 2003), such as the Andean area (Buytaert *et al.*, 2006; Buytaert and De Bièvre, 2012; Salzmann *et al.*, 2009; Vuille *et al.*, 2008; World Bank, 2010). Therefore, the project CLIMANDES (http://www.senamhi.gob.pe/climandes/), a project within the Global Framework for Climate Services (GFCS), aims at providing user-tailored climate services in two pilot regions in Peru. One goal of CLIMANDES is to implement a suitable homogenization method at the national meteorological and hydrological service of Peru SENAMHI (Rosas *et al.*, 2016). To this end, the recently developed semi-automatic homogenization procedure HOMER (Mestre *et al.*, 2013) was chosen since it is state-of-the-art (developed after the COST Action on Homogenization), freely available, and runs on the open-source software R (R Development Core Team, 2014). Until now, only a few studies exist that evaluate HOMER (Freitas *et al.*, 2013; Coll *et al.*, 2014; Vertačnik *et al.*, 2015; Noone *et al.*, 2016). Nevertheless, the approach is currently being implemented in several weather services [e.g. Météo-France, the Norwegian Meteorological Institute, MeteoSwiss, the Irish Meteorological Service Met Éireann, and the Slovenian Environment Agency (Vertačnik *et al.*, 2015)], and it is already applied in countries where station networks are sparse such as Bolivia (Vicente-Serrano *et al.*, 2015; López-Moreno *et al.*, 2016) or Tanzania (Luhunga *et al.*, 2014).

Within the project CLIMANDES, HOMER was applied to station records of the time period 1964 to 2012 from the southern Andes of Peru. Analyses of the homogenized temperature records showed that different approaches (exclusion of stations with quality problems, metadata availability, and different homogenization operators) resulted in differences in the average network trends of 0.06–0.08 °C/decade, while the estimated average network trends range between 0.22 and 0.28 °C/decade for maximum temperature (TX), and between 0.03 and 0.11 °C/decade for minimum temperature (TN) (e.g. Rosas *et al.*, 2016). These considerable differences raised questions on the reliability of breakpoint detection and correction in sparse station networks. Post-analysis of the corrected breakpoints has shown that the standard deviation of the correction amounts is around 0.95 °C for TX and 1.05 °C for TN. This is considerably larger than the standard deviation of detected breakpoints normally encountered in Europe, which is reported to range between 0.6 and 0.8 °C (Auer *et al.*, 2007; Brunetti *et al.*, 2006; Caussinus and Mestre, 2004; Venema *et al.*, 2012). On average, breakpoints were detected every 13–20 years in the Peruvian network, which is comparable to the breakpoint frequency detected in Western European temperature records (Venema *et al.*, 2012). The low station density in the Andes however leads to more noise in the difference series, and thus breakpoints may be less detectable (Auer *et al.*, 2005). Hence, there may actually be a higher number of breakpoints in reality. This ambiguity in breakpoint detection may be a reason for the differing average network trends mentioned above, and it provides motivation for the present study to investigate the influence of station density on homogenization.

Switzerland and the Peruvian Andes share pronounced spatial climate gradients and a complex topography. But in contrast to the station networks in the Peruvian Andes, the Swiss network is dense and contains many high-quality time series back to the 19th century. The Swiss station histories are nearly complete, which allows for a detailed comparison of reported and statistically detected breakpoints (Begert *et al.*, 2003; Kuglitsch *et al.*, 2012). Based on these near ideal conditions for homogenization and the similar topography, a comparison experiment is conducted in this study: the Peruvian network characteristics are mimicked by thinning the Swiss station network to quantify the effects of homogenization in sparse networks. Of course, the mid-latitude climate in Switzerland differs substantially from the tropical climate in Peru. The Peruvian climate is strongly influenced by inter-annual cycles and (sub-)tropical convection. The El Niño Southern Oscillation most dominantly influences the climate in the region, together with the Pacific Decadal Oscillation and the Southern Annular Mode (Seiler *et al.*, 2012). Such variability modes may modulate the performance of homogenization methods and cannot be reflected by the Swiss network. Results from this study are hence not entirely transferable for applications in Peru. However, the study presents a first effort to quantify the influence of low station density on homogenization, trying to closely reflect the Peruvian network characteristics.

2. Data

2.1. Peruvian network

This section introduces the station network in the southern Peruvian Andes, which is used to characterize a *low-density network*. Averaged over the area, the station number in the region of interest corresponds to approximately one station per 10 000 km^2 for temperature (TX and TN), and one per 4000 km^2 for precipitation (P). Meteorological stations that are suitable for climate analyses in the Peruvian Andes typically run since 1964 and have less than 20% missing values.

2.2. Swiss network

The Swiss stations used for this study are 31 for temperature and 55 for precipitation (Figure 1). The station density in Switzerland corresponds to one station per 475 km^2 for temperature and one per 100 km^2 for precipitation. Measurements since 1961 are used to reflect the typical Peruvian time series length of 50 years.

2.3. Correlation analysis

Correlation is the most frequently used measure to identify reference stations for homogenization (Aguilar *et al.*,

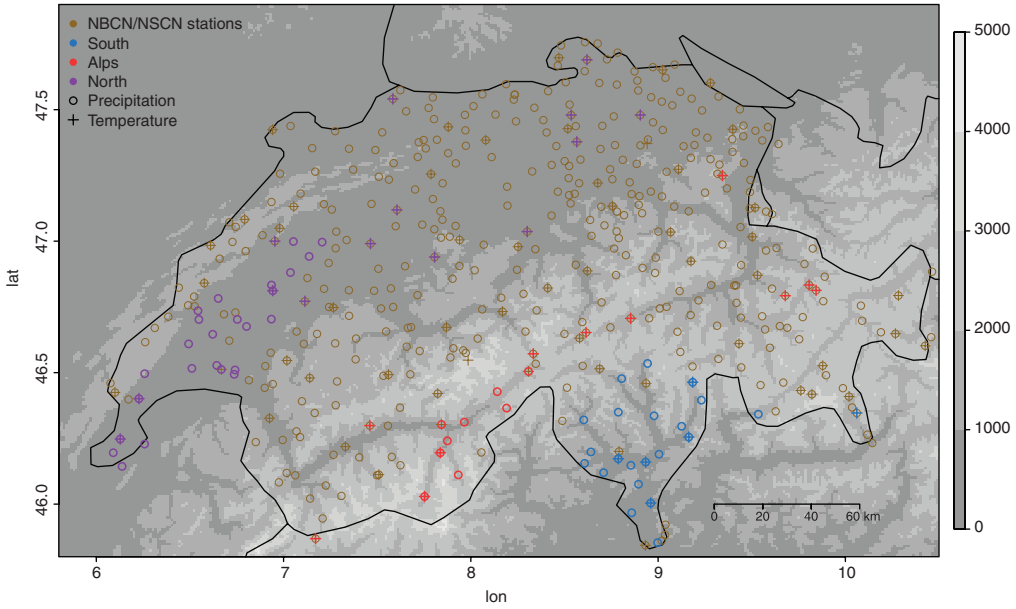

Figure 1. Meteorological stations in Switzerland (all colours). The stations selected for the three clusters are colour-coded: 'Alps' (red), 'South' (blue), and 'North' (purple) for precipitation (circles) and temperature (crosses). The gray background colors represent the elevation in meters above sea level.

2003). For this reason, the correlation structures of the Swiss and the Peruvian networks are analysed. To this end, the Spearman-correlation of the first differences of the de-seasonalized monthly time series was calculated (Peterson and Easterling, 1994).

The structure of the pairwise correlations for the Swiss and the Peruvian networks is shown in Figure 2. We observe that the correlations in the Swiss network range between 0.6 and nearly 1.0 (TX and TN) and between 0.1 and nearly 1.0 (P) for distances up to 100 km. In the Peruvian network, correlations for temperature range between 0.2 and 0.9 (TX and TN) and between 0.1 and 0.8 (P). For homogenization, the correlation of the candidate station to its most closely correlated neighbours is most relevant. Analysis of the Peruvian data shows that typical correlations of the six best correlating neighbours range between 0.60 and 0.80 (TX and TN) and 0.45 and 0.60 (P). These correlations are referred to as the *typical Peruvian correlations* in the following. In Switzerland, these correlations are typically well above 0.90 for both temperature and precipitation.

The best correlations of monthly means (TX and TN) and totals (P) between neighbouring stations of a certain distance in Peru are 0.1–0.2 lower than in Switzerland (Figure 2). This is in accordance with New *et al.* (2000) stating that correlations between stations become insignificant after shorter distances in the tropics (0 to 30°S) than in the subtropics and mid-latitudes (30 to 60°N). New *et al.* (1999) attribute these differences mainly to different large-scale circulation patterns. However, experiences in the field suggest that systematically smaller spatial representativeness of the station sites, lower standard of maintenance, and more frequent measurement and post-processing errors may also contribute to the lower correlations observed in Peru.

3. Methods

3.1. HOMER

The results of the model inter-comparison study (Venema *et al.*, 2012) conducted within the COST Action ES0601 HOME 'Advances in Homogenisation Methods of Climate Series: An Integrated Approach' demonstrate that the performance of some of the widely used methods differ considerably. Based on these results, the method HOMER (Mestre *et al.*, 2013) was designed. HOMER combines two of the best performing methods PRODIGE (Caussinus and Mestre, 2004) and ACMANT (Domonkos *et al.*, 2011a) with a joint-segmentation breakpoint detection method originating from DNA-segmentation (Picard *et al.*, 2011).

Combining different break detection algorithms has been demonstrated to be beneficial. It results in higher confidence when accepting or rejecting breakpoints, especially if a breakpoint cannot be confirmed by metadata (Toreti *et al.*, 2011; Kuglitsch *et al.*, 2012). The correction of the breakpoints in HOMER is done based on a two-factor analysis of variance (ANOVA) model approach. It allows for the correction of a set of stations simultaneously and automatically (Mestre *et al.*, 2013), and was shown to improve breakpoint correction over traditional approaches (Domonkos *et al.*, 2011a; Domonkos, 2013). In this study, both PRODIGE (called 'pairwise detection' in HOMER) and joint segmentation were applied. The ACMANT-component of HOMER was not used since it was constructed for mid and high latitudes and not for use in tropical areas (Domonkos *et al.*, 2011a).

Homogenization with HOMER is an iterative process. Breakpoint detection procedures (pairwise and joint) are alternated with the correction of the breakpoints. The alternating procedure is stopped once every time series is

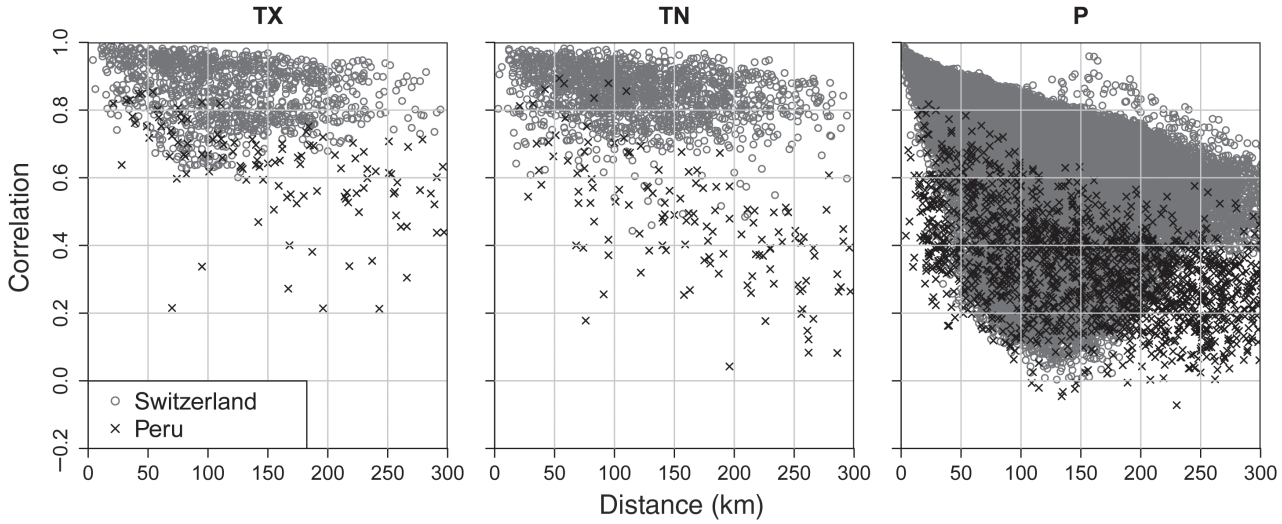

Figure 2. Correlations of station pairs of the Swiss (grey circles) and the Peruvian pilot network (black crosses) as a function of the stations' distance.

considered homogeneous by the homogenization operator. Information from metadata may be included after each detection step. In this study, at most three detection and correction iterations were performed. Following the recommendations of Venema *et al.* (2012), TX and TN data were homogenized on a monthly scale. Precipitation was homogenized on a yearly scale instead, because its signal-to-noise ratio (SNR) is considered too low on the monthly scale (Venema *et al.*, 2012). SNR is defined as the standard deviation of the breakpoints divided by the standard deviation of the noise of the difference series.

3.2. Construction of the dense and sparse networks

To investigate the performance of HOMER in sparse networks, *dense* and *sparse* networks were constructed from the complete Swiss network. They were built such that the dense networks have the highest possible station correlations while the sparse network represents the typical correlations found in Peru.

To this end, three groups were constructed such that: (1) the intra-group correlations are high ($R^2 \geq 0.85$) and (2) the inter-group correlations represent the typical Peruvian correlations ($0.60 \leq R^2 \leq 0.80$ for TN and TX, and $0.45 \leq R^2 \leq 0.60$ for P). The selection of the three groups, with about 6 to 25 members each, was based on hierarchical clustering (Maechler *et al.*, 2013; Kaufman and Rousseeuw, 1990; Struyf *et al.*, 1996, 1997; Lance and Williams, 1966; Begert 2008). They are referred to as 'North', 'Alps', and 'South', according to the location of the stations (Figure 1). Note that the groups are the same for TN and TX, but are partly different for P. In the following, these groups are called the *dense networks.*

In a next step, the *sparse networks* were constructed. Each sparse network is built by randomly sampling one station out of each of the three dense networks. Due to the restraint on the inter-group correlations of the dense networks, each of these sparse networks fulfills the Peruvian correlation condition. Since the number of stations (e.g. three) of each sparse network is not large enough for homogenization with HOMER, each sparse network was complemented with data from additional stations. For temperature, data from the quality controlled European Climate Assessment & Dataset (Van Engelen *et al.*, 2008) were used. For precipitation, records from south-eastern Switzerland were selected to complement the sparse groups. While this is not an optimal setting for homogenization, it reflects the situation in Peru where stations are often clustered. The correlations of the additional stations with the sampled stations fulfill the Peruvian correlation requirements. However, correlations between these additional stations were not restricted.

For each climatological parameter, 30 sparse networks were built. The sparse networks contain between 5 to 25 stations for temperature and between 7 to 14 stations for precipitation. Due to the correlation-based sampling procedure described above, the number of stations in each sparse network varies strongly. This setting well represents the variability of station availability in low density networks such as the one observed in Peru. Due to the random samples, some stations of the dense networks are part of more than one sparse network. On the other hand, some stations of the dense networks were not sampled at all and hence do not appear in any sparse network. For evaluation, only so-called *candidate stations* were considered. These consist of all stations of the dense networks that appear at least once in one of the sparse networks, as a result a total of 30 temperature stations and 40 precipitation stations were analyzed.

3.3. Experiments

The results of data homogenization not only depend on the homogenization algorithms selected but also on available resources such as time or metadata availability. In order to investigate the influence of the availability of resources, four homogenization experiments were conducted: (1) running HOMER in fully automatic mode (*auto*), (2) running HOMER in interactive mode (*manu*) without using metadata, (3) running HOMER in interactive mode and

with the use of metadata for breakpoint corroboration (*meta-post*), and (4) inserting important breakpoints based on metadata before statistical detection (*meta-pre*). These experiments were conducted for each variable and each dense and sparse network, resulting in a total of 396 homogenization experiments.

The experiment *auto* allows assessing the performance of HOMER with regard to the convenience of the procedure, i.e. the low human effort required. Meta information cannot be used in the automatic mode. The experiment *manu* reflects the situation of missing metadata. For both experiments *meta-post* and *meta-pre*, comprehensive metadata are available, but metadata are used differently.

In the sparse network experiments, only metadata of the candidate stations were used. This reflects the situation in countries such as Peru where metadata are mostly lacking and can often only be collected retrospectively, which is very time consuming (revision of original datasheets, observer questioning, etc.). We therefore assume that in reality this is only done for important stations, reflected here by the candidate stations.

3.4. Choice of the reference dataset

To assess the performance of HOMER in sparse networks, all experiments are compared to a *reference dataset*. In this study, meta-post applied to the dense network was chosen as the *reference*. This choice is justified below.

Venema *et al.* (2012) have shown that the algorithm PRODIGE, which is built-in HOMER (Mestre *et al.*, 2013), is amongst the best performing homogenization methods. An analysis of the data used in the inter-comparison study showed that the correlations of the six most closely correlated stations are slightly lower than the correlations between the Swiss stations. Hence, the high correlations of the Swiss network combined with the low number of breakpoints [one breakpoint in 48 years (Kuglitsch *et al.*, 2012) compared to roughly one breakpoint in 15 to 20 years in Europe (Venema *et al.*, 2012)] suggests a good performance of HOMER in the dense Swiss network. Furthermore, the use of comprehensive metadata in *meta-post* increases the reliability of the detected breakpoints. Since *meta-post* takes advantage of all resources, the performance of this approach can be expected to be very high.

It is clear that the truth about the occurrence and magnitude of the breakpoints cannot be known in real observational data. On the other hand, the advantage of using a real dataset for evaluation is evident: the statistical properties of the data and their inhomogeneities are realistic. In the surrogate dataset by Venema *et al.* (2012) for example, the introduced SNR was inadvertently twice as high as real datasets suggest (V. Venema, 2016; personal communication).

3.5. Performance measures

To assess the performance of a homogenization method, different error metrics are usually applied (Venema *et al.*, 2012; Domonkos, 2013). In this study, the temporal consistency of the data series (Section 3.5.1.), the linear trends (Section 3.5.2.), and the measure of efficiency (Section 3.5.3.) are evaluated.

3.5.1. *CRMSE and CRMSF*

The centred root mean squared error (*CRMSE*) is used to measure the temporal consistency of the homogenized or the raw dataset (x) with regard to the reference dataset (y) (e.g. Venema *et al.*, 2012; Domonkos, 2013). It is defined as:

$$CRMSE\left(\tilde{x},\tilde{y}\right)=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\tilde{x}_i-\tilde{y}_i\right)^2}$$

Here, $\tilde{x}=x-m\left(x\right)$ refers to the centred time series, $m(x)$ is the mean of x, and n is the number of time steps. In contrast to the non-centred root mean squared error (RMSE), missed or erroneous breaks are equally penalized with respect to the beginning and end of the time series. Perfect data, corresponding to the reference dataset in this study, result in $CRMSE=0$.

For precipitation, relative changes are considered and therefore the centred root mean squared fraction (CRMSF) is used (Golding, 1998):

$$\begin{aligned}CRMSF\left(\tilde{x},\ \tilde{y}\right)&=\exp\left(CRMSE\left(\log\left(\tilde{x}\right),\log\left(\tilde{y}\right)\right)\right)\\&=\exp\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\log\left(\tilde{x}_i\right)-\log\left(\tilde{y}_i\right)\right)^2}\right)\end{aligned}$$

where $x_i>0$ and $y_i>0$. Here, $\tilde{x}=\frac{x}{gm(x)}$ refers to the centred time series, and $gm(x)$ is the geometric mean of x. Perfect data result in $CRMSF=1$.

3.5.2. *Trend estimation*

Linear trends are estimated based on normalized yearly data using ordinary least squares regression (Frei, 2014; Chambers, 1992; Wilkinson and Rogers, 1973). The yearly precipitation totals are log-transformed to approximate the Gaussian distribution of the residuals more closely. Trends are expressed in unit per 10 years. For evaluation, the RMSE and the bias of the trends are determined. The differences in the average network trends are evaluated using the Wilcoxon signed-rank test (Wilcoxon, 1945, 1949; Bauer, 1972; Hollander and Wolfe, 1973).

3.5.3. *Measures of efficiency*

The measure of efficiency E is defined as the percentage of the RMSE of the homogenized dataset relative to the RMSE of the raw data (Domonkos *et al.*, 2011b):

$$E=\frac{RMSE_{\mathrm{raw}}-RMSE_{\mathrm{hom}}}{RMSE_{\mathrm{raw}}}\times 100.$$

In this manuscript, E_{T} refers to the efficiency measure of the RMSE of the yearly trend estimates, while E_{C} denotes the efficiency measure of both the CRMSE and the CRMSF. A positive efficiency value indicates an

Table 1. Results of the homogeneity (CRMSE and CRMSF) and trend (network average trend and RMSE) analyses of the dense and sparse networks for all experiments, as well as for the raw data.

Variable		Raw	Dense networks				Sparse networks			
			Reference	*manu*	*meta-pre*	*auto*	*meta-post*	*manu*	*meta-pre*	*auto*
Homogeneity										
CRMSE	TX	0.28	0.00	0.11	0.15	0.52	0.25	0.25	0.32	0.30
	TN	0.27	0.00	0.14	0.14	0.32	0.26	0.26	0.31	0.37
CRMSF	P	1.02	1.00	1.01	1.01	1.01	1.02	1.01	1.02	1.08
Trends										
°C/dec [u/decade]	TX	0.31	0.34	0.34	0.37	0.40	0.34	0.32	0.35	0.27
	TN	0.29	0.32	0.29	0.33	0.17	0.29	0.28	0.30	0.20
log(mm)/dec	P	0.007	0.007	0.007	0.007	0.006	0.008	0.008	0.012	0.011
RMSE	TX	0.16	0.00	0.02	0.05	0.13	0.08	0.08	0.06	0.11
	TN	0.13	0.00	0.04	0.03	0.16	0.06	0.07	0.06	0.14
	P	0.017	0.000	0.008	0.008	0.009	0.010	0.013	0.014	0.012

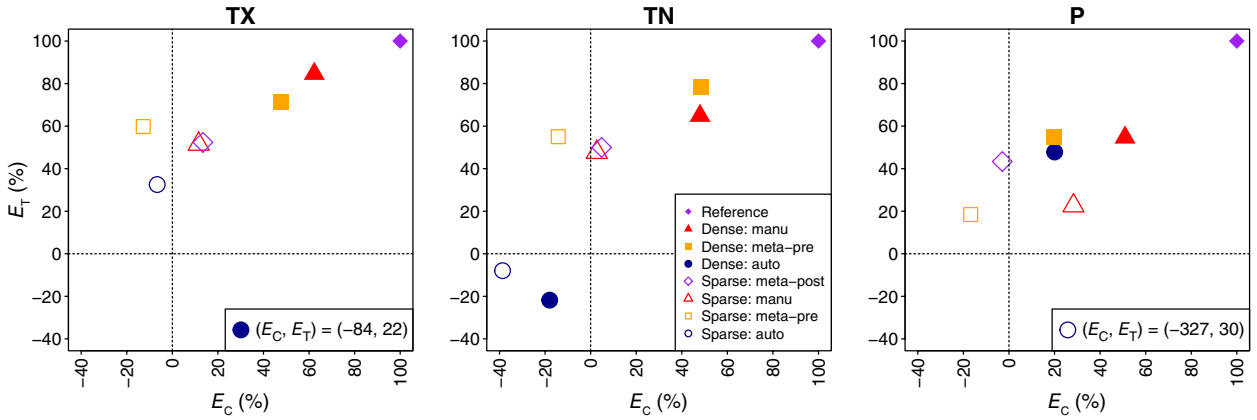

Figure 3. Efficiency measures E_C and E_T for the sparse (open symbols) and the dense networks (closed symbols). Symbols in the first quadrant indicate an improvement in both E_C and E_T over the raw data. By definition, the efficiency measures of the reference are at [100, 100]. Note that the coordinates of E_C for auto dense (left figure) and auto sparse (right figure) lie outside the figure limits. The respective coordinates are provided in the lower right corner of each figure. [Colour figure can be viewed at wileyonlinelibrary.com].

improvement of the homogenized data over the raw data, while a negative value indicates a deterioration of the data after homogenization.

4. Results

4.1. Homogenizing dense networks

4.1.1. Reference dataset

Almost 60 breakpoints were corrected in 30 temperature series and 16 breakpoints were corrected in 40 precipitation series. This corresponds to one breakpoint every 25 years for temperature and one breakpoint every 125 years for precipitation. The low number of detected breakpoints in precipitation may be attributed to the higher noise level of precipitation, as well as to the more robust measurement system (Begert *et al.*, 2005). In the precipitation network, all breakpoints were corroborated with metadata. For temperature, the number of corroborated breakpoints ranges between 70 and 85%.

The CRMSE (CRMSF) between the raw and the reference dataset is 0.27 °C for temperature and 1.02 for precipitation (Table 1 and Figure 3). With regard to trends, the bias of the network average trend is 0.03 °C/decade for temperature, and is close to zero for precipitation. The RMSE of the trends ranges between 0.13 and 0.16 °C/decade for temperature, and is 0.016 log(mm)/decade for precipitation (Figure 4).

4.1.2. Interactive mode

In *meta-pre*, the number of corrected temperature breakpoints is 50% higher than the number of breakpoints of the *reference*. Around 83–89% of the breakpoints were inserted due to a priori information from metadata (Figures 5 and 6). For precipitation, a total of 42 breakpoints were corrected in *meta-pre* (Figure 7), out of which 40 were inserted based on metadata. For *manu*, the number of detected temperature breakpoints is similar to the *reference*, and decreases by around 30–70% for precipitation.

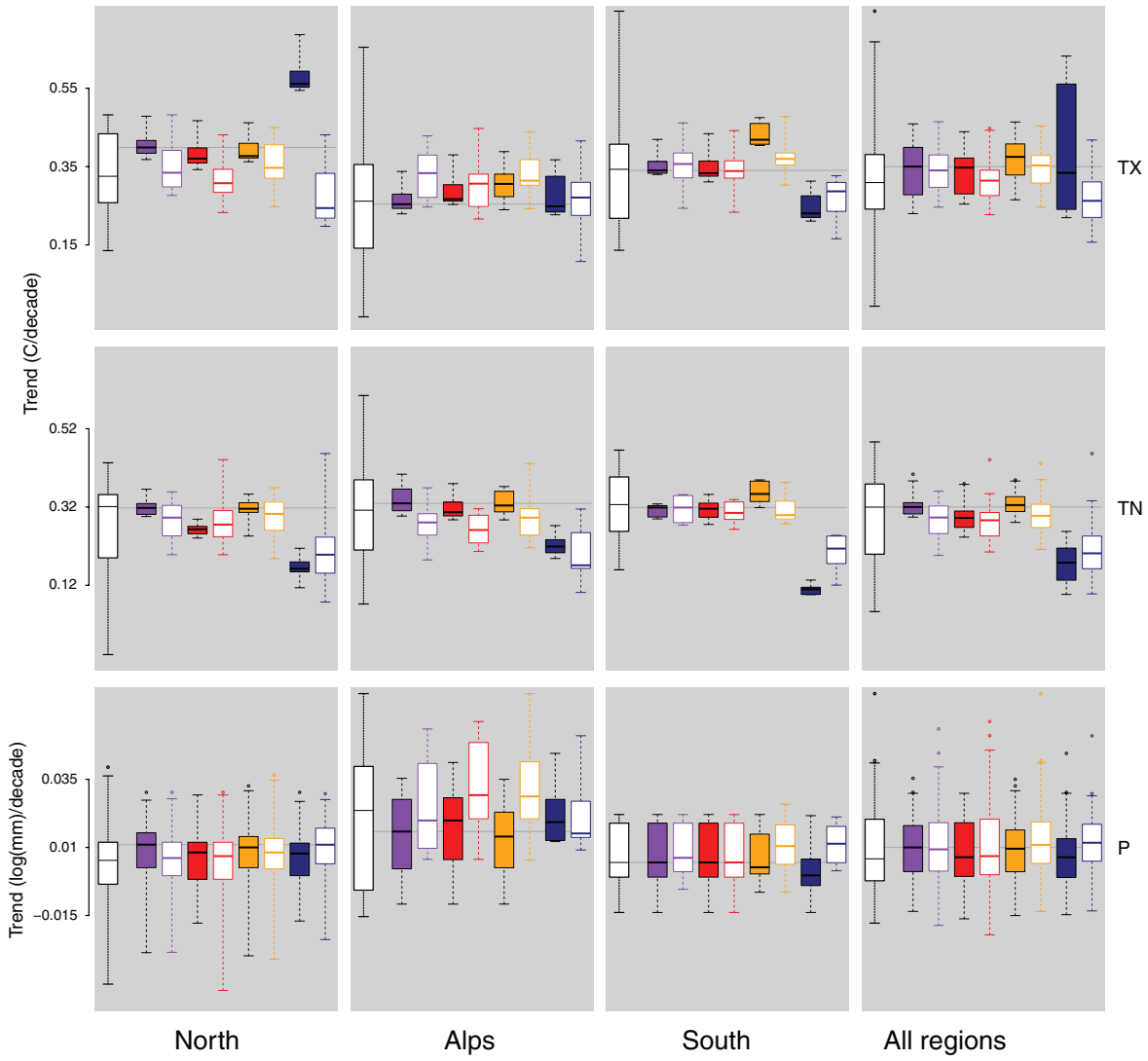

Figure 4. Estimated trends of *raw* (black), *meta-post* (purple), *manu* (red), *meta-pre* (orange), and *auto* (blue) for maximum temperature TX (top row), minimum temperature TN (middle row) and precipitation (bottom row) in the different geographical regions. Filled boxes indicate the dense and non-filled boxes the trends in the sparse networks. The whiskers indicate the 2.5 and 97.5% percentiles, and the boxes the first and third quartiles of the data. The grey line refers to the median of the reference dataset for each region.

Despite the differences in breakpoint number in *manu*, *meta-pre*, and the *reference*, the annual mean correction values between the three experiments do not differ strongly (Figures 5–7). An exception was observed in the precipitation network 'South' which was considered to be completely homogeneous by the *reference*. In contrast in *meta-pre*, a total of 11 breakpoints were inserted into this network.

For *manu* and *meta-pre*, the bias of the average network trend is similar to the bias of the raw data. It ranges between +/− 0.1 to 0.3 °C/decade for temperature, and is almost zero for precipitation. The RMSE of the trends decreases by about 60–80% for temperature compared to the raw data, and by 50% for precipitation (Figure 3). The efficiency measure of the CRMSE (CRMSF) ranges between 48 and 60% for temperature, and between 20 (*meta-pre*) and 40% (*manu*) for precipitation (Figure 3). We observe that *manu* is slightly more efficient in terms of CRMSE than *meta-pre* for TX, but underperforms for TN. In contrast for precipitation, *manu* is more efficient than *meta-pre*.

4.1.3. *Automatic mode*

For precipitation, the results of *auto* in the dense network show an increase in the homogeneity and trend accuracy that is comparable to the experiment *meta-pre* (Table 1 and Figure 3). For temperature, *auto* introduces more error into the data through homogenization (Figure 3). The unsatisfying results for temperature are surprising and can be traced back to the detection of a breakpoint around 1987. The timing of this breakpoint coincides with a sudden large temperature rise in Switzerland (Begert *et al.*, 2005) and Europe in general. This climate shift is also found in independent variables such as spring phenology or snow cover (e.g. Brönnimann, 2015). Since the breakpoint coincides with a climatic feature, these breaks are erroneously introduced by HOMER. Note that the problem was already mentioned by Mestre *et al.* (2013), who

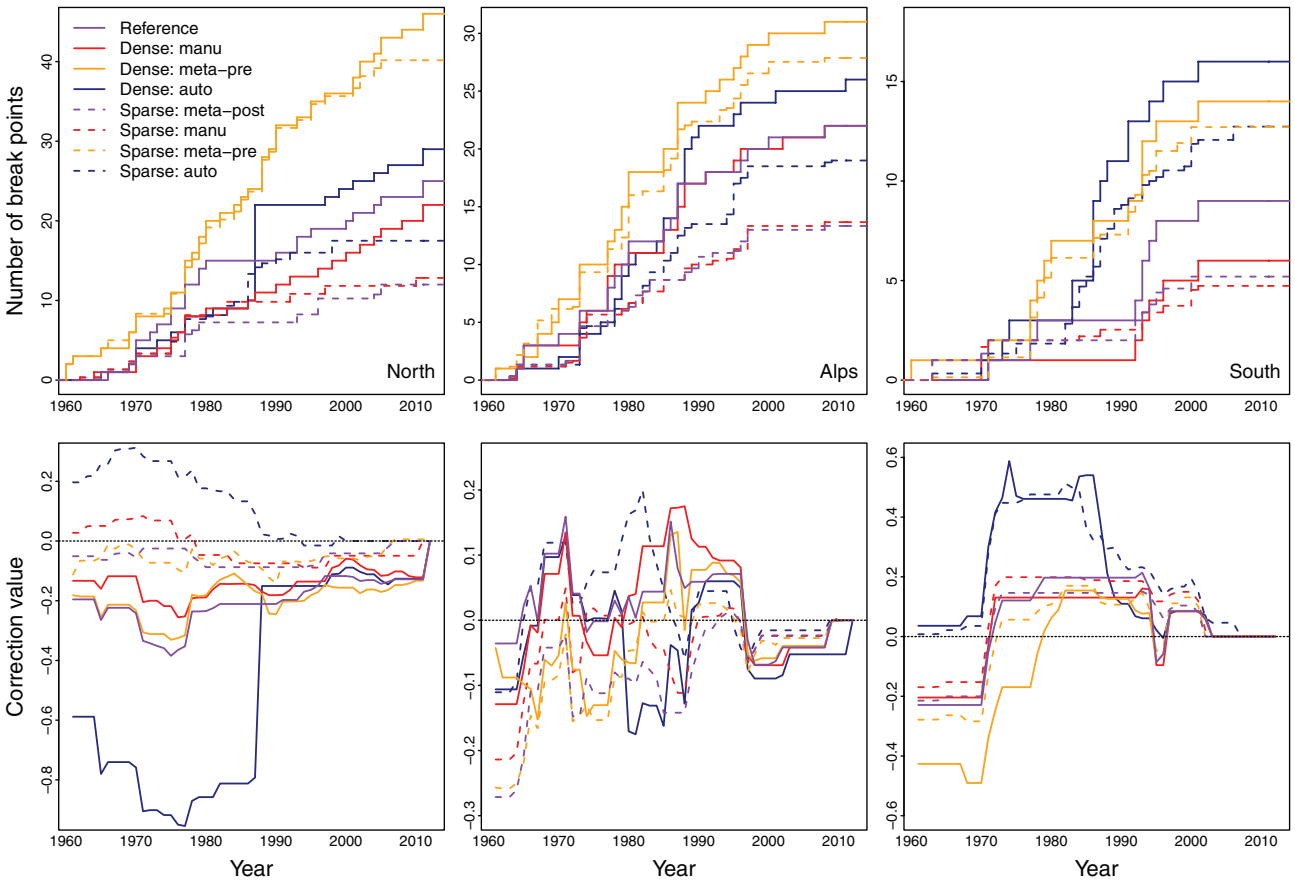

Figure 5. Upper figures: Cumulated number of detected breaks in TX for each experiment for dense (solid) and sparse (dashed) networks for 'North', 'Alps', and 'South' for all experiments (*meta-post* (purple), *manu* (red), *meta-pre* (orange), and *auto* (blue)). Lower figures: Mean correction values of each experiment.

state that ' … , the automatic joint-detection is not perfect' (pp. 60), and that the segmentation of the R-function multiseg 'wrongly attributes a climatic feature' to the breakpoints detected in a time series of Vienna around 1986. The near simultaneous erroneous detection of a breakpoint in 1987 leads to biased corrections, especially in the North network for TX with a bias in the trend of 0.2 °C/decade (Figures 4 and 5).

4.2. Homogenizing sparse networks

4.2.1. Interactive mode

Temperature: Both experiments *manu* and *meta-post* reduce the CRMSE in sparse temperature networks, and improve the accuracy of the estimated trends (Figure 3). Both approaches are therefore useful to improve the homogeneity and trend analyses of temperature data in sparse station networks. In contrast, the experiment *meta-pre* decreases the homogeneity of station data in sparse networks.

The efficiency measure of the CRMSE is slightly positive for both *meta-post* and *manu* ($E_C \approx 13\%$ for TX, and $E_C \approx 4\%$ for TN) (Figure 3). Compared to the dense networks however ($48\% \leq E_C \leq 60\%$ for *manu* and *meta-pre*), these values are rather low. Regarding individual stations, the lowest number of stations that show a decrease of the homogeneity is reached by *manu* (23% for TX; 43% for TN), followed by *meta-post* (37% for TX; 50% for TN). In *meta-pre*, E_C is negative ($E_C \approx -13\%$), implying that *meta-pre* decreases the homogeneity of the temperature series in sparse networks. Further, *meta-pre* decreases the homogeneity of more than 50% of the stations (Table 2 and Figure 8).

The trend bias of the sparse networks lies within +/−0.03 °C/decade for TX (+/−0.04 °C/decade for TN), corresponding to roughly 10% of the estimated average network trend. The bias is not reduced compared to the raw data. However, the trends are more consistent after homogenization (Figure 4). The RMSE of the trends is reduced from 0.13 to 0.16 °C/decade (raw data) to between 0.06 and 0.08 °C/decade for all experiments (Table 1), resulting in an efficiency measure of around 50% (Figure 3).

The number of detected breakpoints for *manu* and *meta-post* is around 25–40% smaller in the sparse networks compared to the dense networks (Figures 5 and 6). Totally, 60–70% of the breakpoints was corroborated by metadata (*meta-post*). Post-analysis of the homogenized data showed that the SNR in the sparse network is almost halved compared to the dense network. The low number of statistically detected breakpoints is hence attributed to the low SNR.

The low SNR also influences the correction of the breakpoints. This is illustrated through a comparison of

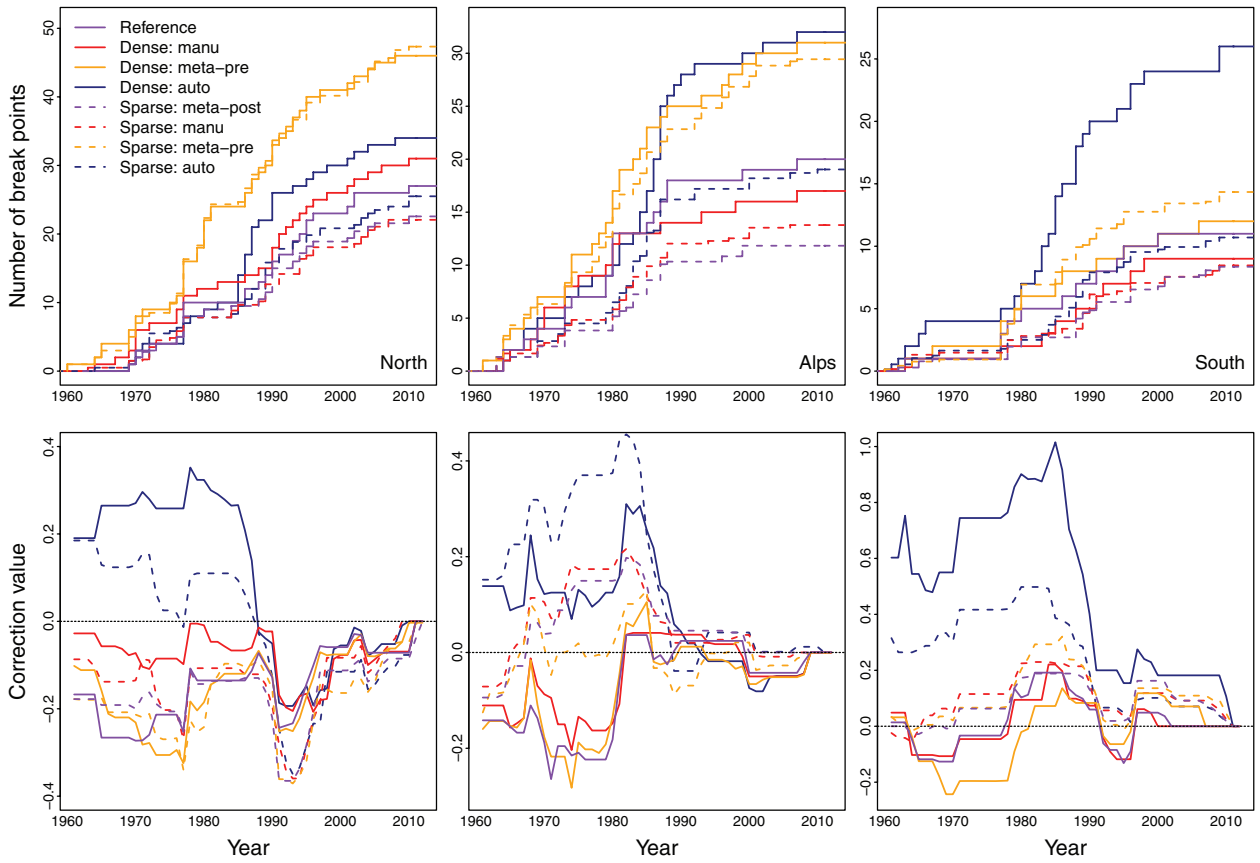

Figure 6. Same as Figure 5 for minimum temperature.

the correction values of *meta-pre* dense and sparse. By definition, (almost) the same breakpoints are inserted in the dense and sparse networks. The correction factors however differ strongly. Examples can be found for TN in the Alps, and for TX in the North and South networks between 1975 and 1980 (Figures 5 and 6). The average correction values differ by around 0.1–0.3 °C, indicating that the low SNR may introduce error into the correction of breakpoints. Additionally we observe that breakpoints, which occur due to the automation of the network around 1980, were often not detected in the sparse networks (Figures 5 and 6).

Precipitation: Around 70% of the stations are considered homogeneous by the reference homogenization. For simplicity, these 70% are called the *homogeneous* time series, while the remaining 30% are referred to as the *inhomogeneous* time series in the following paragraphs.

All sparse experiments improve the homogeneity of 28% of the time series (Table 2). Hence, 93% of the *inhomogeneous* time series are improved after homogenization. However, around one third of the *homogeneous* time series went through a correction by *meta-pre* and *meta-post*, resulting in an adverse effect for these station records. Apparently, breakpoints were inserted too liberally by *meta-pre* and *meta-post* in the sparse network. *Manu* has the best performance for sparse precipitation networks since data homogenization increased the CRMSF of only 5% of the time series. These results are supported by the efficiency measure E_C of the CRMSF (Figure 3): *manu* is the only experiment for which E_C in the sparse network is positive (Figure 3).

With respect to trends, all experiments reduce the RMSE of the trends (Figure 3). The reduction ranges from approximately 20 (*meta-post* and *manu*) to 43% (*meta-pre*). The bias of the trends is smallest for *meta-post* and *meta-pre*. Similar to temperature, homogenization reduces the variability of the trends. However, the trend variability in the investigated networks is similar to the uncertainty of the trend estimated at a station. The results are therefore not significant.

With regard to the number of detected breakpoints, the differences between the reference dataset and the sparse experiments are not consistent among the experiments (Figure 7). For example, the reference homogenization considers the South network to be homogeneous. In contrast, a total of 11 breakpoints were inserted by *meta-pre* in the sparse experiment (4 for *meta-post*). This 'over-detection' by *meta-pre* and *meta-post* is the main reason for the decreased homogeneity mentioned above. *Manu* is the only experiments which does not overestimate the number of breakpoints.

4.2.2. *Automatic mode*

In both the sparse precipitation and temperature networks, the efficiency measures E_C and E_T are negative after applying HOMER in automatic mode (Figure 3). In contrast to

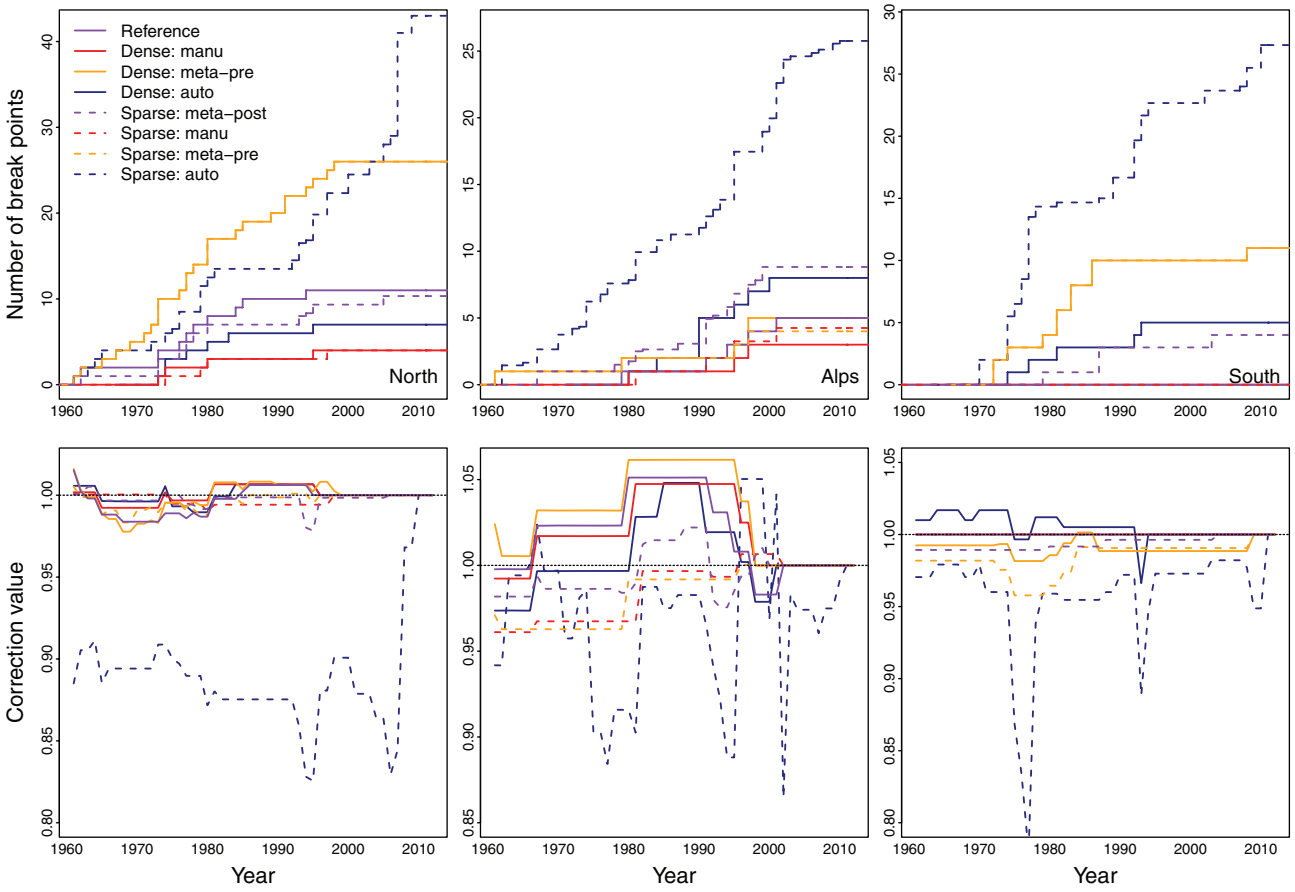

Figure 7. Same as Figure 5 for precipitation.

Table 2. Percentage of stations for which the temporal consistency (measured with the CRMSE or the CRMSF) of the homogenized data is improved (+) after homogenization compared to the raw data, for which the CRMSE (CRMSF) remains the same (○), and for which the data after homogenization is deteriorated (−).

		Dense networks (values in %)				Sparse networks (values in %)			
		Reference	*manu*	*meta-pre*	*auto*	*meta-post*	*manu*	*meta-pre*	*auto*
TX	+	90	83	80	37	63	70	47	47
	○	10	13	0	0	0	7	0	3
	−	0	3	20	63	37	23	53	50
TN	+	97	80	80	33	50	53	37	27
	○	3	7	3	0	0	3	0	0
	−	0	13	17	67	50	43	63	73
P	+	30	15	25	20	28	28	28	2
	○	70	85	48	68	45	68	40	10
	−	0	0	28	12	28	5	32	88

the dense temperature networks, there was no systematic error observed in the sparse networks (Section 4.1.3.), which explains the partly better performance of the experiment applied to the sparse networks. However, due to the large errors in the dense temperature network, the results for *auto* are not further discussed here.

5. Discussion

Homogenization methods are mostly developed in areas of high station densities such as encountered in Europe, and are normally evaluated under these conditions. This is for instance the case for the benchmark study of Venema *et al.* (2012). However in large parts of the world, the available measurements are sparse. In the following, the results of this first investigation on homogenization applied to networks of low station densities are discussed.

5.1. Homogenizing sparse networks with HOMER

This study demonstrates that HOMER may improve the quality of temperature and precipitation data in areas of low station density. The improvements are however small,

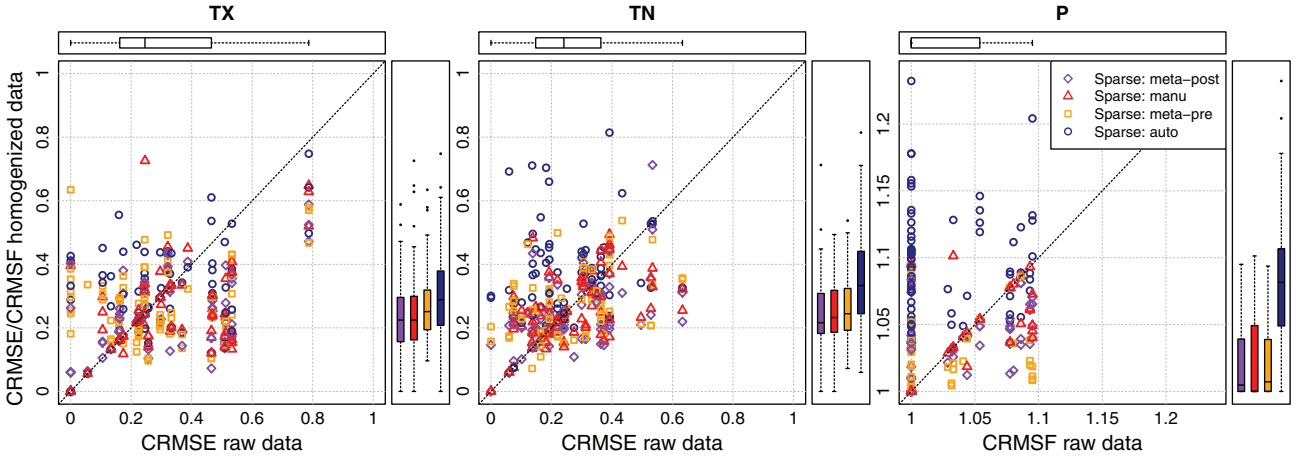

Figure 8. Comparison of the CRMSE (CRMSF for precipitation) of the raw (*x*-axis) and the homogenized data (*y*-axis). The results are displayed for the sparse networks only. Each symbol represents an individual stations. Points below the diagonal line show stations for which the homogeneity is increased after homogenization ($CRMSE_{hom} \leq CRMSE_{raw}$). The boxplots indicate the spread of the CRMSE/CRMSF for the individual experiments. The percentages of stations that improve/deteriorate after homogenization are given in Table 2.

i.e. the CRMSE is reduced by only 5–13% for temperature and by 30% for precipitation data compared to the raw data. Further, the bias of the average temperature and precipitation trend was not significantly altered by homogenization. The main improvement through homogenization with regard to trends resulted in a reduction of the RMSE (50% for temperature and up to 40% for precipitation). For trend analyses, applying HOMER is hence beneficial even in low station density networks. However, the low SNR may impede the detection of relevant breakpoints and may lead to erroneous corrections of the inserted breakpoints. For example, the systematic breakpoint that occurred due to the network automation around 1980 was not coherently detected in the temperature data. This underlines the importance of a careful station selection for homogenization to avoid simultaneous breakpoints (Menne and Williams, 2008; Kuglitsch *et al.*, 2012). In the Peruvian network, systematic changes are not known from the past, and so far this problem might not apply there. However, the issue clearly shows that a careful measuring strategy is essential.

Comparing the three experiments in interactive mode leads to the following conclusions: in contrast to our expectations, the use of metadata (*meta-post*) for breakpoint corroboration does not clearly improve homogenization in sparse networks over purely statistical homogenization. The experiment *manu* resulted in similar results with regard to CRMSE and trends. In contrast, if metadata is introduced a priori, it is likely that corrections become erroneous and may ultimately lead to a deterioration of the data. Such erroneous corrections might also occur if a combination of different variables (e.g. maximum, minimum, and mean temperature, as well as temperature range) are used to determine breakpoints (e.g. Aguilar *et al.*, 2002). We conclude that in sparse temperature station networks metadata should only be used to confirm breakpoints with clear statistical evidence, for instance by adjusting the timing of the occurrence of the breakpoint. This is even more pronounced in precipitation networks, where *manu* is the only experiment that does not introduce breakpoints to homogeneous time series. It must however be mentioned that breakpoint detection is generally ambiguous for precipitation due to the low SNR (e.g. Venema *et al.*, 2012), and even more so in sparse networks.

In addition we found that the use of HOMER in automatic mode may not be recommended for homogenization. The errors obtained for *auto* indicate that this variant in HOMER still requires optimization.

Regarding the interpretation of the results, we must keep in mind that the *reference* used in this study is affected by errors. Domonkos *et al.* (2011b) showed that PRODIGE improves the CRMSE by 70% over the raw data. Compared to the truth (a synthetic dataset), the remaining 30% of the CRMSE could not be corrected for by PRODIGE. Due to these errors in the reference, obtaining the same output by the experiments under evaluation and the reference is more difficult. Therefore, some uncertainty with regard to the presented results remains.

5.2. Strategies to improve data quality in countries of low station density

This study shows that homogenization with HOMER may improve the quality of sparse temperature and precipitation datasets, however only if breakpoints are treated conservatively. The improvements of the homogenized over the raw data are small. These findings are in accordance with Auer *et al.* (2005), who state that inhomogeneities may disappear within the noise for station correlations below 0.5. It is therefore likely that the true number of breakpoints in Peru is considerably larger than the reported number of breaks (Section 1) of one breakpoint every 13–20 years.

In order to enable the generation of homogeneous time series, the issue of the low station density must be addressed in many parts of the world. While installation of new stations is very costly, other measures to maximize benefit from existing resources would be valuable. This could be achieved for example through: (1) a sound

data-quality control and correction system to increase the number of suitable stations for homogenization, (2) the integration of partner networks in and surrounding a region, and hence (3) an international sharing of reliable datasets to improve stations density at the frontiers. Further, to improve the data quality and homogenization in the future, emphasis could be put on (1) regular observer and maintainer instruction to reduce measurement errors, (2) setting-up a real-time quality control systems allowing for immediate intervention in case of measurement errors (for both conventional and automatic stations), (3) analysis of parallel measurements (if existing) to specifically quantify the influence of reported breakpoints, and (4) setting-up a metadata collection and storage systems.

Since metadata gives confidence in breakpoint detection, it would be beneficial to gather information from the past. Possible approaches to recover station histories are station visits, observer questioning, and examination of original data sheets for comments and annotations, and a systematic compilation of information available at the meteorological offices. Some of these approaches (e.g. implementation of a data-quality control system, analysis of parallel measurements) are currently implemented at SENAMHI Peru.

In regions of sparse station networks, homogenization would certainly benefit from an increase in the station density. In the context of installing new stations (nowadays often automatic weather stations), reflections on the network design should be made. This includes: (1) avoiding simultaneous breakpoints for homogenization (Menne and Williams, 2008; Kuglitsch *et al.*, 2012), (2) installation of parallel measurements, (3) ensuring measurement site representativity (Leroy, 1998), and (4) ensuring the better representation of the climatic factors encountered in the regions.

Moreover, in Switzerland it has proven useful to define so-called important climate observing stations (Begert *et al.*, 2007). These stations are climatologically representative for a greater area, and jointly represent the different climates of Switzerland. In regions of low density networks, these stations might consist of the best-quality stations that are representative for a greater area. These stations could be prioritized with respect to station maintenance, observer instruction, and metadata collection. Further, they should be kept as homogeneous as possible (Aguilar *et al.*, 2003) to serve as reliable references for climate studies.

Further, some issues regarding homogenization in general should be addressed. HOMER has not been validated against a synthetic dataset yet. A proper evaluation of HOMER, especially of the joint-detection method, is required. Since many networks worldwide have lower correlation than networks in Western Europe, there is a need for a method inter-comparison focusing on sparse station networks. Other methods could perform better than HOMER in sparse networks. For example, the use of statistical methods not relying on metadata (e.g. PENHOM; Kuglitsch *et al.*, 2009) should be investigated. In addition, there is a need to evaluate homogenization methods under tropical conditions. For example, the mainly convective precipitation regimes in the tropics might have an influence on the performance of relative homogenization methods since these precipitation events are likely to increase the noise in the difference series. In addition, the frequency of breakpoints in Switzerland (Kuglitsch *et al.*, 2012) is clearly lower than in Peru. Small shifts in a time series act as a kind of noise, which substantially lowers the detection skill of larger shifts (Domonkos *et al.*, 2011b). Inclusion of artificial inhomogeneities in the Swiss dataset, or an evaluation of HOMER on a synthetic dataset representing the conditions in Peru could shed light on the influence of such issues on homogenization.

6. Conclusions

Since the performance of homogenization depends to a large degree on factors such as breakpoint magnitudes and frequency, the results of this study cannot so easily be generalized. Nevertheless, the following conclusions may be drawn:

- In sparse networks, potential breakpoints should be inserted conservatively. Otherwise, there is a risk of harmful corrections due to the low signal-to-noise ratio. Metadata should only be used to confirm and adjust the exact timing of breakpoints.
- The performance of homogenization declines sharply in sparse compared to dense station networks. Nevertheless, homogenization may increase the trend accuracy in sparse networks even if the temporal consistency is reduced at the same time.
- The low signal-to-noise ratio in sparse networks reduces the number of statistically detected breakpoints by around 25 to 40% for temperature, and by 50% for precipitation compared to a dense network.
- Application of HOMER in automatic mode is not recommended.
- Low station density and poor correlation are serious hindrances to generating homogeneous series. Other approaches, such as integrating partner networks, a comprehensive quality control, improved station maintenance, among others, should be addressed to increase the quality of climate data in region of low stations densities.

Acknowledgements

We thank three anonymous referees for their review and their constructive comments. Further, we express our gratitude to V. Venema for supporting and encouraging us to publish the presented results. We thank S.C. Scherrer for proof-reading the manuscript. We acknowledge the support of the World Meteorological Organization (WMO) through the project ‘Servicios CLIMáticos con énfasis en los ANdes en apoyo a las DEcisioneS’ (CLIMANDES), Project no. 7F-08453.01 between the Swiss Agency for Development and Cooperation (SDC) and the WMO, and the project ‘Data on climate and Extreme weather for the

Central AnDEs' (DECADE), project no. IZ01Z0_147320 through founding by the Swiss Program for Research on Global Issues for Development (r4d.ch).

References

Aguilar E, Brunet M, Saladié O, Sigró J, López D. 2002. Hacia una aplicación óptima del Standard Normal Homogeneity Test para la homogeneización de series de temperatura. AEG/UNIZAR, 17–34 pp. ISBN: 84-95480-69-7 (in Spanish).

Aguilar E, Auer I, Brunet M, Peterson TC, Wieringe J. 2003. Guidelines on climate metadata and homogenization. WMO/TD No. 1186.

Auer I, Böhm R, Jurkovic A, Orlik A, Potzmann R, Schöner W, Ungersböck M, Brunetti M, Nanni T, Maugeri M, Briffa K, Jones P, Efthymiadis D, Mestre O, Moisselin JM, Begert M, Brazdil R, Bochnicek O, Cegnar T, Gajic-Capka M, Zaninivic K, Majstorovic Z, Szalai S, Szentimrey T. 2005. A new instrumental precipitation dataset in the greater alpine region for the period 1800–2002. *Int. J. Climatol.* **25**: 139–166.

Auer I, Böhm R, Jurkovic A, Lipa W, Orlik A, Potzmann R, Schöner W, Ungersböck M, Matulla Ch, 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. *Int. J. Climatol.* **27**: 17–46. https://doi.org/10.1002/joc.1377.

Bauer DF. 1972. Constructing confidence sets using rank statistics. *J. Am. Stat. Assoc.* **67**: 687–690.

Begert M, Seiz G, Schlegel T, Musa M, Baudraz G, Moesch M. 2003. Homogenisierung von Klimamessreihen der Schweiz und Bestimmung der Normwerte 1961–1990, Schlussbericht des Projekts NORM90. Arbeitsberichte der MeteoSchweiz, No. 67, 170 pp (in German).

Begert M, Schlegel T, Kirchhofer W. 2005. Homogeneous temperature and precipitation series of Switzerland from 1864 to 2000. *Int. J. Climatol.* **25**: 65–80.

Begert M, Seiz G, Foppa N, Schlegel T, Appenzeller C, Müller G. 2007. Die Überführung der klimatologischen Referenzstationen der Schweiz in das Swiss National Basic Climatological Network (Swiss NBCN). Arbeitsberichte der MeteoSchweiz, No. 215, 43 pp (in German).

Begert M. 2008. Die Repräsentativität der Stationen im Swiss National Climatological Network (Swiss NBCN), Arbeitsberichte der MeteoSchweiz, No. 217, 40 pp (in German).

Brönnimann S. 2015. Climatic changes since 1700. Springer International Publishing: Cham, Switzerland. https://doi.org/10.1007/978-3-319-19042-6_4.

Brooks N, Adger WN. 2003. Country level risk measures of climate-related natural disasters and implications for adaptation to climate change. Tyndall Centre Working Paper No. 26, 30 pp.

Brunetti M, Maugeri M, Monti F, Nanni T. 2006. Temperature and precipitation variability in Italy in the last two centuries from homogenized instrumental time series. *Int. J. Climatol.* **26**: 345–381.

Buytaert W, Celleri R, Willems P, Bièvre BD, Wyseure G. 2006. Spatial and temporal rainfall variability in mountainous areas: a case study from the south Ecuadorian Andes. *J. Hydrol.* **329**: 413–421.

Buytaert W, De Bièvre B. 2012. Water for cities: the impact of climate change and demographic growth in the tropical Andes. *Water Resour. Res.* **48**: W08503.

Cao L-J, Yan Z-W. 2012. Progress in research on homogenization of climate data. *Adv. Clim. Change Res.* **3**(2): 59–67.

Caussinus H, Mestre O. 2004. Detection and correction of artificial shifts in climate series. *Appl. Stat.* **53**: 405–425.

Chambers JM. 1992. Linear models. In *Chapter 4 of Statistical Models in S*, Chambers JM, Hastie TJ (eds). Wadsworth & Brooks/Cole: Pacific Grove, CA.

Coll J, Curley C, Walsh S, 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. 84*. WMO: Geneva, Switzerland, 23–45.

Domonkos P, Poza R, Efthymiadis D. 2011a. Newest developments of ACMANT. *Adv. Sci. Res.* **6**: 7–11.

Domonkos P, Venema V, Mestre O. 2011b. Efficiencies of homogenisation methods: our present knowledge and its limitation. In *Proceedings of the 7th Seminar for Homogenization and Quality Control in Climatological Databases*, 11–24.

Domonkos P. 2013. Measuring performances of homogenization methods. *Q. J. Hung. Meteor. Serv.* **117**: 91–112.

Frei C. 2014. *trend: Functions for Trend Estimation and Resting. R Package Version 1.5.2*.

Freitas L, Pereira MG, Caramelo L, Mendes M, Nunes LF. 2013. Homogeneity of monthly air temperature in Portugal with HOMER and MASH. *Q. J. Hung. Meteor. Serv.* **117**: 69–90.

Golding BW. 1998. Nimrod: a system for generating automated very short range forecasts. *Meteorol. Appl.* **5**: 1–16.

Hollander M, Wolfe DA. 1973. *Nonparametric Statistical Methods*. John Wiley & Sons: New York, NY, 27–33 (one-sample), 68–75 (two-sample).

Kaufman L, Rousseeuw PJ. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley: New York, NY.

Kuglitsch F, Toreti A, Xoplaki A, Della-Marta P, Luterbacher J, Wanner H. 2009. Homogenization of daily maximum temperature series in the Mediterranean. *J. Geophys. Res.* **114**: D15108. https://doi.org/10.1029/2008JD011606.

Kuglitsch FG, Auchmann R, Bleisch R, Brönnimann S, Martius O, Stewart M. 2012. Break detection of annual Swiss temperature series. *J. Geophys. Res.* **117**: D13105. https://doi.org/10.1029/2012JD017729.

Lance GN, Williams WT. 1966. A general theory of classificatory sorting strategies. I. Hierarchical Systems. *Comput. J.* **9**: 373–380.

Leroy M. 1998. Meteorological measurement representativity, nearby obstacles influence. Papers Presented at the WMO Technical Conference on Meteorological and Environmental Instruments and Methods of Observation (TECO-1998), Instruments and Observing Methods Report No. 70, 51–54.

López-Moreno JI, Morán-Tejeda E, Vicente-Serrano SM, Bazo J, Azorin-Molina C, Revuelto J, Sánchez-Lorenzo A, Navarro-Serrano F, Aguilar E, Chura O. 2016. Recent temperature variability and change in the Altiplano of Bolivia and Peru. *Int. J. Climatol.* **36**: 1773–1796. https://doi.org/10.1002/joc.4459.

Luhunga P, Mutayoba E, Ngongolo H. 2014. Homogeneity of monthly mean air temperature of the United Republic of Tanzania with HOMER. *Atmos. Clim. Sci.* **4**: 70–77. https://doi.org/10.4236/acs.2014.41010.

New M, Hulme M, Jones P. 1999. Representing twentieth-century space–time climate variability. Part I. Development of a 1961–90 mean monthly terrestrial climatology. *J. Clim.* **12**: 829–856.

New M, Hulme M, Jones P. 2000. Representing twentieth-century space–time climate variability. Part II. Development of 1901–96 monthly grids of terrestrial surface climate. *J. Clim.* **13**: 2217–2238.

Noone S, Murphy C, Coll J, Matthews T, Mullan D, Wilby RL, Walsh S. 2016. Homogenization and analysis of an expanded long-term monthly rainfall network for the Island of Ireland (1850–2010). *Int. J. Climatol.* **36**: 2837–2853. https://doi.org/10.1002/joc.4522.

Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K. 2013. *cluster: Cluster Analysis Basics and Extensions. R package version 1.14.4*.

Menne MJ, Williams CN. 2008. Homogenization of temperature series via pairwise comparisons. *J. Clim.* **22**: 1700–1717.

Mestre O, Domonkos P, Picard F, Auer I, Robin S, Lebarbier E, Böhm R, Aguilar E, Guijarro J, Vertačnik G, Klancar M, Dubuisson B, Stepanek P. 2013. HOMER: a homogenization software – methods and applications. *Q. J. Hung. Meteor. Serv.* **117**: 47–67.

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

Peterson TC, Easterling DR. 1994. Creation of homogeneous composite climatological benchmark series. *Int. J. Climatol.* **14**: 671–679.

Peterson TC, Easterling DR, Karl TR, 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, Forland EJ, Hassen-Bauer I, Alexandersson H, Jones P, Parker D. 1998. Homogeneity adjustments of in situ atmospheric climate data: a review. *Int. J. Climatol.* **18**: 1493–1517.

R Development Core Team. 2014. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria, http://www.Rproject.org/.

Reeves J, Chen J, Wang XL, Lund R, Lu Q. 2007. A review and comparison of changepoint detection techniques for climate data. *J. Appl. Meteor. Climatol.* **46**: 900–915.

Rosas G, Gubler S, Oria C, Acuña D, van Geijtenbeek D, Jacques M, Konzelmann T, Lavado W, Matos A, Mauchle F, Rohrer M, Rossa A, Scherrer SC, Valdez M, Valverde M, Villar G, Villegas E. 2016. Towards implementing climate services in Peru – The project CLIMANDES. *Clim. Serv.* **4**: 30–41.

Salzmann N, Huggel C, Calanca P, Díaz A, Jonas T, Jurt C, Konzelmann T, Lagos P, Rohrer M, Silverio W, Zappa M. 2009. Integrated assessment and adaptation to climate change impacts in the Peruvian Andes. *Adv. Geosci.* **22**: 35–39.

Seiler C, Hutjes RWA, Kabat P. 2012. Climate variability and trends in Bolivia. *J. Appl. Meteorol. Climatol.* **52**: 130–146.

Struyf A, Hubert M, Rousseeuw PJ. 1996. Clustering in an object-oriented environment. *J. Stat. Softw.* **1**: 33.

Struyf A, Hubert M, Rousseeuw PJ. 1997. Integrating robust clustering techniques in S-PLUS. *Comput. Stat. Data Anal.* **26**: 17–37.

Toreti A, Kuglitsch F, Xoplaki E, Luterbacher J, Wanner H. 2010. A novel method for the homogenization of daily temperature series and its relevance for climate change analysis. *J. Clim.* **25**: 5325–5331.

Toreti A, Kuglitsch FG, Xoplaki E, Luterbacher J. 2011. A novel approach for the detection of inhomogeneities affecting climate time series. *J. Appl. Meteorol. Climatol.* **51**: 317–326.

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

Van Engelen A, Klein Tank A, Van de Schrier G, Klok L. 2008. European Climate Assessment & Dataset (ECA&D), Towards an operational system for assessing observed changes in climate extremes. KNMI Report.

Venema VKC, Mestre O, Aguilar E, Auer I, Guijarro JA, Domonkos P, Vertačnik G, Szentimrey T, Stepanek P, Zahradnicek P, Viarre J, Müller-Westermeier G, Lakatos M, Williams CN, Menne MJ, Lindau R, Rasol D, Rustemeier E, Kolokythas K, Marinova T, Andresen L, Acquaotta F, Fratianni S, Cheval S, Klancar M, Brunetti M, Gruber C, Prohom Duran M, Likso T, Esteban P, Brandsma T. 2012. Benchmarking homogenization algorithms for monthly data. *Clim. Past* **8**: 89–115.

Vertačnik G, Dolinar M, Bertalanič R, Klančar M, Dvoršek D, Nadbath M. 2015. Ensemble homogenization of Slovenian monthly air temperature series. *Int. J. Climatol.* **35**: 4015–4026. https://doi.org/10.1002/joc.4265.

Vicente-Serrano SM, El Kenawy A, Azorin-Molina C, Chura O, Trujillo F, Aguilar E, Martín-Hernández N, López-Moreno JI, Sanchez-Lorenzo A, Moran-Tejeda E, Revuelto J, Ycaza P, Friend F. 2015. Average monthly and annual climate maps for Bolivia. *J. Maps* **12**(2): 295–310. https://doi.org/10.1080/17445647.2015.1014940.

Vuille M, Francou B, Wagnon P, Juen I, Kaser G, Mark BG, Bradley RS. 2008. Climate change and tropical Andean glaciers: past, present and future. *Earth Sci. Rev.* **89**: 79–96.

Wilcoxon F. 1945. Individual comparisons by ranking methods. *Biometrics Bull.* **1**(6): 80–83.

Wilcoxon F. 1949. Some rapid approximate statistical procedures. *Ann. N. Y. Acad. Sci.* **52**(6): 808–814.

Wilkinson GN, Rogers CE. 1973. Symbolic descriptions of factorial models for analysis of variance. *Appl. Stat.* **22**: 392–9.

World Bank. 2010. Adaptation to Climate Change – Vulnerability Assessment and Economic Aspects: Plurinational State of Bolivia, Washington, DC. World Bank. https://openknowledge.worldbank.org/handle/10986/12744 License: CC BY 3.0 Unported (accessed 25 April 2017).